Information Processing and Management 63 (2026) 104880

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm

Reasoning based on symbolic and parametric knowledge bases: A survey☆

Mayi Xu[1], Yunfeng Ning[1], Yongqi Li, Jianhao Chen, Jintao Wen, Yao Xiao, Shen Zhou, Birong Pan, Zepeng Bao, Xin Miao, Hankun Kang, Ke Sun, Tieyun Qian*

School of Computer Science, Wuhan University, Wuhan, 430072, Hubei, China

ARTICLE INFO

Keywords:
Reasoning
Symbolic knowledge base
Parametric knowledge base
Pre-trained language models
Knowledge graphs

ABSTRACT

Reasoning refers to drawing new conclusions based on existing knowledge, which can support various applications. While existing surveys provide comprehensive overviews within their perspectives, they mainly organize methods by task, algorithmic paradigm, or application scenario. As a result, they offer limited discussion and explanation on how the types, properties, and scales of the underlying knowledge bases constrain or strength different reasoning strategies. To bridge this gap, this survey summarizes reasoning methods from the view of their dependent knowledge base. First, this perspective covers a broader range of knowledge bases than previous surveys. This allows for a more comprehensive comparison and understanding of their differences when employed in reasoning. Second, this perspective introduces a data-centric taxonomy that directly reflecting how the distinct characteristics of knowledge bases, such as information density and representational nature, explicitly constrain or enhance reasoning strategies. For instance, high-stakes systems require transparent and verifiable knowledge bases, whereas creative applications prioritize flexibility. Finally, this perspective outlines future directions to guide more advanced reasoning research. It advocates for shifting the focus from purely performance toward more comprehensive metrics such as verifiability, consistency, transparency, and safety

1. Introduction

Reasoning refers to inferring new conclusions from existing knowledge (F. Yu et al., 2024). This ability is fundamental to human intelligence and essential for complex tasks such as problem-solving, decision-making, and critical thinking. The cognitive process of reasoning involves using evidence, arguments, and logic to draw conclusions or make judgments (Huang & Chang, 2023). This process is essential for many real-world applications, including clinical diagnosis (Banning, 2008; Montgomery Jr, 2018; Zack et al., 2023), basic education (W. Chen et al., 2023; Chen et al., 2024; Lewkowycz et al., 2022), and financial analysis (Son et al., 2023; Yuan et al., 2025; Zhao et al., 2024). Reasoning ability is central to human intelligence, yet modern natural language processing systems still struggle to reason based on the information they are given or have already learned (Bhargava & Ng, 2022; Duan et al., 2020; Qiao et al., 2023; Wang et al., 2022). The study of reasoning is essential in fields like neuroscience (Krawczyk, 2012), psychology (Wason, 1968), philosophy (Passmore, 1961; Rescher, 2001), and computer science (Huth, 2004). It is a key

☆ This article is part of a Special issue entitled: 'Employing Surveys' published in Information Processing and Management.

* Corresponding author.
E-mail address: qty@whu.edu.cn (T. Qian).

[1] Equal contribution.

https://doi.org/10.1016/j.ipm.2026.104880
Received 24 February 2025; Received in revised form 29 April 2026; Accepted 29 April 2026

Table 1
Coverage of dependent knowledge bases in different reasoning surveys. Abbreviation: knowledge graph (KG), knowledge base (KB), pre-trained language models (PLMs).

Surveys	Static KG	Temporal KG	Multi-modal KG	Structured table	Unstructured text	Heterogeneous KB	PLMs
X. Chen et al. (2020)	●	○	○	○	○	○	○
Ji et al. (2021)	●	●	○	○	○	○	○
Lan et al. (2021)	●	○	○	○	○	○	●
Zhu et al. (2021b)	○	○	○	○	●	○	●
Shen et al. (2022b)	○	○	○	○	●	○	●
Jin et al. (2022)	○	○	○	●	○	○	○
Zhang et al. (2023)	○	○	○	○	●	○	●
Huang and Chang (2023)	○	○	○	○	○	○	●
Qiao et al. (2023)	○	○	○	○	○	○	●
Su et al. (2024)	○	●	○	○	○	○	●
F. Yu et al. (2024)	○	○	○	○	○	○	●
Liang et al. (2024)	●	●	●	○	○	○	○
Meng et al. (2024)	●	○	○	○	○	○	○
Plaat et al. (2024)	○	○	○	○	○	○	●
DeLong et al. (2025)	●	○	○	○	○	○	○
Bandyopadhyay et al. (2025)	○	○	○	○	○	○	●
Ours	●	●	●	●	●	●	●

area of research for bridging the gap between human and machine intelligence (Qiao et al., 2023). Therefore, developing artificial intelligence systems with strong reasoning abilities is both a key research goal and a means to enhance complex applications. Boisvert et al. (2024), Lucas et al. (2024), Shi et al. (2024), Sun et al. (2025), F. Yu et al. (2024).

Previous surveys review reasoning methods from different perspectives. For instance, a survey focuses on natural language reasoning (F. Yu et al., 2024). Some surveys highlight reasoning on structured knowledge graphs (X. Chen et al., 2020; DeLong et al., 2025; Liang et al., 2024; Meng et al., 2024). There are also surveys that focus on reasoning with specific knowledge sources, such as Wikidata and Wikipedia, for question answering (Lan et al., 2021; Shen et al., 2022b; Su et al., 2024; Zhang et al., 2023; Zhu et al., 2021b). Recent works also summarize reasoning methods for prompting large language models (Huang & Chang, 2023; Plaat et al., 2024; Qiao et al., 2023). While existing surveys provide comprehensive overviews within their perspectives, they mainly organize methods by task, algorithmic paradigm, or application scenario. As a result, they offer limited discussion and explanation on how the types, properties, and scales of the underlying knowledge bases fundamentally constrain or strength different reasoning strategies.

As previous work points out, reasoning is a process of integrating existing knowledge to derive new conclusions about the world (Fagin et al., 2004; Huang & Chang, 2023; F. Yu et al., 2024). Hence, current reasoning methods rely heavily on knowledge bases (Chen et al., 2024b; Gu et al., 2025; Huang et al., 2025; Jiang et al., 2025; Zhang et al., 2025). However, both the scenarios to which the knowledge bases are applied and their storage formats are significantly different (J. Lee et al., 2024; Xin et al., 2025; Zhang et al., 2025). Therefore, investigating reasoning from the perspective of the knowledge base helps us gain a deeper understanding of the challenges and future directions of reasoning-related works.

To bridge the critical gap, this survey proposes to summarize reasoning methods from the view of their dependent knowledge base. First, as shown by the statistics in Table 1, this perspective encompasses a wider range of dependent knowledge bases than previous reasoning surveys. Such extensive coverage enables a more comprehensive comparison and a deeper understanding of their differences when employed in reasoning. Second, this perspective introduces a data-centric taxonomy, highlighting the intrinsic interaction between knowledge representation and reasoning mechanisms. It directly reflect how the distinct characteristics of knowledge bases, such as information density and representational nature, explicitly constrain or enhance reasoning strategies. Third, this perspective highlights several promising future directions that may be overlooked by other perspectives. These directions can help guide the development of more effective and robust reasoning systems.

Building on this view, our survey enables a more direct comparison of different approaches, clarifying their respective challenges and advantages, and providing practical guidance for future research and system design. Based on their knowledge storage formats, knowledge bases can be categorized into symbolic and parametric types. Symbolic knowledge bases represent information explicitly through human-interpretable symbols, whereas parametric knowledge bases encode knowledge implicitly within their parameters. We investigate the reasoning methods based on symbolic knowledge bases, parametric knowledge bases, and both of them, respectively. The development timeline of reasoning paradigms is shown in Fig. 1. Finally, we explore the challenges and potential future directions for reasoning with both symbolic and parametric knowledge.

In summary, the main contributions of this survey are:

- To the best of our knowledge, we are the first to provide a comprehensive survey on reasoning studies from the perspective of their dependent knowledge bases.
- We conduct a thorough investigation of various types of reasoning methods that utilize symbolic knowledge bases, parametric knowledge bases, and their combination, whereas previous reviews only focused on one of them.
- We have meticulously summarized the challenges and future research directions related to reasoning, which will contribute to advancing the development of this field.

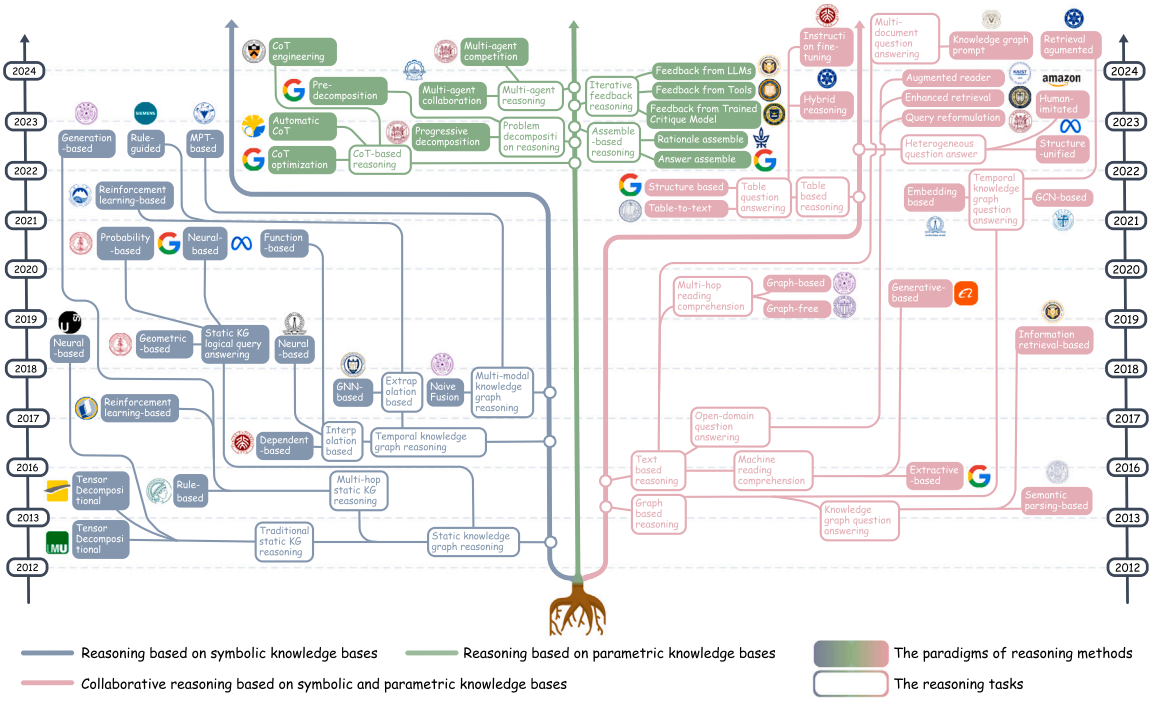

Fig. 1. The development timeline of reasoning paradigms. Abbreviation: knowledge graph (KG), multi-modal pre-trained transformer (MPT), chain-of-thought (CoT), and graph convolutional network (GCN).

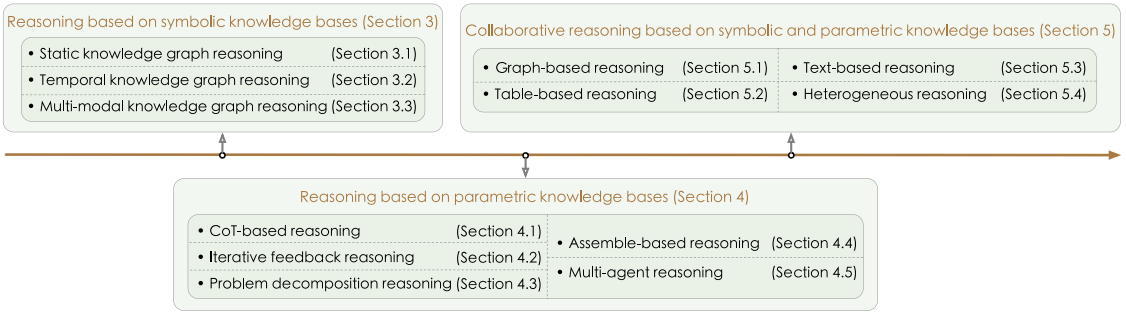

Fig. 2. Overall framework of reasoning-related methods.

Organization of this survey: we first introduce the background in Section 2. Then, we systematically introduce different reasoning tasks in Sections 3–5, each of which encompasses multiple types of methods. The overall Organization is shown in Fig. 2. At each method section, we illustrate the working principles and mechanisms of this type of methods at the beginning. In the middle part, we analyze some representative methods according to the development history. At the end, we discuss the advantages and disadvantages of this type of methods. Additionally, at the end of each task section, there is a dataset subsection and a method comparison subsection. The former summarizes commonly used datasets in the corresponding task and provides statistics on their features across different dimensions. It also analyzes how these datasets might influence the development of reasoning methods. In the latter, we compare the performance, efficiency, interpretability, and robustness of different types of reasoning methods. Based on the review, analysis, and summary of existing reasoning methods, we suggest several promising directions for future research in Section 6. Finally, we conclude this paper in Section 7.

2. Background

In this section, we first introduce the concepts of symbolic and parametric knowledge bases. Then, we introduce the taxonomy of reasoning in detail.

In artificial intelligence, symbolic and parametric knowledge bases represent different paradigms of knowledge representation. These paradigms align with symbolism (Simon & Newell, 1976) and connectionism (Rumelhart et al., 1986), respectively. The

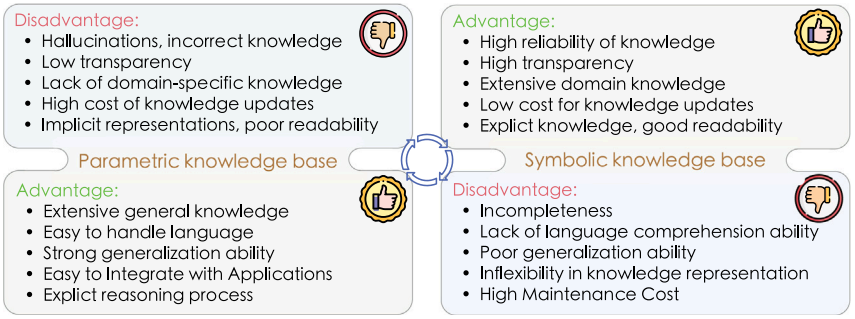

Fig. 3. The comparison between parametric and symbolic knowledge bases.

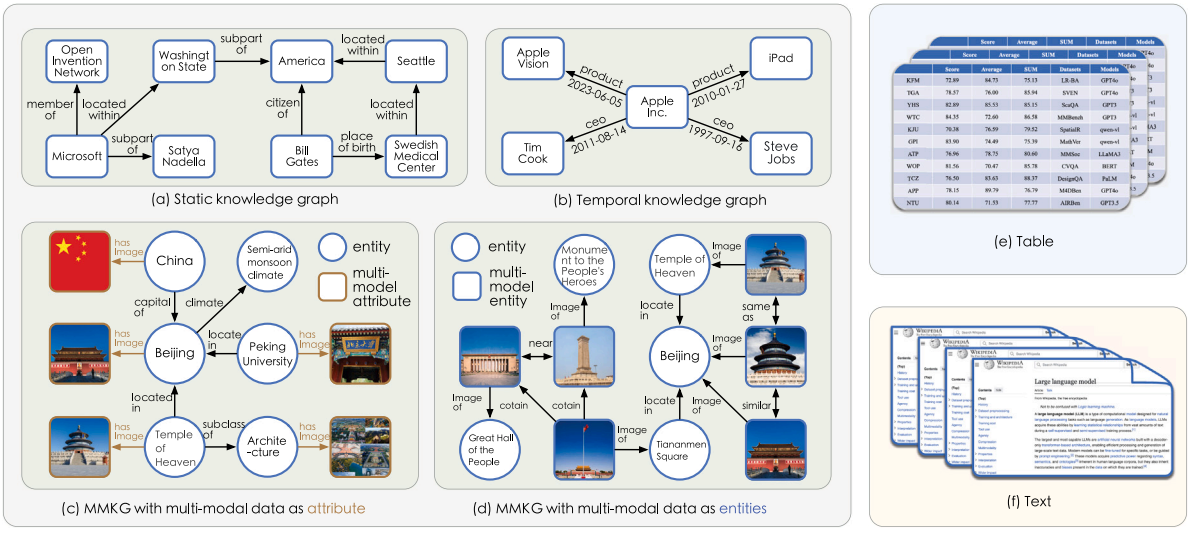

Fig. 4. Different symbolic knowledge bases. MMKG: multi-modal knowledge graph.

symbolic knowledge base involves explicit knowledge and logical structures for reasoning. It is fundamental to symbolic AI, which focuses on rule-based manipulation of symbols (Davis et al., 1993). In contrast, the parametric knowledge base captures knowledge implicitly through learned parameters, emphasizing adaptability and pattern recognition (LeCun et al., 2015). As depicted in Fig. 3, both types of knowledge bases have their own advantages and disadvantages, and in certain aspects, they can complement each other well (Cheng et al., 2024; Pan et al., 2024; Zhang et al., 2024).

2.1. Symbolic knowledge bases

Symbolic knowledge bases include KGs, tables, and text. Among these, KGs and tables are structured, while text is unstructured. The structured KGs can be partitioned into static knowledge graphs (SKGs), temporal knowledge graphs (TKGs), and multi-modal knowledge graphs (MMKGs). The following part of this subsection introduces the details of these symbolic knowledge bases.

Static knowledge graph: As shown in Fig. 4(a), an SKG is a structured semantic knowledge base that can express various associations between entities in a graphical manner (Liang et al., 2024). An SKG contains many factual triplets (h, r, t), where h, r, and t represent the head entity, the relation, and the tail entity, respectively.

For example, (*Microsoft*, *located within*, *Washington State*) represents "Microsoft is located within Washington State". An SKG is represented as ($\mathcal{E}$, $\mathcal{R}$, $\mathcal{F}$), where $\mathcal{E}$, $\mathcal{R}$, and $\mathcal{F}$ denote the entity set, the relation set, and the triplet set $\{(h, r, t)\} \subseteq \mathcal{E} \times \mathcal{R} \times \mathcal{E}$, respectively.

Temporal knowledge graph: As shown in Fig. 4(b), TKGs can store temporal information such as events that evolve over time. The quadruples (h, r, t, *time*) are the basic unit of TKGs, where h, r, t, and *time* represent the head entity, the relation, the tail entity, and the timestamp, respectively.

For example, the quadruple (*Apple Inc.*, *product*, *iPad*, *2010–01–27*) means "On January 27, 2012, Apple Inc. produced iPad". A TKG is represented as ($\mathcal{E}$, $\mathcal{R}$, $\mathcal{T}$, $\mathcal{F}$), where $\mathcal{E}$, $\mathcal{R}$, $\mathcal{T}$, $\mathcal{F}$ denote the entity set, relation set, the timestamp set, and the quadruple set $\{(h, r, t, time)\} \subseteq \mathcal{E} \times \mathcal{R} \times \mathcal{E} \times \mathcal{T}$, respectively. Another representation way of TKGs is (G_{t1}, G_{t2}, $\cdots$, G_{tn}), where G_{ti} denotes the SKG containing all triplets happened in timestamp t_i.

Multi-modal knowledge graph: The MMKG integrates various multi-modal data into one SKG, such as text and images. The MMKGs are generally divided into two types.

As illustrated in Fig. 4(c), the first type of MMKG models multi-modal data as attributes attached to existing entities in the KG. Here, modalities are not separate nodes. Instead, they are embedded as attribute values (e.g., image URLs or audio clips) directly within entity nodes. This design focuses on enriching entities with multi-modal features, streamlining the structure and facilitating attribute-centric queries.

As illustrated in Fig. 4(d), the second type of MMKG represents multi-modal data as distinct entities within the KG. These modality-specific objects are directly connected to relevant real-world entities via relational edges. For example, an image related to a person entity would exist as a separate node, linked via an "hasImage" relation. This method treats each piece of multi-modal data as an independent entity, enabling fine-grained representation and reasoning over the modalities themselves. Notably, MMKGs with multi-modal data as attribute are more prevalent in current research and applications within the semantic web community due to their accessibility and similarity to SKGs.

Structured table: As shown in Fig. 4(e), a structured table contains n records and m attributes. Each record can be represented as a vector $\mathbf{r}_i = (a_{i1}, a_{i2}, \dots, a_{im})$, where a_{ij} denotes the value of the jth attribute in the ith record. The entire table can be viewed as a set of n records $\{\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_n\}$. The format of table makes data easy to store, manage, and analyze (Vanoirbeek, 1992). Hence, it is widely used in various fields such as financial statements, scientific research data, and inventory management (Dagdelen et al., 2024; Sui et al., 2024).

Unstructured text: As shown in Fig. 4(f), unstructured text primarily refers the text information without a standardized format or organization, such as books and articles. Unstructured text can also store knowledge and be widely used in various applications with the help of current large language models.

Heterogeneous Knowledge base: This type of knowledge base can simultaneously contain multiple symbolic knowledge bases with different structures, such as knowledge graphs, tables, and text.

2.2. Parametric knowledge bases

In recent natural language processing (NLP) literature (Y. Lee et al., 2024; Lee et al., 2022; Namburi et al., 2023; Saito et al., 2025; Wang et al., 2024), the parametric knowledge is frequently used to describe the internalized knowledge captured by pre-trained language models (PLMs) through their large number of parameters. Although in a broader scope, the parametric knowledge base could also encompass parametric rules, parametric constraints, or mathematical models. However, in modern NLP literature, defining them as knowledge bases is relatively rare, as they typically store very limited knowledge. In practice, knowledge bases are generally expected to store a substantial amount of information in either implicit or explicit form, supporting a wide range of reasoning tasks. Hence, in this survey, parametric knowledge bases mainly refer to PLMs. PLMs are pre-trained on large-scale text corpora via self-supervised learning and store abundant knowledge in parameters, enabling them to reason with implicit knowledge in parameters (Talmor et al., 2020a). Based on architecture differences, current PLMs can be divided into encoder-only, decoder-only, and encoder–decoder.

Encoder-only PLMs: The representative encoder-only PLMs mainly include BERT (Devlin et al., 2019) and its variants (Clark et al., 2020; He et al., 2021). Different pre-training strategies are designed to incorporate pre-training knowledge into their parameters. For instance, BERT is pre-trained through the application of masked language modeling. ELECTRA (Clark et al., 2020) introduces the replaced token detection pre-training task, which enhances the learning efficiency and performance of the model. To effectively utilize parametric knowledge in encoder-only PLMs, a direct approach is to extract semantic representations from the input text. This method is widely used in reasoning tasks, such as open-domain question answering (Borgeaud et al., 2022; Yamada et al., 2021), to incorporate pre-trained knowledge.

Decoder-only PLMs: decoder-only PLMs leverage an autoregressive mechanism to generate coherent text by sequentially predicting the next token based on preceding context. Representative decoder-only PLMs encompass GPT (Achiam et al., 2023), LLaMA (Touvron et al., 2023), and Qwen (Bai et al., 2023). Most PLMs with large parameter scales (also known as large language models, LLMs) use a decoder-only architecture. During reasoning, we can elicit knowledge from LLMs through prompting methods such as chain-of-thought (Wei et al., 2022). Conversely, reasoning with Small Language Models (SLMs) often requires task-specific fine-tuning to effectively activate their parametric knowledge (Hendrycks et al., 2021; Lewis et al., 2020).

Encoder–decoder PLMs: Encoder–decoder PLMs, such as T5 (Raffel et al., 2020) and BART (Lewis et al., 2020), decoupling text understanding and generation into separate vector spaces. While this architecture facilitates knowledge reasoning by leveraging encoder-derived semantic features alongside decoder-generated targets (Izacard & Grave, 2021a), it suffers from high architectural complexity and intensive pre-training costs.

Both symbolic and parametric knowledge bases face significant scalability challenges as the size and complexity of stored knowledge increase. For symbolic knowledge bases, the main difficulty lies in handling the explosive growth of entities, relations, records, and texts. As symbolic knowledge expands, reasoning and query operations become computationally expensive. Furthermore, keeping the symbolic information consistent or up to date becomes increasingly difficult. Integrating diverse data sources, such as multi-modal and temporal KGs, also introduces schema alignment and ontology management challenges. For parametric knowledge bases, scalability problems arise mainly from the computational cost of training and maintaining large language models. Although increasing model parameters improves knowledge capacity, it also leads to higher memory and inference costs. Updating large language models with new knowledge remains inefficient. In short, symbolic knowledge bases struggle with the scalability of explicit reasoning, while parametric ones encounter limits in implicit storage and efficient Updating.

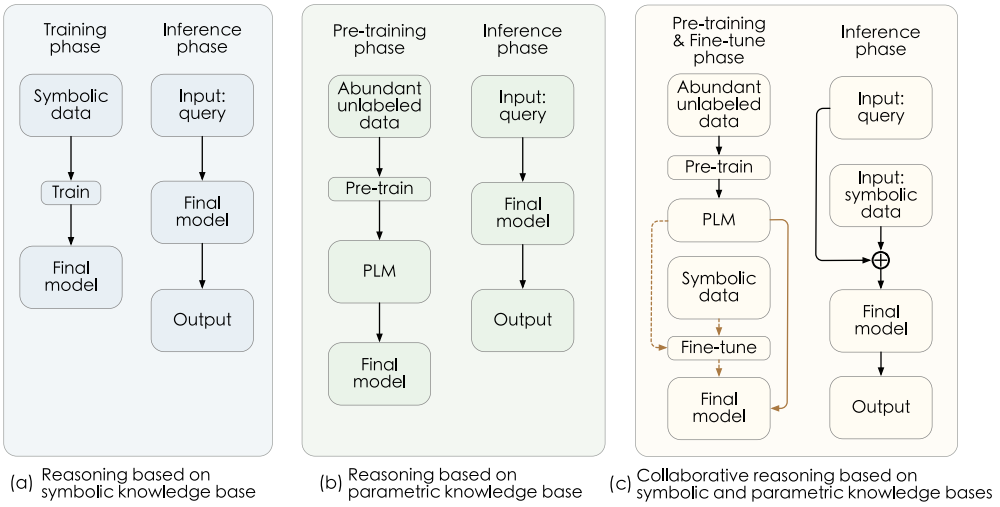

Fig. 5. Reasoning methods based on different knowledge bases. The contents of parts (a), (b), and (c) correspond to Section 3, Section 4, and Section 5, respectively. Abbreviation: pre-trained language model (PLM).

2.3. Taxonomy of reasoning

Reasoning methods are categorized into symbolic, parametric, and collaborative paradigms based on their underlying knowledge bases. Symbolic reasoning relies on explicit structures like knowledge graphs to ensure traceability within the logic (Yu, Yang, et al., 2023). Conversely, parametric reasoning leverages implicit knowledge distributed within large language models to enable flexible inference (H. Yu et al., 2024), but this approach often faces challenges regarding verifiability and outdated information. Collaborative reasoning addresses these limitations by integrating structured symbolic bases with learned parametric representations. Frameworks such as retrieval-augmented generation (Lewis et al., 2020) and neuro-symbolic architectures (DeLong et al., 2025) exemplify this bidirectional interaction. They provide a unified lens for understanding the evolving landscape of reasoning systems.

As shown in Fig. 5(a), reasoning methods based on symbolic knowledge bases train a model to learn the knowledge in the symbolic knowledge bases during the training phase. During inference, only a query is input to the model. The model's reasoning capability comes from the knowledge modeled in the training phase. Typical tasks for this type of method, such as static knowledge graph reasoning, require reasoning for an unknown element in an incomplete triplet or quadruple. Bordes et al. (2013), Jiang et al. (2016)

As shown in Fig. 5(b), reasoning methods based on parametric knowledge bases leverage the knowledge in PLMs' parameters. During inference, only a query is input. The model's reasoning capability comes from the parametric knowledge modeled during pre-training. Compared to the training phase in Fig. 5(a), the pre-training phase is usually task-independent. Typical tasks for this type of method, such as mathematical reasoning, require reasoning for a natural language answer given a knowledge-intensive question (Liang et al., 2024; Madaan et al., 2023; Nye et al., 2021; Wang et al., 2023; Wei et al., 2022).

As shown in Fig. 5(c), collaborative reasoning methods based on symbolic and parametric knowledge bases also model parametric knowledge during the pre-training phase. Some methods further fine-tune the PLMs using knowledge from symbolic knowledge bases to enhance the domain knowledge learning. During inference, these methods retrieve relevant knowledge from symbolic knowledge bases. Then, they integrate this information with the parametric knowledge stored in PLMs to enhance reasoning performance. Similar to reasoning methods based on parametric knowledge bases, this type of method, such as knowledge graph questions answering, also performs reasoning in the form of natural language question answering (Cao et al., 2022; Jiang et al., 2024; Lehmann et al., 2024; Zheng et al., 2023).

We add two brief examples to show how this taxonomy leads to different system designs. First, when a system is intended for high-stakes domains such as medical or legal reasoning, the source information transparency is critical. Our taxonomy suggests prioritizing symbolic knowledge bases. This choice ensures that each reasoning step is clearly grounded in verifiable and human-readable facts. In contrast, some reasoning applications emphasize interpretive flexibility, such as creative brainstorming. In these cases, natural language explanations and flexible reasoning processes are often more important than strict answer optimality. The taxonomy therefore points designers toward parametric knowledge bases. Parametric knowledge allows the system to produce more fluid and human-like explanations. It can also capture complex and latent relationships that are difficult to formalize within rigid symbolic structures. This design better supports interpretive flexibility.

In this survey, we introduce, analyze, and summarize each type of reasoning method according to its development history and the specific content it encompasses. Reasoning methods based on symbolic knowledge bases have a longer history than the other two types. Consequently, they also encompass a broader range of research branches. As a result, the corresponding section contains more detailed material and includes a greater amount of content.

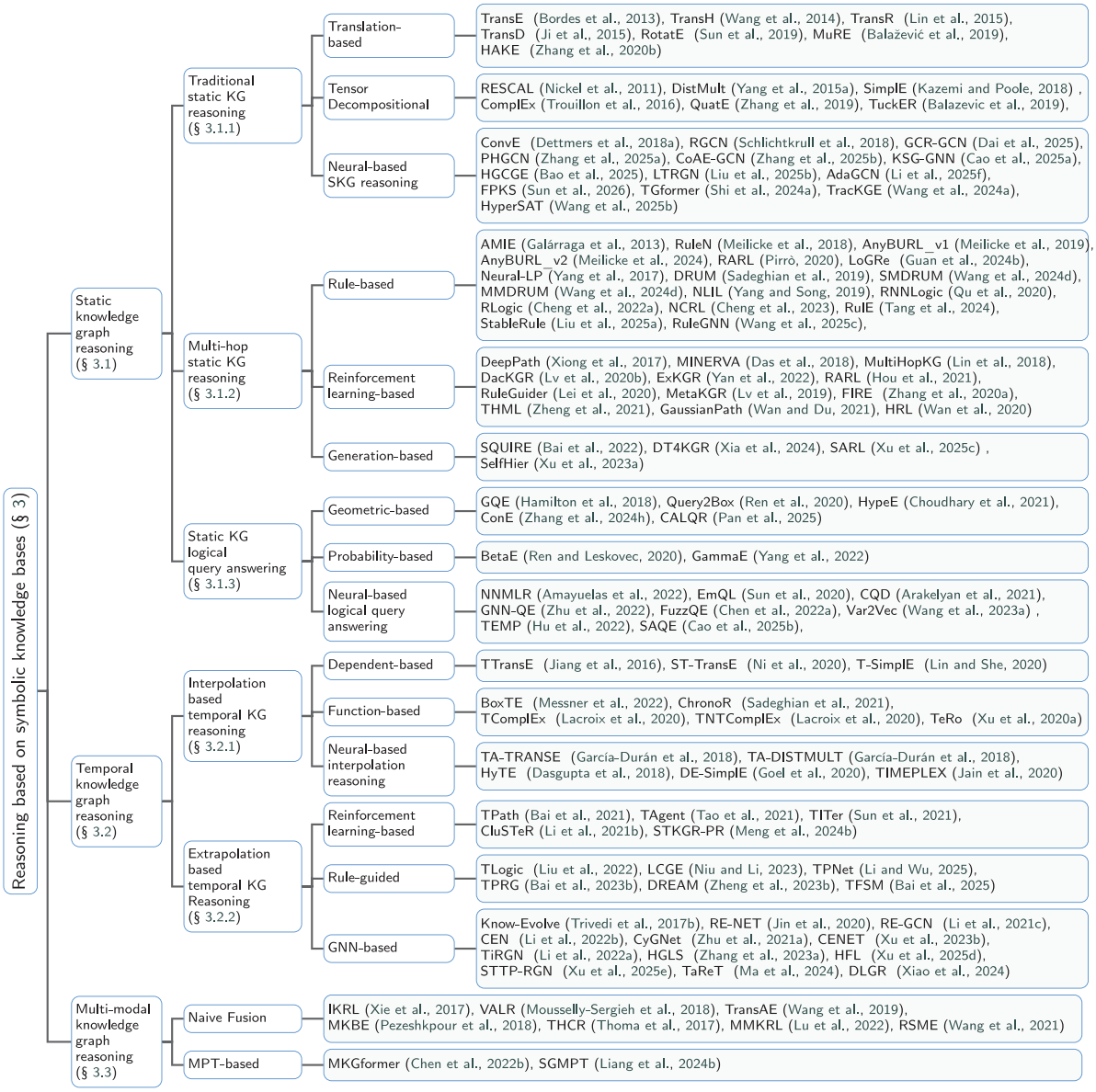

Fig. 6. Taxonomy of symbolic reasoning across static, temporal, and multi-modal KG reasoning. Representative methods for each task or sub-task are shown in the green box. Abbreviation: Multi-modal Pre-trained Transformer (MPT). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3. Reasoning based on symbolic knowledge bases

In this section, we investigate the reasoning methods based on symbolic knowledge bases. Based on the structural types of KGs, we investigate static knowledge graph reasoning, temporal knowledge graph reasoning, and multi-modal knowledge graph reasoning. The overall taxonomy of reasoning methods based on symbolic knowledge bases is shown in Fig. 6. Furthermore, Fig. 7 depicts the core technology and pipeline of each type of reasoning method based on symbolic knowledge bases.

3.1. Static knowledge graph reasoning

Static knowledge graph reasoning refers to the completion of incomplete triplets (incomplete knowledge) based on the given fact triplets (existing knowledge) in the SKG, thereby obtaining new complete factual triplets (new knowledge). According to the differences in query form and output form, static knowledge graph reasoning tasks can be divided into three sub-tasks: traditional

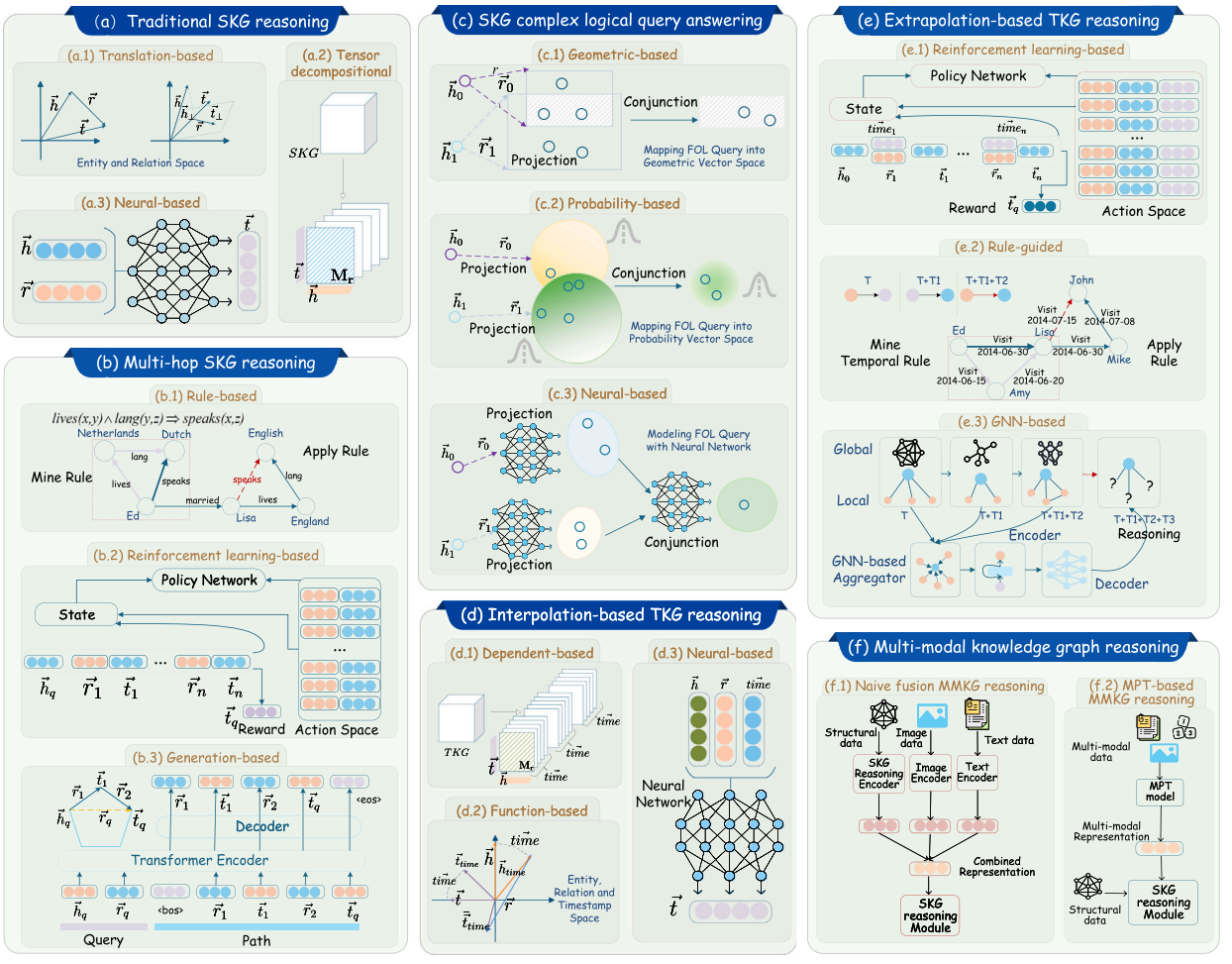

Fig. 7. Symbolic reasoning pipelines for Static (SKG), Temporal (TKG), and Multi-modal (MMKG) knowledge graphs reasoning tasks. SKG reasoning tasks including (a) Traditional (Section 3.1.1), (b) Multi-hop (Section 3.1.2), and (c) Complex logical query answering (Section 3.1.3); TKG tasks covering (d) Interpolation (Section 3.2.2) and (e) Extrapolation (Section 3.2.2); and (f) MMKG reasoning (Section 3.3). Abbreviations: first-order logical (FOL), multi-modal pre-trained transformer (MPT).

SKG reasoning, multi-hop SKG reasoning, and SKG complex logical query answering. In general, multi-hop SKG reasoning requires providing an explicit reasoning path while completing the triplets. SKG complex logical query answering requires modeling the complex logical symbols.

3.1.1. Traditional SKG reasoning

Task definition: Given the query $(h, r, ?)$, $(?, r, t)$, or $(h, ?, t)$, traditional SKG reasoning methods aim to predict the missing entity or relation directly. Generally, the traditional SKG reasoning methods first learn the representations of relations and entities in SKGs. Then, it constructs a scoring function to calculate the validity of possible triplets. According to the representation approaches, traditional SKG reasoning methods can be categorized into three types: translation-based, tensor decompositional, and neural-based (Ali et al., 2022; Liang et al., 2024), as illustrated in Fig. 7(a).

(1) Translation-based methods: As shown in Fig. 7 (a.1), translation-based methods first learn low-dimensional embeddings for entities and relations in a continuous vector space. Then, they compute the distance of these embeddings to reason about the final answer entity. The structural commonalities of translation-based methods lie in their unified framework. In this framework, relations function as geometric transformations that operate on head entity embeddings to approximate tail entity embeddings within vector spaces.

The foundational model, TransE (Bordes et al., 2013), regards the relation as a translation operation. Specifically, it posits that $h + r \approx t$, where h, r, and t reside in the same vector feature space. Although efficient, it struggles to accurately represent complex relations such as one-to-many, many-to-one, and symmetric relations. This fundamental shortcoming spurs a wave of innovations aimed at enhancing the model's expressiveness.

Building on TransE, several methods introduce additional parameters to allow entities and relations to exhibit different semantic properties. For instance, TransH (Wang et al., 2014) interprets relations as transformation operations on hyperplanes. This approach enables a single entity to have distinct representations when involved in different relations. As a result, it alleviates the one-to-many and many-to-one issues. Similarly, TransR (Lin et al., 2015) models h, r, and t in different vector spaces with the projection matrix M_r for each relation r. Building on TransR, TransD (Ji et al., 2015) further dynamically constructs mapping matrices for every entity-relation pair.

Furthermore, several methods transcend the limitations of traditional Euclidean geometry by exploring alternative vector spaces whose intrinsic properties are better suited for specific relational patterns. For example, to effectively model symmetric and antisymmetric relations, RotatE (Sun et al., 2019) considers relations as rotation operations in a complex space. This geometric operation naturally captures the desired properties: a relation and its inverse correspond to rotations in opposite directions. Similarly, MuRE (Balažević et al., 2019) utilizes the Poincaré ball in the hyperbolic space to model the tree-like structures of relations. HAKE (Zhang, Cai , et al., 2020) employs a polar coordinate system to explicitly model semantic hierarchies, using the modulus to represent hierarchical level and the phase to model semantic similarity.

Overall, the primary advantage of translation-based methods is their efficiency and flexibility in modeling relations as translation operations from head to tail entities, enabling the handling of large-scale SKGs through straightforward mathematical operations. However, these methods come with disadvantages: the simplicity of the translation assumption limits their ability to model more intricate relations without complex extensions, reducing their effectiveness on SKGs with diverse or multi-faceted relations.

(2) Tensor decompositional methods: As depicted in Fig. 7 (a.2), tensor decompositional methods first represent the atoms in an SKG with latent embeddings and model their high-order interactions through tensor decomposition. Then, they score candidate tail entities based on the decomposed tensor representations to reason about the final answer. The structural commonalities of tensor decompositional methods center on their use of matrix operations to capture pairwise interactions between entity embeddings, where an SKG is represented as a three-dimensional tensor with each relation forming a distinct matrix slice M_r.

The foundational model, RESCAL (Nickel et al., 2011), utilizes a bilinear score function $f_r(h,t) = \mathbf{h}^\top \mathbf{M}_r \mathbf{t}$ to measure the trustworthiness of the triple, where h and t are vector representations of the entities and M_r is a matrix associated with the relation. Although it can deeply capture all pairwise interactions between the latent features of head and tail entities, it comes at a significant cost: the matrix operations lead to high computational complexity.

Building on RESCAL, several methods aim at improving computational efficiency by simplifying the interaction model. For example, DistMult (B. Yang et al., 2015) constrains the relation matrices to be diagonal, effectively reducing the bilinear form to a simple trilinear dot product. Similarly, SimplE (Kazemi & Poole, 2018) cleverly represents each relation as two distinct vectors (for forward and backward directions) and averages their scores, thus enabling the modeling of asymmetric relations while maintaining simplicity and efficiency.

Furthermore, several methods aim at enhancing expressive power to model asymmetric relations. This is achieved by moving beyond the real number space into more expressive algebraic domains. For example, ComplEx (Trouillon et al., 2016) introduces complex-valued embeddings to naturally and effectively model asymmetric relations. Similarly, QuatE (Zhang et al., 2019) extends the embedding space to quaternions, where the Quaternion embeddings are capable of capturing more intricate latent feature interactions, leading to improved performance on complex relational patterns. Additionally, TuckER (Balazevic et al., 2019) employs the more expressive Tucker decomposition, which can capture a wide range of relational patterns.

Overall, the primary advantage of tensor decompositional methods is their natural suitability for modeling complex relations. However, these methods come with disadvantages: they involve significantly more complex computations than translation-based methods, making them unsuitable for large-scale SKGs, and thereby limiting their practical applicability in real-world scenarios.

(3) Neural-based SKG reasoning methods: While translation-based and tensor decompositional methods have laid a strong foundation for embedding SKGs, they typically require storing a distinct vector or matrix for every entity and relation, leading to a parameter count that grows linearly or even quadratically with the size of the SKG. To overcome this limitation, neural-based SKG reasoning methods have emerged as a powerful alternative. As presented in Fig. 7 (a.3), this approach leverages the non-linear modeling capabilities of neural networks to learn the complex correlations between head entities, relations, and tail entities. After training, the neural networks reason about the final tail entities by scoring the candidate tail entities given a head entity and relation. The structural commonalities of this type of method center on their ability to capture complex, non-linear interactions through specialized neural network architectures.

Some approaches utilize classical neural network architectures, such as convolutional neural networks (Gu et al., 2018), to model entities and relations. For instance, ConvE (Dettmers et al., 2018a) utilizes 2D convolutional layers over reshaped entity and relation embeddings to extract complex, non-linear features. By capturing cross-dimensional correlations missed by linear scoring functions, ConvE achieves high expressiveness with fewer parameters through convolutional weight sharing. While ConvE takes advantage of the great nonlinearity fitting ability of neural networks, it overlooks translational properties. Hence, ParamE (Che et al., 2020) integrates translational properties with neural network modeling. It treats relations as neural network parameters that map head entity embeddings to tail entity embeddings, thereby improving expressiveness for reasoning. To address the limited interactions captured by ConvE, InteractE (Vashishth et al., 2020) is proposed, incorporating feature permutation, novel feature reshaping, and circular convolution to enhance interaction capacity.

Recently, some studies use graph neural networks (GNNs) or graph convolutional networks (GCNs) (Zhou et al., 2020) to effectively leverage the graph structure of the SKG itself. In these methods, the representation of an entity is informed by its neighborhood for reasoning. For example, RGCN (Schlichtkrull et al., 2018) operates by having each entity (node) aggregate feature

information from its immediate neighbors. By stacking multiple layers, RGCN enables entities to incorporate information from multi-hop neighbors, effectively capturing the rich topological structure of the graph and leading to significant performance improvements. To address the limitations of RGCN in capturing wide-range relational information, GCR-GCN (Dai et al., 2025) integrates Formal Concept Analysis to redistribute relational parameter weights based on the similarity of granular concepts. It simulates human-like data conceptualization to enhance both interpretability and computational efficiency in reasoning. PHGCN (Zhang et al., 2025) addresses over-smoothing and the neglect of high-order structures in KG embedding by integrating pair-wise and simplicial complex features. This self-adaptive model achieves state-of-the-art performance, demonstrating the critical value of high-order topological information in knowledge graph completion. CoAE-GCN (D. Zhang et al., 2025) enhances knowledge graph reasoning by integrating entity-relation co-occurrence, attention-driven local structural features, and approximate entailment. This framework effectively bridges semantic correlation gaps and improves representational capacity, delivering superior performance. KSG-GNN (Cao et al., 2025) transforms knowledge graphs into similarity-based homogeneous structures, allowing standard GNNs to be applied directly through the principle of relational homophily. This framework simplifies embedding learning and consistently outperforms strong baselines. Existing GCNs for knowledge graph embedding face representation over-smoothing in deep layers and high structural distortion when modeling hierarchical graph data. Hence, HGCGE (Bao et al., 2025) utilizes hyperbolic convolutions to capture deep non-Euclidean relationships, achieving superior performance even in low-dimensional settings. To address the limitations of GCNs in capturing long-range dependencies within reasoning, LTRGN (Liu et al., 2025) integrates linear self-attention with multi-relational graph networks to efficiently encode global node interactions. AdaGCN (Li et al., 2025) addresses the limitations of indiscriminate neighbor aggregation in GCNs by introducing an adaptive message-passing mechanism that selectively weights or separates neighbor features during knowledge graph reasoning. FPKS (Sun et al., 2026) secures decentralized KGs by integrating federated learning with GCN-based bidirectional aggregation. This privacy-preserving framework effectively overcomes structural fragmentation to deliver high-quality, distributed knowledge representations for reasoning.

In addition, there is a small body of work that employs transformer networks (Vaswani et al., 2017) to boost KGR performance. GNN-based or GCN-based methods consider graph structure information but ignore the contextual information of nodes in the knowledge graph, making them unable to discern valuable entity and relation information. In response to this limitation, TGformer (F. Shi et al., 2024) use a graph transformer to build knowledge embeddings with triplet-level and graph-level structural features in the KGs. TracKGE (M. Wang et al., 2024) enhances reasoning by combining sequence-based transformers with adaptive contrastive learning to capture complex relations. By addressing data incompleteness through mask nodes, it consistently outperforms state-of-the-art baselines in reasoning. HyperSAT (J. Wang et al., 2025) integrates subgraph sampling and heterogeneous attention into a transformer framework to capture the complex structural and directional information of KGs, which delivers superior performance.

Overall, the primary advantage of neural-based SKG reasoning methods is their ability to utilize neural network architectures to effectively model the similarity between entities and relations, as neural networks excel at capturing semantic features and can store representations of large-scale SKGs with fewer parameters. However, these methods come with disadvantages: the introduction of neural networks significantly reduces model interpretability, making it difficult to understand the reasoning process behind predictions.

3.1.2. Multi-hop SKG reasoning

Task definition: Given a query $(h, r, ?)$, multi-hop SKG reasoning methods aim to predict the target tail entity t through an n-hop reasoning path $\tau : h \xrightarrow{r_1} e_1 \xrightarrow{r_2} e_2 \cdots \xrightarrow{r_n} e_n$, where e_i and r_i represent the entity and the relation in the path τ. The last entity e_n in τ is treated as the predicted target tail entity t. Compared to traditional SKG reasoning methods, multi-hop SKG reasoning methods provide better interpretability by providing explicit reasoning paths. The multi-hop SKG reasoning methods can be categorized into rule-based, reinforcement learning-based (RL-based), and generation-based, as demonstrated in Fig. 7(b).

(1) Rule-based methods: As indicated in Fig. 7 (b.1), rule-based methods first mine logical rules from SKGs. Then, this type of method reason about missing entities or relations by matching queries to the rules. The structural commonalities of rule-based methods center on their symbolic approach to multi-hop SKG reasoning, where they consistently focus on automatically discovering symbolic logical patterns expressed as horn clauses. These methods operate directly on the symbolic structure of the SKG, with a core philosophy of identifying human-understandable relationships that govern entity connections. For instance, rules like $mother\_of(m, c) \wedge married\_to(m, f) \Rightarrow father\_of(f, c)$ demonstrate how facts can be inferred from other facts, where the logical formula $r(s, o)$ represents the triple (s, r, o).

Early rule-based methods like AMIE (Galárraga et al., 2013) laid the groundwork by introducing association rule mining techniques to SKGs. It emphasizes learning highly reliable rules from SKGs quickly and effectively through a novel rule confidence measure and automated rule learning methods. However, AMIE is frequently limited by its inefficiency, and it often suffers from low coverage. To address these issues, RuleN (Meilicke et al., 2018) extends AMIE's rule types to learn rules of the form $\exists r(x, y) \Rightarrow r(x, a)$, where r is a relation, x and y are entity variables, and a is a constant entity. Similarly, AnyBURL_v1 (Meilicke et al., 2019) not only supports a wider range of rule types but also uses a bottom-up approach for rule generalization from ground triplets. It even introduces a symbolic reinforcement learning framework to guide the rule mining process, making rule mining more effective and efficient. AnyBURL_v2 (Meilicke et al., 2024) introduces four key extensions to AnyBURL_v1 designed to improve its scalability and performance on larger KGs. These enhancements include the integration of reinforcement learning, inequality constraints, and a multithreaded architecture to optimize rule discovery and confidence estimation efficiency. RARL (Pirrò, 2020) discovers logical rules by focusing on the semantic relationships between concepts within KGs. By separating rule generation from quality checking, it provides a more flexible and efficient way to process data. Recently, LoGRe (S. Guan et al., 2024) has analyzed sparse SKGs to

automatically construct and aggregate relation-path rules for reasoning. This approach eliminates two major drawbacks: the need for manually crafted rules and the low coverage typically associated with generated rules.

Some rule-based methods extend the above association rule mining by learning logic rules and their corresponding weights in a differentiable manner. To achieve this, most methods leverage neural logic programming, which extends the rule mining from counting to learning. For instance, Neural-LP (Yang et al., 2017) performs soft counting of rule instances through sequences of differentiable tensor multiplications. It employs an attention-based neural controller to learn scores for specific logical rules. A limitation of Neural-LP is that it may assign a disproportionately high score to a meaningless rule if it shares an atom with a useful one. To mitigate this, DRUM (Sadeghian et al., 2019) leverages bidirectional RNNs to prune potentially incorrect rule bodies. To address the lack of formal guarantees in rules extracted by DRUM, which may be unsound or incomplete, SMDRUM and MMDRUM (X. Wang et al., 2024) formalize the rule extraction process within an extended Datalog framework. By integrating specific model constraints into the learning objective, they ensure the soundness and completeness of the derived rules while maintaining competitive performance. Furthermore, methods such as Neural-LP are limited to chain-like rules and query-specific relational paths. To address these limitations, NLIL (Yang & Song, 2019) learns globally consistent logical rules through a divide-and-conquer search strategy, enabling scalable rule learning and interpretable explanations. Since neural logic programming requires heavy matrix computations with the joint optimization of logical rules and weights, the training process is computationally demanding. As a result, these methods are computationally inefficient when handling large KGs.

To address the efficiency issue, RNNLogic (Qu et al., 2020) decouples rule generation from rule weight learning by introducing a rule generator and a reasoning predictor, respectively. While RNNLogic relies on observed rule instances to define the scoring function for rule evaluation, it suffers from poor generalization. To overcome this, RLogic (K. Cheng et al., 2022) introduces a scoring model based on predicate representation learning. In addition, RLogic incorporates the deductive nature of logical rules into rule learning, which is particularly important when supporting evidence is limited. It further strengthens deduction by recursively decomposing a large sequential model into smaller atomic models. Recently, NCRL (Cheng et al., 2023) is proposed as an end-to-end neural model for learning compositional logical rules. It identifies the optimal compositional structure of a rule body and decomposes it into smaller components to infer the rule head. By iteratively merging these components through a recurrent attention unit, NCRL ultimately predicts a single rule head.

More recently, RulE (Tang et al., 2024) learns embeddings for logical rules and jointly represents entities, relations, and rules in a unified space. Based on these rule embeddings, RulE enables soft rule inference and enhances reasoning through rule-aware regularization. To address the limitations of rule learning methods under distribution shifts, StableRule (Liu et al., 2025) is proposed for out-of-distribution KG reasoning. It integrates feature decorrelation with a rule learning network to mitigate covariate shifts and improve the robustness and generalization of logical rule learning. Rules extracted from explainable GNNs often lack statistical plausibility and fail to faithfully align with the model's actual reasoning. To resolve this, RuleGNN (Z. Wang et al., 2025) establishes a direct correspondence between model parameters and symbolic rules, enabling rule-guided training and the efficient extraction of high-quality, sound explanations.

Overall, the primary advantage of rule-based methods is their strong interpretability and generalization capabilities, allowing for transparent reasoning processes that can be easily understood and verified by human experts. However, these methods also have some disadvantages, as the rule mining process is often constrained by strict formal requirements, making it difficult to adapt to diverse and evolving knowledge patterns. Specifically, because these methods do not learn latent representations of entities and relations, they often suffer from limited coverage and suboptimal performance.

(2) Reinforcement learning-based methods As displayed in Fig. 7 (b.2), reinforcement learning (RL)-based methods model multi-hop reasoning as a Markov Decision Process (MDP) (Puterman, 1990). They first train a policy network using reward signals. After training, the policy network explores multi-hop paths to reason about the final answer. The structural commonalities of RL-based methods center on their utilization of a reinforcement learning framework, where agents systematically navigate through SKGs to discover relational paths. These methods have evolved from basic path-finding frameworks to more sophisticated approaches that address key challenges such as sparse rewards and few-shot scenarios.

DeepPath (Xiong et al., 2017) establishes a foundational framework by adopting reinforcement learning to search for reasoning paths and identify target relations between given head and tail entities. Similarly, MINERVA (Das et al., 2018) addresses a more complex and practical scenario: identifying target tail entities through reinforcement learning given specific relations and head entities. Following MINERVA, most RL-based methods are devoted to tackling the sparse rewards problem and to designing a more efficient policy network in incomplete SKGs. For instance, MultiHopKG (Lin et al., 2018) is one of the first to use traditional SKG reasoning models to estimate the rewards of unobserved target entities, thereby reducing the impact of incorrect negative samples, which is similar to DacKGR (Lv et al., 2020), ExKGR (Yan et al., 2022), RARL (Hou et al., 2021), and RuleGuider (Lei et al., 2020).

Moreover, several RL-based methods have integrated meta-learning to address the decreased reasoning capability of reinforcement learning in few-shot relation scenarios. For instance, Meta-KGR (Lv et al., 2019) is the first to apply meta-learning to multi-hop KG reasoning, using the meta-learning algorithm MAML (Finn et al., 2017) to learn meta-parameters from high-frequency relations, which can then be quickly adapted to sparse relations. FIRE (C. Zhang et al., 2020) uses heterogeneous neighbor information to enhance entity embeddings, thereby reducing the search space. THML (Zheng et al., 2021) proposes a difficulty-aware meta-reinforcement learning method, greatly enhancing the generalization capability.

Recently, there have been some quite creative methods. For example, GaussianPath (Wan & Du, 2021) proposes a Bayesian multi-hop reasoning paradigm to capture the uncertainty of reasoning paths and thus explore a broader range of them. Additionally, HRL (Wan et al., 2020) proposes a high-level policy to learn historical information and a low-level policy to recognize relation clusters and efficiently express the multiple semantics of relations.

Overall, the primary advantage of RL-based methods is their ability to learn representations of entities and relations while simultaneously acquiring path elements, which gives them stronger modeling capabilities compared to rule-based methods. However, these methods have notable disadvantages: they are particularly susceptible to the sparsity, incompleteness, and noise present in SKGs, as their effectiveness fundamentally depends on path quality.

(3) Generation-based methods: As described in Fig. 7 (b.3), generation-based methods adopt a generative framework to generate the reasoning paths step by step. This type of method first collects multi-hop reasoning data by rule-based methods. Then they train the encoder–decoder generation framework using the data. After learning the reasoning patterns from the data, the framework reasons about the paths and final answers in a generative manner. The structural commonalities of generation-based methods center on their encoder–decoder architecture that transforms multi-hop reasoning into a path generation task, representing a paradigm shift from explicit rule matching or reinforcement learning processes. These methods maintain a core focus on flexible reasoning while integrating various advanced techniques. By framing reasoning as generation, these methods implicitly learn complex inference patterns and adapt to diverse query structures without hand-crafted features or reward engineering.

For example, SQUIRE (Bai et al., 2022) employs an encoder–decoder model to directly translate a query into a reasoning path. Its key innovation lies in a rule-enhanced, iterative training strategy that allows it to effectively leverage the strengths of symbolic logic within a neural framework, achieving superior performance over both traditional rule-based baselines and RL-based baselines. DT4KGR (Xia et al., 2024) presents a similar idea but employs a pretrained SKG embedding model (e.g., TransE, ComplEx, or ConvE) as the encoder, in contrast to the Transformer used in SQUIRE. Similarly, SARL (K. Xu et al., 2025) introduces a structure-aware graph transformer for rule learning. SARL learns more generalizable and transferable rules, making it easily adaptable to other models. Additionally, cold-start multi-hop reasoning presents a specific challenge: the model often lacks precise guidance and explicit paths. To address this, SelfHier (M. Xu et al., 2023) proposes a hierarchical guidance mechanism coupled with a self-verification strategy. This approach significantly improves performance in low-resource scenarios.

Overall, the primary advantage of generation-based methods is their function as a powerful and stable approximation of the reinforcement learning process. In this process, the generative model implicitly learns a policy to maximize the likelihood of a correct path, enabling the methods to more fully capture the nuanced features of entities and relations. However, these methods come with disadvantages: the very nature of the end-to-end, black-box generative process significantly reduces model interpretability, making it difficult to understand or trace the reasoning behind specific path generation decisions.

3.1.3. SKG complex logical query answering

Task definition: In a KG, an atomic query can be represented by a single incomplete triple, such as $(?, r, t)$, $(h, ?r, t)$, and $(h, r, ?)$. Connecting multiple atomic queries with first-order logical (FOL) operators leads to the formulation of a FOL query (Nguyen et al., 2025) with complex logic. The FOL operators include conjunction ($\wedge$), disjunction ($\vee$), negation ($\neg$), existential quantifier ($\exists$), and universal quantifier ($\forall$). Given an incomplete FOL query, complex logical query answering (CLQA) methods aim to reason about a target entity set that satisfies the logical constraints of the query. For instance, the FOL query $q = \{C_? \mid \exists P : \text{assoc}(d_1, P) \wedge \text{assoc}(d_2, P) \wedge \text{target}(P, C_?)\}$ represents the complex logic question: "Identify potential drugs $C_?$ that can act on the proteins P associated with both diseases d_1 and d_2".

To efficiently process such complex queries, CLQA methods often represent them in standardized logical forms, such as conjunctive normal form (CNF) and disjunctive normal form (DNF). A query in DNF is structured as a disjunction of multiple conjunction queries. It can be mathematically formulated as $q_{DNF} = \bigvee_{i=1}^{n} \left(\bigwedge_{j=1}^{m_i} a_{ij} \right)$, where each a_{ij} represents an atomic query or its negation. DNF allows the reasoning system to decompose a complex query into several independent conjunctive sub-queries. On the other hand, CNF represents the query as a conjunction of disjunctions, formulated as $q_{CNF} = \bigwedge_{i=1}^{n} \left(\bigvee_{j=1}^{m_i} a_{ij} \right)$. This form is often used to enforce complex constraints and filter out entities that fail to meet any of the mandatory logical clauses.

In the field of CLQA, DNF is generally more widely adopted than CNF. This preference arises because the DNF structure naturally aligns with the path-based reasoning paradigm and geometric embedding operations. Specifically, most state-of-the-art embedding models, such as Query2Box (Ren et al., 2020) and BetaE (Ren & Leskovec, 2020), are optimized to handle conjunctive queries as intersections in a latent space. By transforming a complex logical query into DNF, the system can compute multiple conjunctive paths independently and then take their union as the final result. In contrast, CNF remains the standard for symbolic logic reasoning and constraint satisfaction. However, its "conjunction of disjunctions" structure is significantly harder to model geometrically. In continuous vector representations, performing union operations before intersections often leads to fragmented or less stable search spaces.

By transforming arbitrary FOL queries into these normal forms, embedding-based CLQA frameworks can map logical operators into embedding spaces, ensuring that the reasoning process remains both computationally tractable and logically consistent. Based on the type of embedding space, the CLQA methods can be categorized into geometric-based, probability-based, and neural-based, as illustrated in Fig. 7(c).

(1) Geometric-based methods: As shown in Fig. 7 (c.1), geometric-based methods project entities to points in a continuous vector space and represent complex logical queries as geometric regions (e.g., boxes, cones). The FOL operators are modeled as spatial operations on these regions. Finally, the answer set is retrieved by identifying the entity points that fall inside the resulting query region. The structural commonalities of geometric-based methods center on their unified paradigm of embedding entities and relations into geometric spaces. Within this space, they model logical operations as geometric transformations.

GQE (Hamilton et al., 2018) establishes a foundational framework by representing entities as points and modeling logical conjunctions via geometric intersections. While elegant, its reliance on single-point embeddings limits its ability to capture the inherent uncertainty and set-based nature of query answers.

To address this limitation, subsequent work has explored more sophisticated geometric primitives that can represent sets of entities in a more natural way. For example, Query2Box (Ren et al., 2020) shifts the paradigm from points to hyper-rectangles (or "boxes"). Query2Box represents a query as a box, which is defined by a center point and an offset vector. This geometric approach allows it to effectively model the set of candidate answers. Consequently, it offers a significant boost in expressiveness and robustness compared to point-based methods. Similarly, HypeE (Choudhary et al., 2021) leverages the properties of hyperbolic space to model hierarchical relationships more effectively, while ConE (Zhang, Wang, et al., 2024) employs cone embeddings in a polar coordinate system, where the logical intersection is modeled as the intersection of cones. Additionally, CALQR (Pan et al., 2025) introduces a plug-and-play module designed to enhance existing geometric models, thereby improving accuracy without sacrificing the core geometric intuition.

Overall, the primary advantage of geometric-based methods is their computationally efficient framework for the SKG complex logical query answering task, achieving efficient computation by predicting the region where the target entity is located. However, these methods face significant limitations: their performance is inherently bounded by the representational capacity of their chosen geometric primitives (e.g., boxes, cones), which may struggle to accurately capture the full complexity of logical relations. This constraint becomes particularly evident when dealing with highly intricate or nested logical structures that cannot be easily mapped to simple geometric shapes.

(2) Probability-based methods: A fundamental limitation of geometric-based methods lies in their reliance on hard boundaries, as they assume that a target entity must strictly reside within a crisp and well-defined geometric region. This deterministic paradigm fails to capture the inherent semantic ambiguity and uncertainty prevalent in real-world SKGs, where entities often exhibit partial or graded membership to a query's intent. To bridge this gap, probability-based methods introduce a paradigm shift, moving from rigid geometric primitives to flexible probability distributions. As presented in Fig. 7 (c.2), probability-based methods first project queries as parameterized distributions and execute logical operators via probabilistic mathematical transformations. Finally, candidate entities are ranked based on their probability density within the derived complex query distribution. The structural commonalities of probability-based methods emerge from a shared representational framework that unifies symbolic logic and continuous probability.

For example, BetaE (Ren & Leskovec, 2020) pioneered to introduce the use of the Beta distribution for embedding. The Beta distribution is particularly well-suited for this task due to its bounded support on the interval $[0, 1]$, which naturally aligns with the probabilistic interpretation of an entity's relevance to a query. Therefore, BetaE can faithfully model the uncertainty and noise inherent in SKGs, offering a more nuanced and realistic representation than its geometric predecessors. Building upon this foundation, GammaE (Yang et al., 2022) further refines the modeling of uncertainty. It employs the Gamma distribution, chosen for its distinct mathematical properties. The Gamma distribution's linear property under addition and its strong boundary support make it exceptionally effective at avoiding ambiguous answers, particularly in scenarios involving long-tail entity distributions or complex query chains. GammaE thus represents a significant evolution, addressing specific representational challenges and enhancing the robustness of probabilistic embeddings.

Overall, the primary advantage of probability-based methods is that they extend geometric-based approaches by transforming sharply defined geometric regions into probability distributions, which can naturally capture incompleteness and noise in SKGs. However, these methods face significant trade-offs: they suffer from weaker interpretability compared to their geometric counterparts, as the probabilistic nature of their representations makes the reasoning process more challenging to understand and visualize.

(3) Neural-based logical query answering methods: Neural-based logical query answering methods have emerged as a powerful and flexible paradigm. As shown in Fig. 7 (c.3), they first learn the correspondence between complex queries and target entities using neural networks. Then, the neural networks are used to reason about the final answers. The structural commonalities of these methods center on a unified, end-to-end learning paradigm, which transforms complex logical query answering into a neural network optimization problem. These methods have consistently evolved from generic feed-forward networks to architectures with stronger inductive biases tailored to specific data characteristics. Throughout this progression, they maintain a core focus on automatically discovering and modeling complex patterns in data through high-capacity, flexible neural architectures.

A foundational approach is demonstrated by NNMLR (Amayuelas et al., 2022), which employs multilayer perceptrons (MLPs) and MLP-Mixers to model both atomic queries and their logical combinations. By computing the distance between query and entity embeddings, NNMLR establishes a baseline for the expressive power of generic feed-forward networks in this task. Building on this, subsequent works have integrated networks with stronger inductive biases tailored to specific data characteristics. For example, EmQL (Sun et al., 2020) represents entity sets using k-hot vectors, which facilitates straightforward set operations and enables an exactly emulate logical, and hence faithful, reasoning systems. EmQL employs a BERT model to train entity representations, ensuring that entities to be grouped into sets are positioned closely in the embedding space. Similar methods include CQD (Arakelyan et al., 2021), GNN-QE (Zhu et al., 2022), FuzzQE (Chen, Hu, & Sun, 2022), TEMP (Hu et al., 2022), and Var2Vec (D. Wang et al., 2023). Recently, SAQE (Cao et al., 2025) has introduced a novel approach that addresses the limitations of previous methods by explicitly modeling both triple-level and path-level knowledge within a unified GNN framework, enhancing generalizations for complex logical query answering.

Overall, the primary advantage of neural-based logical query answering methods is their significant shift towards a unified, end-to-end learning paradigm for SKG complex logical query answering. They excel at automatically discovering and modeling complex patterns from data, often achieving state-of-the-art performance through their high model capacity and flexibility. However, this enhanced performance comes with notable trade-offs: the black-box nature of neural networks significantly sacrifices interpretability, making it difficult to understand how specific query results are derived.

Table 2
Dataset statistics of static knowledge graph reasoning.

Datasets	# Ent.	# Rel.	# Tri.	Domain	Source	Links
ATOMIC (Sap et al., 2019)	304,388	9	785,937	Commonsense	Crowdsourcing	https://
FB15K-237 (Toutanova et al., 2015)	14,541	237	310,116	General	Freebase	https://
ConceptNet100K (Li et al., 2016)	78,527	34	100,000	General	ConceptNet	https://
ConceptNet (Speer et al., 2017)	28,370,083	50	34,074,917	General	Wikipedia,OpenCyc,WordNet	https://
NELL-995 (Xiong et al., 2017)	75,492	200	154,213	General	Web	https://
DBpedia50K (Shi & Weninger, 2018)	49,000	654	43,756	General	Wikipedia	https://
YAGO3-10 (Dettmers et al., 2018b)	123,182	37	1,079,040	General	YAGO	https://
CoDEX_S (Safavi & Koutra, 2020)	2034	42	36,543	General	Wikidata	https://
CoDEX_M (Safavi & Koutra, 2020)	17,050	51	206,205	General	Wikidata	https://
CoDEX_L (Safavi & Koutra, 2020)	77,951	69	612,437	General	Wikidata	https://
UMLS (Kemp et al., 2006)	135	49	6752	Biomedical	human experts	https://
Hetionet (Himmelstein et al., 2017)	47,031	24	2,250,197	Biomedical	public datasets	https://
OpenBioLink (Breit et al., 2020)	180,992	28	4,192,002	Biomedical	public datasets	https://
Nation (Kemp et al., 2006)	14	55	1592	Social Sciences	human experts	https://
WN18 (Bordes et al., 2013)	40,943	18	151,442	Semantics	WordNet	https://

3.1.4. The datasets of SKG reasoning methods

In this section, we select, summarize, and analyze datasets related to the static knowledge graph reasoning task. While striving to cover all relevant datasets for the task, we follow specific selection criteria to ensure their high quality, providing better guidance for researchers in this field. The criteria including (1) scale diversity: ensuring coverage from small to large datasets to balance experimental efficiency, (2) source diversity: incorporating data constructed through various collection and annotation methods to enhance robustness, (3) domain diversity: encompassing a wide range of knowledge domains to assess model generalization, and (4) usage frequency: emphasizing datasets that are widely recognized and frequently employed in the research community to ensure comparability and relevance. These criteria are also applied in the subsequent dataset sections and will not be repeated.

We statistically summarize the information of the related dataset from multiple dimensions, including (1) # Ent.: Entity number; (2) # Rel.: Relation number; (3) # Tri.: Fact triplets number; (4) Domain: The domain of knowledge stored in the SKGs; (5) Source: The source of the SKGs, and (6) Links: The storage address of the SKGs. The statistical results are shown in Table 2.

In static knowledge graph reasoning, the characteristics of the dataset, particularly its size and sparsity, strongly influence the selection and performance of reasoning methods. When the dataset is large, traditional translation-based and tensor decomposition approaches often face scalability issues because their limited representational capacity requires expanding feature vector dimensions, which rapidly increases the number of parameters. The neural-based SKG reasoning methods, in contrast, can handle large datasets more efficiently. Their network architectures enable the extraction of deep semantic features, allowing them to achieve comparable or better results with fewer parameters.

3.1.5. The comparison of SKG reasoning methods

In this section, we conduct a comparison of various static knowledge graph reasoning methods, highlighting their respective strengths and limitations across four key dimensions: Performance (reasoning accuracy), Efficiency (computational efficiency and resource consumption), Interpretability (reasoning transparency), Robustness (stability under out-of-distribution, data noise, or data sparsity scenarios). These dimensions are selected to provide a comprehensive evaluation framework, encompassing core aspects critical for assessing the practical utility of these methods. Furthermore, we add a brief rationale for each evaluation. Please note that all subsequent similar sections will utilize the dimensions specified above, and this description will not be repeated. The corresponding results are presented in Table 3.

The comparison reveals a consistent trade-off among these metrics across static knowledge graph (SKG) reasoning sub-tasks. Complicated methods, such as tensor decompositional (Balazevic et al., 2019), neural-based (Cao et al., 2025), logical query (Cao et al., 2025), and generation-based (K. Xu et al., 2025) methods, generally achieve strong modeling capabilities and high performance. However, their reliance on complex parameterization and latent representations results in high computational demands, lower interpretability due to their black-box nature, and moderate robustness caused by data sensitivity. Conversely, explicit approaches, including rule-based (S. Guan et al., 2024) and geometric-based (Pan et al., 2025) methods, offer highly transparent reasoning logic and superior efficiency, but are limited by constrained overall performance and robustness.

In summary, these findings highlight that approaches excelling in accuracy and adaptability tend to sacrifice transparency and computational efficiency, while highly interpretable or efficient methods face bottlenecks in achieving robust reasoning.

3.2. Temporal knowledge graph reasoning

The objective of temporal knowledge graph reasoning is to leverage existing events and knowledge to reason about unseen events or predict future events. Previous research (Liang et al., 2024) primarily categorizes reasoning tasks into two scenarios: interpolation and extrapolation, depending on whether the model has seen the timestamps in the query. From a reasoning perspective, interpolation generally completes missing facts by analyzing known knowledge in TKGs, while extrapolation focuses on predicting unknown events by learning embeddings of entities and relations from historical facts on continuous TKGs.

Table 3
Comparison of static knowledge graph reasoning methods. Key dimensions: Performance (reasoning accuracy), Efficiency (computational efficiency and resource consumption), Interpretability (reasoning transparency), Robustness (stability under out-of-distribution, data noise, or data sparsity scenarios).

Sub-task	Method	Performance	Efficiency	Interpretability	Robustness
Traditional SKG reasoning	Translation-based	Moderate *Straightforward translation pattern*	High *Lightweight operations*	Moderate *Implict geometric translation*	Moderate *Sensitive to complexity*
	Tensor decompositional	High *Rich tensor interaction patterns*	Low *Heavy matrix operations*	Moderate *Latent decomposition*	High *Robust representation*
	Neural-based SKG reasoning	High *Effective non-linear interactions*	Moderate *Adequate parameter sharing*	Low *Opaquea modeling process*	Moderate *Conditional stability*
Multi-hop SKG reasoning	Rule-based	Moderate *Limited rule coverage*	Moderate *Steady mining overhead*	High *Transparent symbolic logic*	Moderate *Strict formalism*
	Reinforcement learning-based	High *Flexible path learning capability*	Low *Excessive exploration cost*	Moderate *Partial path visibility*	Low *Poor noise tolerance*
	Generation-based	High *Adaptive modeling capability*	Moderate *Moderate generation cost*	Low *Opaquea generation process*	Moderate *Ordinary reliability*
SKG complex logical query answering	Geometric-based	Moderate *Fair constraint expressivity*	High *Efficient region prediction*	High *Intuitive geometric regions*	Low *Fragile boundary management*
	Probability-based	High *Effective uncertainty modeling*	Moderate *Complex distribution computation*	Moderate *Probabilistic obscurity*	High *Superior noise tolerance*
	Neural-based logical query answering	High *Complex pattern modeling capability*	Moderate *Unified training process*	Low *Opaquea reasoning process*	Moderate *Reasonable data resilience*

3.2.1. Interpolation-based TKG reasoning

Task definition: Given a TKG with facts from $time_0$ to $time_T$, the Interpolation-based TKG reasoning method aims to complete missing quadruple $(h, r, ?, time_i)$ or $(?, r, t, time_i)$ in history $(time_0 \leq time_i \leq time_T)$. Interpolation-based TKG reasoning methods can be divided into dependent-based, function-based, and neural-based, as illustrated in Fig. 7(d).

(1) Dependent-based methods: As demonstrated in Fig. 7 (d.1), dependent-based methods are founded on a philosophy of indirect temporal modeling. Instead of treating time as a primary, explicit dimension to be directly manipulated, these methods associate timestamps with entities or relations. This approach preserves a fundamental design choice that constrains temporal information to be dependent on entity or relation representations, effectively capturing evolution through time associations. By embedding temporal context within the structural elements themselves, these methods maintain a more compact representation while still capturing the dynamic nature of temporal relationships.

For example, TTransE (Jiang et al., 2016) pioneers the idea of jointly embedding relations and their associated timestamps into a unified vector space. By extending the classic TransE (Bordes et al., 2013) model, TTransE demonstrates the feasibility of incorporating temporal cues without fundamentally altering the underlying framework of traditional SKG reasoning methods. However, its treatment of time is relatively coarse-grained.

To address this limitation, ST-TransE (Ni et al., 2020) introduces a specialized time embedding method that constrains the representation learning of entities and relations. A more significant architectural shift is seen in T-SimplE (Lin & She, 2020), which departs from simple vector operations. Additionally, by leveraging a fourth-order tensor to explicitly model the interactions within a quadruple, T-SimplE (Lin & She, 2020) achieves a substantial improvement in capturing complex temporal associations.

Overall, the primary advantage of dependent-based methods is their fundamental design choice of attaching temporal information directly to entities or relations, which endows them with significant computational efficiency and simplicity, making them well-suited for interpolation-based TKG reasoning scenario. However, this indirect approach utilizes time information in a coarse-grained manner, resulting in limited temporal expressiveness that fails to capture nuanced temporal relationships. Consequently, these methods tend to underperform on TKGs exhibiting complex and irregular evolutionary patterns, where fine-grained temporal modeling is essential for accurately representing the dynamic nature of TKGs over time.

(2) Function-based methods: As shown in Fig. 7 (d.2), function-based methods represent a paradigm shift from the indirect time association seen in dependent-based methods. The structural commonalities of function-based methods center on their use of specialized functions (e.g., transfer matrices, rotations, and transformations in complex space) to directly encode temporal dynamics into the embedding space, rather than merely attaching timestamps to entities or relations. The core philosophy is to directly model how entities and relations evolve over time, enabling the capture of more nuanced and fine-grained temporal patterns, such as periodicity and trends.

For example, BoxTE (Messner et al., 2022) extends the static BoxE (Abboud et al., 2020) model by incorporating temporal information through a relation-specific transfer matrix, which facilitates the exploration of more complex inference patterns over time. ChronoR (Sadeghian et al., 2021) associates timestamps with relations, considering each relation-timestamp pair as a rotation that maps the head entity to the tail entity. TComplEx and TNTComplEx (Lacroix et al., 2020) extend the third-order tensor to a fourth-order tensor in complex space to enable reasoning. Notably, TNTComplEx assumes that certain facts remain static over time, separating the TKG into temporal and non-temporal components. Similarly, TeRo (C. Xu et al., 2020) incorporates timestamps into the embeddings of head and tail entities in complex space to capture their temporal evolution, and it represents the relation as a rotation that maps the head entity to the tail entity.

Overall, the primary advantage of function-based methods is their superior modeling precision and expressiveness, achieved by directly encoding temporal dynamics through sophisticated mathematical functions. This approach enables them to capture complex temporal patterns effectively, making them particularly powerful for interpolation-based TKG reasoning scenarios. However, this expressive power comes at a significant cost. The reliance on carefully designed functions often leads to high model complexity and substantial computational overhead, making these methods computationally intensive. Moreover, these models can be prone to overfitting on the specific temporal patterns present in the training data, which may limit their generalization capabilities to unseen or more complex temporal dynamics, reducing their effectiveness in real-world applications where temporal patterns may evolve or differ from those observed during training.

(3) Neural-based interpolation methods: As shown in Fig. 7 (d.3), unlike the function-based paradigm that models time with mathematical formulas, neural-based interpolation methods adopt a data-driven approach. They treat timestamps as sequential data and use neural networks to implicitly learn the temporal evolution of entities and relations. The structural commonalities of neural-based temporal methods center on employing architectures like Long Short-Term Memory (LSTM) networks or Convolutional Neural Networks (CNNs) to encode timestamps into rich vector representations that capture complex temporal dependencies and correlations.

For instance, TA-TRANSE (García-Durán et al., 2018) is a temporal-aware version of TrasnE (Bordes et al., 2013). It utilizes LSTM to learn time-aware representations of relations, and represents quadruples as a set of triplets in the form of (h, r_{seq}, t), where r_{seq} means a relation that may include a timestamp suffix. Similarly, TA-DISTMULT (García-Durán et al., 2018) is a temporal extension of DistMult (B. Yang et al., 2015), considering the relation with temporal information as a sequence. Additionally, HyTE (Dasgupta et al., 2018) is an extension of TransH (Wang et al., 2014), DE-SimplE (Goel et al., 2020) is an extension of SimplE (Kazemi & Poole, 2018). These methods often consider temporal constraints to enhance temporal reasoning capabilities. For example, TIMEPLEX (Jain et al., 2020) leverages the recurrent nature of certain facts and the temporal interactions between pairs of relations during expansion. These additional temporal constraints can help assess a quadruple's validity better.

Overall, the primary advantage of neural-based interpolation methods is their flexible and powerful framework for learning complex and non-linear temporal patterns. This is achieved by leveraging neural networks such as LSTMs and CNNs to thoroughly model time information. However, this expressive power comes with notable drawbacks. The neural-based interpolation methods suffer from low interpretability, as the internal workings of neural networks function as "black boxes", making it difficult to understand how specific temporal patterns are being captured.

3.2.2. Extrapolation-based TKG reasoning

Task definition: Given a TKG with facts from $time_0$ to $time_T$, the Extrapolation-based TKG reasoning method aims to predict unknown facts $(h, r, ?, time_j)$ or $(?, r, t, time_j)$ that occur in the future $(time_j > time_T)$. Extrapolation-based TKG reasoning methods can be categorized into RL-based, rule-guided, and GNN-based, as illustrated in Fig. 7(e).

Compared to the interpolation-based TKG reasoning, the extrapolation-based TKG reasoning places higher demands on modeling sequential information in TKGs, and shares similarities with Multi-hop SKG reasoning tasks. Therefore, many methods have adopted reinforcement learning techniques and rule application, forming RL-based and rule-guided methods. Additionally, for accurately learning entity node information under long-term temporal evolution, while naive neural-based interpolation methods can learn time-aware representations of entities or relations, they cannot learn time-aware representations of the continuously changing graph structure. Consequently, methods combining GNNs have been proposed to simultaneously learn structural features and temporal information, thereby enhancing reasoning capabilities.

(1) Reinforcement learning-based methods: As shown in Fig. 7 (e.1), reinforcement learning (RL)-based methods frame the extrapolation reasoning task as a sequential decision-making or path-finding problem. Initially, a policy network is trained using reward signals. Once trained, the policy network explores the reasoning paths to deduce the final answer. The structural commonality among these methods lies in their treatment of temporal information, either by representing it as an additional state vector or as a gating mechanism to guide agent decisions. This framework enables the agent to navigate through temporal knowledge graphs, identifying valid reasoning paths from head entities to target tail entities.

A primary line of research concentrates on embedding temporal information directly into the RL framework, in order to guide the agent's decisions. For instance, TPath (Bai et al., 2021) pioneers this by treating time as an additional vector that evolves with the agent's state, making the environment explicitly time-aware. Building on this, TAgent (Tao et al., 2021) introduces a novel temporal gate mechanism to filter candidate actions, which allows the model to selectively attend to relations that are more relevant given the current temporal context. TITer (Sun et al., 2021) further refines the reward design by defining a relative time encoding function and a time-shaped reward based on the Dirichlet distribution.

Another significant direction focuses on enhancing the efficiency and robustness of the reasoning strategy itself, particularly to overcome the challenges of sparse rewards and incomplete information. CluSTeR (Li et al., 2021) shifts from a single-path search to a multi-clue elicitation strategy. It employs a beam search to gather diverse evidence from historical facts and then uses Graph Convolutional Networks (GCNs) (Kipf & Welling, 2017) to synthesize these clues into a final prediction, which significantly improves the model's ability to handle complex queries. Similarly, STKGR-PR (X. Meng et al., 2024) directly addresses the sparsity problem by dynamically generating and filling in missing intermediate paths using a temporal embedding model. This proactive approach ensures that the agent is not hindered by a lack of direct connections, thereby improving its reasoning capability in sparse TKGs.

Overall, the primary advantage of RL-based methods is their powerful and interpretable framework for extrapolation-based TKG reasoning. This approach allows for transparent decision-making where each step in the reasoning path can be examined and understood, providing valuable insights into how conclusions are reached. However, this explicit path-finding approach may

incur significant computational overhead, especially when dealing with TKGs with complex structures where the number of possible paths grows exponentially.

(2) Rule-guided methods: As displayed in Fig. 7 (e.2), rule-guided methods derive temporal logical rules (extensions of the aforementioned logical rules) from TKGs and utilize them to predict facts. The structural commonalities of these methods center on bridging connectionist learning and symbolic reasoning, achieved through the extraction and application of temporal logical rules from TKGs. These methods maintain a core philosophy: leveraging human-understandable logical patterns for temporal reasoning. To this end, they treat temporal rules as either direct predictors or guiding mechanisms within learning loops, rather than as mere auxiliary components. Operating on the premise that the evolution of facts in a TKG often follows discernible logical patterns, these approaches explicitly mine and leverage such patterns to make more robust and explainable predictions.

A primary line of research focuses on the mining and direct application of temporal logical rules. For example, TLogic (Liu et al., 2022) automatically mines cyclic temporal logical rules by extracting temporal random walks from the graph. LCGE (Niu & Li, 2023) mines the temporal rules with several time constraint patterns to construct a rule-guided predicate embedding regularization strategy for learning the causality among events. Traditional rule-guided methods fail to address the continuously emerging unseen entities over time and ignore the historical dependencies between entities and relations. To overcome these limitations, TPNet (Li & Wu, 2025) extracts reliable temporal logical paths instead of solely logical rules from historical subgraphs, achieving significant performance improvements.

Another significant direction focuses on involving the deep fusion of rules with learning frameworks, especially reinforcement learning. Here, rules are not merely an external tool but are instead integrated into the core learning loop to address fundamental challenges like sparse rewards and semantic noise. For example, TPRG (L. Bai et al., 2023) proposes a similar concept of temporal rules and has made improvements based on TPath (Bai et al., 2021), achieving improvements. DREAM (S. Zheng et al., 2023) proposes a reinforcement learning framework where the agent can receive adaptive rewards by imitating demonstrations at both the semantic and rule levels to eliminate the issue of sparse rewards. Additionally, TFSM (Bai et al., 2025) considers the impact of prior events on the current entity embedding and employs a strategy based on path search to prune the search space.

The main advantage of rule-guided methods is their exceptional interpretability and strong ability to generalize. Because they are built on symbolic logic, they provide clear, human-understandable reasoning paths that can be easily verified and trusted. However, their primary weakness is limited coverage. Any finite set of rules will inevitably struggle to capture the full complexity and novelty of the real world. This can lead to poor performance when faced with situations that do not fit established patterns or involve new, unseen relations.

(3) GNN-based methods: As presented in Fig. 7 (e.3), the integration of Graph Neural Networks (GNNs) has marked a pivotal evolution in extrapolation-based TKG reasoning, giving rise to methods that we term GNN-based methods. First, GNNs are used to learn and enhance the representations of nodes in a TKG. These enhanced representations are then used to predict the facts. The structural commonality among these approaches centers on their joint modeling of structural dependencies and temporal dynamics through GNNs, usually along with recurrent networks. They preserve a fundamental design choice that treats temporal knowledge graphs as evolving structures requiring both spatial and temporal encoding. These methods maintain a consistent architectural pattern that combines recurrent mechanisms with neighborhood aggregation, creating flexible frameworks well-suited for extrapolation-based reasoning.

Know-Evolve (R.S. Trivedi et al., 2017) pioneers this approach by modeling the occurrence of facts as a multivariate point process over time, using a deep recurrent network to learn non-linearly evolving entity representations. Building on this, RE-NET (Jin et al., 2020) framed the problem as predicting sequences of subgraphs, employing a GNN-based autoregressive model with neighborhood aggregation to enhance interpretability. These models established the core principle of using GNNs to encode the local structural context of entities at or around a given timestamp, thereby laying the groundwork for more sophisticated analyses.

Moving beyond localized views, models such as RE-GCN (Li et al., 2021) and its advanced variant CEN (Li et al., 2022) treat the history of the graph as a holistic sequence. By processing the entire TKG sequence, they can effectively pinpoint long-term dependencies and track the evolutionary trajectories of entities and relations. Concurrently, CyGNet (Zhu et al., 2021) leverages a copy-generation mechanism to explicitly model and exploit high-frequency, repetitive events, while CENET (Y. Xu et al., 2023) introduces a novel differentiation between historical and non-historical dependencies, allowing for more precise query-specific reasoning.

Despite these advances, a critical limitation persists: many of these models still struggle to capture complex, high-order connectivity and the nuanced, long-range sequential patterns essential for a truly deep understanding. To address these issues, TiRGN (Y. Li et al., 2022) develops different structural encoders to capture sequential and recurring patterns within historical data. Furthermore, HGLS (M. Zhang et al., 2023) designs a hierarchical graph framework to model long-term dependencies of entities across different timestamps. Additionally, HFL (Xu et al., 2025) models time intervals using tokens to incorporate both short-term and long-term historical facts into the reasoning process, while also utilizing the inherent reasoning capabilities of PLMs. STTP-RGN (X. Xu et al., 2025) designs an encoder based on graph neural networks that features a dual recurrent mechanism. This mechanism allows it to simultaneously evolve both entity and relationship representations across adjacent timestamps and spatial stamps. Recently, TaReT (Ma et al., 2024) utilizes a pioneering attention-based relation graph model to capture both deep and shallow historical temporal information. DLGR (Xiao et al., 2024) designs a novel two-level relation-aware attention network based on RGCN to comprehensively characterize each relation-specific entity representation for reducing noise.

Overall, the primary advantage of GNN-based methods is their significant adeptness at capturing structural information and temporal dynamics with strong robustness. Their ability to model intricate connections between entities while maintaining resilience to noise and incomplete data positions them as valuable tools for extrapolation-based TKG reasoning scenario. However, they face significant issues of high computational complexity, which can limit their scalability to large-scale TKGs.

Table 4
Dataset statistics of temporal knowledge graph reasoning.

Datasets	# Ent.	# Rel.	# Time.	# Events	Domain	Source	Links
YAGO11k (Dasgupta et al., 2018)	10,623	10	189	161,540	General	YAGO	https://
Wikidata12k (Dasgupta et al., 2018)	12,554	24	232	3,419,607	General	Wikidata	https://
YAGO-3SP (Wu et al., 2022)	27,009	37	3	130,757	General	YAGO	https://
DBpedia-3SP (Wu et al., 2022)	66,967	968	3	201,089	General	DBpedia	https://
IMDB-13-3SP (Wu et al., 2022)	3,244,455	14	30	627,096	Movie	IMDB	https://
IMDB-30SP (Wu et al., 2022)	243,148	14	3	7,923,773	Movie	IMDB	https://
GDELT-small (Zhang et al., 2022)	500	20	366	3,419,607	Social Science	GDELT	https://
ICEWS14 (Zhang et al., 2022)	7128	230	365	90,730	Social Science	ICEWS	https://
ICEWS05-15 (Zhang et al., 2022)	10,488	251	4017	479,329	Social Science	ICEWS	https://
ICEWS14 (Zhang et al., 2022)	7128	230	365	90,730	Social Science	ICEWS	https://

3.2.3. The datasets of TKG reasoning methods

In this section, we select, summarize, and analyze datasets related to the temporal knowledge graph reasoning task. We statistically summarize the information of the related dataset from multiple dimensions, including (1) # Ent.: Entity number; (2) # Rel.: Relation number; (3) # Time.: Timestamp number; (4) # Events: Eventual quadruples number; (5) Domain: The domain of knowledge stored in the TKGs; (6) Source: The source of the TKGs, and (7) Links: The storage address of the TKGs. The statistical results are shown in Table 4. Notably, we have not strictly limited these datasets to either interpolation-based TKG reasoning or extrapolation-based TKG reasoning, as this mainly depends on how the corresponding method segments the dataset.

The characteristics of temporal knowledge graph (TKG) datasets, such as their scale, temporal structure, and event sparsity, strongly influence the choice of reasoning methods. Compared with static knowledge graphs, TKGs often exhibit complex forms of sparsity. This sparsity arises not only from missing relations among entities but also from temporal discontinuities where certain entities have no interactions within particular time periods. Such uneven temporal distributions can cause biased representations and unstable inference. In addition, inconsistencies or noise in temporal annotations, including overlapping or delayed timestamps, introduce further uncertainty into reasoning processes. These dataset properties motivate the use of reasoning methods that can model latent temporal dependencies and compensate for missing contextual information. For example, approaches like ANEL (Y. Zhang et al., 2025) are designed to infer implicit temporal links by exploiting cross-time relational patterns, enriching sparse entities with information from neighboring timestamps. Neural architectures that incorporate temporal attention or memory components extend this strategy by capturing long-range dependencies. As a result, these methods are particularly suitable for datasets characterized by high temporal sparsity and noise, improving both robustness and generalization in diverse temporal settings.

3.2.4. The comparison of TKG reasoning methods

In this section, we conduct a comparison of various temporal knowledge graph reasoning methods, highlighting their respective strengths and limitations. The corresponding results are presented in Table 5.

For interpolation-based TKG reasoning methods, performance and efficiency are inversely related. Dependent-based models (Lin & She, 2020) prioritize high efficiency and moderate interpretability over performance and robustness. Function-based methods (Messner et al., 2022) enhance performance via fine-grained modeling but suffer in efficiency due to complex architecture. Neural-based approaches (Jain et al., 2020) achieve strong performance but incur heavy computational cost and lack interpretability.

For extrapolation-based TKG reasoning methods, GNN-based (X. Xu et al., 2025) and reinforcement learning-based (X. Meng et al., 2024) methods both attain high performance alongside low efficiency. They differ mainly in robustness, as the former benefits from superior noise tolerance, whereas the latter is limited by their sensitivity on high path quality. Meanwhile, rule-guided methods (Bai et al., 2025) offer a balanced strategy with superior interpretability driven by transparent symbolic rules.

Overall, the data reveal a clear trade-off across methods: achieving higher performance or robustness often means reduced efficiency and interpretability, making it challenging to optimize for all objectives at once.

3.3. Multi-modal knowledge graph reasoning

Task definition: The goal of MMKG reasoning methods is similar to SKG reasoning methods, which aims to complete the triplet (h, r, t) when one of h, r, or t is missing. In particular, the entity (h and r) could be text or images or has attributes of text and images. The MMKG reasoning methods can be divided into naive fusion MMKG reasoning and multi-modal pre-trained transformer-based (MPT-based) MMKG reasoning methods, as illustrated in Fig. 7(f).

(1) Naive fusion methods: As indicated in Fig. 7 (f.1), the naive fusion methods follow a two-stage pipeline. The modality encoding phase uses the corresponding encoder to obtain the embeddings of different modality data. Then, the integration phase fuses these embeddings to reason about the final answers. These methods maintain a modular design philosophy where visual, linguistic, and structural information are initially processed independently before being integrated through increasingly refined operations.

Table 5
Comparison of temporal knowledge graph reasoning methods. Key dimensions: Performance (reasoning accuracy), Efficiency (computational efficiency and resource consumption), Interpretability (reasoning transparency), Robustness (stability under out-of-distribution, data noise, or data sparsity scenarios).

Sub-task	Method	Performance	Efficiency	Interpretability	Robustness
Interpolation-based TKG reasoning	Dependent-based	Moderate *Coarse-grained modeling*	High *Simple architecture*	Moderate *Implicit temporality*	Low *Limited expressiveness*
	Function-based	High *Fine-grained modeling*	Moderate *Complex architecture*	Moderate *Complex mathematical functions*	Low *Significant overfitting risk*
	Neural-based interpolation	High *Strong nonlinear modeling capability*	Low *Heavy computational cost*	Low *Opaque modeling process*	Moderate *Reasonable Data adaptation*
Extrapolation-based TKG reasoning	Reinforcement learning-based	High *Adaptive path search*	Low *Excessive path exploration overhead*	Moderate *Step-wise transparency*	Low *High path quality sensitivity*
	Rule-guided	Moderate *Limited rule coverage*	Moderate *Reasonable inference speed*	High *Transparent symbolic foundations*	Moderate *High rule coverage reliance*
	GNN-based	High *Structural-temporal modeling*	Low *High graph computational complexity*	Moderate *Sufficient local aggregation visibility*	High *Superior noise tolerance*

The earliest work, IKRL (Xie et al., 2017), uses a neural image encoder to construct representations for all images of an entity, and combines the multiple image representations with the original structure-based representations derived from models like TransE (Bordes et al., 2013). This simple yet effective combination allows for the joint training of a model capable of leveraging both visual and structural information. Building directly on this concept, VALR (Mousselly-Sergieh et al., 2018) refines the integration strategy by defining the overall energy of a triple as the sum of sub-energy functions, each corresponding to a distinct modality (visual, linguistic, structural). This energy-based formulation provides a more principled way to balance the contributions of different information sources.

Building upon the foundational approach of combining pre-computed features, subsequent research seeks deeper integration within the model architecture itself, moving beyond simple concatenation. For example, TransAE (Wang et al., 2019) combines multi-modal autoencoder with TransE (Bordes et al., 2013) model, where the hidden layer of the autoencoder is used to encode multi-modal data. MKBE (Pezeshkpour et al., 2018) advocates for a more specialized design by developing modality-specific encoding, scoring, and decoding layers, acknowledging that different data types might benefit from tailored processing. THCR (Lu et al., 2022) further explores the creation of a shared latent space where information from different modalities is fused to complement and enrich the relational knowledge, aiming for a more holistic entity representation.

To address the limitations of static fusion and the potential for noisy or irrelevant multi-modal data to degrade performance, several methods have been proposed to enhance the robustness and adaptability of the integration process. For example, MMKRL (Lu et al., 2022) proposes a joint learning framework that is easily extendable to any modality and adopts an adversarial strategy to enhance model robustness. RSME (Wang et al., 2021) automatically adjusts the influence of visual context, encouraging useful information and filtering out irrelevant information to avoid encoding excessive noise.

Overall, the primary advantage of naive fusion MMKG reasoning methods is their systematic approach to combining disparate data sources. This approach involves a clear progression: starting from simple feature concatenation, moving to deeper architectural integration, and finally achieving more robust and adaptive fusion processes. This modular design and intuitive evolution of fusion strategies make the reasoning process more transparent, enhancing interpretability. However, these methods may be considered "naive" compared to modern end-to-end multi-modal pre-trained transformers, which often achieve superior performance at the expense of explainability.

(2) MPT-based methods: As shown in Fig. 7 (f.2), in contrast to naive fusion MMKG reasoning methods, MPT-based MMKG reasoning methods integrate Multi-modal Pre-trained Transformers (MPTs). This enables the models to harness the rich, contextualized representations from both textual and visual data to reason about final answers. The structural commonalities of these methods lie in the progressive refinement of fusion paradigms, which integrate MPTs with graph structural knowledge. They maintain a core focus on capturing nuanced, contextually relevant representations by fostering dynamic interactions between textual tokens and visual objects.

For example, MKGformer (Chen, Zhang, et al., 2022) introduces a novel transformer architecture that performs token-level multi-modal fusion. By enabling fine-grained interactions between textual tokens and visual objects, the model can dynamically filter out irrelevant visual noise, leading to cleaner and more contextually relevant unified representations. Similarly, SGMPT (Liang et al., 2024) designs a structure-guided fusion module that uses weighted summation and alignment constraint to inject the structural information into both textual and visual features.

Overall, MPT-based MMKG reasoning methods can capture more nuanced and contextually relevant representations. They achieve this by fostering fine-grained interactions between modalities and explicitly leveraging the graph's structural knowledge. This leads to significantly enhanced reasoning performance and robustness against noise. However, this increased sophistication comes at a cost. The intricate fusion mechanisms and complex architectures introduce substantially higher model complexity and computational overhead, posing challenges for training efficiency and practical deployment in resource-constrained environments. This complexity reduces interpretability, making it difficult to understand the model's decision-making process.

Table 6
Dataset statistics of multi-modal knowledge graph reasoning.

Datasets	# Ent.	# Rel.	# Facts	Type	Source	Links
IMGpedia (Ferrada et al., 2017)	14,765,300	44,295,900	3,119,207,705	KG+TXT+IMG	DBPedia,Wikimedia	https://
FB-IMG (Mousselly-Sergieh et al., 2018)	11,757	1231	350,293	KG+TXT+IMG	Freebase	https://
MMKG-DB15K (Ye, Hui, et al., 2019)	14,777	279	99,028	KG+Numeric+IMG	Freebase,DBpedia	https://
MMKG-Yago15k (Ye, Hui, et al., 2019)	15,283	32	122,886	KG+Numeric+IMG	Freebase,YAGO	https://
MKG-W (Xu et al., 2022)	15,000	169	42,746	KG+TXT+IMG	Wikipedia	https://
MKG-Y (Xu et al., 2022)	15,000	28	26,638	KG+TXT+IMG	YAGO	https://
RichPedia (Wang et al., 2020)	29,985	3	119,669,570	KG+IMG	WikiPedia	https://
FB15k-237-IMG (Chen, Zhang, et al., 2022)	14,541	237	310,116	KG+IMG	Freebase	https://
WN18-IMG (Chen, Zhang, et al., 2022)	14,541	18	151,442	KG+IMG	WordNet	https://

Table 7
Comparison of multi-modal knowledge graph reasoning methods. Key dimensions: Performance (reasoning accuracy), Efficiency (computational efficiency and resource consumption), Interpretability (reasoning transparency), Robustness (stability under out-of-distribution, data noise, or data sparsity scenarios).

Method	Performance	Efficiency	Interpretability	Robustness
Naive Fusion MMKG reasoning	Low *Shallow modality integration*	High *Efficient modular architecture*	High *Transparent modular logic*	Low *Vulnerable noise tolerance*
MPT-based MMKG reasoning	High *Fine-grained modality fusion*	Low *High model complexity*	Low *Opaque fusion mechanics*	High *Enhanced noise filtration*

3.3.1. The datasets of MMKG reasoning methods

In this section, we select, summarize, and analyze datasets related to the multi-modal knowledge graph reasoning task. We statistically summarize the information of the related dataset from multiple dimensions, including (1) # Ent.: Entity number; (2) # Rel.: Relation number; (3) # Facts: Fact number; (4) Type: MMKG type. A MMKG is represented by specific combinations of modalities. Taking "KG+TXT+IMG" for example, "KG" means the entity has a simple name or ID, "TXT" means the entity has a textual description as attributes, "IMG" means the entity has single or multiple corresponding images as attributes; (5) Source: The source of the MMKGs, and (6) Links: The storage address of MMKGs. The statistical results are shown in Table 6.

The characteristics of a dataset, particularly its multi-modal structure and noise patterns, strongly influence how reasoning methods are designed and selected. In contrast to static or temporal knowledge graphs, multi-modal knowledge graphs (MMKGs) introduce additional complexity because reasoning models must operate over heterogeneous and often imperfect modalities. Noise in MMKGs typically appears in four forms. Modality missing occurs when entities lack certain modalities. Modality ambiguity arises when a modality contains low-quality or imprecise information. Modality inconsistency appears when different modalities provide conflicting signals. Modality redundancy refers to excessive or repetitive information that limits the efficiency of representation learning. These issues increase data heterogeneity and make it harder to design reasoning approaches that achieve stable cross-modality fusion. As a result, researchers often choose reasoning strategies that explicitly account for noise characteristics. For example, Multimodal Boosting (Mai et al., 2024) selects weakly supervised weighting to reduce noisy modality effects while preserving complementary information, whereas SNAG (Chen et al., 2025) adopts Transformer-based masking to handle modality-level noise. More generally, models that integrate structure-aware reasoning with cross-modal attention are preferred for datasets with high variability and noise, since they offer improved robustness and generalization across diverse multi-modal environments.

3.3.2. The comparison of MMKG reasoning methods

In this section, we compare various multi-modal knowledge graph reasoning methods, highlighting their respective strengths and limitations. The corresponding results are presented in Table 7.

For multi-modal knowledge graph reasoning, naive fusion (Lu et al., 2022) and MPT-based (Liang et al., 2024) methods demonstrate a clear trade-off between efficiency and reasoning capability. The former prioritizes efficiency and transparent logic through modular design but struggles with complex reasoning and is easily disrupted by noisy data. Conversely, the latter achieves superior performance and robustness via fine-grained fusion and noise filtration at the expense of increased computational overhead and reduced interpretability.

3.4. Heterogeneous information networks, knowledge graphs, and reasoning

The concept of heterogeneous information networks (HINs) is highly related to KGs. However, existing reasoning surveys lack sufficient clarification between the relations among HINs and KGs in the era of reasoning. In this section, we discuss and analyze the relationship between HINs, KGs, and reasoning to help researchers better understand knowledge-driven reasoning.

The concept of HINs was first proposed by Sun et al. (2009). A HIN is defined as $\mathcal{H} = (\mathcal{G}_I, \mathcal{G}_s)$ and represents a directed graph that integrates two distinct viewpoints (Liu et al., 2024; Shi et al., 2021). These viewpoints incorporate a *schema graph* $\mathcal{G}_s$ for meta-level abstraction and an *instance graph* $\mathcal{G}_I$ for instance-level instantiations.

The instance graph $\mathcal{G}_I = (V, E)$ is composed of a set of entities V and a set of links $E \subseteq V \times V$ that connect those entities. Each entity is associated with a specific subset of entity types via the type mapping function $\tau : V \to 2^T$, in which T represents the entity type space. Similarly, the relation mapping $\phi : E \to 2^R$ assigns a link to various relation types within the defined relation set R.

The schema graph $\mathcal{G}_S = (T, R)$ for the HIN $\mathcal{H}$ is defined as a directed graph over the entity types T where the links represent sets of relation types from R. A meta-path is a path defined on the schema graph of a HIN. It functions as a vital tool for network analysis by capturing the semantic relationships between different entity types. A meta-path M of length l is defined as $M = t_1 \xrightarrow{r_1} t_2 \xrightarrow{r_2} \cdots \xrightarrow{r_{l-1}} t_l$, where $t_i \in T$ indicates an entity type and $r_i \in R$ represents a relation. An instance-level path $P = v_1 \xrightarrow{r_1} v_2 \xrightarrow{r_2} \cdots \xrightarrow{r_{l-1}} v_l$ satisfies M if $\forall i \in \{1, \ldots, l\}, t_i \in \tau(v_i)$ and $\forall i \in \{1, \ldots, l-1\}, r_i \in \phi(v_i, v_{i+1})$. In this context, P is regarded as a concrete path instance derived from M. These path instances are extensively employed to evaluate the reliability and significance of their corresponding meta-paths

In practice, many networks are represented as HINs, including bibliographic networks (Sun & Han, 2013; Sun et al., 2011; Tang et al., 2008), social networks (Zhong et al., 2013), and user-item networks (Jamali & Lakshmanan, 2013). For instance, a bibliographic network of computer science researchers derived from DBLP (Ley, 2002) is a typical HIN. It comprises three primary entity types, including papers, venues, and authors. In this network, each paper is connected to its respective authors and venue through a predefined set of relation types. However, these networks typically contain only a very limited number of entity types and relation types. Hence, such HINs are commonly referred to as schema-simple HINs (Shi et al., 2021, 2016; Sun et al., 2022). Based on these HINs, a diverse array of tasks has been developed, including similarity measurement (Sun et al., 2011; Zhou et al., 2019), clustering (Sun et al., 2009; B. Zhang et al., 2021), classification (Dong et al., 2017; Fu et al., 2017), link prediction (Zhang et al., 2013, 2015), ranking (Ng et al., 2011; Z. Xu et al., 2020), recommendation (Shi et al., 2018; Yu et al., 2013), and information fusion (Koutra et al., 2013; Umeyama, 2002).

However, while these tasks are highly valuable for practical applications, they do not inherently involve reasoning. Reasoning mainly refers to inferring new conclusions from existing knowledge (F. Yu et al., 2024). These works focus primarily on data mining (Shi et al., 2016) rather than the derivation of new knowledge through reasoning. This limitation stems largely from their relatively simple schemas, which lack the structural complexity and knowledge richness required to support sophisticated reasoning.

Recently, KGs have emerged as a primary paradigm for storing vast amounts of structured information within heterogeneous networks. Due to the large scale and high diversity of their entity types and relation types, KGs are commonly regarded as schema-complex HINs (Shi et al., 2021, 2016). As noted by Sun et al. (2022), KGs extend beyond simple data storage by integrating structured knowledge, which facilitates robust symbolic reasoning in practical applications. As detailed in Section 3, a variety of reasoning tasks are built upon different types of KGs. These approaches focus on leveraging semantic relations, topological structures, or logical patterns to facilitate the reasoning of new knowledge.

In summary, while HINs focus on the structural modeling of multi-typed data, KGs extend this framework with rich semantics and complex schemas. This transition from schema-simple networks to schema-complex knowledge bases drives a shift from traditional data mining toward sophisticated, knowledge-driven reasoning. This evolution provides the essential foundation for advanced AI systems that require robust reasoning capabilities.

3.5. The summary of reasoning based on symbolic knowledge bases

In summary, reasoning based on symbolic knowledge bases presents a diverse set of methodologies. Each method reflects distinct trade-offs among performance, efficiency, interpretability, and robustness. Static knowledge graph reasoning methods show strong modeling and inference capabilities, especially those based on neural and tensor architectures. However, these approaches often sacrifice computational efficiency and interpretability. Temporal knowledge graph reasoning adds another layer of complexity by introducing time-dependent information. Methods such as GNN or RL based approaches achieve high performance, but they usually suffer from low efficiency and limited transparency. Rule-guided approaches provide better interpretability, yet their accuracy and scalability remain constrained. Multi-modal knowledge graph reasoning faces further challenges due to the integration of heterogeneous modalities. Methods focusing on fine-grained fusion and robustness are powerful but computationally expensive and less interpretable. In contrast, simpler fusion strategies are more efficient but show weaker adaptability and reasoning depth.

Overall, across all symbolic knowledge base reasoning paradigms, a fundamental tension remains. Methods that emphasize higher reasoning accuracy and adaptability often do so at the expense of interpretability and computational efficiency. Future research may focus on developing adaptive frameworks that achieve a better balance among these competing objectives, enhancing both the practicality and the explanatory power of symbolic reasoning systems.

4. Reasoning based on parametric knowledge bases

Since reasoning methods based on parametric knowledge bases are often task-independent, this section reviews them from the perspective of their equipped techniques rather than tasks. These methods mainly perform reasoning in the form of question answering.

Task definition: Given a knowledge-intensive question, which requires deep understanding and reasoning capability to answer correctly, this type of method is required to leverage the knowledge encoded in parametric knowledge bases to reason for the final answer.

Recent NLP studies (Berkovitch et al., 2025; Cheng et al., 2024; Zhang et al., 2024) widely use parametric knowledge to describe the knowledge stored in the parameters of PLMs. Although parametric knowledge bases can broadly include rules, constraints,

or mathematical models, they are rarely referred to as knowledge bases in NLP field due to their limited knowledge. Generally, knowledge bases are expected to hold substantial information. Therefore, this survey focuses on PLMs as parametric knowledge bases, since they are pre-trained on large-scale corpora and store abundant knowledge in their parameters.

Based on the size of their parameter scales, PLMs can be divided into small language models (SLMs) and large language models (LLMs). SLMs-based reasoning methods (Rajani et al., 2019) need to fine-tune on task-oriented or domain-specific data to enhance their performance. During this process, the model's parameters are updated so that it can better recognize patterns and relationships relevant to reasoning within the given domain. For instance, in arithmetic reasoning or symbolic reasoning tasks, the fine-tuning data often consists of carefully crafted question-solution pairs or annotated explanations. By adjusting model parameters to minimize the prediction error on these examples, SLMs learn to associate specific linguistic cues with correct reasoning steps, effectively transferring some of the abstract knowledge encoded in their parameters into actionable strategies tailored for the task. For example, LoP (Talmor et al., 2020b) trains RoBERTa (Liu et al., 2019) on both implicit pre-trained knowledge and explicit free-text statements to symbolic reasoning. The method (Hendrycks et al., 2021) fine-tunes the GPT-2 (Radford et al., 2019) to generate full step-by-step solutions to arithmetic reasoning.

Although SLMs-based approaches have demonstrated good performance compared to traditional rule-based (Fletcher, 1985; Yuhui et al., 2010), symbolic-based (Liguda & Pfeiffer, 2012; Shi et al., 2015), and statistical-based (Koncel-Kedziorski et al., 2015a; Upadhyay et al., 2016) methods, they also face some significant challenges. First, fine-tuning SLMs generally relies on large-scale and high-quality training data, and collecting such datasets is often time-consuming and resource-intensive. Specifically, their effectiveness is closely related to the diversity and coverage of the fine-tuning data. When training samples are too limited or lack variety, SLMs may exhibit poor generalization and fail to perform well on reasoning tasks that differ from those seen during training. Second, the data modeling capacity of SLMs is relatively weak compared with LLMs due to their small parameter scale. This limited capacity restricts their ability to capture complex patterns and nuanced relationships within data. Additionally, SLMs may struggle with tasks that require deeper contextual understanding or long-term dependencies in text. As a result, their performance often lags behind current powerful LLMs, particularly when addressing sophisticated reasoning or multi-step problem-solving tasks. Overcoming these limitations may require not only improved training strategies but also the development of more robust and scalable architectures capable of handling a broader array of reasoning scenarios.

Recently, LLMs-based reasoning methods have shown impressive abilities, and prompting is the primary way to interact with LLMs. Compared with SLMs, LLMs possess strong generalization and in-context learning capabilities by providing a few demonstrations (i.e., few-shot learning) or instruction to solve new problems without any demonstrations (i.e., zero-shot learning). Therefore, we mainly investigate reasoning methods based on LLMs.

Many prompting methods have been proposed for reasoning problems (Shinn et al., 2023; M. Xu et al., 2024). The first attempt is made by Brown et al. (2020), which developed a zero-shot prompting method by adding a natural language description of the task in the prompt. Some follow-up methods try to optimize the prompting strategy from the perspective of Chain-of-Thought (CoT) (Wei et al., 2022), Iterative feedback (Madaan et al., 2023), Problem decomposition (Nye et al., 2021), Assemble (Wang et al., 2023), and multi-agent (Liang et al., 2024). These methods are widely applied to mathematical, commonsense, and symbolic reasoning tasks.

Despite the significant progress achieved with LLM-based prompting methods, several important challenges remain. One of the most critical issues is hallucination, where LLMs generate plausible yet factually incorrect or unsupported statements during the reasoning process. This is particularly problematic in high-stakes domains such as mathematical problem solving and symbolic reasoning, where even minor inaccuracies can lead to completely incorrect conclusions. A prominent approach to addressing hallucination is incorporating external symbolic knowledge bases for knowledge-augmented reasoning. By appending external knowledge from reliable sources, such as verified knowledge graphs, LLMs tend to produce factually accurate and logically consistent outputs. Furthermore, other solutions to mitigate hallucination include implementing assemble methods, where multiple model outputs are aggregated, and introducing iterative feedback to further refine and validate the generated results.

In addition to hallucination, reproducibility is another challenging aspect of LLM-based reasoning. The outputs of LLMs can be highly sensitive to prompt phrasing, the specific demonstrations chosen in few-shot settings, model randomness (e.g., temperature), and other contextual factors. As a result, it is difficult to guarantee that the same prompt will consistently produce identical or even similar reasoning paths. This randomness undermines the reliability of LLM-driven solutions, especially when used in scientific or real-world applications that demand deterministic and transparent decision-making processes. To improve reproducibility, researchers commonly fix random seeds, strictly document and share experimental settings (including prompts and hyperparameters), and evaluate models on standardized benchmark datasets. Additionally, open-sourcing complete implementations and developing unified evaluation frameworks or automated experiment management tools are practices that help standardize experimental procedures and enhance the reproducibility of LLM-based reasoning methods.

The overall taxonomy of reasoning methods based on parametric knowledge bases is shown in Fig. 8. Furthermore, Fig. 9 illustrates the core technology and pipeline of each type of reasoning method based on parametric knowledge bases.

4.1. CoT-based reasoning

Recent studies (Kojima et al., 2022; Wei et al., 2022; Zhou et al., 2023) find that generating a series of intermediate reasoning steps (also known as CoT and rationale) significantly improves the ability of LLMs to perform complex reasoning. The intermediate reasoning steps of CoT-series methods contribute to enhancing the logical consistency of the reasoning processes before reaching a conclusion. In this way, CoT-series methods significantly improve the LLMs' ability to perform tasks that require multi-step reasoning and deep understanding. There are three types of CoT-series methods: CoT optimization, CoT engineering, and automatic CoT methods, as illustrated in Fig. 9(a).

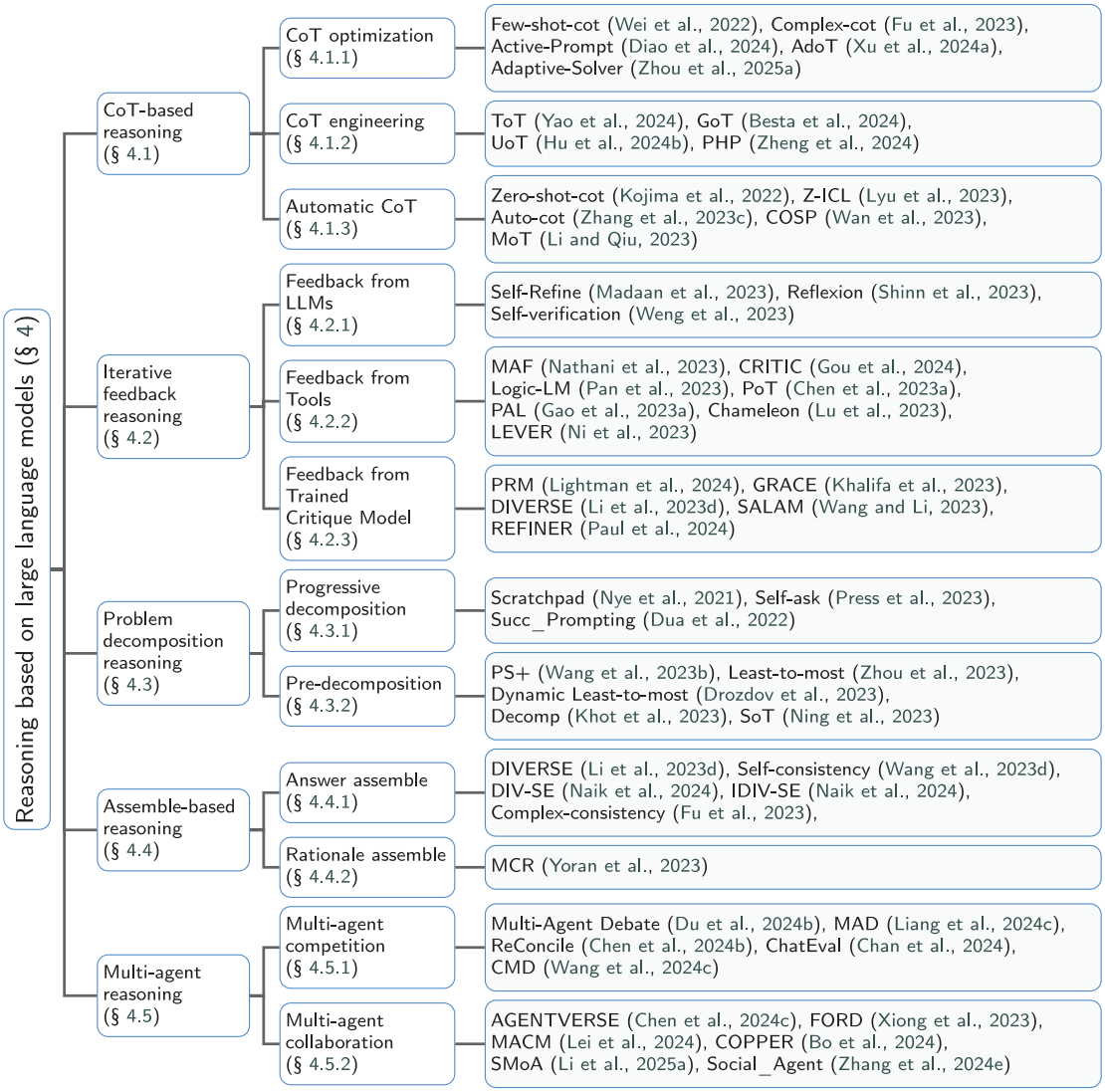

Fig. 8. Taxonomy of parametric reasoning across five prompting paradigms: CoT-based, iterative feedback, problem decomposition, assemble-based, and multi-agent reasoning. Representative methods for each task or sub-task are shown in the green box. Abbreviation: Chain-of-Thought (CoT), large language models (LLMs). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

4.1.1. CoT optimization

As shown in Fig. 9 (a.1), CoT optimization methods first construct the question-rationale pairs as demonstrations. Then, they concate the demonstrations and target question to guide the LLMs to reason step by step toward final answer. Instead of providing just the correct answer, these demonstrations explicitly show the step-by-step reasoning process leading to that answer. The LLM learns to mimic this reasoning pattern when faced with similar new questions.

For example, Few-shot-cot (Wei et al., 2022) adopts some questions and manually constructs CoT as demonstrations for the first time. Following this line, the CoT optimization methods try to optimize the CoT in demonstrations from different perspectives. For instance, Complex-cot (Fu et al., 2023) finds that demonstrations with higher reasoning complexity achieve substantially better performance on multi-step reasoning. Hence, it constructs CoT with more reasoning steps in demonstrations. To determine which questions are the most important and helpful to annotate from a pool of task-specific questions, Active-Prompt (Diao et al., 2024) proposes an uncertainty-based annotation strategy, which can reduce the model's uncertainty and help elicit the reasoning ability of LLMs. To solve the mismatch between the question difficulty and the methods' complexity, AdoT (M. Xu et al., 2024) first presents a difficulty measuring approach for questions that computes the syntactic and semantic complexity of their rationales. Then, it proposes a demonstration set construction and a difficulty-adapted retrieval strategy to adaptively construct reasonable demonstrations based on the difficulty of the questions. Similarly, Adaptive-Solver (J. Zhou et al., 2025) has also identified that a

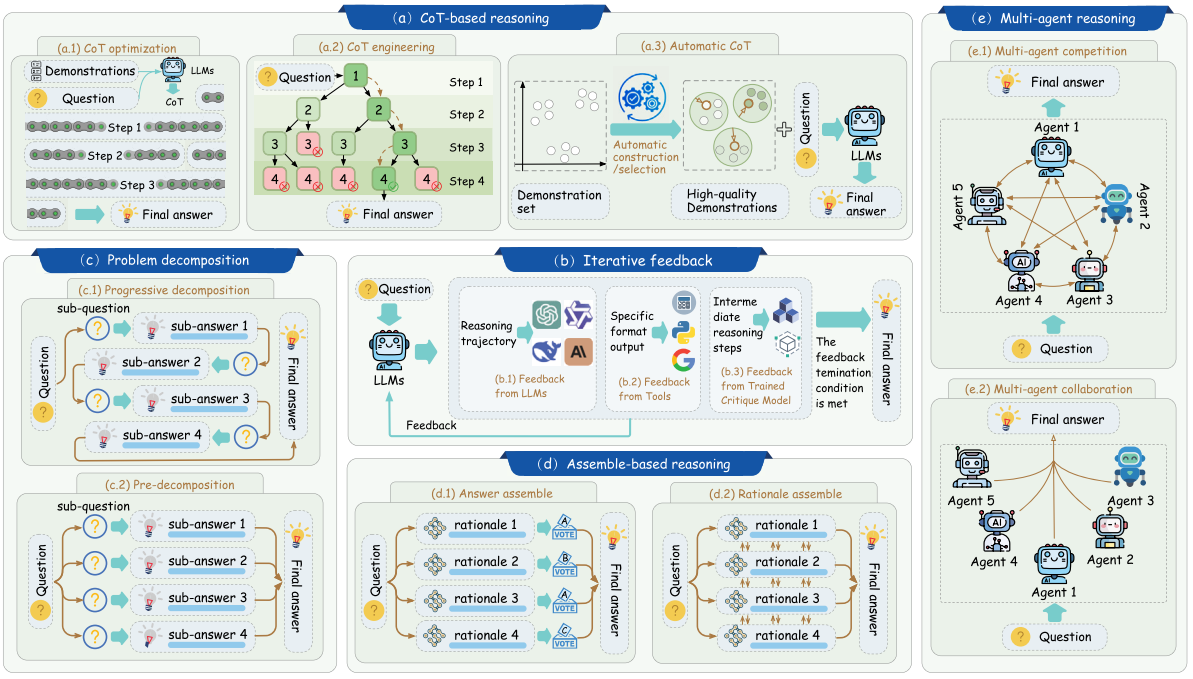

Fig. 9. Parametric reasoning methods across five categories: (a) CoT-based (Section 4.1); (b) Iterative feedback (Section 4.2); (c) Problem decomposition (Section 4.3); (d) Assemble-based (Section 4.4); and (e) Multi-agent reasoning (Section 4.5). Abbreviations: chain-of-thought (CoT).

one-size-fits-all strategy leads to unnecessary overhead or suboptimal performance, and therefore adapts strategies to the specific problems to achieve a balance between cost and efficiency.

Overall, the primary advantage of CoT optimization is its ability to significantly boost LLM performance on complex reasoning tasks. They enhance interpretability by making the LLM's reasoning steps visible, aiding in error diagnosis. However, CoT also has drawbacks: performance heavily depends on the quality and relevance of the provided demonstrations, and crafting effective examples can be labor-intensive. It increases computational cost due to longer generated outputs and may propagate errors if an intermediate step is incorrect. Additionally, the approach might not generalize well to tasks dissimilar to the demonstrations.

4.1.2. CoT engineering

As demonstrated in Fig. 9 (a.2), CoT engineering methods fundamentally extend basic CoT optimization methods by structuring the reasoning process into more sophisticated, often non-linear, frameworks. The core practice involves decomposing complex problems into manageable intermediate steps and explicitly managing the exploration and combination of these steps. Drawing from cognitive theories of problem-solving as heuristic search through a problem space, these methods create explicit representations of possible reasoning states (like nodes in a tree or graph) and paths between them. Techniques often incorporate mechanisms for generating diverse candidate thoughts, evaluating their promise, and strategically searching through the resulting structure to find the optimal solution path. This represents a shift from simple step-by-step reasoning to actively planning, exploring, and backtracking within a defined reasoning space.

Inspired by cognitive science, which characterizes problem-solving as a search through a combinatorial problem space, ToT (Yao et al., 2024) actively maintains a tree of thoughts, where each thought is a coherent language sequence that serves as an intermediate step toward problem-solving. GoT (Besta et al., 2024) models the information generated by LLMs as an arbitrary graph, where information units are vertices, and edges correspond to dependencies between these vertices. GoT enables combining arbitrary LLM thoughts into synergistic outcomes, distilling the essence of whole networks of thoughts, or enhancing thoughts using feedback loops. The information needed to solve the task is not initially given in many reasoning-related applications. To enhance LLMs in actively seeking information, UoT (Hu et al., 2024) incentivizes a model to seek information in a way that maximally reduces the amount of information it does not know. To optimize the generated answer progressively, PHP (Zheng et al., 2024) performs automatic multiple interactions between queries and LLMs by using previously generated answers as hints.

Overall, the primary advantage of CoT engineering is its ability to tackle significantly more complex and ambiguous problems than basic CoT, leading to higher accuracy and robustness. However, these methods come with disadvantages: they introduce substantial computational overhead due to multiple LLM calls and complex state tracking. Designing effective structures (trees/graphs) and aggregation rules requires significant expertise and task-specific tuning, increasing implementation complexity. The approach may also become impractical for real-time applications due to latency. Balancing sophistication with efficiency remains a key challenge.

4.1.3. Automatic CoT

The aforementioned two types of methods achieve excellent performance, but they rely on manually constructed demonstrations, which may generalize poorly between data from different domains and tasks. Hence, as shown in Fig. 9 (a.3), automatic CoT methods try to construct pseudo-demonstrations to guide LLMs under a zero-shot setting. They first automatically generate a large number of demonstrations by sourcing them from raw text corpora or the LLM's own outputs, bypassing the need for human annotation. Then, they design a method to select a few high-quality and representative demonstrations. Besides, few approaches use universal, task-agnostic instructions appended to queries to activate an LLM's inherent step-by-step reasoning capability without needing demonstrations.

For instance, Zero-shot-cot (Kojima et al., 2022) concatenates a simple but effective instruction "Let's think step by step" after the given question, which can activate the inherent multi-step reasoning capability of LLMs. To solve the problem that performance drops significantly when no demonstration is available, Z-ICL (Lyu et al., 2023) constructs pseudo-demonstrations from a raw text corpus. It retrieves relevant text from the corpus using the nearest neighbor search and then adjusts the pseudo-demonstrations with physical neighbor and synonym labeling to avoid the copying effect. Auto-cot (Z. Zhang et al., 2023) samples questions with diversity and automatically generates rationales to construct demonstrations. COSP (Wan et al., 2023) constructs demonstrations from the LLM zero-shot outputs via carefully designed criteria that combine consistency, diversity, and repetition. MoT (Li & Qiu, 2023) pre-thinks on the unlabeled dataset and saves the high-confidence thoughts through answer entropy as external memory. During inference, MoT lets the LLM recall relevant memory to help itself reason and answer.

Overall, the primary advantage of automatic CoT methods is their significant reduction in human effort and expertise required compared to manual demonstration construction, making them highly scalable and adaptable across diverse domains without needing specific examples. However, these methods also present challenges. The quality and relevance of automatically generated instructions or pseudo-demonstrations can be inconsistent and unpredictable, sometimes leading to irrelevant, erroneous, or poorly structured reasoning chains that negatively impact final answer accuracy. Furthermore, methods that rely on the model's own outputs risk propagating and even amplifying any initial errors or biases present in the zero-shot responses or the underlying source data. Ensuring the robustness and reliability of the automatically triggered or constructed reasoning remains a key challenge.

4.2. Iterative feedback reasoning

Inspired by the human behavior of trial, checking errors, and correcting them during reasoning, some researches focus on utilizing iterative feedback to correct mistakes in the reasoning steps to enhance the reasoning capabilities of LLMs (L. Pan et al., 2024). Using iterative feedback to improve reasoning typically involves three steps: (1) reasoning, (2) critique and feedback, and (3) reasoning refinement. The sources of feedback in these methods mainly include LLMs themselves, various tools (including calculators, search engines, logic tools, and code interpreters), and trained critique models, as described in Fig. 9(b).

4.2.1. Feedback from LLMs

As shown in Fig. 9 (b.1), the pipeline of this type of method first generates an initial reasoning trajectory. This is followed by an evaluation or verification phase, carried out by the LLM itself. Based on the feedback or identified errors, the model iteratively refines its reasoning, either by re-generating specific reasoning steps or adjusting the overall trajectory. This process can be repeated multiple times, aiming to incrementally enhance the accuracy and coherence of the final answer through self-supervision and feedback.

Specifically, Self-Refine (Madaan et al., 2023) explores how to achieve iterative feedback and refinement based on the LLM itself to improve the quality of reasoning. Reflexion (Shinn et al., 2023) employs an iterative process of "Trajectory→Evaluation→Reflection→Next Trajectory" to iteratively enhance the reasoning process. Self-verification (Weng et al., 2023) implements "Forward Reasoning" and "Backward Verification" to validate the reasoning process and selects the highest-quality reasoning results based on evaluation scores. However, it should be noted that recent studies (Huang et al., 2024; Kamoi et al., 2024) have suggested that this "self-reflection" approach may be constrained by the model's inherent reasoning capabilities, potentially hindering improvements in reasoning quality.

The main advantage of these methods is their ability to improve LLM reasoning without requiring external supervision or additional labeled data, making them scalable and cost-effective. However, a key disadvantage is that their effectiveness largely depends on the model's existing reasoning capabilities. If the LLM struggles with complex logic or systematic errors, its self-evaluations and corrections may be limited, potentially resulting in marginal improvements.

4.2.2. Feedback from tools

The general framework for these tool-integrated feedback methods typically begins by identifying the type of reasoning task in the LLM's output. As displayed in Fig. 9 (b.2), the output text is first transformed into an specific format, such as arithmetic expressions, logical statements, search queries, or structured code. It is then converted into a representation that is compatible with the chosen external tool. The corresponding tool processes this input and generates precise feedback, which is subsequently used to correct or refine the LLM's original response. This modular framework allows for flexible integration of different external tools, enabling targeted error correction based on the task at hand.

Research has explored integrating various tools into feedback modules to help correct reasoning errors. Different types of reasoning errors necessitate different tools. For example, calculators are often used to provide precise arithmetic results as feedback for arithmetic tasks (Gou et al., 2024; Nathani et al., 2023), while search engines are employed to verify factual errors (Gou et al., 2024; Lu et al., 2023). For logical reasoning tasks, Logic-LM (Pan et al., 2023) suggests using logical tools like First-order

Logic Provers to identify logic errors and provide feedback. Moreover, considering that the pre-training corpora of LLMs contain a substantial amount of structured code, some studies suggest transforming reasoning tasks into code form and then using code interpreters to provide feedback (W. Chen et al., 2023; Gao et al., 2023; Lu et al., 2023; Ni et al., 2023; Nye et al., 2021). All these methods focus on transforming the output text of LLMs into formats suitable for inputting into various tools, thus obtaining precise external feedback and enhancing the quality of reasoning.

The main advantage of this type of approach is that it leverages specialized external systems to provide highly accurate feedback, which can significantly improve the reasoning abilities of LLMs and reduce error rates. However, these methods also introduce increased complexity, require careful design to ensure correct task-tool matching and input transformation, and may suffer from latency or reliability issues if the external tools are slow or unavailable. Additionally, maintaining seamless interaction between the LLM and various tools remains a technical challenge.

4.2.3. Feedback from trained critique model

As shown in Fig. 9 (b.3), this type of methods generally first generate intermediate reasoning steps using an initial LLM, followed by employing a specialized critic model to evaluate or provide feedback on each step. This feedback can take the form of selecting the most accurate intermediates, identifying errors, or offering natural language critique. Theories underpinning this approach suggest that providing step-level supervision or critique enables finer-grained error correction and targeted guidance, leading to more effective learning and improved performance on complex reasoning tasks.

Earlier studies (Lightman et al., 2024) demonstrated that process-based reward models outperform outcome-based reward models when providing feedback for mathematical reasoning tasks. Building on this, GRACE (Khalifa et al., 2023) and DIVERSE (Li et al., 2023) propose training critic models capable of selecting optimal intermediate reasoning steps as feedback. Furthermore, SALAM (Wang & Li, 2023) and REFINER (Paul et al., 2024) explore the idea of training models that generate error analyses in natural language form as feedback, which can be used to refine reasoning steps iteratively. All these approaches involve training task-specific critique models, enabling them to fully leverage the available training data for specific reasoning tasks. As a result, they achieve significantly improved feedback quality and corrective effectiveness compared to relying on general-purpose LLMs for feedback.

The advantages of this type of method include the ability to deliver highly targeted and actionable feedback, resulting in improved reasoning accuracy and reduced error propagation. Additionally, tailoring critic models to specific tasks allows for better utilization of task-relevant data and context. However, such approaches also present disadvantages, including the need for large amounts of high-quality annotated data and increased training complexity, as well as reduced generalizability to new tasks or domains due to their task-specific nature.

4.3. Problem decomposition reasoning

When solving complex problems, decomposing a problem into multiple simpler or more detailed sub-problems is an important strategy employed by human. The general process of problem decomposition-related methods is decomposing a complex problem into several simpler sub-problems. These sub-problems are then solved one by one. Finally, the answers to the sub-problems are combined to obtain the answers to the original problem. When the task is complex or the individual reasoning steps are hard to learn, this method often yields superior results. However, when dealing with simple problems, further decomposing them may not be very meaningful and increase time overhead. There are two types of problem decomposition methods: progressive decomposition and pre-decomposition methods, as illustrated in Fig. 9(c).

4.3.1. Progressive decomposition

As shown in Fig. 9 (c.1), progressive decomposition methods alternately decompose the questions and reason for the sub-questions step-by-step. At each step, either the LLMs or an external controller identifies the next sub-problem to solve, generates intermediate answers or reasoning steps, and iteratively updates the context until all necessary components for the final answer are complete. This approach leverages recursive or stepwise reasoning, allowing LLMs to tackle complicated tasks by focusing on one manageable part at a time.

For instance, Succ_Prompting (Dua et al., 2022) iteratively decomposes the complex question into the following simple question to answer and then repeats until the complex question is answered. Scratchpad (Nye et al., 2021) allows the model to produce an arbitrary sequence of intermediate tokens, which it calls a scratchpad, before producing the final answer. For example, on addition problems, the scratchpad contains the intermediate results from a standard long addition algorithm. Self-ask (Press et al., 2023) asks itself follow-up questions before answering the initial question, which will narrow the compositionality gap that models can correctly answer all sub-problems but not generate the overall solution.

The main advantage of progressive decomposition methods lies in their interpretability and improved performance on complex, multi-step reasoning tasks, as they make the reasoning process explicit and modular. However, they can be less efficient due to increased computational overhead, risk of error propagation across steps, and sensitivity to the quality of each intermediate sub-question and answer. These trade-offs make them powerful but sometimes less practical for real-time or resource-constrained applications.

4.3.2. Pre-decomposition

As displayed in Fig. 9 (c.2), unlike the progressive decomposition methods, the pre-decomposition methods decompose the questions before reasoning. these methods generally involve an initial step where the complex question is broken down into multiple easy sub-questions or a structured plan before reasoning. This pre-processing can be performed via expert-curated templates, parsing algorithms, or prompting strategies. Once decomposition is complete, each sub-task or sub-question is addressed either sequentially or in parallel.

Specifically, PS+ (Wang et al., 2023a) devises a plan to divide the entire task into smaller subtasks and then carries out the subtasks according to the plan. Least-to-most (Zhou et al., 2023) breaks down a complex problem into a series of simpler subproblems and then solves them in sequence. To promote the practicality of Least-to-most, Dynamic Least-to-most (Drozdov et al., 2023) obtain the problem reduction via a multi-step syntactic parse of the input. Furthermore, it dynamically selects exemplars from a fixed pool such that they collectively demonstrate as many parts of the decomposition as possible. Decomp (Khot et al., 2023) argues that few demonstrations of the complex task are not sufficient for current models to learn to perform all necessary reasoning steps as tasks become more complicated. Hence, it solves complex tasks by instead decomposing them into simpler sub-tasks and delegating these to sub-task specific LLMs, with both the decomposer and the sub-task LLMs having their own few-shot prompts. SoT (Ning et al., 2023) guides the LLM to derive a skeleton first by itself. Based on the skeleton, the LLMs then complete each point in parallel.

The main advantage of pre-decomposition methods is their ability to simplify complex reasoning tasks, often improving accuracy and model interpretability by breaking down and clarifying the problem structure up front. They can also facilitate modularity and reuse of solutions for recurring sub-tasks. However, disadvantages include reliance on the quality of the initial decomposition; errors at this stage can propagate or confound downstream reasoning. Additionally, some complex or ambiguous problems may be difficult to decompose effectively without significant domain knowledge or sophisticated heuristic rules.

4.4. Assemble-based reasoning

The core idea of assemble-related methods is that a complex reasoning problem typically admits multiple different ways of thinking, leading to its unique correct answer (Wang et al., 2023). As shown in Fig. 9(d), typical answer assemble methods first generate multiple different rationales with answers and then choose the most consistent one as the final answer. Furthermore, a few rationale assemble methods try to leverage the difference between multiple reasoning processes to enhance reasoning performance. The assemble-related methods demonstrate excellent performance. Moreover, they can be easily integrated with other classes of methods, such as CoT-series. However, due to the need to generate multiple reasoning processes, the overhead of this class of methods is relatively high.

4.4.1. Answer assemble

As shown in Fig. 9 (d.1), methods in this category first generate multiple rationales. These rationales can be either explicitly solicited from the LLMs through creative prompting or generated via independent inference calls. Then, aggregation strategies such as voting are then employed to select the most reliable or correct answer from the pool of generated rationales. Underlying these approaches is the theory that increasing diversity among rationales exposes the LLMs to a broader space of solutions, reducing the likelihood of systematic errors and improving overall answer quality.

Self-consistency (Wang et al., 2023) first samples a diverse set of rationales instead of only taking the greedy one and then selects the most consistent answer by marginalizing out the sampled reasoning paths. Instead of voting among all rationales, Complex-consistency (Fu et al., 2023) votes among top-K complex rationales with more steps. To leverage variations of the input prompt to introduce the diversity needed for assembling, DIV-SE (Naik et al., 2024) automatically improves prompt diversity by soliciting feedback from the LLM to ideate approaches that are apt for the problem. Then, it assembles the diverse prompts across multiple inference calls. To reduce the inference costs of DIV-SE, IDIV-SE (Naik et al., 2024) combines all approaches within the same prompt and aggregates all resulting outputs to leverage diversity. Similar to DIV-SE (Naik et al., 2024), DIVERSE (Li et al., 2023) proposes to increase the diversity of rationales by sampling from a single prompt and varying the prompt itself. It first uses a verifier to score the quality of each rationale and guide the voting mechanism. Then, it assigns a fine-grained label to each step of the reasoning path and uses a step-aware verifier to attribute the correctness or wrongness of the final answer to each step.

The primary advantage of these methods is their resilience to individual errors and biases, as aggregating diverse answers tends to yield more robust and accurate answers. However, a significant disadvantage is the increased computational and inference cost associated with generating answers repeatedly. Additionally, the effectiveness of the aggregation mechanism relies heavily on the quality and diversity of the sampled rationales.

4.4.2. Rationale assemble

As shown in Fig. 9 (d.2), answer assemble methods generate multiple rationales and aggregate the final answers using a voting mechanism that disregards the information contained in the intermediate steps. In contrast, rationale assemble methods focus on leveraging these intermediate steps to enhance reasoning. Specifically, they first generate multiple rationales or reasoning chains for a given question, extract their intermediate reasoning steps, and then aggregate these steps to form a comprehensive pool of evidence.

To provide a unified explanation for the predicted answer, MCR (Yoran et al., 2023) focuses on rationale assemble, which leverages the relations between intermediate steps across multiple rationales. MCR mixes information between multiple relations and selects the most relevant facts to generate an explanation and predict the answer. Unlike answer assemble methods, sampled

rationales are used not for their predictions but to collect evidence from multiple rationales. MCR concatenates the intermediate steps from each rationale into a unified context, passed to a meta-reasoner model along with the original question. The meta-reasoner model prompts to meta-reason on multiple rationales and produces a final answer with an explanation. In this way, MCR could combine facts from multiple chains to produce the final answer with an explanation of the answer's validity.

The main advantage of rationale assemble methods is their robustness, as aggregating over multiple rationales often leads to better answer accuracy and reduced risk of individual path errors. However, a significant drawback is their lack of interpretability, as they do not provide unified explanations, making it difficult for users to trust or understand the prediction process. Furthermore, this type of method is also quite costly, as it requires multiple generations of rationales.

4.5. Multi-agent reasoning

Multi-agent reasoning draws inspiration from *society of minds* concepts (Minsky, 1988) found in multi-agent systems. In contrast to single-agent methods, such as CoT and ToT, multi-agent reasoning methods emphasize the diversity of ideas and the importance of communication, adversarial interaction, and collaboration among multiple agents, as illustrated in Fig. 9(e). In the reasoning stage, multi-agents express their individual viewpoints and interact in various ways (such as through debate, collaboration, and community communication) to arrive at a final solution. The divergent thinking of multi-agent reasoning determines that (i) the distorted thinking of one agent can be rectified by other agents, (ii) the supplementation of one agent's resistance to change by others, and (iii) the reception of external feedback by each agent from others.

A limitation of multi-agent reasoning is that it requires more time cost, as agents often need to participate in multiple rounds of interaction to present and counter arguments. Additionally, current LLM-based agents may struggle to maintain coherence and relevance in long-context scenarios, leading to potential misunderstandings and context loss. Enhancing the long-text modeling capabilities of large language models remains challenging for future research. Multi-agent reasoning methods can be categorized into multi-agent competition and multi-agent collaboration.

4.5.1. Multi-agent competition

As demonstrated in Fig. 9 (e.1), multiple LLM-based agents conduct independent thinking and reasoning in a multi-agent competition way. When agents hold differing opinions, they examine each other's responses and adapt their answers accordingly. Through several rounds of adversarial interaction, the multi-agent system ultimately reaches a final answer that satisfies the internal logic of each agent while aligning with the feedback from other agents. This structured communication iteratively drives the agents toward a consensus or a more robust solution, leveraging diverse agent perspectives and the ability to self-correct through dialogue.

Recently, abundant multi-agent competition methods are designed to explored how to unleash the potential of multi-agent systems. For instance, Multi-Agent Debate (Y. Du et al., 2024) is a role-symmetric multi-agent competition framework where different agents engage in spontaneous discussion. MAD (Liang et al., 2024) introduced different roles, such as judges and debaters, into the debate process. This represents a role-asymmetric multi-agent debate architecture. RECONCILE (J. Chen et al., 2024) facilitates deeper multi-agent discussions by introducing confidence assessments and persuasive explanations in the form of roundtable meetings. ChatEval (Chan et al., 2024) adopts three distinct communication strategies within its diversified role communication process: one-on-one, simultaneous-talk, and simultaneous-talk-with-summarizer. Additionally, research evaluating multi-agent competition has shown that, as demonstrated by CMD (Q. Wang et al., 2024), effective prompt engineering can enable a single agent to achieve performance comparable to multi-agent discussions. However, multi-agent discussions have a distinct advantage in contexts lacking examples, and discussions involving multiple LLMs can enhance the performance of weaker LLMs.

A key advantage of multi-agent competition methods is their ability to enhance the reliability, robustness, and creativity of generated answers by harnessing collective intelligence and adversarial scrutiny. These methods can mitigate individual agent biases and reduce hallucinations, thereby improving overall performance. This advantage is particularly evident in scenarios where labeled examples are scarce or strong single models are unavailable. However, disadvantages include increased computational cost, greater complexity in coordination, and challenges in designing effective interaction protocols. There also remains a risk of convergence on suboptimal answers if agents overly align or fail to challenge flawed reasoning.

4.5.2. Multi-agent collaboration

As shown in Fig. 9 (e.2), multi-agent LLM collaboration involves agents working together cooperatively to solve a given problem. In general, multi-agent LLM collaboration methods follow a pipeline consisting of problem abstraction, agent role assignment, and iterative interaction among agents via communication protocols or decision-making mechanisms. Theories underlying these methods are often inspired by social or cognitive sciences, such as group problem-solving, debate, and self-reflection. Collaboration may involve expert recruitment, structured debates, or reflective feedback loops, and can be guided by optimization algorithms to improve cooperation and solution quality.

AGENTVERSE (Chen et al., 2024) simulates the problem-solving process of human groups through mechanisms such as expert recruitment, collaborative decision-making, and tool utilization. FORD (Xiong et al., 2023) explores the issue of mutual consistency among multiple LLMs by introducing a formalized debate mechanism, illuminating both the potential and challenges inherent in LLM collaboration. MACM (Lei et al., 2024) first abstracts the conditions and objectives of a problem, then employs a multi-agent interaction system to iteratively uncover new conditions that facilitate the achievement of the goals, ultimately solving the problem. COPPER (Bo et al., 2024) enhances the collaborative capabilities of multi-agent systems based on LLMs through a self-reflection mechanism. This framework involves training a shared reflector and utilizes a counterfactual proximal policy optimization (PPO)

mechanism to optimize the quality of reflections. SMoA (Li et al., 2025) introduces sparse to optimize the fully connected structures commonly found in traditional multi-agent methods, thereby balancing performance and computational cost. Social_Agent (J. Zhang et al., 2024) explores the collaboration mechanisms among agents and analyzes these mechanisms from a social psychology perspective.

The advantages of multi-agent LLM collaboration include enhanced problem-solving abilities through specialization and diverse perspectives, as well as increased robustness via collective intelligence. However, these methods also face significant challenges, such as increased computational cost, potential for inconsistency or conflict among agents, and the added complexity of coordination and communication, which demands careful design to ensure efficiency and reliability.

4.6. Recent breakthroughs in reasoning-enhanced large language models and their implications

In recent years, the field of artificial intelligence has witnessed a paradigm shift from large language models to large reasoning models (LRMs), such as OpenAI-o1 (OpenAI, 2024) and DeepSeek-R1 (Guo et al., 2025). While traditional LLMs primarily focused on pattern recognition and linguistic coherence, LRMs aim to achieve structured and goal-directed thinking. This transformation reflects a broader ambition to create reasoning systems that not only generate text fluently but also reason, plan, and adapt in a human-like manner.

Recent breakthroughs in LRMs have emerged from several architectural and methodological innovations. One key development is the integration of modular thinking processes that allow the model to decompose complex tasks into intermediate steps. Instead of producing answers directly or eliciting rationale by prompting (Kojima et al., 2022), LRMs constructs internal reasoning chains that can be verified or refined through self-evaluation mechanisms. Another major improvement comes from reinforcement learning through reasoning feedback (Zhang et al., 2025). Unlike previous reinforcement learning from human feedback, which focused on surface-level responses, the reasoning-based variant rewards the internal consistency and logical validity of intermediate steps. This approach significantly enhances the ability of models to solve complex and multi-hop questions. As a result, LRMs demonstrate higher performance in complex tasks such as scientific question answering, logical inference, and theorem proving.

The shift toward reasoning-centric paradigms has profound implications for the generalization capabilities of AI systems. Early LLMs often failed to generalize beyond their training distribution, especially in domains that required explicit reasoning or multi-step planning. In contrast, LRMs exhibit improved transfer learning capabilities. When faced with novel problems, they are able to reconstruct solution pathways based on abstract reasoning principles rather than relying solely on memorized patterns. Furthermore, recent LRMs have shown the ability to engage in strategic reasoning and decision making (Li et al., 2025). This progress suggests a growing potential for models to function as autonomous agents capable of performing tasks that require foresight, such as scientific hypothesis generation, software debugging, or complex negotiation.

Although impressive in reasoning ability, recent studies (Wang et al., 2025; Zhou et al., 2025) find that LRMs face serious safety challenges. The study finds that open-source models like DeepSeek-R1 are less safe than proprietary ones, with weaker defenses against adversarial attacks and higher risks of producing harmful or misleading content. Their reasoning process itself can generate unsafe thoughts even when the final answer appears safe. Moreover, training that enhances reasoning ability or distills models often reduces safety alignment, making stronger safeguards and safety-focused training essential. Hence, how to balance the reasoning capability and safety is a promising and urgent research direction for LRMs.

The recent breakthroughs in LRMs mark a significant milestone in the evolution of artificial intelligence. They are not merely extensions of LLMs but represent a new class of cognitive architectures designed for reasoning, planning, and self-correction. Their implications extend far beyond text generation, influencing the future of research, automation, and human-machine collaboration. Continued progress in this domain will depend on advances in model interpretability, alignment, and evaluation methodologies, ensuring that reasoning capabilities evolve responsibly and beneficially.

4.7. The datasets of reasoning based on parametric knowledge bases

In this section, we select, summarize, and analyze datasets related to the reasoning task based on parametric knowledge bases. We statistically summarize the information of the related dataset from multiple dimensions, including (1) Task Type: The specific task associated with each dataset; (2) # Ques.: Question number; (3) Q&A source: The main construction methods of questions and answers. They are mainly divided into three categories: “Generate”, “Expert”, and “Crowdsourcing”. “Generate” refers to the design of programs for automated generation, “Expert” refers to direct crawling from professional websites or carefully designed by domain experts, and “Crowdsourcing” refers to completion by crowdsourcing workers with general cultural levels; (4) Rationale: Whether they contain rationales; (5) Answer type: The form of answers; (6) Links: The storage address of the datasets. The statistical results are shown in Table 8.

In contrast to reasoning methods based on symbolic knowledge bases, reasoning methods that rely on LLMs over parametric knowledge bases reveal unique dependencies on dataset characteristics. The effectiveness of these methods depends less on the overall volume of training data and more on the quality, diversity, and representativeness of the examples that shape in-context reasoning. As a result, dataset size plays only a minor role in determining performance, whereas the breadth of coverage across question types and reasoning patterns becomes the key factor influencing generalization. When models are trained or prompted with narrowly distributed or task-specific data, their reasoning ability tends to remain confined to limited domains. Conversely, datasets that integrate examples from multiple reasoning contexts such as arithmetic, symbolic, and commonsense reasoning foster greater robustness and adaptability. Recent research further suggests that balanced inclusion of examples from diverse domains enhances reasoning consistency and reduces overfitting to repetitive question structures. These findings indicate that thoughtful control of dataset composition is central to the effective selection and development of parametric reasoning methods.

Table 8
Dataset statistics of reasoning based on parametric knowledge bases.

Datasets	Task Type	# Ques.	Q&A source	Rationale	Answer type	Links
AQUA (Ling et al., 2017)	Arithmetic Reasoning	254	Generate	✓	Option	https://
GSM8K (Cobbe et al., 2026)	Arithmetic Reasoning	1319	Crowdsourcing	×	Number	https://
SVAMP (Patel et al., 2021)	Arithmetic Reasoning	1000	Generate	✓	Number	https://
AddSub (Hosseini et al., 2014)	Arithmetic Reasoning	395	Expert	✓	Number	https://
MultiArith (Roy & Roth, 2015)	Arithmetic Reasoning	600	Expert	✓	Number	https://
SingleEq (Koncel-Kedziorski et al., 2015b)	Arithmetic Reasoning	508	Expert	✓	Number	https://
Last Letters (Wei et al., 2022)	Symbolic Reasoning	500	Generate	×	String	https://
Coin Flip (Wei et al., 2022)	Symbolic Reasoning	500	Generate	×	Yes/No	https://
CommonsenseQA (Talmor et al., 2019)	Commonsense Reasoning	1221	Crowdsourcing	×	Option	https://
StrategyQA (Geva et al., 2021)	Commonsense Reasoning	2290	Crowdsourcing	✓	Yes/No	https://

Table 9
Comparison of methods for reasoning based on parametric knowledge bases. Key dimensions: Performance (reasoning accuracy), Efficiency (computational efficiency and resource consumption), Interpretability (reasoning transparency), Robustness (stability under out-of-distribution, data noise, or data sparsity scenarios).

Method	Performance	Efficiency	Interpretability	Robustness
CoT Optimization	Moderate *Detailed reasoning process*	Moderate *High context overhaed*	High *Clear reasoning steps*	Low *Heavy demonstration dependence*
CoT Engineering	High *Strong complex problem solving ability*	Low *Excessive LLM calls*	High *Transparent reasoning paths*	High *Broad thought exploration*
Automatic CoT	Low *Unstable demonstration*	High *Efficient automated annotation*	Moderate *Variable reasoning paths*	Low *Severe error amplification*
Feedback from LLMs	Moderate *Slef-limited performance*	Low *Substantial iterative overhead*	High *Explicit correction steps*	Moderate *Moderate Error propagation risk*
Feedback from tools	High *Specialized tool feedback*	Low *High tool call overhead*	High *Explicit feedback mechanisms*	Moderate *Moderate tool reliability sensitivity*
Feedback from trained critique model	High *Effective targeted refinement*	Low *High training cost*	Low *Opaque critique logic*	Low *Limited task applicability*
Progressive Decomposition	High *Effective step-wise decomposition*	Low *Significant iterative overhead*	High *Excellent modular transparency*	Low *Pronounced error propagation*
Pre-decomposition	Moderate *Reasonable decomposition effectiveness*	Moderate *Scalable parallel processing*	High *Logical sub-question structure*	Low *Significant error propagation*
Answer Assemble	High *Diversity-enhanced performance*	Low *Redundant generation costs*	High *Comprehensive explanation sets*	High *Robust bias mitigation*
Rationale Assemble	High *Compelling aggregated evidence*	Low *Demanding rationale generation*	Low *Fragmented explanation structure*	High *Strong error tolerance*
Multi-agent Competition	High *Synergistic collective intelligence*	Low *Comprehensive agent interactions*	High *Transparent dialogue logs*	High *Powerful adversarial correction*
Multi-agent Collaboration	High *Expertise-driven complementation*	Low *High communication costs*	High *Explicit collaborative dialogues*	High *Reliable mutual correction*

4.8. The comparison of reasoning methods based on parametric knowledge bases

In this section, we compare various reasoning methods based on parametric knowledge bases. The corresponding results are presented in Table 9.

Most methods demonstrate strong performance, particularly those utilizing external feedback (Lightman et al., 2024), decomposition (Khot et al., 2023), or collaborative agents (Y. Du et al., 2024). However, this often results in low efficiency caused by multiple iterations and communication overhead. While interpretability remains a shared advantage through transparent reasoning steps for most methods, robustness varies significantly. Approaches like CoT engineering (Zheng et al., 2024) and collaborative systems (Lei et al., 2024) ensure high robustness via mutual error correction, whereas others (Li & Qiu, 2023; J. Zhou et al., 2025) face error propagation and limited generalizability.

Overall, systems leveraging agent collaboration achieve a favorable balance of performance, interpretability, and robustness at the expense of computational efficiency. Simpler techniques improve efficiency but are susceptible to instability. This highlights a fundamental dilemma where robust and interpretable methods require substantial computational resources.

4.9. The summary of reasoning based on parametric knowledge bases

In summary, parametric reasoning methods show strong performance in various reasoning tasks, including mathematical, commonsense, and symbolic reasoning. Despite these achievements, important challenges remain. One challenge is reproducibility. The results of LLMs often vary with prompt design, randomness, and context. Researchers address this by fixing random seeds, standardizing benchmarks, and sharing full experimental settings. Furthermore, hallucination is a major issue, as LLMs sometimes

generate out-of-data yet incorrect statements. Integrating external symbolic knowledge sources can help reduce such errors. In the next section, we provide a comprehensive discussion of this topic.

Overall, LLM-based reasoning demonstrates strong potential for general and interpretable reasoning. Methods that incorporate feedback, collaboration, or decomposition tend to achieve high performance and robustness, though they require more computational resources. Balancing accuracy, interpretability, and efficiency remains a key direction for future work in reasoning based on parametric knowledge bases.

4.10. Comparative analysis between symbolic and parametric reasoning

Across the symbolic and parametric reasoning paradigms, each exhibits distinct advantages and shortcomings shaped by their underlying representations of knowledge. Symbolic reasoning, grounded in structured and explicit knowledge bases such as knowledge graphs, excels in tasks that demand logical consistency, transparent rule application, and fine-grained control over reasoning processes. Conversely, parametric reasoning is anchored in the distributed representations of pre-trained language models. It excels in tasks that require contextual understanding and abstraction. Moreover, it can generalize beyond explicitly encoded facts.

In particular, symbolic reasoning methods perform strongly in domains where knowledge is explicit, structured, and logically interdependent. However, their reliance on predefined symbolic structures constrains adaptability. When confronted with incomplete or noisy data, symbolic systems often struggle to infer plausible conclusions beyond their formalized scope. For example, static or temporal knowledge graph reasoning systems can struggle when new entities or relations appear that were absent during schema construction, leading to degraded robustness and poor out-of-distribution generalization. The deterministic nature that ensures transparency also restricts flexibility, making symbolic reasoning brittle in open-ended or ambiguous linguistic contexts.

Parametric reasoning methods, by contrast, excel in capturing implicit associations and statistical regularities from large-scale text corpora. LLMs, equipped with massive parameter spaces, can infer relations that are not explicitly defined but statistically implied, such as analogical reasoning or contextual interpretation in commonsense and mathematical reasoning. This allows them to generalize well in scenarios where symbolic systems lack coverage, particularly in open-domain question answering or multi-step reasoning with incomplete data. However, LLMs frequently produce hallucinations, where plausible but factually incorrect statements emerge because their reasoning relies on probabilistic token prediction rather than strict logical inference. Moreover, the parametric knowledge encoded within LLMs may become outdated over time, as it is implicitly learned from static pre-training data. Updating such knowledge is far more costly than refreshing symbolic knowledge bases. This limitation makes parametric reasoning less reliable in domains that evolve rapidly or demand real-time factual accuracy.

Concrete examples further clarify these distinctions. In arithmetic reasoning, fine-tuned small language models or rule-based symbolic solvers outperform LLMs due to their explicit procedural reasoning and exact logical grounding. Conversely, in commonsense reasoning or narrative comprehension, LLMs significantly outperform symbolic methods because they can infer unstated background knowledge from learned textual patterns. Similarly, in temporal reasoning, symbolic models with rule-based temporal constraints (e.g., temporal KGs) ensure consistent causal and chronological order, while LLMs may produce temporally inconsistent outputs despite their linguistic fluency. In multi-modal reasoning, symbolic fusion methods achieve clarity and efficiency by explicitly aligning entity relations across modalities. However, they fall short when modality interactions are subtle or context-dependent. LLMs with multimodal extensions (e.g., vision-language transformers) better capture such nuanced dependencies, though with higher computational cost.

On closed-world knowledge graph benchmarks such as FB15k-237 and WN18RR, symbolic or rule-augmented methods consistently outperform embedding-only models, achieving absolute gains of 6%–12% in Hits@1 (Han et al., 2018; Lv et al., 2020). However, their performance degrades by more than 20% in inductive settings with unseen entities or relations, reflecting limited adaptability to open-world scenarios (Teru et al., 2020). In contrast, powerful LLMs such as gpt-4o substantially outperform symbolic pipelines on open-domain and commonsense benchmarks, obtaining approximately 80%–85% accuracy on CommonsenseQA and StrategyQA, compared to accuracy levels typically below 70% for symbolic approaches (Geva et al., 2021; Talmor et al., 2019). Despite parametric reasoning being considered efficient for mathematical reasoning, analyses on GSM-Symbolic (Mirzadeh et al., 2025) show that LLMs remain fragile in this task. Simple numerical perturbations or the addition of irrelevant clauses can reduce accuracy by up to 65%, with performance dropping from about 87% on GSM8K to below 80%, and further to around 42% on harder variants, indicating reliance on surface pattern matching rather than robust reasoning. These findings motivate the combination of symbolic reasoning and parametric reasoning jointly improve robustness and generalization.

Overall, the two paradigms reveal a complementary relationship rather than a competitive one. Symbolic reasoning contributes precision, transparency, and formal validity, whereas parametric reasoning contributes flexibility, generalization, and semantic richness. Symbolic approaches are preferable in scenarios requiring traceable inference, such as compliance checking, scientific discovery, or structured decision-making. Parametric approaches are advantageous in open-domain understanding, adaptive reasoning, and natural language interaction, where context and uncertainty dominate. Yet, neither paradigm alone achieves the balance of accuracy, interpretability, and efficiency demanded by real-world reasoning systems. This motivates the rise of collaborative reasoning frameworks, which can integrate symbolic rigor with parametric adaptability. In this way, they combine the verifiability of structured reasoning with the fluency and coverage of neural representations.

5. Collaborative reasoning based on symbolic and parametric knowledge bases

In this section, we are devoted to investigating collaborative reasoning methods based on symbolic and parametric knowledge bases. These methods mainly perform reasoning in the form of question answering. Generally, given a knowledge-intensive question, this type of method is required to leverage the knowledge in symbolic and parametric knowledge bases collaboratively to reason for the correct answer.

Based on the structure of the symbolic knowledge bases, these question answering tasks can be categorized into graph-based reasoning, table-based reasoning, text-based reasoning, and heterogeneous reasoning. The symbolic knowledge bases store knowledge in the form of structured graphs, structured tables, and unstructured text in graph-based, table-based, and text-based reasoning tasks, respectively. In particular, heterogeneous reasoning investigates how to leverage symbolic knowledge from multiple heterogeneous symbolic knowledge bases, such as KGs, tables, and text. The overall taxonomy of collaborative reasoning methods based on symbolic and parametric knowledge bases is shown in Fig. 10. Furthermore, Fig. 11 depicts the core technology and pipeline of each type of collaborative reasoning method based on symbolic and parametric knowledge bases.

5.1. Graph-based reasoning

The tasks of graph-based reasoning include knowledge graph question answering (KGQA) and temporal knowledge graph question answering (TKGQA), where symbolic knowledge is stored in SKGs and TKGs, respectively.

5.1.1. Knowledge graph question answering

Task definition: Given a question and an SKG, KGQA methods are required to understand the intent of the question via the parametric knowledge bases and retrieve the entity nodes from the SKG as answers.

KGQA methods can be categorized into two classes: semantic parsing-based (SP-based) methods and information retrieval-based (IR-based) methods (Liang et al., 2024), as illustrated in Fig. 11(a). SP-based methods aim to parse the questions into the logical forms (such as SPARQL, S-expression and query graph) to yield the correct answer. IR-based methods construct a question-specific subgraph of the SKG and retrieve the most matching answers. In recent years, most methods utilize PLMs to integrate a substantial amount of external knowledge. Due to the introduction of parametric knowledge, the level of intelligence has been significantly enhanced, leading to considerable improvements in accuracy and task versatility.

(1) Semantic parsing-based methods: As shown in Fig. 11 (a.1), Semantic parsing (SP)-based methods first convert a natural language question into a logical form, which aligns with the existing knowledge in the given SKG. Then, they executing the logical form on a SKG to derive the final answer. Early SP-based methods are limited to independent and identically distributed scenarios and exhibited poor performance in solving problems requiring commonsense knowledge, demonstrating relatively limited intelligence. Consequently, recent SP-based methods focus on the paradigm shift from traditional independent and identically distributed scenarios to leveraging PLMs for enhanced semantic understanding and logical form generation.

The structural commonality among SP-based methods is the transformation of the rigid question-to-logical-form conversion process into flexible approaches that either fine-tune PLMs or directly utilize LLMs. The general pipelines of SP-based methods typically follow a structured approach: question comprehension, logical form generation, knowledge alignment, and execution. This structural evolution enables SP-based methods to overcome limitations in handling commonsense knowledge and understanding diversified natural language questions, while effectively bridging the gap between parametric general knowledge and symbolic professional knowledge.

For example, KQA Pro (Cao et al., 2022) introduces a compositional and interpretable programming language KoPL to represent the reasoning process of complex questions. It fine-tunes BART (Lewis et al., 2020) to achieve compositional reasoning. CBR-KBQA (Das et al., 2021) uses ROBERTA-base (Liu et al., 2019) to encode each question independently and generate a logical form for a new question by retrieving cases that are relevant to it with the pre-trained ROBERTA-base weights. RnG-KBQA (Ye et al., 2022) introduces T5 (Raffel et al., 2020) to construct the final logical form based on the questions and the high-ranked candidate logical forms, demonstrating excellent performance even when dealing with questions involving unseen schema items. TrackerQA (Gao et al., 2024) encodes knowledge through graph transformation layers with directed message-passing control and employs a question-aware attention mechanism to predict the exact BGP paths.

Several SP-based methods utilize LLMs to parse natural language questions into logical forms in a few-shot in-context learning setting. For example, KB-BINDER (Li et al., 2023) generates drafts with LLM as preliminary logical forms and then binds the entities, relations, and schema items of the drafts to SKG iteratively. FlexKBQA (Li et al., 2024) introduces a self-training approach with execution guidance, using the LLM to convert logical forms into natural language questions and utilizing unlabeled user questions iteratively.

In addition, fine-tuning LLM also makes sense. For example, ChatKBQA (Luo et al., 2024) proposes generating the logical form with fine-tuned LLMs first, then retrieving and replacing entities and relations through an unsupervised retrieval method. GAIL-Finetune (Zhang, Wen, & Zhao, 2024) fine-tunes Llama-2-7B to produce expert-level sample and evaluate the authenticity and relevance of the sequences to tackle the challenges in low-resource KGQA scenario. Triad (Zong et al., 2024) utilizes an LLM-based agent with three different roles for KBQA tasks.

Overall, the primary advantage of SP-based methods is their ability to leverage PLMs and LLMs, which enhances semantic understanding and improves their handling of commonsense reasoning, leading to superior performance on complex questions. However, these methods come with disadvantages: they introduce a significant knowledge gap between parametric general

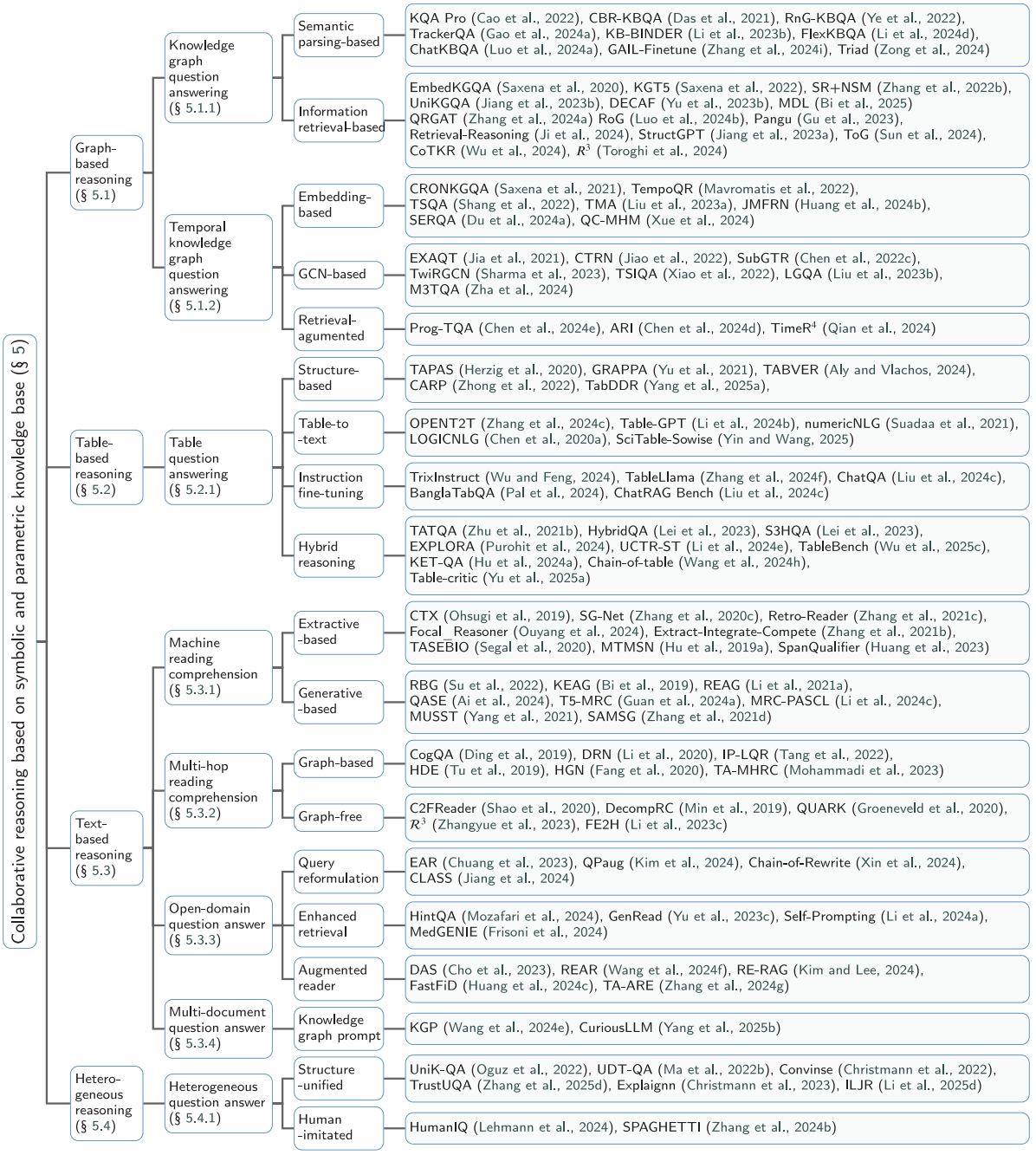

Fig. 10. Taxonomy of collaborative reasoning across four tasks: graph-based, table-based, text-based reasoning, and heterogeneous reasoning. Each task may comprise several sub-tasks, and the representative methods corresponding to each task or sub-task are displayed in the green box. Abbreviation: graph convolutional network (GCN). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

knowledge and symbolic professional knowledge. This gap can result in reasoning errors. The potential for mistakes in logical form generation and difficulties in handling unseen schema items create additional reliability concerns. Furthermore, the substantial computational cost of fine-tuning large models and the need for extensive training data pose practical limitations that may hinder deployment in resource-constrained environments. Balancing semantic sophistication with operational efficiency remains a key challenge.

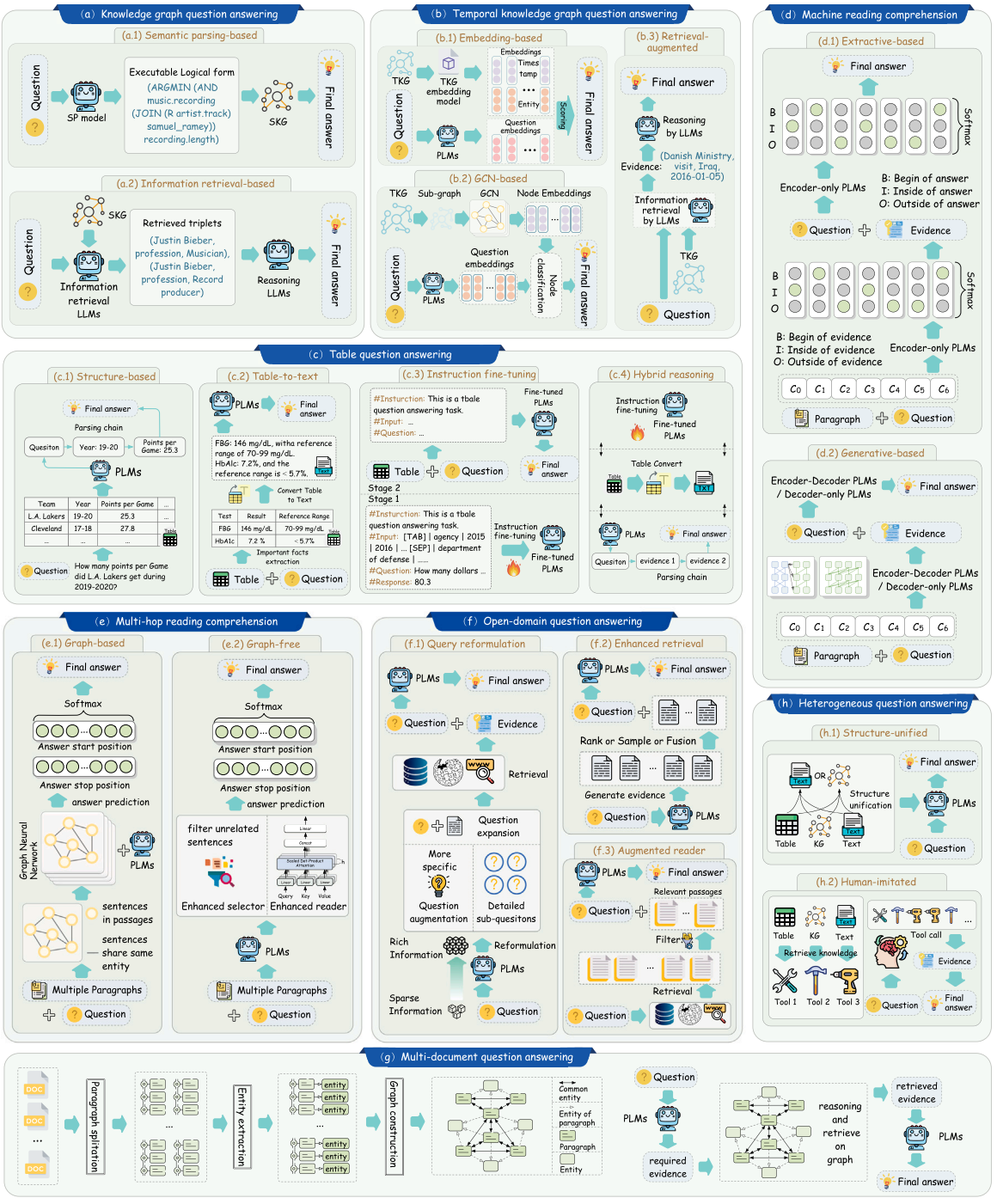

Fig. 11. Collaborative reasoning pipeline for seven tasks: (a) KGQA (Section 5.1.1), (b) TKGQA (Section 5.1.2), (c) Table QA (Section 5.2.1), (d) MRC (Section 5.3.1), (e) MHRC (Section 5.3.2), (f) ODQA (Section 5.3.3), and (g) Heterogeneous QA (Section 5.4.1). Abbreviation: semantic parsing (SP), static knowledge graph (SKG), information retrieval (IR), pre-trained language models (PLMs), temporal knowledge graph (TKG), and graph convolutional network (GCN).

(2) Information retrieval-based methods: As shown in Fig. 11 (a.2), information retrieval (IR)-based methods first retrieve question-relevant triplets from the SKG. These triplets, along with the question, are then fed into reasoning LLMs to reason about the final answer. Early IR-based methods perform poorly because they lack sufficient understanding of user questions and the guidance of common-sense knowledge during the reasoning process. Hence, recent advanced IR-based methods generally apply PLMs to model

questions and subgraphs of the SKG. The structural commonalities of these methods center on their paradigm shift from traditional node classification to leveraging PLMs and integrating with retrieval-augmented generation (RAG) frameworks. The general pipelines of IR-based methods typically involve three key stages: question understanding, evidence retrieval, and answer extraction.

For example, EmbedKGQA (Saxena et al., 2020) has been proposed to handle SKG sparsity, where ComplEx (Trouillon et al., 2016) embeddings are trained to represent SKG elements and RoBERTa (Liu et al., 2019) embedding is used to represent the question. KGT5 (Saxena et al., 2022) considers both SKG reasoning and KGQA as sequence-to-sequence tasks, where a simple Transformer that has the same architecture as T5-small (Raffel et al., 2020) has been trained to achieve excellent performance. SR+NSM (Zhang et al., 2022) utilizes RoBERTa (Liu et al., 2019) to encode the question and relations in SKG iteratively to expand paths. Then it could construct subgraph with low size but high answer coverage to find answers. UniKGQA (Jiang et al., 2023b) combines a PLM with an ultra-simple GNN to transfer the retrieved knowledge to the reasoning phase. DECAF (Yu, Zhang. et al., 2023) constructs retriever and reader based on FiD-large retriever (Izacard & Grave, 2021b) to generate both logical forms and direct answers jointly. Recently, MDL (Bi et al., 2025) introduce a modular dual learning framework to significantly boosts the performance of both knowledge graph question answering and question generation. Moreover, several methods, such as QRGAT (Zhang et al., 2024), leverage PLMs to encode questions while employing GNNs to model SKGs, thereby enabling a unified graph reasoning process for improved evidence retrieval and answer extraction.

Recently, several methods have adopted the retrieval-augmented generation (RAG) paradigm because LLMs can simultaneously model both user questions and SKG elements to perform simple deductive reasoning. For example, RoG (L. Luo et al., 2024) proposes a planning-retrieval-reasoning framework, which fine-tunes LLaMA2-Chat-7B with relation paths and valid reasoning paths in SKGs. Then it can generate reasoning paths for faithful reasoning. Similarly, Retrieval-Reasoning (Ji et al., 2024) decomposes the problem into retrieval and reasoning modules and then fine-tunes LLM at three levels: entity, relation, and graph. Pangu (Gu et al., 2023) leverages the discriminative capabilities of the LLM for context-based language understanding. The symbolic agent explores SKG to construct effective plans incrementally, and the LLM agent evaluates the reasonableness of candidate plans to guide the search process. StructGPT (Jiang et al., 2023a) solves KGQA based on structured data, where the facts in SKG could be linearized into LLM to reason naturally. ToG (Sun et al., 2024) treats LLM as an agent capable of exploring SKG and performing reasoning with retrieved knowledge. The agent iteratively executes beam search on the KG, discovers the most promising reasoning paths, and returns the most likely reasoning results. CoTKR (Wu et al., 2024) rewrites retrieved subgraphs into natural language formats comprehensible to LLMs. R^3 (Toroghi et al., 2024) surfaces the commonsense knowledge relevant to the question from LLMs and uses it to guide the SKG pruning to find answers.

Overall, the primary advantage of IR-based methods is their strong performance and robust few-shot/zero-shot capabilities through tight integration of parametric and symbolic knowledge, enabling effective question answering even with limited training examples. However, these methods come with disadvantages: they introduce significant challenges in balancing retrieval precision with computational cost, as improving accuracy often requires more resource-intensive processing. The potential for semantic noise in retrieved subgraphs can lead to irrelevant or misleading information being incorporated into the reasoning process. Additionally, difficulties in capturing long-range dependencies within the structure of SKGs limit their ability to fully understand complex relationships that span multiple knowledge connections, potentially reducing the quality of final answers.

5.1.2. Temporal knowledge graph question answering

Task definition: Given a temporal question and a TKG, TKGQA methods are required to understand the intent and temporal constraints of the question via the parametric knowledge bases and retrieve the entity nodes or timestamps from the TKG as answers.

As illustrated in Fig. 11(b), the primary solution for TKGQA involves integrating the knowledge of TKGs and PLMs. Some methods use PLM and pre-trained TKG embeddings to match questions and entities and timestamps, known as embedding-based methods. Some other methods utilize PLM and Graph Convolutional Networks (GCNs) (Kipf & Welling, 2017) to learn the features of TKG elements, known as GCN-based methods. In recent years, there have also been methods that utilize LLMs by converting the facts from TKGs into text, which are named retrieval-augmented methods because they primarily enhance the reasoning capabilities of LLMs by accurately retrieving useful knowledge from TKGs.

(1) Embedding-based methods: As shown in Fig. 11 (b.1), embedding-based methods first employ embedding models to obtain representations of TKG elements, such as timestamps, entities, and relations. Then, they leverage PLMs to encode questions into embeddings. Finally, by computing matching scores between question embeddings and TKG element embeddings, the answers can be inferred. The structural commonalities of these methods center on their dual-embedding framework, which separately encodes TKG elements and natural language questions before computing their semantic alignment.

A classic embedding-based method is CRONKGQA (Saxena et al., 2021), which builds on EmbedKGQA (Saxena et al., 2020) by employing an advanced temporal knowledge graph embedding model to derive embeddings of entities and timestamps. It uses PLM to obtain embeddings of entity/time mentioned in the question, models the question as a "virtual relation", and predicts missing entities and timestamps in the TKG. Similarly, QC-MHM (Xue et al., 2024) is more refined in handling questions and TKGs. It first injects temporal order information into timestamp embeddings, modifying TComplEx (Lacroix et al., 2020) to obtain the embeddings of the entity, relation, and timestamp. It then inputs the sentence and SPO (subject, predicate or relation, and object) into Sentence-BERT (Reimers & Gurevych, 2019) to obtain the embedding vectors and model the matching between questions and SPO . Other similar methods include TempoQR (Mavromatis et al., 2022), TSQA (Shang et al., 2022), TMA (Liu, Liang, Fang, et al., 2023), JMFRN (R. Huang et al., 2024), and SERQA (C. Du et al., 2024).

Overall, the primary advantage of embedding-based methods is their more natural reasoning process and lighter model architecture, enabling better modeling of questions and improved matching with TKGs. However, these methods come with disadvantages:

they struggle significantly with complex temporal constraints and temporal relational terms, limiting their effectiveness in time-sensitive scenarios. Complex temporal constraints require the use of multiple factual quadruples, which increases computational complexity and can overwhelm the model's capacity. Additionally, temporal relational terms are highly sensitive to the model, making it difficult to maintain consistent performance across different temporal expressions and potentially leading to incorrect reasoning when dealing with time-based queries.

(2) GCN-based methods: As described in Fig. 11 (b.2), GCN-based methods first leverage GCNs (Kipf & Welling, 2017) to learn the embeddings of nodes in subgraphs of TKG. Then, they encode embeddings of question by PLMs. Subsequently, this type of method formulates answer prediction as a node classification task based on the learned embeddings. The structural commonalities of these methods center on their dual-pathway architecture that separately processes natural language questions and TKG subgraphs before integrating them for answer prediction. Compared to embedding-based methods, GCN-based methods can model various constraints within complex problems and retain a more complete subgraph of the TKG.

For example, EXAQT (Jia et al., 2021) utilizes fine-tuned BERT models and GCN (Kipf & Welling, 2017) to identify relevant facts. It specifically employs Group Steiner Trees to compute question-relevant compact subgraphs within the KG. Additionally, relational graph convolutional network (Schlichtkrull et al., 2018) has been constructed to predict answers. Similarly, GenTKGQA (Gao et al., 2024) first leverages an LLM and a pre-trained temporal graph neural network to model question and extract information from the subgraph, respectively. Then, it performs instruction tuning to enable complex temporal reasoning. Other similar methods include SubGTR (Z. Chen et al., 2022), TSIQA (Xiao et al., 2022), CTRN (Jiao et al., 2022), TwiRGCN (Sharma et al., 2023), M3TQA (Zha et al., 2024) and LGQA (Liu, Liang, Li, et al., 2023).

Overall, the primary advantage of GCN-based methods is their exceptional ability to handle complex temporal queries and preserve structural information within TKGs, offering superior constraint modeling compared to embedding-based approaches. However, these methods come with disadvantages: they face significant computational complexity when processing large subgraphs, making them less efficient for real-time applications. Additionally, they struggle to capture long-range temporal dependencies, which can limit their reasoning capabilities across extended time periods. Furthermore, GCN-based approaches typically require substantial training data and often perform poorly in few-shot scenarios, limiting their applicability in data-constrained environments.

(3) Retrieval-augmented methods: As illustrated in Fig. 11 (b.3), retrieval-augmented methods first leverage LLMs to retrieve question-relevant evidence from the TKG. The retrieved evidence is then fed back into the LLMs. Finally, by jointly exploiting the parametric knowledge encoded in LLMs and the symbolic knowledge contained in the retrieved evidence, the LLMs reason about the final answer. The structural commonalities of these methods center on their retrieval-generation architecture that first extracts relevant temporal and factual knowledge from TKGs before integrating it with LLM capabilities.

For example, Prog-TQA (Chen et al., 2024a) uses an LLM to understand questions and generate corresponding program drafts with symbolic operators as logical forms, given a few examples. With a self-improvement strategy, the quality of these logical forms is enhanced to yield the final answers. Similarly, ARI (Z. Chen et al., 2024) improves the LLM's capacity to integrate abstract methodologies derived from historical experience. TimeR^4 (Qian et al., 2024) differentiates the knowledge of TKG into temporal knowledge and factual knowledge and then improves retrieval accuracy through modules such as retrieval, rewriting, and ranking.

Overall, the primary advantage of retrieval-augmented methods is their emphasis on retrieving TKG elements, which by textualizing the knowledge within the TKG, effectively reduces hallucinations in LLMs and enhances their reasoning capabilities, leading to significant performance breakthroughs compared to embedding-based and GCN-based methods. However, these methods come with disadvantages: due to LLMs' insufficient sensitivity to temporal knowledge, they still face significant challenges regarding knowledge integration, making it difficult to seamlessly combine retrieved information with the model's internal knowledge. Additionally, they struggle with complex temporal reasoning, particularly when dealing with multi-hop temporal relationships or intricate temporal constraints that require precise understanding of time-based dependencies, limiting their effectiveness on temporally complex queries.

5.1.3. The datasets of graph-based reasoning methods

In this section, we select, summarize, and analyze datasets related to the graph-based reasoning task. We statistically summarize the information of the related dataset from multiple dimensions, including (1) Domain: The domain of knowledge corresponding to the datasets; (2) Question type: The type of questions in the datasets; (3) # Ques.: Question number; (4) Q&A source: The main construction methods of questions and answers. They are mainly divided into three categories: "Generate", "Expert", and "Crowdsourcing". "Generate" refers to the design of programs for automated generation, "Expert" refers to direct crawling from professional websites or carefully designed by domain experts, and "Crowdsourcing" refers to completion by crowdsourcing workers with general cultural levels; (5) KG: Specific KG (SKG or TKG) they use; (6) Links: The storage address of the datasets. The statistical results are shown in Table 10.

The characteristics of a dataset strongly influence the choice of reasoning methods in graph-based reasoning tasks. Dataset size determines how effectively different approaches can learn and generalize. Semantic parsing methods typically require large labeled datasets to learn precise mappings from questions to logical forms, whereas information retrieval methods rely more on pre-trained knowledge and can perform well even with limited supervision. In temporal knowledge graph question answering, GCN-based models tend to be highly sensitive to data volume, while embedding-based and retrieval-augmented methods are often more adaptable when data is scarce. Structural properties of the dataset also play a critical role. Factors such as triplet density and connection patterns shape the availability of reasoning paths. Densely connected graphs enable richer multi-hop reasoning, whereas sparse graphs constrain reasoning depth. In temporal datasets, the distribution of timestamps and the complexity of relational dynamics further affect which reasoning strategies achieve the best performance.

Table 10
Dataset statistics of graph-based reasoning.

Datasets	Domain	Question type	# Ques.	Q&A source	KG	Links
Free917 (Cai & Yates, 2013)	General	Static	917	Expert	Freebase	https://
WebQuestions (Berant et al., 2013)	General	Static	5810	Crowdsourcing	Freebase	https://
WebQuestionsSP (Yih et al., 2016)	General	Static	4737	Crowdsourcing	Freebase	https://
ComplexQuestions (Bao et al., 2016)	General	Static	2100	Expert	Freebase	https://
MetaQA/1-hop (Zhang et al., 2018)	Movie	Static	116,045	Generate	Wikipedia	https://
MetaQA/2-hop (Zhang et al., 2018)	Movie	Static	148,724	Generate	Wikipedia	https://
MetaQA/3-hop (Zhang et al., 2018)	Movie	Static	142,744	Generate	Wikipedia	https://
QALD (Usbeck et al., 2024)	General	Static	806	Expert	DBpedia	https://
LC-QuAD (P. Trivedi et al., 2017)	General	Static	5000	Generate	DBpedia	https://
LC-QuAD2.0 (Dubey et al., 2019)	General	Static	5000	Generate	DBpedia	https://
TempQuestions (Jia et al., 2018)	General	Temporal	1271	Expert	Freebase	https://
TimeQuestions (Jia et al., 2021)	General	Temporal	16,181	Expert	Wikidata	https://
CRONQUESTIONS (Saxena et al., 2021)	General	Temporal	410,000	Expert	Wikidata	https://
Complex-CRONQUESTIONS (Z. Chen et al., 2022)	General	Temporal	45,821	Expert	Wikidata	https://
MultiTQ (Z. Chen et al., 2023)	Social Science	Temporal	500,000	Expert	ICEWS	https://

Table 11
Comparison of graph-based reasoning methods. Key dimensions: Performance (reasoning accuracy), Efficiency (computational efficiency and resource consumption), Interpretability (reasoning transparency), Robustness (stability under out-of-distribution, data noise, or data sparsity scenarios).

Sub-task	Method	Performance	Efficiency	Interpretability	Robustness
Knowledge graph question answering	Semantic parsing-based	High *Enhanced semantic understanding*	Moderate *Moderate resource consumption*	High *Explicit logical forms*	Moderate *Limited schema generalization*
	Information retrieval-based	High *Superior parametric-symbolic fusion*	Moderate *Manageable retrieval overhead*	Moderate *Inspectable retrieval logic*	High *Strong few-shot adaptability*
Temporal knowledge graph question answering	Embedding-based	Moderate *Limited temporal constraints*	High *Lightweight architecture*	Low *Opaque representation*	Moderate *Fragile temporal expressions*
	GCN-based	High *Effective structure modeling*	Low *Heavy subgraph computation*	Moderate *Traceable structural patterns*	Low *Heavy data dependence*
	Retrieval-augmented	High *Significant hallucination reduction*	Moderate *Reasonable retrieval overhead*	Moderate *Explicit retrieved knowledge*	Moderate *Incomplete temporal integration*

5.1.4. The comparison of graph-based reasoning methods

In this section, we conduct a comparison of various graph-based reasoning methods. The corresponding results are presented in Table 11.

For knowledge graph question answering, semantic parsing-based (Zong et al., 2024) and information retrieval-based (Toroghi et al., 2024) methods achieve high performance and comparable efficiency but differ in interpretability and robustness. The former method emphasizes interpretability via explicit logical forms, yet struggles with robustness due to limited schema generalization. Conversely, the latter show strong robustness through strong few-shot adaptability but weak at interpretability due to inspectable retrieval logic.

For temporal knowledge graph question answering, similar trade-offs emerge across different paradigms. Embedding-based methods (Xue et al., 2024) are efficient but highly lack interpretability. GCN-based methods (Zha et al., 2024) enhance performance through effective structure modeling, yet incur high computational cost and depend heavily on data quality. Retrieval-augmented methods (Qian et al., 2024) offer a more balanced solution, achieving strong performance while remaining reasonable on other metrics.

Overall, no single paradigm excels in all aspects, and each comes with its own trade-offs. Therefore, method selection should be driven by the specific priorities of the application. Future work could explore adaptive approaches that automatically adjust these trade-offs in practice.

5.2. Table-based reasoning

The task of table-based reasoning mainly refers to table question answering (Table QA), where symbolic knowledge is stored in structured tables.

5.2.1. Table question answering

Task definition: Given a question and a table, Table QA methods are required to understand the intent of the question via the parametric knowledge bases and find the correct answer from the table. The table QA methods can be divided into: structure-based, table-to-text, instruction fine-tuning, and hybrid reasoning methods, as illustrated in Fig. 11(c).

(1) Structure-based methods: As shown in Fig. 11 (c.1), structure-based methods first parse the table structure. Then, they utilize the relationships between rows, columns, and cells to find the final answers. They convert table elements into graphs or embeddings to capture structural relationships, then use neural networks, such as GNNs or Transformers, to understand how elements interact. These methods are often pre-trained or fine-tuned on large text-table datasets, enabling them to directly map questions to relevant table components. Some approaches also generate intermediate logical forms, such as SQL queries or reasoning chains, to bridge the gap between natural language input and table-based answer extraction.

Many methods primarily focus on understanding the semantic association between table headers and data areas. For example, Müller (Mueller et al., 2019) proposes encoding a table into a graph and employing GNNs and pointer networks to select answers and address sequential questions directly from the table. Similarly, TAPAS (Herzig et al., 2020) flattens tables, encodes structure through multiple positional embeddings, and uses pre-training on text-table pairs to predict cell selections and aggregation operators for Table QA. GRAPPA (Yu et al., 2021) takes a different approach by inducing a synchronous context-free grammar to generate synthetic data, combining pre-training on masked language modeling and SQL semantic prediction. This enables GRAPPA to utilize structural knowledge effectively for semantic parsing, converting user inputs into executable programs, and enhancing performance across various datasets.

However, while these methods excel in querying structured data accurately, they face significant limitations when dealing with complex tables, such as nested structures, multi-table setups, or untitled data. They also struggle with flexible questions or generative tasks, highlighting a gap in adaptability. To address these challenges, a few methods have been introduced. For example, TABVER (Aly & Vlachos, 2024) integrates arithmetic reasoning with natural logic reasoning systems, enabling tabular fact-checking tasks to overcome the limitations of symbolic reasoning models and natural logic systems in handling arithmetic operations. Similarly, CARP (Zhong et al., 2022) employs mixed-modal reasoning chains to explicitly model intermediate reasoning steps, improving the interpretability of the model's reasoning process. Besides, TabDDR (S.-X. Yang et al., 2025) introduces a dual-denoiser-reasoner architecture that specifically targets both structural and textual noise simultaneously, offering a balanced approach to noise reduction without introducing additional noise. Despite these advancements, studies reveal that when rows and columns are rearranged to create new examples, the performance of large models declines significantly. This suggests that current approaches to table structure understanding lack robustness and fail to adapt effectively to structural changes.

The main advantage of structure-based methods is their strong performance in extracting answers from well-organized tables due to their explicit use of table structure and relational information. However, they are sensitive to variations in table layouts and struggle with more unstructured or complex data, such as nested tables or those lacking clear headers. Furthermore, they can be limited in handling open-ended or generative tasks, as their design is inherently focused on structured answer retrieval rather than flexible or context-aware reasoning.

(2) Table-to-text methods: As indicated in Fig. 11 (c.2), table-to-text methods first extract important facts from the table. Then, these extracted elements are transformed into fluent natural language evidence. Finally, based on the given question and the natural language evidence, PLMs reason about final answers.

For instance, OPENT2T (Zhang et al., 2024b), an open-source toolkit, facilitates the replication and comparison of existing systems, driving innovation in developing new models. Another example is Table-GPT (Li et al., 2024), which proposes a "table-tuning" paradigm to improve language models' performance on table-related tasks. Additionally, methods for summarizing table contents enable users to complete question-answering tasks without needing to browse individual entries in the table (Moosavi et al., 2021; Suadaa et al., 2021; Wiseman et al., 2017). Despite these advancements, current research primarily focuses on surface-level achievements, with limited attention to logical reasoning. While existing methods address issues of surface authenticity, they often restate data facts without demonstrating robust reasoning or generalization capabilities. LOGICNLG (W. Chen et al., 2020) aims to bridge this gap by enabling models to generate logically inferred natural language statements from facts in open-domain, structured tables. Recently, SciTable-Sowise (Yin & Wang, 2025) integrates a fine-tuned DeBERTa model with Type-Guided CoT Prompts to enhance the automation and accuracy of table content processing in scientific papers, wherein table-to-text generation constitutes the central component of this task.

The main advantage of table-to-text methods lies in their flexibility and accessibility, enabling users to comprehend complex tabular information quickly without manual inspection. They are also adept at handling unstructured queries and generating context-aware answers. However, these methods can suffer from reduced transparency and limited interpretability, making it hard to verify outputs, especially when logical reasoning is required. Furthermore, their reliance on substantial annotated data and heavy computational requirements can hinder scalability and practical deployment in resource-limited settings.

(3) Instruction fine-tuning methods: Recently, efforts have been made to enhance LLMs' ability to process table data by developing specialized instruction fine-tuning datasets. As shown in Fig. 11 (c.3), in stage 1, this type of method first collect richly annotated instruction-based datasets containing instruction, table data with corresponding questions and responses. The PLMs are then fine-tuned with these data to enhance the table reasoning capability. In stage 2, the same instruction used in fine-tuned, table, question are inputted to the PLMs to reason about final answers.

Noteworthy examples include TrixInstruct (Wu & Feng, 2024) and TableLlama (T. Zhang et al., 2024), which utilize datasets that cover diverse, realistic tables and related tasks. After fine-tuning on these datasets, LLMs show significant improvements in handling table-based questions. Additionally, ChatQA (Liu et al., 2024) employs a two-stage instruction fine-tuning strategy, yielding substantial gains in table-related tasks. The first stage involves supervised fine-tuning on diverse instruction datasets, while the second stage, context-enhanced instruction fine-tuning, incorporates table QA and other high-quality QA datasets to further refine the model's conversational QA capabilities in context-specific scenarios.

Table 12
Dataset statistics of table-based reasoning.

Datasets	Domain	# Ques.	Q&A source	Table source	Links
HiTab (Cheng et al., 2022)	General	14 K	Statistical reports, Wikipedia	Crowdsourcing	https://
FeTaQA (Nan et al., 2022)	General	10 K	Crowdsourcing	Wikipedia	https://
TableInstruct (T. Zhang et al., 2024)	General	1.24 M	Expert	Wikipedia, statistical scientific reports	https://
FEVEROUS (Aly et al., 2021)	General	87 K	Expert	Wikipedia	https://
TableBench (Wu et al., 2025b)	General	0.8 K	Crowdsourcing	existing datasets	https://
BanglaTabQA (Pal et al., 2024)	General	19 K	Generate	Wikipedia	https://
ChatRAG Bench (Liu et al., 2024)	General	29.2 K	Expert	Internet	https://
KET-QA (M. Hu et al., 2024)	General	13 K	Crowdsourcing	Wikidata	https://
IM-TQA (Zheng et al., 2023)	General	1.2 K	Crowdsourcing	published studies	https://

A key advantage of this type of method is the boosted performance and contextual relevance in table-oriented tasks, achieved through targeted supervision and staged refinement. These methods become more adept at understanding the semantics and structures inherent in tabular data. However, these approaches can be limited by the scale and diversity of available fine-tuning datasets, potentially introducing biases or coverage gaps. Additionally, the fine-tuning process can be resource-intensive, requiring significant computational costs and careful dataset engineering.

(4) Hybrid reasoning methods: As shown in Fig. 11 (c.4), hybrid reasoning methods typically follow a multi-phase pipeline that leverages both traditional structured reasoning (e.g., rule-based and retrieval mechanisms) and advanced deep learning models such as PLMs. Common stages include pre-processing and retrieval of relevant data from tables and unstructured texts, followed by prompt construction tailored to the question and retrieved content. Afterwards, multi-step reasoning is iteratively performed to synthesize evidence across modalities and generate answers. The integration of retrieval-augmented generation, in-context learning, and step-wise evidence aggregation is fundamental to handling complex reasoning tasks that require both factual correctness and contextual understanding.

Building on this foundation, S3HQA (Lei et al., 2023) proposes a three-stage framework, comprising retrieval, selector, and reasoner, where LLMs are employed as generative components in the reasoning stage. While the hybrid reasoning method is effective in solving complex table question answering tasks, it comes with drawbacks such as being complex to implement, requiring high computing resources, and in certain situations, needing manual rule design or model fine-tuning, EXPLORA (Purohit et al., 2024) offers a distinct perspective which can save resources by reducing the number of LLM calls. As a static subset selection method, it uses a scoring function to select examples, bypassing reliance on LLM parameters or outputs. Recently, to address the challenge of insufficient annotations, UCTR-ST (Z. Li et al., 2024) is proposed to generate diverse synthetic data. It produces three types of programs, including SQL queries, logical forms, and arithmetic expressions, and achieves strong performance in both unsupervised and few-shot settings.

The main advantage of hybrid reasoning methods lies in their superior ability to handle complex, multi-hop, and cross-modal table question answering by effectively combining structured and unstructured evidence. They tend to outperform purely neural or rule-based approaches, especially in scenarios demanding nuanced understanding and logical inference. However, these benefits come at the cost of increased system complexity, higher computational requirements, and often reliance on carefully engineered prompts or additional components. Additionally, their implementation may require significant manual effort for model customization and continuous maintenance to ensure robustness across diverse question types and domains.

5.2.2. The datasets of table-based reasoning methods

In this section, we select, summarize, and analyze datasets related to the table-based reasoning task. We statistically summarize the information of the related dataset from multiple dimensions, including (1) Domain: The domain of knowledge corresponding to the datasets; (2) # Ques.: Question number; (3) Q&A source: The main construction methods of questions and answers. They are mainly divided into three categories: "Generate", "Expert", and "Crowdsourcing". "Generate" refers to the design of programs for automated generation, "Expert" refers to direct crawling from professional websites or carefully designed by domain experts, and "Crowdsourcing" refers to completion by crowdsourcing workers with general cultural levels; (4) Table source: The source of tables; (5) Links: The storage address of the datasets. The statistical results are shown in Table 12.

The characteristics of a dataset, including its size and structural organization, play a central role in determining which reasoning methods are most effective. When datasets are large, models often achieve better pattern recognition and greater robustness, but these advantages may level off once the dataset reaches a certain size. The structural clarity of tables also matters. Well-organized tables with consistent layouts and accurate mappings, as seen in resources like WikiSQL, tend to support reasoning methods that rely on precise relational understanding. In contrast, datasets with noisy or ambiguous structures often require approaches that can tolerate uncertainty. Hybrid reasoning methods are usually more capable of adapting to diverse table structures, yet their success depends on having enough data to support their complexity.

Table 13
Comparison of table-based reasoning methods. Key dimensions: Performance (reasoning accuracy), Efficiency (computational efficiency and resource consumption), Interpretability (reasoning transparency), and Robustness (stability under out-of-distribution, data noise, or data sparsity scenarios).

Method	Performance	Efficiency	Interpretability	Robustness
Structure-based	High *Precise structural alignment*	Moderate *Tolerable graph construction cost*	High *Transparent reasoning steps*	Low *Significant layout sensitivity*
Table-to-text	Moderate *Shallow fact extraction capability*	Low *complex multi-stage pipeline*	Low *Black-box output*	Moderate *Decent context awareness*
Instruction fine-tuning	High *Effective task-specific supervision*	Low *Resource-intensive training*	Low *Black-box adaptation*	Low *Excessive data dependency*
Hybrid reasoning	High *Comprehensive multi-evidence synthesis*	Low *High system complexity*	Moderate *Mixed evidence*	High *Reliable integrated verification*

5.2.3. The comparison of table-based reasoning methods

In this section, we compare various table-based reasoning methods, and the corresponding results are presented in Table 13.

For table-based reasoning, structure-based methods (S.-X. Yang et al., 2025) deliver high performance and interpretability but lack robustness against layout changes. Both table-to-text approaches (Yin & Wang, 2025) and instruction fine-tuning methods (T. Zhang et al., 2024) severely lack efficiency and interpretability due to their resource intensive and opaque nature. Conversely, hybrid reasoning models (P. Yu et al., 2025) overcome robustness limitations and maintain high performance, yet they incur the highest computational cost driven by high system complexity.

In summary, method selection relies on specific application requirements. A promising future research is integrating the strengths of existing paradigms to advance reliable and explainable table reasoning.

5.3. Text-based reasoning

The tasks of text-based reasoning include Machine reading comprehension (MRC), Multi-hop reading comprehension (MHRC), also known as multi-hop question answering, Open-domain question answering (ODQA), and Multi-document question answering (MDQA), where symbolic knowledge is stored in unstructured text.

5.3.1. Machine reading comprehension

Task definition: Machine reading comprehension aims to evaluate a machine's ability to understand language effectively. MRC methods are presented with one or more text passages and are then required to answer questions based on the provided passages (Chen et al., 2016; Sachan & Xing, 2016; Zhang, Wu, et al., 2020; Z. Zhang et al., 2021). As highlighted by Li et al. (2023), Ouyang et al. (2024), improving MRC performance involves developing several skills, including numerical reasoning, commonsense reasoning, and logical reasoning.

The research on machine reading comprehension has garnered significant interest over the past decade. The MRC tasks have progressed from the initial cloze-style tests (Hermann et al., 2015; Hill et al., 2016) to span-based answer extraction from passages (Joshi et al., 2017; Rajpurkar et al., 2016), as well as to multiple-choice (Lai et al., 2017) and free answering formats (Nguyen et al., 2016). In the early years, rule-based methods (Hirschman et al., 1999; Riloff & Thelen, 2000) focus on designing heuristic algorithms. These algorithms are specifically tailored to the grammar of a language and are used to assist in learning, discovery, or problem-solving. By employing trial-and-error techniques, they aim to find evidence within a given sentence to answer a question. Statistic-based methods (Charniak et al., 2000) involves quantifying occurrences of words and utilizing these numerical representations to infer potential answers. Recently, numerous studies (Baradaran & Amirkhani, 2021; Hu, Wei, et al., 2019) focus on leveraging machine learning methods to extract the features in the questions and passages, which enhance the machine's ability to understand and process text automatically.

With the development of deep learning, various attention-based methods (Cui et al., 2017; Dhingra et al., 2017; Kadlec et al., 2016; Seo et al., 2017) are proposed to facilitate interactions between passages and questions. Recently, PLMs have achieved significant success in MRC tasks. The PLMs-based models demonstrate a strong ability to capture contextual and sentence-level language representations, which notably improve the benchmark performance of current MRC systems (Ouyang et al., 2024; Zhang, Wu, et al., 2020; Z. Zhang et al., 2021). In line with this trend, our focus is primarily on PLMs-based MRC methods, which combine symbolic knowledge in text passages and parameter knowledge in PLMs. The PLMs-based MRC methods can be further divided into Extractive MRC and Generative MRC (Ai et al., 2024). As illustrated in Fig. 11(d), the Extractive MRC methods try to predict the start and end positions of answers directly from the context, while the Generative MRC methods are devoted to generating answers by reformulating information across the context.

(1) Extractive MRC methods: As shown in Fig. 11 (d.1), recent MRC research primarily concentrates on extractive question answering using encoder-only PLMs, which directly predict the start and end positions of answers within the context. The standard workflow begins by encoding the question and passage together to produce detailed contextual representations. The representations are then passed through task-specific modules that identify the answer span's start and end positions within the passage. More

advanced methods may refine this process by integrating auxiliary strategies, such as syntactic cues, iterative evidence selection, or sequence tagging, to better capture complex relationships and address scenarios with multiple answer spans.

For instance, CTX (Ohsugi et al., 2019) adopt PLMs to independently obtain paragraph representations conditioned with the current question, previous questions, and previous answers. To extract the answer span, the start and end positions of the current answer are predicted based on the concatenation of the paragraph representations encoded in the previous step.

In this category of methods, PLMs are mainly used as encoders to extract general features from text paragraphs and questions. The focus is on designing downstream models to extract task-oriented features. For instance, SG-Net (Zhang, Wu, et al., 2020) leverages syntactic guidance in text modeling, achieving substantial performance gains in MRC by introducing explicit syntactic constraints in the attention mechanism. Inspired by how humans solve reading comprehension questions, Retro-Reader (Z. Zhang et al., 2021) integrates two reading and verification strategies stages. First, it uses sketchy reading to quickly grasp the relationship between the passage and the question, forming an initial judgment. Then, it conducts intensive reading to verify the answer and provide the final prediction. Focal_Reasoner (Ouyang et al., 2024) extracts fact units from raw texts via syntactic processing and constructs a supergraph. Then, it performs reasoning over the supergraph and a logical fact regularization and aggregates the learned representation to decode the correct answer. Extract-Integrate-Compete (C. Zhang et al., 2021) iteratively selects complementary evidence with a novel query updating mechanism and adaptively distills supportive evidence, followed by a pairwise competition to push models to learn the subtle difference among similar text pieces.

In the methods mentioned above, the system is mainly expected to extract a single answer from the passage for a given question. However, in many scenarios, questions may have multiple answers scattered in the passages, and all the answers should be found to answer the questions completely. For extracting answers with multi-span, TASEBIO (Segal et al., 2020) transfers MRC to a sequence tagging task, predicting whether each token is part of the answer. MTMSN (Hu, Peng,et al., 2019) combines a multi-type answer predictor designed to support various answer types (e.g., span, count, negation, and arithmetic expression) with a multi-span extraction method for dynamically producing one or multiple text spans. SpanQualifier (Huang et al., 2023) presents a novel span-centric scheme to generate representations for all spans in the context and predicts a qualification threshold. Furthermore, it designs a global loss function to jointly optimize overall spans instead of independently optimizing loss on each individual span, which avoids the influence of label imbalance on training the proposed span-centric scheme.

The key advantage of extractive MRC methods lies in their lightweight architectures and clear, span-based outputs. Their designs are straightforward and often interpretable. However, because they rely on precise span extraction, they are less effective for questions that require multi-span or synthesized answers from different text segments. Additionally, these models are typically limited when answers that are not explicitly present in the input context.

(2) Generative MRC methods: Recently, significant progress has been made in controllable text generation. As indicated in Fig. 11 (d.2), beyond extractive methods, there is also growing interest in applying generative language models for MRC, which generate answers by reformulating information across the context. Typically, generative MRC methods encode the input question and context, optionally integrate extraction modules to identify supporting evidence, and then generate free-form answers through a decoder. Some approaches further enhance answer quality by combining generative loss with extractive supervision or by adaptively incorporating external knowledge sources and symbolic reasoning during generation.

For instance, RBG (Su et al., 2022) combines a Seq2Seq language model-based generator with a machine reading comprehension module. The reader produces an evidence probability score for each sentence, which will be integrated with the generator for final distribution prediction. KEAG (Bi et al., 2019) composes a natural answer by exploiting and aggregating evidence from all four information sources available: question, passage, vocabulary, and knowledge. During the process of answer generation, KEAG adaptively determines when to utilize symbolic knowledge and which fact from the knowledge is useful. REAG (C. Li et al., 2021) incorporates an extractive mechanism into a generative model to leverage relevant information to a given question in the contextual passage. Specifically, REAG adds an extraction task on the encoder to obtain the rationale for an answer, which is the most relevant piece of text in an input document to the given question. T5-MRC (B. Guan et al., 2024) uses the STS model to label training evidence more accurately and proposes a threshold-based method to filter evidence during model training. QASE (Ai et al., 2024) proposes a novel adaptation of controlled text generation tailored to the specific challenges of MRC, focusing on the precision and relevance of generated answers. Unlike methods that modify the overall generative process through complex architectural alterations or additional learning mechanisms, QASE directly utilizes the question and context to guide inferences. MRC remains a challenging task in few-shot settings and low-resource scenarios. To address this issue, MRC-PASCL (R. Li et al., 2024) introduces a novel noun-entity-aware data selection and generation strategy tailored to the characteristics of the MRC task and data, with a particular focus on masking nouns and named entities in the context.

For multi-span answers, MUSST (Yang et al., 2021) combines the benefits of span extraction and the simplicity of a multi-span approach to generate free-form answers. It also provides a comprehensive framework for multi-passage generative MRC, which consists of a passage ranker, a multi-span answer annotator, and a question-answering module. Unlike the studies that mainly focus on introducing generative mechanisms, SAMSG (Zhang et al., 2021) focuses on handling the writing form of the answer and proposes a novel non-generative decoder to exploit the results from the extractive decoder fully. It learns to score every word in the given passage for how likely they are in the expected answer, then calculates the score of a candidate span from the words' scores.

The key advantage of generative MRC methods is their flexibility in handling complex question contexts and producing natural, free-form answers that may require aggregation or rephrasing beyond simple span selection. They are also well-suited for multi-span and multi-passage settings. However, generative-extractive methods are often more computationally intensive, require larger training datasets, and can be more difficult to interpret than purely extractive models. Moreover, they may suffer from errors in answer synthesis, such as generating hallucinated or incomplete information if not well-aligned with the evidence.

5.3.2. Multi-hop reading comprehension

Task definition: Multi-hop reading comprehension methods focus on integrating and reasoning over multiple pieces of evidence to answer complex questions. Unlike single-hop MRC, where questions are typically straightforward, and answers can be derived from one or a few nearby sentences, MHRC involves reasoning chains that traverse multiple sentences or even passages. This requires a deep text understanding and reasoning capability, making it more akin to real-world scenarios.

The key challenge of MHRC lies in its demand for multi-step reasoning, where a model must identify and connect intermediate information to form a coherent reasoning path. This reasoning chain culminates in the extraction of the correct answer, supported by a series of evidence sentences. Therefore, MHRC not only tests a model's ability to find the answer but also its capacity to justify the reasoning process with a clear rationale, presenting evidence as proof of the multi-hop reasoning process. Datasets like HotpotQA (Yang et al., 2018a) have been specifically designed to evaluate such multi-hop reasoning capabilities. They include tasks that expect the model to extract and present evidence sentences, thereby showing a clear reasoning trail. Consequently, MHRC better aligns with real-world scenarios where information is dispersed across long passages or multiple documents, necessitating more comprehensive models than those used for single-hop MRC tasks. In MHRC, the given passages are considered as symbolic knowledge bases, and the parametric knowledge bases (PLMs) are adopted to encode the questions and the passages. In MHRC, methods are primarily divided into two categories: graph-based and graph-free methods (Mohammadi et al., 2023; Zhangyue et al., 2023), as illustrated in Fig. 11(e). Graph-based methods must construct and reason over a graph structure created from the input data. However, graph-free methods do not rely on such structures and often utilize techniques like self-attention for reasoning.

(1) Graph-based MHRC methods: As shown in Fig. 11 (e.1), the main idea behind graph-based approaches is to represent the input data, including both context passages and the question, as a graph structure. In this graph, nodes correspond to entities, sentences, or significant text spans, while edges represent semantic relationships, co-occurrences, or other relevant connections between these elements. Through iterative message passing or information propagation over the graph, models dynamically aggregate and integrate evidence from multiple nodes. Finally, the learning node representations are used to predict answer start and stop position. This graph-based framework not only captures rich interactions among entities but also provides a flexible basis for incorporating additional mechanisms such as attention or node/edge filtering to enhance reasoning capabilities.

Most current research relies on entity graphs where nodes are formed from the context and question entities. For instance, Entity-GCN (De Cao et al., 2019) compiles scattered information from one or more documents by building an entity graph. In this structure, nodes represent entity mentions, while edges illustrate relationships between these mentions within and across multiple documents. BAG (Cao et al., 2019) transforms documents into a graph in which nodes are entities and edges are relationships between them. The graph is then imported into graph convolutional networks to learn relation-aware representations of nodes. Furthermore, BAG introduces bi-directional attention between the graph and a query with multi-level features to derive the mutual information for the final prediction. Many studies assume that all contexts are pertinent and ignore the negative impact of irrelevant contexts. To filter out unrelated context, DFGN (Qiu et al., 2019) designs a paragraph-selection module to eliminate unrelated paragraphs. It dynamically builds an entity graph from the question entities to locate relevant supporting entities and text spans. Based on DFGN, DFGN_Dual (Cao et al., 2021) introduces dual reasoning channels to predict the final answer and supporting facts, respectively, which gain better step-by-step reasoning compared to a single-channel approach. Similarly, SAE (Tu et al., 2020) incorporates a paragraph-selection step to filter out irrelevant context segments, thereby shrinking the problem space while utilizing sentences as graph nodes.

To further improve the performance, some methods try to emulate the human brain's cognitive processes for multi-hop MRC. For example, inspired by the dual process theory of human (Evans, 1984, 2003, 2008; Sloman, 1996), CogQA (Ding et al., 2019) builds a cognitive graph in an iterative process by coordinating an implicit extraction module and an explicit reasoning module. The extraction module extracts question-relevant entities to construct the cognitive graph. Then, the reasoning module conducts the reasoning procedure over the graph and collects clues to guide the extraction module in extracting next-hop entities better. DRN (Li et al., 2020) designs a query reshaping mechanism that visits a query repeatedly to mimic people's reading habits. It dynamically reasons over an entity graph with graph attention and the query reshaping mechanism to promote its comprehension and reasoning ability. IP-LQR (Tang et al., 2022) incorporates phrases in the latent query reformulation to improve the cognitive ability of the proposed method for MHRC.

Some studies try to construct more intricate graphs using multiple node types to encompass the available contextual information in the graph constructions fully. For instance, HDE (Tu et al., 2019) proposes a heterogeneous document-entity graph, which contains different granularity levels of information, including candidates, documents, and entities in specific document contexts. To aggregate clues from scattered texts across multiple paragraphs, HGN (Fang et al., 2020) creates a hierarchical graph by constructing nodes on different granularity levels, including questions, paragraphs, sentences, and entities. Furthermore, TA-MHRC (Mohammadi et al., 2023) uses more helpful information about the context, such as the topic of sentences, the topic of relationships, and the importance and strength of relationships, when filtering paragraphs and constructing the graph. Thus, the proposed graph is a weighted graph with four types of nodes and six types of edges to cover the complete information of the context.

The key advantage of graph-based methods is their effectiveness in modeling complex relationships and reasoning paths, as they can explicitly capture interactions between multiple entities or facts scattered across the text. This explicit structure often improves model interpretability and can support more accurate multi-hop reasoning. However, constructing high-quality graphs typically relies on accurate entity recognition and relationship extraction, which can introduce errors. Furthermore, building and processing large graphs can be computationally demanding and may not scale well to very long documents or large datasets. Despite their strengths, graph-based approaches may struggle in noisy or poorly structured texts where relationships are ambiguous or missing.

(2) Graph-free MHRC methods: Compared with the graph-based methods, graph-free methods avoid the explicit construction of graph structures. Instead, as shown in Fig. 11 (e.2), they typically rely on more straightforward architectures, such as using PLMs and self-attention mechanisms, to reason about final answer. It is worth noting that C2FReader (Shao et al., 2020) finds that graph structure can play an important role only when the PLMs are used in a feature-based manner. When the PLMs are used in the fine-tuning approach, the graph structure may not be helpful. Building upon this, recent graph-free methods often decompose complex multi-hop questions into simpler sub-questions or use relevance-based mechanisms to select and prioritize key information. These approaches integrate end-to-end learning, allowing the model to implicitly capture cross-sentence reasoning and dependencies without explicitly modeling relations as a graph, making the overall pipeline simpler and more flexible.

Although graph-free approaches suffer from a performance gap compared to the best graph-based models, numerous methods try to merge the gap by designing powerful mechanisms. DecompRC (Min et al., 2019) first decomposes the multi-hop question into several single-hop sub-questions according to a few reasoning types in parallel. Then, DecompRC leverages a single-hop reading comprehension model for every reasoning type to answer each sub-question and combines the answers according to the reasoning type. Finally, DecompRC leverages a decomposition scorer to judge which decomposition is the most suitable and outputs the answer from that decomposition as the final answer. QUARK (Groeneveld et al., 2020) scores individual sentences from an input set of paragraphs based on their relevance to the question. Then, it feeds the highest-scoring sentences to a span prediction model to produce an answer to the question. Finally, it scores sentences from the input set of paragraphs again to identify the supporting sentences using the answer. Inspired by the F1 score, $\mathcal{R}^3$ (Zhangyue et al., 2023) develops an F1 Smoothing mechanism to calculate the significance of each token within the smooth distribution. Furthermore, it incorporates curriculum learning (Bengio et al., 2009) and devises the linear decay label smoothing algorithm, gradually reducing the smoothing weight and allowing the model to focus on more challenging samples during training. FE2H (X. Li et al., 2023) introduces a document selection module that iteratively performs binary classification tasks to select relevant documents by simply adding a prediction layer on a PLM. Then, it trains the reader module on a single-hop QA dataset and transfers it into the multi-hop QA task inspired by humans' progressive learning process.

The key advantage of graph-free methods is that they are generally easier to implement and extend, as they avoid complex graph construction and instead leverage the strengths of PLMs. This leads to high efficiency and scalability to a large number of inputs. However, they can be less effective at capturing explicit multi-hop reasoning or relational information inherent to the question, which can create a performance gap compared to the best graph-based models, especially in scenarios where structured reasoning over multiple entities is crucial.

5.3.3. Open-domain question answering

Task definition: Open-domain question answering methods are required to retrieve relevant passages from a large-scale corpus and generate the final answer based on the retrieved passages. The ODQA task is more challenging than MRC and MHRC, which search the support facts within a smaller set of candidate passages. Recently, ODQA has been widely used to test the retrieval augmented generation (RAG) systems (Kim & Lee, 2024; Wang, Ren, et al., 2024).

Most ODQA methods follow a retrieve-and-read pipeline (Chen et al., 2017; Zhang et al., 2023; Zhu et al., 2021b). The objective of the retrieval phase is to retrieve evidence-related passages from a large symbolic knowledge corpus, such as Wikipedia.[2] The retriever can be divided into sparse retrieval and dense retrieval. Sparse retrieval methods rely on word-level matching to link vocabulary with documents. Notable methods include Boolean Retrieval (Salton et al., 1983), BM25 (Robertson & Zaragoza, 2009), and SPLADE (Formal et al., 2021). Dense retrieval methods capture deep semantic information to comprehend the underlying semantics of documents, thereby enhancing retrieval accuracy. Key examples include DPR (Karpukhin et al., 2020), ANCE (Xiong et al., 2021), RocketQA (Qu et al., 2021), DrBoost (Lewis et al., 2022), and SimLM (Wang et al., 2023b). The goal of the reading phase is comprehension and reasoning, akin to MRC, to derive answers based on the retrieved passages. Generally, existing readers can be categorized into extractive readers and generative readers. Extractive readers predict an answer span from the retrieved passages. Notable methods include REALM (Guu et al., 2020), Skylinebuilderretro (Wu et al., 2020), RETRO (Borgeaud et al., 2022), and BPR (Yamada et al., 2021). Generative Readers generate answers in natural language using sequence-to-sequence models. Key examples include RAG (Lewis et al., 2020), Fusion-in-Decoder (Izacard & Grave, 2021a), MDR (Xiong et al., 2021), and RALM (Yoran et al., 2024).

Recently, the emergence of LLMs has demonstrated their potential for open-domain question answering (Xin et al., 2024). As illustrated in Fig. 11(f), this section mainly investigates how to leverage the LLMs to optimize the retrieve-and-read pipeline. At the retrieval stage, the LLMs can be utilized for query reformulation and Enhanced retrieval. At the reader stage, LLMs can serve as augmented readers.

(1) Query reformulation methods: As shown in Fig. 11 (f.1), query reformulation methods focus on refining input questions to convey user intent more accurately. These methods generate alternative or expanded versions of a query, and then use a selection or ranking process to identify the most effective reformulations for retrieving evidence. These methods frequently utilize PLMs or heuristic strategies to automatically generate a variety of query candidates. Furthermore, they may also incorporate feedback loops or re-ranking mechanisms to enhance the relevance and utility of the retrieved information. By reformulating ambiguous, under-specified, or overly broad queries, these methods can better align the search process with the user's true information needs, ultimately enhancing retrieval accuracy and improving the quality of the answers provided.

[2] https://www.wikipedia.org

For instance, EAR (Chuang et al., 2023) first applies a query expansion model to generate a diverse set of queries and then uses a query reranker to select the ones that could lead to better retrieval results. QPaug (Kim et al., 2024) decomposes the original questions into multiple-step sub-questions. By augmenting the original question with detailed sub-questions and planning, QPaug can make the query more specific on what needs to be retrieved, improving the retrieval performance. Chain-of-Rewrite (Xin et al., 2024) finds that current methods face challenges stemming from term mismatch and limited interaction between information retrieval systems and LLMs. Hence, it leverages the guidance and feedback gained from the analysis to provide faithful and consistent extensions for effective question answering. Specifically, CLASS (Jiang et al., 2024) employs LLMs for query transformation via in-context learning in Cross-lingual ODQA tasks.

The advantages of query reformulation methods include enhanced retrieval performance and improved handling of ambiguous or complex user inputs. However, they may introduce additional computational overhead from generating and ranking multiple queries. Furthermore, poorly constructed reformulations can introduce noise or bias, potentially leading to irrelevant retrieval or increased system complexity, which can be challenging to manage and evaluate in practice.

(2) Enhanced retrieval methods: Enhanced retrieval methods adopt LLMs as knowledge sources to provide relevant contextual documents, thereby increasing the likelihood of uncovering the correct answer. As demonstrated in Fig. 11 (f.2), these methods typically follow a pipeline where the initial user query prompts LLMs to generate supplementary, knowledge-rich content, like hints and contextual passages, rather than retrieving existing documents from an external source. This generated text is then used by answer-generation models, either directly for in-context learning or as input for downstream reasoning modules. These methods often leverage the LLM's parametric knowledge and strong generative capabilities, combining them with question-focused prompting strategies to synthesize context that is tightly aligned with the given task. Although they do not directly utilize the knowledge from a symbolic knowledge base, we still introduce it due to its advanced features.

For instance, HintQA (Mozafari et al., 2024) produces multiple hints for each question. Then, it substitutes the retrieved passages and generated contexts with the generated hints. GenRead (Yu et al., 2023) prompts an LLM to generate contextual documents based on a given question and then reads the generated documents to produce the final answer. Self-Prompting (J. Li et al., 2024) prompts LLMs step by step to generate multiple pseudo QA pairs with background passages and explanations entirely from scratch. These generated elements are then utilized for in-context learning. MedGENIE (Frisoni et al., 2024) prompts a medical LLM to furnish multi-view background contexts for a given question. Then, it designs two readers for prompting LLMs and fine-tuning SLMs, respectively.

The key advantage of this type of approach is its ability to generate highly relevant, customized contexts, which can enhance the answer accuracy even in scenarios where explicit, symbolic knowledge retrieval struggles. These methods are flexible and can compensate for unavailable external databases. However, they also have notable downsides: generated evidence may sometimes be factually incorrect or hallucinated, lacking the verifiability that comes with retrieving real-world documents. Additionally, this approach may require significant computational resources, as it often involves multiple rounds of large-scale inference using LLMs.

(3) Augmented reader methods: Augmented reader methods are typically built upon Retrieval-Augmented Generation (RAG) frameworks and focus on refining the quality of retrieved documents before they are provided to the LLM. As shown in Fig. 11 (f.3), the standard pipeline involves retrieving a set of candidate passages and evaluating their relevance using a dedicated module. Only the most relevant context is then provided to the language model for answer generation. The core idea is that improving input relevance directly enhances the model's answer quality, robustness, and reasoning ability.

To overcome the challenges posed by irrelevant retrieved documents and overconfident scores, DAS (Cho et al., 2023) proposes a negation-based instruction to allow LLMs to abstain from answering. Then, it designs a score adjustment strategy to adjust the answer scores by reflecting the query generation score as the relevance between the given query-document pairs. Considering that LLMs cannot precisely assess the relevance of retrieved documents, thus likely leading to misleading or even incorrect utilization of external knowledge, REAR (Wang, Ren, et al., 2024) incorporates an assessment module that precisely assesses the relevance of retrieved documents and proposes an improved training method based on bi-granularity relevance fusion and noise-resistant training. RE-RAG (Kim & Lee, 2024) introduces a relevance estimator that not only provides relative relevance between contexts as previous rerankers did but also provides confidence, which can be used to classify whether the given context is helpful in answering the given question. FastFiD (Y. Huang et al., 2024) performs sentence selection post the output of the reader's encoder and maintains only the essential sentences as references for the reader's decoder, thereby significantly reducing the inference time for each query. Considering that the retrieved context may contain noise and irrelevant information and augmenting noisy context can potentially distract LLMs, TA-ARE (Z. Zhang et al., 2024) dynamically determines retrieval necessity and relies only on LLMs' parametric knowledge when deemed unnecessary.

The main advantage of these approaches is their ability to improve answer accuracy and robustness by effectively filtering out noise and irrelevant information, thus allowing LLMs to focus on truly helpful context. They also enhance efficiency by minimizing unnecessary computations on irrelevant input data. However, a key disadvantage is the added model complexity and potential for information loss if relevant passages are mistakenly filtered out. Additionally, designing effective relevance adjudication and integrating it with LLMs can increase development and training costs.

5.3.4. Multi-document question answering

Task definition: Multi-document question answering methods aims to find the supporting facts from multiple entire documents. MDQA demands a thorough understanding of the logical associations among the contents and structures of documents.

Although some methods also claim to perform document-based QA, they typically focus on paragraphs with key information, not the entire document. An entire document is usually much longer than a paragraph and contains more distracting information.

Table 14
Dataset statistics of text-based reasoning.

Datasets	Domain	# Ques.	# Pas.	Q&A source	Passages source	Links
CNN/DailyMail (Hermann et al., 2015)	News	1.38 M	312 K	Generate	CNN and DailyMail websites	https://
NewsQA (Trischler et al., 2017)	News	120K	12.7 K	Crowdsourcing	CNN websites	https://
PeopleDaily/CFT (Trischler et al., 2017)	News	880 K	60K	Generate	People Daily websites	https://
TriviaQA (Joshi et al., 2017)	News	95.9K	663 K	Crowdsourcing	Bing Search	https://
RACE (Lai et al., 2017)	Science	97.6 K	28 K	Expert	English Exam	https://
SQuAD1.1 (Rajpurkar et al., 2016)	General	107.8 K	536	Crowdsourcing	Wikipedia	https://
SQuAD2.0 (Rajpurkar et al., 2018)	General	150 K	536	Crowdsourcing	Wikipedia	https://
WikiQA (Y. Yang et al., 2015)	General	3,047	29.3 K	Crowdsourcing	Wikipedia	https://
HotpotQA (Yang et al., 2018b)	General	113 K	-	Crowdsourcing	Wikipedia	https://
Natural Questions (Kwiatkowski et al., 2019)	General	3.09 M	323 K	Crowdsourcing	Google Search	https://
2WikiMultiHopQA (Ho et al., 2020)	General	192.6 K	-	Crowdsourcing	Wikipedia and Wikidata	https://
IIRC (Ferguson et al., 2020)	General	13 K	-	Crowdsourcing	Wikipedia	https://
FanOutQA (Zhu et al., 2024)	General	8,339	-	Crowdsourcing	Wikipedia	https://

MDQA requires methods to identify support facts from the entire document, which is challenging. First, an entire document can be very lengthy, and supporting facts may comprise only a tiny part. Moreover, the text within a document is often on a single topic, making different passages highly related and difficult to distinguish.

As shown in Fig. 11(g), MDQA methods follow a pipeline that first segments the document into smaller passages or units for easier processing. Next, they employ information retrieval or graph-based techniques to identify candidate passages likely to contain supporting facts. Some approaches then use LLM to aggregate contextual evidence and generate answers. The multi-hop reasoning and structured traversal of this type of method can more effectively locate and integrate the disjoint pieces of relevant information scattered throughout long texts.

The first MDQA method is KGP (Wang, Lipka, et al., 2024), it formulates the proper context in prompting LLMs, consisting of a graph construction module and a graph traversal module. For graph construction, KGP creates a KG over multiple documents with nodes symbolizing passages or document structures (e.g., pages/tables) and edges denoting the semantic/lexical similarity between passages or document structural relations. For graph traversal, KGP designs an LLM-based graph traversal agent that navigates across nodes and gathers supporting passages to assist LLMs in MD-QA. Considering that some questions often require synthesizing information from multiple frequently unrelated documents, CuriousLLM (Z. Yang et al., 2025) fine-tunes a decoder-only LLM to emulate the curious nature of a human researcher to generate follow-up questions based on both the initial user query and passages retrieved in previous steps. These questions serve as a guide to identify the most relevant neighboring passages for the subsequent hops in the search process.

The main advantage of these methods is their ability to handle complex queries that require aggregating dispersed facts from extensive and potentially heterogeneous documents, enhancing answer accuracy and robustness. However, these methods often demand substantial computational resources and may struggle with efficiency due to the sheer size and complexity of document graphs. Additionally, the reliance on accurate passage selection or graph construction introduces points of failure that could affect overall performance, especially when the document's structure is ambiguous or poorly defined.

5.3.5. The datasets of text-based reasoning methods

In this section, we select, summarize, and analyze datasets related to the text-based reasoning task. We statistically summarize the information of the related dataset from multiple dimensions, including (1) Domain: the domain of knowledge corresponding to the datasets; (2)# Ques.: Question number; (3) # Pas.: Passage number; (4) Q&A source: The main construction methods of questions and answers. They are mainly divided into three categories: "Generate", "Expert", and "Crowdsourcing". "Generate" refers to the design of programs for automated generation, "Expert" refers to direct crawling from professional websites or carefully designed by domain experts, and "Crowdsourcing" refers to completion by crowdsourcing workers with general cultural levels; (5) Passage source: The source of passages; (6) Links: The storage address of the datasets. The statistical results are shown in Table 14.

The characteristics of a dataset play a central role in determining which reasoning methods are most effective for text-based reasoning tasks. When the dataset is large, methods that rely on statistical learning or complex pattern extraction tend to perform better, as they can exploit the abundance of training examples. However, once the dataset reaches a certain size, additional data often yields diminishing improvements, which may prompt the selection of methods that emphasize reasoning efficiency or structural interpretability instead. The internal structure of the dataset is equally important. Datasets containing varied reasoning chains, multiple question types, and wide domain coverage generally favor methods capable of flexible inference and transfer learning. In particular, datasets that require reasoning across multiple pieces of evidence encourage the use of approaches designed for multi-hop reasoning, since these methods are better suited to integrate dispersed information and construct coherent logical connections.

5.3.6. The comparison of text-based reasoning methods

In this section, we conduct a comparison of various text-based reasoning methods. The corresponding results are presented in Table 15.

Table 15
Comparison of text-based reasoning methods. Key dimensions: Performance (reasoning accuracy), Efficiency (computational efficiency and resource consumption), Interpretability (reasoning transparency), and Robustness (stability under out-of-distribution, data noise, or data sparsity scenarios).

Sub-task	Method	Performance	Efficiency	Interpretability	Robustness
Machine reading comprehension	Extractive MRC	Moderate *Adequate span supervision*	High *Lightweight architecture*	High *Explicit span evidence*	Low *Significant training sensitivity*
	Generative MRC	High *Powerful, flexible generator*	Low *Complex generative architecture*	Low *Opaque decision-making*	Moderate *Standard data dependency*
Multi-hop reading comprehension	Graph-based	High *Superior structural enhancement*	Low *Substantial construction overhead*	High *Explicit reasoning paths*	Low *Structural Fragility*
	Graph-free	Moderate *Limited relational modeling*	High *Simple architecture*	Low *Inexistent reasoning paths*	Low *Heavy model dependency*
Open-domain question answering	Query reformulation	Moderate *Sufficient query details*	Low *Redundant query generation*	Moderate *Semi-transparent reformulation*	Moderate *Tolerable reformulation noise*
	Enhanced retrieval	High *Customized, rich context*	Low *Lengthy generation overhead*	Low *Inadequate verifiable evidence*	Low *Heavy model dependency*
	Augmented reader	High *Relevant context selection*	Moderate *Additional filtering module*	Moderate *Partial assessment opacity*	High *Robust noise filtering*
Multi-document question answering	Knowledge graph prompt	High *Effective information integration*	Low *Substantial construction overhead*	High *Explicit evidence*	Moderate *Noticeable graph dependency*

In machine reading comprehension, extractive MRC methods (Huang et al., 2023) prioritize efficiency and interpretability but lack robustness due to the significant training sensitivity. In contrast, generative MRC models (R. Li et al., 2024) provide high performance and flexibility, though at the expense of efficiency and interpretability. For multi-hop reading comprehension, graph-based methods (Fang et al., 2020) yield high performance and explicit reasoning paths, but suffer from construction overhead and structural fragility. Graph-free methods (X. Li et al., 2023) improve efficiency but heavily depend on PLMs, limiting their transparency and robustness. For open-domain question answering, query reformulation (Jiang et al., 2024) achieves competitive performance metrics, though it suffers from reduced efficiency due to the overhead of multiple queries. Enhanced retrieval (Frisoni et al., 2024) and augmented reader (Y. Huang et al., 2024) techniques prioritize performance over efficiency and interpretability, though augmented readers provide better robustness against noise. Multi-document question answering methods (Wang, Lipka, et al., 2024) excel in performance, interpretability, and robustness, but complex graph construction significantly reduces their efficiency.

In summary, while developing a comprehensive text-based reasoning system is challenging, it is of great significance. Achieving this may require introducing entirely new text-processing architectures to better balance various metrics.

5.4. Heterogeneous reasoning

Graph-based reasoning, table-based reasoning, and text-based reasoning have all been individually studied extensively. However, reasoning based on two or more heterogeneous symbolic knowledge bases, known as heterogeneous question answering (Heterogeneous QA), is under-studied (Zhang et al., 2024a). Exploring how to explore the knowledge from multiple heterogeneous symbolic knowledge bases fully is extremely important for enhancing the practicality of reasoning methods.

5.4.1. Heterogeneous question answering

Task definition: Heterogeneous question answering methods aim to find the evidence from heterogeneous knowledge bases to answer a knowledge-intensive question.

Some methods investigate how to leverage symbolic knowledge from different sources, including those on closed domain Chen et al. (2020), Lei et al. (2023), Liu et al. (2024), Miller et al. (2016), Pramanik et al. (2024) and open domain W. Chen et al. (2021), Han and Gardent (2023), Ma et al. (2022a), Zhao et al. (2024), but very limited existing work experiments on graph, table, and text, simultaneously. To promote relevant research, CONVMIX (Christmann et al., 2022) and COMPMIX (Christmann et al., 2024) collect the heterogeneous QA datasets that require knowledge from all three heterogeneous sources. A simple solution for handling heterogeneous QA is to assemble several specialized systems. In this approach, the input question is dispatched to multiple sub-systems, and one of them is chosen to provide the final answer. Although this method can leverage state-of-the-art models optimized for various information sources, it significantly increases the complexity of the entire system. Additionally, it poses challenges in addressing questions that require reasoning across multiple sources of information (Oguz et al., 2022). Hence, constructing an integrated system compatible with multiple heterogeneous symbolic knowledge bases is essential and promising. Current methods with integrated systems can be divided into structured-unified and human-imitated methods, as illustrated in Fig. 11(h).

(1) Structure-unified methods: As indicated in Fig. 11 (h.1), the structured-unified methods convert multiple heterogeneous bases into one type to reason about final answer. The first type of structured-unified method converts different structures to unstructured text, and the second one converts different structures to structured graphs. This unification enables the downstream application of powerful models, such as PLMs or GNNs, for reasoning and answer generation. This type of method highlights the

Table 16
Dataset statistics of heterogeneous reasoning.

Datasets	Domain	# Ques.	Q&A source	KG	KG size	Text	# Pas.	Table	# Table	OR	HQ	OD	Links
WIKIMOVIES (Miller et al., 2016)	Movie	100 K	Generate	✓	–	✓	17 K	✗	–	✓	✗	✗	https://
HYBRIDQA (Chen et al., 2020)	General	70 K	Crowdsourcing	✗	–	✓	293 K	✓	13 K	✗	✓	✓	https://
MULTIMODALQA (Talmor et al., 2021)	General	30 K	Generate	✗	–	✓	218 K	✓	10 K	✓	✗	✓	https://
OTT-QA (W. Chen et al., 2021)	General	45 K	Crowdsourcing	✗	–	✓	–	✓	–	✓	✓	✓	https://
MANYMODALQA (Hannan et al., 2020)	General	10 K	Crowdsourcing	✗	–	✓	3,789	✓	528	✗	✓	✓	https://
TAT-QA (Zhu et al., 2021a)	General	17 K	Crowdsourcing	✗	–	✓	3,902	✓	7,431	✗	✓	✗	https://
FINQA (Chen et al., 2021)	Finance	8,281	Crowdsourcing	✗	–	✓	–	✓	–	✗	✓	✗	https://
HETPQA (Shen et al., 2022a)	Product	6,000	Crowdsourcing	✗	–	✓	–	✓	–	✗	✓	✗	https://
COMPMIX (Christmann et al., 2024)	General	9,410	Crowdsourcing	✓	–	✓	–	✓	–	✓	✓	✓	https://

importance of aligning representations to enhance cross-source reasoning, simplify the management of diverse data formats, and leverage a unified model architecture for better knowledge integration and QA performance.

In the first type of method, UniK-QA (Oguz et al., 2022) flattens the lists, tables, and KGs to text using simple heuristics methods. Then, it adopts a text-based QA method as the solution to make full use of the powerful PLMs. UDT-QA (Ma et al., 2022b) unifies both representation and model for ODQA over structured data and unstructured text. The key idea is to augment the retriever with a data-to-text verbalizer for accessing heterogeneous knowledge bases, i.e., KGs from WikiData, tables and texts from Wikipedia. Convinse (Christmann et al., 2022) learns an explicit and structured representation of an incoming question and its conversational context. It harnesses this frame-like representation to uniformly capture relevant evidence from KB, text, and tables. Finally, it adopts a fusion-in-decoder model to generate the answer.

In the second type of method, TrustUQA (Zhang et al., 2025) designs a condition graph to unify tables and KGs and uses an LLM and demonstration-based two-level method for reasoning on the condition graph. Explaignn (Christmann et al., 2023) constructs a heterogeneous graph from entities and evidence snippets retrieved from a KG, a text corpus, web tables, and infoboxes. This large graph is then iteratively reduced via GNNs incorporating question-level attention until the best answers and explanations are distilled out. The former gives up the advantage of using formal query languages on structured data, which can support operations such as ranking and averaging. The latter gives up the advantage of the expressiveness and versatility of free-text knowledge representation. As Zhang et al. (2024a) points out, the first type of method sacrifices the benefits of using formal query languages on structured data, which can support operations like ranking and averaging. The second type of method relinquishes the expressiveness and versatility offered by free-text knowledge representation. Recently, ILJR (S. Li et al., 2025) leverages multi-source knowledge prompting to enhance interpretable legal judgment prediction. By combining chain prompt reasoning with contrastive knowledge fusion, it establishes a solid foundation for judgment that incorporates both factual logic and pertinent legal knowledge.

A key advantage of structured-unified methods is their ability to integrate and reason over diverse knowledge sources using a unified format, which greatly simplifies subsequent model design and allows full utilization of advanced language or graph-based models. This unified approach enhances flexibility and enables more consistent answer generation across different data types. However, an inherent drawback is the potential loss of important information during the conversion process, as transforming structured data into text may strip away explicit relationships, and graph unification can miss subtle semantic details found in free-text. Moreover, this extra unification step can introduce complexity, additional pre-processing overhead, and possible information distortion, ultimately affecting the precision and completeness of the generated answers.

(2) Human-imitated methods: As shown in Fig. 11 (h.2), the human-imitated methods integrate reasoning steps over heterogeneous knowledge bases by mimicking how humans find responses to questions, which break down QA solution processes as tool calls and thoughts. This type of method generally involves sequentially identifying relevant resources, utilizing external tools like retrieval systems or databases, and iteratively refining interim findings. By structuring the reasoning process in stages, with each stage governed by tool selection and evidence integration, these approaches enable LLMs to systematically collect, verify, and synthesize diverse information sources to support the reasoning of final answer. This process closely resembles the step-by-step way humans handle complex questions.

For instance, HumanIQ (Lehmann et al., 2024) proposes a human-like approach that teaches LLMs to gather heterogeneous information by imitating how humans use retrieval tools. During the preparation stage, the method is required to identify suitable tools and solution processes using those tools. Then, it leverages an LLM to replicate the solution processes at the inference stage. Similarly, SPAGHETTI (Zhang et al., 2024a) obtains evidence from heterogeneous sources in parallel, including structured KG, plain text, linearized tables, infoboxes, and LLM-generated claims that are verified, and gathers that evidence to generate the final answer using a few-shot LLM.

The key advantage of human-imitated methods is that they offer improved transparency and flexibility by explicitly modeling reasoning paths and integrating various tools, leading to more interpretable and accurate answers. However, they can introduce additional system complexity, higher computational overhead, and potential error propagation across steps. Furthermore, designing robust tool selection and coordination mechanisms remains challenging, especially as the diversity and scale of available tools grow.

Table 17
Comparison of heterogeneous reasoning methods. Key dimensions: Performance (reasoning accuracy), Efficiency (computational efficiency and resource consumption), Interpretability (reasoning transparency), and Robustness (stability under out-of-distribution, data noise, or data sparsity scenarios).

Method	Performance	Efficiency	Interpretability	Robustness
Structure-unified	Moderate *Partial information loss*	Low *Substantial pre-processing overhead*	High *Explicit evidence*	Moderate *Intrinsic unification sensitivity*
Human-imitated	High *Stepwise evidence integration*	Low *Heavy multi-tool overhead*	High *Explicit reasoning paths*	Low *Significant error propagation*

5.4.2. The datasets of heterogeneous reasoning methods

In this section, we select, summarize, and analyze datasets related to the heterogeneous reasoning task. We statistically summarize the information of the related dataset from multiple dimensions, including (1) Domain: the domain of knowledge corresponding to the datasets; (2)# Ques.: Question number; (3) Q&A source: The main construction methods of questions and answers. They are mainly divided into three categories: "Generate", "Expert", and "Crowdsourcing". "Generate" refers to the design of programs for automated generation, "Expert" refers to direct crawling from professional websites or carefully designed by domain experts, and "Crowdsourcing" refers to completion by crowdsourcing workers with general cultural levels; (4) KG: Whether they use KG as knowledge bases; (5) KG size: the number of triplets in KG used; (6) Text: Whether they use Text as knowledge bases; (7) # Pas.: the number of passages; (8) Table: Whether they use Table as knowledge bases; (8) # Table: The number of tables; (9) OR: Whether they support open retrieval; (10) HQ: Whether the questions are constructed by human; (11) OD: Whether the answers could be found in open domain; (12) Links: The storage address of the datasets. The statistical results are shown in Table 16.

The choice of a reasoning method is closely influenced by the characteristics of the dataset. The quality and diversity of samples determine whether a few-shot heterogeneous reasoning approach can effectively integrate information from graphs, tables, and text. When fine-tuning is involved, the dataset size becomes a deciding factor. Larger and more varied datasets support more stable training, reduce overfitting, and enable the model to capture complex cross-modal relationships. The distribution of the dataset also affects which reasoning strategy is most appropriate. Datasets limited to a single domain may favor specialized methods but often result in poor performance on multi-domain tasks. In contrast, datasets that cover a wide range of domains encourage the selection of reasoning methods designed for adaptability. Such diversity helps the model develop the capacity to generalize across heterogeneous data types and unseen reasoning contexts.

5.4.3. The comparison of heterogeneous reasoning methods

In this section, we compare two heterogeneous reasoning methods. The corresponding results are presented in Table 17.

The structure-unified (S. Li et al., 2025) and human-imitated methods (Zhang et al., 2024a) both excel in interpretability through explicit evidence or reasoning paths but struggle with low efficiency due to preprocessing and tool coordination overheads respectively. They are primarily different in performance and robustness. The former yield moderate performance and robustness caused by partial information loss and intrinsic unification sensitivity. Conversely, the latter achieve strong performance via stepwise evidence integration but face limited robustness due to error propagation across reasoning steps.

In summary, addressing the efficiency and robustness of heterogeneous reasoning methods remains a pressing challenge that warrants further attention. Furthermore, the establishment of comprehensive evaluation standards is also essential for the development of next-generation reasoning frameworks. First, the evaluation criteria should evolve from a single dimension to multiple dimensions (F. Xu et al., 2025). It should comprehensively take performance, efficiency, interpretability, and robustness into consideration. Second, for each dimension, there should also be consideration of multiple metrics for more comprehensive evaluation. For performance, beyond standard exact match or F1 scores, it should be measured by step-wise correctness in multi-hop reasoning (Lee & Hockenmaier, 2025; Xia et al., 2025). For efficiency, the inference latency, memory usage, and token consumption are all important metrics in practical deployment for reasoning systems (Feng et al., 2025; Liu et al., 2025; Qu et al., 2025). For interpretability, the focus should be on faithfulness (Li et al., 2025; Tutek et al., 2025) and transparency (Y. Chen et al., 2025). Faithfulness ensures the reasoning steps truly reflect the model's internal logic. Transparency assesses whether the reasoning prcess is clear and human-understandable. For robustness, the criteria must include stability (Yu et al., 2025) and noise-resistance (Gan et al., 2024). The system should maintain stability across different prompt variations. It must also remain accurate when faced with irrelevant information or adversarial distractors.

5.5. The summary of collaborative reasoning based on symbolic and parametric knowledge bases

In summary, collaborative reasoning makes excellent use of the verifiability of symbolic knowledge and the flexibility of parametric knowledge (Sun et al., 2024; Wu et al., 2024). However, it also introduces multiple trade-offs between performance and efficiency. Heterogeneous reasoning methods (Zhang et al., 2025, 2024a) offer high performance but suffer from complexity, while single source reasoning methods (Toroghi et al., 2024; Z. Yang et al., 2025; P. Yu et al., 2025) are efficient but less effective. Striking a balance between computational cost and performance is essential for practical collaborative reasoning systems. Furthermore, the combination of symbolic and parametric knowledge may introduce knowledge conflicts that damage system robustness (Xu et al., 2024). Future research must therefore focus on techniques that can better balance the interaction and combination between symbolic and parametric knowledge during reasoning.

6. Future directions

Building on the comprehensive review of reasoning methods from the knowledge base perspective, this section outlines nine promising future research directions that are essential for advancing reasoning systems in complex, real-world applications.

6.1. Addressing data sparsity and enhancing generalization

To build robust reasoning systems, overcoming the sparsity of high-quality data and ensuring generalization are critical. These two challenges are deeply interconnected, as data sparsity often leads to overfitting on seen data and poor generalization on unseen out-of-distribution (OOD) instance (Opedal et al., 2025). Future research should focus on the following directions:

- **Scalable and diverse data synthesis:** When domain-specific annotated data are limited, using LLMs to generate synthetic reasoning samples can significantly expand training sets and reduce manual costs (Maharana & Bansal, 2022). In this direction, future research should move beyond data quantity and focus on the diversity and quality of synthetic data to enhance OOD generalization in practical deployment. Sachdeva et al. (2024).
- **Transfer learning from related domains:** Leveraging reasoning skills and logical structures from data-rich domains can bridge the gap in data-scarce specialized tasks (Wenzel et al., 2022). Research should investigate how to effectively transfer these capabilities while avoiding catastrophic forgetting of in-domain knowledge.

6.2. Cost efficient reasoning

Existing advanced reasoning strategies often become inefficient by generating excessive tokens (Y. Du et al., 2024; Wang et al., 2023). Practical applications require reasoning systems that balance high performance with affordable time and cost. Future research should explore the following directions:

- **Incorporating symbolic reasoning strategies:** Although rule-based and statistical-based reasoning methods are less effective than LLMs in overall effectiveness, they can efficiently handle basic or easy reasoning steps within complex tasks. Integrating these models with LLMs for collaborative reasoning can achieve cost efficient.
- **Efficiency adaptive reasoning:** Humans naturally select different reasoning strategies depending on problem complexity. Developing systems that adaptively apply fast thinking for simple queries and detailed thinking for difficult ones offers a promising path to cost efficient reasoning (M. Xu et al., 2024).

6.3. Ensuring the safety of reasoning

Ensuring reasoning safety involves two aspects: preventing harmful content in reasoning processes and results, and protecting private data in local symbolic knowledge bases from leaking to cloud LLMs.

- **Preventing toxicity while reasoning:** Current toxicity detection primarily targets hate speech or bias in general texts (Fortuna & Nunes, 2018; Haber et al., 2023; Kebriaei et al., 2024; Liang et al., 2021). Unlike general texts, reasoning processes often involve specialized knowledge from domains like chemistry and biology. For example, harmful molecular formulas designed for illegal drug are difficult to detect using only text semantics. Hence, augmenting traditional toxicity detection with domain expertise is essential.
- **Protecting data privacy while reasoning:** Many applications integrate local symbolic knowledge bases with third-party LLMs for retrieval augmented generation to improve reasoning (Gao et al., 2023). However, this raises privacy risks, as local knowledge bases may contain sensitive information and third-party LLMs often operate as black boxes. To protect privacy, sensitive information can be anonymized before sending to LLMs (Ning et al., 2026). Hence, exploring the anonymization-compatible reasoning techniques is a practical direction.

6.4. Developing reasoning systems across disciplines

Most current reasoning systems focus on a single discipline (Ahn et al., 2024; Merenda et al., 2026; Qiu et al., 2025; Wu et al., 2025; Z. Zhang et al., 2025), limiting their ability to generalize across different disciplines with interconnected knowledge. Many real-world problems are inherently interdisciplinary and require reasoning systems that integrate knowledge from multiple disciplines (Huang et al., 2024; Yue et al., 2024; Zhou et al., 2026). To advance interdisciplinary reasoning, several directions merit further exploration:

- **Cross-discipline knowledge representation and fusion:** Developing mechanisms to encode discipline-specific knowledge (e.g., mathematical equations and molecular structures) into a unified representation space (Lee et al., 2026; Wang et al., 2026). By dynamically integrating these representations during reasoning, models can synergistically leverage cross-discipline knowledge to solve complex problems.
- **Cross-discipline transfer learning:** Exploring methods for effective transfer learning and discipline adaptation, enabling reasoning systems to leverage relevant knowledge and skills acquired in one discipline to assist with tasks in another (Z. Wang et al., 2024). This may involve meta-learning or multi-task learning strategies.

6.5. Enhancing reasoning explainability

Although advanced reasoning systems perform well, their decision-making processes often remain opaque (Sun et al., 2026; Zhang et al., 2025). Clear and faithful explanations for reasoning process are critical for high-stakes domains such as medical (Hossain et al., 2025) and legal (Li et al., 2025). Potential research directions include:

- **Intrinsic and faithful explainability:** Research should focus on architectures that generate intermediate reasoning steps intrinsically, ensuring that explanations faithfully reflect the model's internal reasoning process (Lyu et al., 2024). For example, integrating symbolic programs into the reasoning process can improve explainability by aligning each step with both model behavior and domain-specific logic.
- **Evaluation metrics and benchmarks:** Establishing standardized metrics and comprehensive datasets for assessing the quality of explanations in reasoning systems (Shen et al., 2025), which will facilitate systematic progress in this area.

6.6. Personalized reasoning system

As AI assistants become more integrated into daily life, their reasoning systems must tailor outputs to users with diverse backgrounds, goals, and knowledge levels (Kim et al., 2025; Luo et al., 2025). For instance, experts may require concise technical explanations while beginners benefit from detailed guidance and contextual information. Future research should focus on developing reasoning systems that dynamically adjust their behavior through user modeling and interactive design. Potential directions include:

- **User modeling for adaptive reasoning:** One promising direction involves constructing user models that capture individual expertise, experience, and communication styles (Kim et al., 2026). These profiles allow systems to select appropriate explanation styles and highlight relevant background information for the user.
- **Interactive feedback for personalized learning:** Another direction involves designing systems that encourage or simulate users to provide feedback on responses (Wu et al., 2025). The feedback loop allows the system to adjust its responses to evolving user needs, which enhances both satisfaction and accuracy.

6.7. Addressing biases and fairness in reasoning

Reasoning systems can unintentionally propagate biases, such as gender bias and cultural bias, present in their training data (Gupta et al., 2024; Wu et al., 2025a). When deployed in real-world applications, these systems risk reinforcing and amplifying existing inequities. As reasoning systems play increasingly influential roles in socially sensitive domains (Shaikh et al., 2023), mitigating these risks becomes critically important. Potential research directions include:

- **Bias detection and correction mechanisms:** Creating algorithms that automatically identify and correct biased reasoning patterns to ensure fair outcomes (Lin et al., 2025; Shrestha & Srinivasan, 2025). This involves monitoring model outputs and intermediate steps for systematic bias against specific groups.
- **Fairness-aware reasoning frameworks:** Developing systems that explicitly model fairness during both learning and inference (Kabra et al., 2025). Possible approaches include incorporating ethical guidelines as constraints and employing multi-objective optimization to balance accuracy and fairness.

6.8. Integrating retrieval augmented generation with symbolic reasoning

RAG mitigates hallucinations and outdated information by retrieving external knowledge to support reasoning (Hu et al., 2025; QianyiHu et al., 2025). Current RAG frameworks often directly feed retrieved information into LLMs to generate final results (Wang, Lipka, et al., 2024; Z. Yang et al., 2025). In such pipelines, the utilization and fusion of retrieved knowledge lack explicit interpretability and rely heavily on the internal mechanisms of LLMs. While symbolic reasoning excels in explicit logic and verifiable inference. The integration of RAG and symbolic reasoning presents several research challenges and opportunities.

- **Symbolic reasoning over retrieved knowledge:** While symbolic reasoning can operate over the retrieved information to further explore important knowledge (Saxena & Gaur, 2026), such as rule chaining, it may enhance both reasoning precision and interpretability.
- **RAG with iterative symbolic verification:** Another direction involves using symbolic modules to verify the logical validity of model outputs (Petruzzellis et al., 2025). This verification can extend to iterative loops where symbolic feedback guides further retrieval or reasoning.

6.9. Addressing knowledge conflicts in collaborative reasoning

Collaborative reasoning that integrates symbolic and parametric knowledge bases has shown great potential for improving reasoning accuracy and interpretability (Zhang et al., 2025). However, a critical challenge lies in the knowledge conflict between these two forms of knowledge bases (Liu et al., 2026; Xu et al., 2024). These conflicts can cause inconsistent results and reduce the reliability of collaborative systems. Future research should focus on detecting and reconciling knowledge conflicts in these systems. Several promising directions include the following:

- **Conflict detection and representation:** Developing effective mechanisms to identify conflicts between symbolic and parametric knowledge (Jin et al., 2024). Researchers could develop alignment models to measure semantic and logical consistency between different knowledge types.
- **Conflict-aware reasoning strategies:** Detected knowledge conflicts should be managed dynamically instead of being avoided. Methods like conflict-aware prompting (Choi et al., 2025) and adaptive decoding (Khandelwal et al., 2025) can help LLMs reconcile symbolic constraints with contextual flexibility.
- **Harmonizing conflicts among heterogeneous knowledge:** Since knowledge is often stored in diverse formats (Christmann et al., 2024), future systems must learn to harmonize conflict knowledge among heterogeneous sources, such as KGs and tables. This process may refer to multi-objective training strategies designed to harmonize inconsistencies across different formats.

7. Conclusion

In this paper, we provide a comprehensive survey on reasoning methods with a specific focus on the usage of knowledge bases, addressing a gap in existing literature. By categorizing knowledge bases into symbolic and parametric types, we offer a novel perspective on how reasoning can be enhanced when leveraging different formats of stored information. Symbolic knowledge bases, such as KGs and tables, offer explicit and human-readable knowledge, while parametric knowledge bases encode knowledge implicitly within parameters, such as large language models. Then, we investigate how these knowledge bases, individually and in combination, support reasoning processes. Our comprehensive taxonomy and investigation highlights that the next evolution of reasoning systems must transcend simple proficiency to address three critical frontiers: (1) robustness and efficiency, overcoming instability in OOD scenarios and high computational costs, (2) trustworthiness, ensuring safety against toxicity, privacy leakage, and resolving conflicts between symbolic and parametric knowledge, and (3) versatility, expanding capabilities into interdisciplinary, and personalized contexts. Ultimately, we conclude that advancing reasoning capabilities requires deeply integrating the explicit structure of symbols with the flexibility of parameters. Furthermore, a shift toward more comprehensive design and evaluation standards is necessitated for next-generation reasoning frameworks, where verifiability, knowledge consistency, and structural transparency are prioritized over a mere emphasis on raw performance. We hope these insights inspire further exploration and advancements in the field, ultimately contributing to developing artificial intelligence systems with more robust reasoning capabilities.

CRediT authorship contribution statement

Mayi Xu: Writing – original draft, Conceptualization, Investigation, Methodology, Formal analysis. **Yunfeng Ning:** Writing – original draft. **Yongqi Li:** Writing – original draft. **Jianhao Chen:** Writing – original draft. **Jintao Wen:** Writing – original draft. **Yao Xiao:** Writing – original draft. **Shen Zhou:** Writing – original draft. **Birong Pan:** Writing – original draft. **Zepeng Bao:** Writing – original draft. **Xin Miao:** Writing – original draft. **Hankun Kang:** Writing – original draft. **Ke Sun:** Writing – original draft. **Tieyun Qian:** Writing – original draft, Project administration, Funding acquisition, Supervision.

Acknowledgments

This work was supported by the grants from the National Natural Science Foundation of China (NSFC) project (Grant No. 62276193, 62576256) and the Fundamental Research Funds for the Central Universities, China (Grant No. 2042022dx0001).

Data availability

No data was used for the research described in the article.

References

Abboud, R., Ceylan, İ. İ., Lukasiewicz, T., & Salvatori, T. (2020). BoxE: A box embedding model for knowledge base completion. In *Advances in neural information processing systems 33: annual conference on neural information processing systems 2020, neurIPS 2020, December 6-12, 2020, virtual*. URL: https://proceedings.neurips.cc/paper/2020/hash/6dbbe6abe5f14af882ff977fc3f35501-Abstract.html.

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. (2023). Gpt-4 technical report. http://dx.doi.org/10.48550/ARXIV.2303.08774, arXiv Preprint arXiv:2303.08774.

Ahn, J., Verma, R., Lou, R., Liu, D., Zhang, R., & Yin, W. (2024). Large language models for mathematical reasoning: Progresses and challenges. In *Proceedings of the 18th conference of the European chapter of the association for computational linguistics: student research workshop* (pp. 225–237). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2024.eacl-srw.17, URL: https://aclanthology.org/2024.eacl-srw.17/.

Ai, L., Hui, Z., Liu, Z., & Hirschberg, J. (2024). Enhancing pre-trained generative language models with question attended span extraction on machine reading comprehension. In *Proceedings of the 2024 conference on empirical methods in natural language processing* (pp. 10046–10063). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2024.emnlp-main.560.

Ali, M., Berrendorf, M., Hoyt, C. T., Vermue, L., Galkin, M., Sharifzadeh, S., Fischer, A., Tresp, V., & Lehmann, J. (2022). Bringing light into the dark: A large-scale evaluation of knowledge graph embedding models under a unified framework. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(12), 8825–8845. http://dx.doi.org/10.1109/TPAMI.2021.3124805.

Aly, R., Guo, Z., Schlichtkrull, M. S., Thorne, J., Vlachos, A., Christodoulopoulos, C., Cocarascu, O., & Mittal, A. (2021). The fact extraction and VERification over unstructured and structured information (FEVEROUS) shared task. In *Proceedings of the fourth workshop on fact extraction and vERification* (pp. 1–13). Dominican Republic: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2021.fever-1.1, URL: https://aclanthology.org/2021.fever-1.1.

Aly, R., & Vlachos, A. (2024). Tabver: Tabular fact verification with natural logic. *Transactions of the Association for Computational Linguistics*, *12*, 1648–1671, URL: https://doi.org/10.1162/tacl_a_00722.

Amayuelas, A., Zhang, S., Rao, S. X., & Zhang, C. (2022). Neural methods for logical reasoning over knowledge graphs. In *The tenth international conference on learning representations (ICLR 2022)*. OpenReview, URL: https://openreview.net/forum?id=tgcAoUVHRIB.

Arakelyan, E., Daza, D., Minervini, P., & Cochez, M. (2021). Complex query answering with neural link predictors. In *9th international conference on learning representations, ICLR 2021, virtual event, Austria, May 3-7, 2021*. OpenReview.net, URL: https://openreview.net/forum?id=Mos9F9kDwkz.

Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., et al. (2023). Qwen technical report. http://dx.doi.org/10.48550/ARXIV.2309.16609, arXiv preprint arXiv:2309.16609.

Bai, L., Chen, M., Zhu, L., & Meng, X. (2023). Multi-hop temporal knowledge graph reasoning with temporal path rules guidance. *Expert Systems with Applications*, *223*, Article 119804. http://dx.doi.org/10.1016/J.ESWA.2023.119804.

Bai, Y., Lv, X., Li, J., Hou, L., Qu, Y., Dai, Z., & Xiong, F. (2022). SQUIRE: a sequence-to-sequence framework for multi-hop knowledge graph reasoning. In *Proceedings of the 2022 conference on empirical methods in natural language processing, EMNLP 2022, abu dhabi, United arab emirates, December 7-11, 2022* (pp. 1649–1662). Association for Computational Linguistics, http://dx.doi.org/10.18653/V1/2022.EMNLP-MAIN.107.

Bai, L., Yu, W., Chen, M., & Ma, X. (2021). Multi-hop reasoning over paths in temporal knowledge graphs using reinforcement learning. *Appl. Soft Comput.*, *103*, Article 107144. http://dx.doi.org/10.1016/J.ASOC.2021.107144.

Bai, L., Zhang, H., An, X., & Zhu, L. (2025). Few-shot multi-hop reasoning via reinforcement learning and path search strategy over temporal knowledge graphs. *Information Processing & Management*, *62*(3), Article 104001. http://dx.doi.org/10.1016/j.ipm.2024.104001, URL: https://www.sciencedirect.com/science/article/pii/S0306457324003601.

Balažević, I., Allen, C., & Hospedales, T. (2019). Multi-relational poincaré graph embeddings. In *Proceedings of the 33rd international conference on neural information processing systems*. Red Hook, NY, USA: Curran Associates Inc., URL: https://proceedings.neurips.cc/paper/2019/hash/f8b932c70d0b2e6bf071729a4fa68dfc-Abstract.html.

Balazevic, I., Allen, C., & Hospedales, T. (2019). TuckER: Tensor factorization for knowledge graph completion. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)* (pp. 5185–5194). Hong Kong, China: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/D19-1522, URL: https://aclanthology.org/D19-1522.

Bandyopadhyay, D., Bhattacharjee, S., & Ekbal, A. (2025). Thinking machines: A survey of llm based reasoning strategies. http://dx.doi.org/10.48550/ARXIV.2503.10814, arXiv preprint arXiv:2503.10814.

Banning, M. (2008). Clinical reasoning and its application to nursing: concepts and research studies. *Nurse Education in Practice*, *8*(3), 177–183. http://dx.doi.org/10.1016/j.nepr.2007.06.004, URL: https://www.sciencedirect.com/science/article/pii/S1471595307000595.

Bao, J., Duan, N., Yan, Z., Zhou, M., & Zhao, T. (2016). Constraint-based question answering with knowledge graph. In *Proceedings of COLING 2016, the 26th international conference on computational linguistics: technical papers* (pp. 2503–2514). Osaka, Japan: The COLING 2016 Organizing Committee, http://dx.doi.org/10.18653/V1/2022.NAACL-INDUSTRY.31.

Bao, L., Wang, Y., Song, X., & Sun, T. (2025). HGCGE: hyperbolic graph convolutional networks-based knowledge graph embedding for link prediction. *Knowledge and Information Systems*, *67*(1), 661–687. http://dx.doi.org/10.1007/S10115-024-02247-8.

Baradaran, R., & Amirkhani, H. (2021). Ensemble learning-based approach for improving generalization capability of machine reading comprehension systems. *Neurocomputing*, *466*, 229–242. http://dx.doi.org/10.1016/J.NEUCOM.2021.08.095.

Bengio, Y., Louradour, J., Collobert, R., & Weston, J. (2009). Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning* (pp. 41–48). http://dx.doi.org/10.1145/1553374.1553380, URL: https://dl.acm.org/doi/abs/10.1145/1553374.1553380.

Berant, J., Chou, A., Frostig, R., & Liang, P. (2013). Semantic parsing on freebase from question-answer pairs. In *Proceedings of the 2013 conference on empirical methods in natural language processing* (pp. 1533–1544). Seattle, Washington, USA: Association for Computational Linguistics, http://dx.doi.org/10.18653/V1/D13-1160.

Berkovitch, Y., Glickman, O., Somech, A., & Wolfson, T. (2025). Generating tables from the parametric knowledge of language models. In *Proceedings of the 4th international workshop on knowledge-augmented methods for natural language processing* (pp. 50–65). http://dx.doi.org/10.18653/v1/2025.knowledgenlp-1.4, URL: https://aclanthology.org/2025.knowledgenlp-1.4/.

Besta, M., Blach, N., Kubicek, A., Gerstenberger, R., Podstawski, M., Gianinazzi, L., Gajda, J., Lehmann, T., Niewiadomski, H., Nyczyk, P., et al. (2024). Graph of thoughts: Solving elaborate problems with large language models. *38*, In *Proceedings of the AAAI conference on artificial intelligence* (16), (pp. 17682–17690). http://dx.doi.org/10.1609/AAAI.V38I16.29720.

Bhargava, P., & Ng, V. (2022). Commonsense knowledge reasoning and generation with pre-trained language models: A survey. In *Proceedings of the AAAI conference on artificial intelligence*: *vol. 36*, (11), (pp. 12317–12325). http://dx.doi.org/10.1609/AAAI.V38I16.29720.

Bi, S., Miao, Z., & Min, Q. (2025). A modular dual learning for improving question answering and generation over knowledge graphs. *IEEE Transactions on Audio, Speech and Language Processing*, *33*, 401–417. http://dx.doi.org/10.1109/TASLPRO.2025.3527218, URL: https://ieeexplore.ieee.org/document/10833763.

Bi, B., Wu, C., Yan, M., Wang, W., Xia, J., & Li, C. (2019). Incorporating external knowledge into machine reading for generative question answering. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)* (pp. 2521–2530). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/D19-1255, URL: https://aclanthology.org/D19-1255/.

Bo, X., Zhang, Z., Dai, Q., Feng, X., Wang, L., Li, R., Chen, X., & Wen, J.-R. (2024). Reflective multi-agent collaboration based on large language models. *Advances in Neural Information Processing Systems*, *37*, 138595–138631, URL: http://papers.nips.cc/paper_files/paper/2024/hash/fa54b0edce5eef0bb07654e8ee800cb4-Abstract-Conference.html.

Boisvert, L., Thakkar, M., Gasse, M., Caccia, M., de Chezelles, T., Cappart, Q., Chapados, N., Lacoste, A., & Drouin, A. (2024). Workarena++: Towards compositional planning and reasoning-based common knowledge work tasks. *Advances in Neural Information Processing Systems*, *37*, 5996–6051, URL: https://proceedings.neurips.cc/paper_files/paper/2024/file/0b82662b6c32e887bb252a74d8cb2d5e-Paper-Datasets_and_Benchmarks_Track.pdf.

Bordes, A., Usunier, N., Garcia-Durán, A., Weston, J., & Yakhnenko, O. (2013). Translating embeddings for modeling multi-relational data. In *Proceedings of the 26th international conference on neural information processing systems - volume 2* (pp. 2787–2795). Red Hook, NY, USA: Curran Associates Inc., URL: https://proceedings.neurips.cc/paper/2013/hash/1cecc7a77928ca8133fa24680a88d2f9-Abstract.html.

Borgeaud, S., Mensch, A., Hoffmann, J., Cai, T., Rutherford, E., Millican, K., Van Den Driessche, G. B., Lespiau, J.-B., Damoc, B., Clark, A., et al. (2022). Improving language models by retrieving from trillions of tokens. In *International conference on machine learning* (pp. 2206–2240). PMLR, URL: https://proceedings.mlr.press/v162/borgeaud22a.html.

Breit, A., Ott, S., Agibetov, A., & Samwald, M. (2020). OpenBioLink: a benchmarking framework for large-scale biomedical link prediction. *Bioinformatics*, *36*(13), 4097–4098. http://dx.doi.org/10.1093/BIOINFORMATICS/BTAA274.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901, URL: https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

Cai, Q., & Yates, A. (2013). Large-scale semantic parsing via schema matching and lexicon extension. In *Proceedings of the 51st annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 423–433). Sofia, Bulgaria: Association for Computational Linguistics, URL: https://aclanthology.org/P13-1042.

Cao, Y., Fang, M., & Tao, D. (2019). BAG: Bi-directional attention entity graph convolutional network for multi-hop reasoning question answering. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)* (pp. 357–362). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/N19-1032, URL: https://aclanthology.org/N19-1032/.

Cao, Y., Lin, X., Wu, Y., Shi, F., Shang, Y., Tan, Q., Zhou, C., & Zhang, P. (2025). A data-centric framework of improving graph neural networks for knowledge graph embedding. *World Wide Web*, *28*(1), 2. http://dx.doi.org/10.1007/S11280-024-01320-0.

Cao, X., Liu, Y., Hu, B., & Zhang, Y. (2021). Dual-channel reasoning model for complex question answering. *Complexity*, *2021*(1), Article 7367181. http://dx.doi.org/10.1155/2021/7367181.

Cao, S., Shi, J., Pan, L., Nie, L., Xiang, Y., Hou, L., Li, J., He, B., & Zhang, H. (2022). KQA pro: A dataset with explicit compositional programs for complex question answering over knowledge base. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 6101–6119). Dublin, Ireland: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2022.acl-long.422, URL: https://aclanthology.org/2022.acl-long.422.

Cao, Z., Xu, Q., Yang, Z., He, Y., Cao, X., & Huang, Q. (2025). SAQE: Complex logical query answering via semantic-aware representation learning. *IEEE Transactions on Knowledge and Data Engineering*, *37*(11), 6651–6665. http://dx.doi.org/10.1109/TKDE.2025.3603877, URL: https://ieeexplore.ieee.org/document/11151822.

Chan, C.-M., Chen, W., Su, Y., Yu, J., Xue, W., Zhang, S., Fu, J., & Liu, Z. (2024). ChatEval: Towards better LLM-based evaluators through multi-agent debate. In *The twelfth international conference on learning representations*. OpenReview.net, URL: https://openreview.net/forum?id=FQepisCUWu.

Charniak, E., Altun, Y., de Salvo Braz, R., Garrett, B., Kosmala, M., Moscovich, T., Pang, L., Pyo, C., Sun, Y., Wy, W., et al. (2000). Reading comprehension programs in a statistical-language-processing class. In *ANLP-NAACL 2000 workshop: reading comprehension tests as evaluation for computer-based language understanding systems*. URL: https://aclanthology.org/W00-0601/.

Che, F., Zhang, D., Tao, J., Niu, M., & Zhao, B. (2020). ParamE: Regarding neural network parameters as relation embeddings for knowledge graph completion. In *Proceedings of the AAAI conference on artificial intelligence*: *vol. 34*, (03), (pp. 2774–2781). AAAI Press, http://dx.doi.org/10.1609/AAAI.V34I03.5665.

Chen, D., Bolton, J., & Manning, C. D. (2016). A thorough examination of the CNN/Daily mail reading comprehension task. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 2358–2367). The Association for Computer Linguistics, http://dx.doi.org/10.18653/V1/P16-1223.

Chen, W., Chang, M., Schlinger, E., Wang, W. Y., & Cohen, W. W. (2021). Open question answering over tables and text. In *9th international conference on learning representations, ICLR 2021, virtual event, Austria, May 3-7, 2021*. OpenReview.net, URL: https://openreview.net/forum?id=MmCRswl1UYl.

Chen, Z., Chen, W., Smiley, C., Shah, S., Borova, I., Langdon, D., Moussa, R., Beane, M., Huang, T.-H., Routledge, B., & Wang, W. Y. (2021). FinQA: A dataset of numerical reasoning over financial data. In *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 3697–3711). Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2021.emnlp-main.300, URL: https://aclanthology.org/2021.emnlp-main.300.

Chen, W., Chen, J., Su, Y., Chen, Z., & Wang, W. Y. (2020). Logical natural language generation from open-domain tables. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 7929–7942). Online: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2020.acl-main.708, URL: https://aclanthology.org/2020.acl-main.708.

Chen, Y., Deng, H., Han, K., & Zhao, Q. (2025). Policy frameworks for transparent chain-of-thought reasoning in large language models. http://dx.doi.org/10.48550/arXiv.2503.14521, arXiv preprint arXiv:2503.14521.

Chen, Z., Fang, Y., Zhang, Y., Guo, L., Chen, J., Pan, J. Z., Chen, H., & Zhang, W. (2025). Noise-powered multi-modal knowledge graph representation framework. In *Proceedings of the 31st international conference on computational linguistics* (pp. 141–155). Abu Dhabi, UAE: Association for Computational Linguistics, URL: https://aclanthology.org/2025.coling-main.11/.

Chen, D., Fisch, A., Weston, J., & Bordes, A. (2017). Reading wikipedia to answer open-domain questions. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 1870–1879). Vancouver, Canada: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/P17-1171, URL: https://aclanthology.org/P17-1171/.

Chen, X., Hu, Z., & Sun, Y. (2022). Fuzzy logic based logical query answering on knowledge graphs. In *Thirty-sixth AAAI conference on artificial intelligence, AAAI 2022, thirty-fourth conference on innovative applications of artificial intelligence, IAAI 2022, the twelveth symposium on educational advances in artificial intelligence, EAAI 2022 virtual event, February 22 - March 1, 2022* (pp. 3939–3948). AAAI Press, http://dx.doi.org/10.1609/AAAI.V36I4.20310.

Chen, X., Jia, S., & Xiang, Y. (2020). A review: Knowledge reasoning over knowledge graph. *Expert Systems with Applications*, *141*, Article 112948. http://dx.doi.org/10.1016/j.eswa.2019.112948, URL: https://www.sciencedirect.com/science/article/pii/S0957417419306669.

Chen, Z., Li, D., Zhao, X., Hu, B., & Zhang, M. (2024). Temporal knowledge question answering via abstract reasoning induction. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 4872–4889). Bangkok, Thailand: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2024.acl-long.267, URL: https://aclanthology.org/2024.acl-long.267.

Chen, Z., Liao, J., & Zhao, X. (2023). Multi-granularity temporal question answering over knowledge graphs. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 11378–11392). Toronto, Canada: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2023.acl-long.637, URL: https://aclanthology.org/2023.acl-long.637.

Chen, W., Ma, X., Wang, X., & Cohen, W. W. (2023). Prompting: Disentangling computation from reasoning for numerical reasoning tasks. *Trans. Mach. Learn. Res.*, *2023*, URL: https://openreview.net/forum?id=YfZ4ZPt8zd.

Chen, J., Saha, S., & Bansal, M. (2024). Reconcile: Round-table conference improves reasoning via consensus among diverse LLMs. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 7066–7085). Bangkok, Thailand: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2024.acl-long.381, URL: https://aclanthology.org/2024.acl-long.381.

Chen, W., Su, Y., Zuo, J., Yang, C., Yuan, C., Chan, C., Yu, H., Lu, Y., Hung, Y., Qian, C., Qin, Y., Cong, X., Xie, R., Liu, Z., Sun, M., & Zhou, J. (2024). AgentVerse: Facilitating multi-agent collaboration and exploring emergent behaviors. In *The twelfth international conference on learning representations, ICLR 2024, vienna, Austria, May 7-11, 2024*. OpenReview.net, URL: https://openreview.net/forum?id=EHg5GDnyq1.

Chen, C., Wang, X., Lin, T.-E., Lv, A., Wu, Y., Gao, X., Wen, J.-R., Yan, R., & Li, Y. (2024). Masked thought: Simply masking partial reasoning steps can improve mathematical reasoning learning of language models. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 5872–5900). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2024.acl-long.320, URL: https://aclanthology.org/2024.acl-long.320/.

Chen, W., Zha, H., Chen, Z., Xiong, W., Wang, H., & Wang, W. Y. (2020). Hybridqa: A dataset of multi-hop question answering over tabular and textual data. In *Findings of the association for computational linguistics: EMNLP 2020* (pp. 1026–1036). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2020.findings-emnlp.91, URL: https://aclanthology.org/2020.findings-emnlp.91/.

Chen, X., Zhang, N., Li, L., Deng, S., Tan, C., Xu, C., Huang, F., Si, L., & Chen, H. (2022). Hybrid transformer with multi-level fusion for multimodal knowledge graph completion. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval* (pp. 904–915). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3477495.3531992.

Chen, Z., Zhang, Z., Li, Z., Wang, F., Zeng, Y., Jin, X., & Xu, Y. (2024a). Self-improvement programming for temporal knowledge graph question answering. In *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (LREC-cOLING 2024)* (pp. 14579–14594). Torino, Italia: ELRA and ICCL, URL: https://aclanthology.org/2024.lrec-main.1270.

Chen, Z., Zhao, X., Liao, J., Li, X., & Kanoulas, E. (2022). Temporal knowledge graph question answering via subgraph reasoning. *Knowledge-Based Systems*, *251*, Article 109134. http://dx.doi.org/10.1016/j.knosys.2022.109134, URL: https://www.sciencedirect.com/science/article/pii/S0950705122005603.

Chen, Z., Zhou, Q., Shen, Y., Hong, Y., Sun, Z., Gutfreund, D., & Gan, C. (2024b). Visual chain-of-thought prompting for knowledge-based visual reasoning. In *Proceedings of the AAAI conference on artificial intelligence*: *vol. 38*, (2), (pp. 1254–1262). AAAI Press, http://dx.doi.org/10.1609/AAAI.V38I2.27888.

Cheng, K., Amed, N. K., & Sun, Y. (2023). Neural compositional rule learning for knowledge graph reasoning. In *The eleventh international conference on learning representations, ICLR 2023, kigali, rwanda, May 1-5, 2023*. OpenReview.net, URL: https://openreview.net/forum?id=F8VKQyDgRVj.

Cheng, Z., Dong, H., Wang, Z., Jia, R., Guo, J., Gao, Y., Han, S., Lou, J.-G., & Zhang, D. (2022). Hitab: A hierarchical table dataset for question answering and natural language generation. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 1094–1110). Dublin, Ireland: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2022.acl-long.78, URL: https://aclanthology.org/2022.acl-long.78.

Cheng, K., Liu, J., Wang, W., & Sun, Y. (2022). Rlogic: Recursive logical rule learning from knowledge graphs. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining* (pp. 179–189). ACM, http://dx.doi.org/10.1145/3534678.3539421.

Cheng, S., Pan, L., Yin, X., Wang, X., & Wang, W. Y. (2024). Understanding the interplay between parametric and contextual knowledge for large language models. http://dx.doi.org/10.48550/ARXIV.2410.08414, CoRR. arXiv:2410.08414.

Cho, S., Seo, J., Jeong, S., & Park, J. C. (2023). Improving zero-shot reader by reducing distractions from irrelevant documents in open-domain question answering. In *Findings of the association for computational linguistics: EMNLP 2023* (pp. 3145–3157). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2023.findings-emnlp.207, URL: https://aclanthology.org/2023.findings-emnlp.207/.

Choi, E., Park, J., Lee, H., & Lee, J. (2025). Conflict-aware soft prompting for retrieval-augmented generation. In *Proceedings of the 2025 conference on empirical methods in natural language processing* (pp. 26969–26983). http://dx.doi.org/10.18653/v1/2025.emnlp-main.1371, URL: https://aclanthology.org/2025.emnlp-main.1371/.

Choudhary, N., Rao, N., Katariya, S., Subbian, K., & Reddy, C. K. (2021). Self-supervised hyperboloid representations from logical queries over knowledge graphs. In *Proceedings of the web conference 2021* (pp. 1373–1384). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3442381.3449974.

Christmann, P., Saha Roy, R., & Weikum, G. (2022). Conversational question answering on heterogeneous sources. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval* (pp. 144–154). ACM, http://dx.doi.org/10.1145/3477495.3531815.

Christmann, P., Saha Roy, R., & Weikum, G. (2023). Explainable conversational question answering over heterogeneous sources via iterative graph neural networks. In *Proceedings of the 46th international ACM SIGIR conference on research and development in information retrieval* (pp. 643–653). ACM, http://dx.doi.org/10.1145/3539618.3591682.

Christmann, P., Saha Roy, R., & Weikum, G. (2024). CompMix: A benchmark for heterogeneous question answering. In *Companion proceedings of the ACM web conference 2024* (pp. 1091–1094). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3589335.3651444.

Chuang, Y.-S., Fang, W., Li, S.-W., Yih, W.-t., & Glass, J. (2023). Expand, rerank, and retrieve: Query reranking for open-domain question answering. In *Findings of the association for computational linguistics: ACL 2023* (pp. 12131–12147). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2023.findings-acl.768, URL: https://aclanthology.org/2023.findings-acl.768/.

Clark, K., Luong, M., Le, Q. V., & Manning, C. D. (2020). ELECTRA: pre-training text encoders as discriminators rather than generators. In *8th international conference on learning representations, ICLR 2020, addis ababa, ethiopia, April 26-30, 2020*. OpenReview.net, URL: https://openreview.net/forum?id=r1xMH1BtvB.

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., et al. (2026). Training verifiers to solve math word problems. http://dx.doi.org/10.48550/arXiv.2110.14168, arXiv preprint arXiv:2110.14168.

Cui, Y., Chen, Z., Wei, S., Wang, S., Liu, T., & Hu, G. (2017). Attention-over-attention neural networks for reading comprehension. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 593–602). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/P17-1055, URL: https://aclanthology.org/P17-1055/.

Dagdelen, J., Dunn, A., Lee, S., Walker, N., Rosen, A. S., Ceder, G., Persson, K. A., & Jain, A. (2024). Structured information extraction from scientific text with large language models. *Nature Communications*, *15*(1), 1418. http://dx.doi.org/10.1038/s41467-024-45563-x, URL: https://www.nature.com/articles/s41467-024-45563-x#citeas.

Dai, Y., Yan, M., & Li, J. (2025). Granular concept-enhanced relational graph convolution networks for link prediction in knowledge graph. *Information Sciences*, *694*, Article 121698. http://dx.doi.org/10.1016/J.INS.2024.121698.

Das, R., Dhuliawala, S., Zaheer, M., Vilnis, L., Durugkar, I., Krishnamurthy, A., Smola, A., & McCallum, A. (2018). Go for a walk and arrive at the answer: Reasoning over paths in knowledge bases using reinforcement learning. In *6th international conference on learning representations, ICLR 2018, vancouver, BC, Canada, April 30 - May 3, 2018, conference track proceedings*. OpenReview.net, URL: https://openreview.net/forum?id=Syg-YfWCW.

Das, R., Zaheer, M., Thai, D., Godbole, A., Perez, E., Lee, J. Y., Tan, L., Polymenakos, L., & McCallum, A. (2021). Case-based reasoning for natural language queries over knowledge bases. In *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 9594–9611). Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2021.emnlp-main.755, URL: https://aclanthology.org/2021.emnlp-main.755.

Dasgupta, S. S., Ray, S. N., & Talukdar, P. (2018). HyTE: Hyperplane-based temporally aware knowledge graph embedding. In *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 2001–2011). Brussels, Belgium: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/D18-1225, URL: https://aclanthology.org/D18-1225.

Davis, R., Shrobe, H., & Szolovits, P. (1993). What is a knowledge representation? *AI Magazine*, *14*(1), 17. http://dx.doi.org/10.1609/AIMAG.V14I1.1029.

De Cao, N., Aziz, W., & Titov, I. (2019). Question answering by reasoning across documents with graph convolutional networks. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)* (pp. 2306–2317). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/N19-1240, URL: https://aclanthology.org/N19-1240/.

DeLong, L. N., Mir, R. F., & Fleuriot, J. D. (2025). Neurosymbolic AI for reasoning over knowledge graphs: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, *36*(5), 7822–7842. http://dx.doi.org/10.1109/TNNLS.2024.3420218.

Dettmers, T., Minervini, P., Stenetorp, P., & Riedel, S. (2018a). Convolutional 2D knowledge graph embeddings. In *Proceedings of the thirty-second AAAI conference on artificial intelligence, (AAAI-18), the 30th innovative applications of artificial intelligence (IAAI-18), and the 8th AAAI symposium on educational advances in artificial intelligence (EAAI-18), new orleans, louisiana, USA, February 2-7, 2018* (pp. 1811–1818). AAAI Press, http://dx.doi.org/10.1609/AAAI.V32I1.11573.

Dettmers, T., Minervini, P., Stenetorp, P., & Riedel, S. (2018b). Convolutional 2D knowledge graph embeddings. In *Proceedings of the thirty-second AAAI conference on artificial intelligence, (AAAI-18)* (pp. 1811–1818). AAAI Press, http://dx.doi.org/10.1609/AAAI.V32I1.11573.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)* (pp. 4171–4186). Minneapolis, Minnesota: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/N19-1423, URL: https://aclanthology.org/N19-1423.

Dhingra, B., Liu, H., Yang, Z., Cohen, W., & Salakhutdinov, R. (2017). Gated-attention readers for text comprehension. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 1832–1846). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/P17-1168, URL: https://aclanthology.org/P17-1168/.

Diao, S., Wang, P., Lin, Y., Pan, R., Liu, X., & Zhang, T. (2024). Active prompting with chain-of-thought for large language models. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: long papers), ACL 2024, bangkok, thailand, August 11-16, 2024* (pp. 1330–1350). Association for Computational Linguistics, http://dx.doi.org/10.18653/V1/2024.ACL-LONG.73.

Ding, M., Zhou, C., Chen, Q., Yang, H., & Tang, J. (2019). Cognitive graph for multi-hop reading comprehension at scale. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 2694–2703). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/P19-1259, URL: https://aclanthology.org/P19-1259/.

Dong, Y., Chawla, N. V., & Swami, A. (2017). Metapath2vec: Scalable representation learning for heterogeneous networks. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 135–144). http://dx.doi.org/10.1145/3097983.3098036.

Drozdov, A., Schärli, N., Akyürek, E., Scales, N., Song, X., Chen, X., Bousquet, O., & Zhou, D. (2023). Compositional semantic parsing with large language models. In *The eleventh international conference on learning representations*. OpenReview.net, URL: https://openreview.net/forum?id=gJW8hSGBys8.

Du, C., Li, X., & Li, Z. (2024). Semantic-enhanced reasoning question answering over temporal knowledge graphs. *Journal of Intelligent Information Systems*, *62*(3), 859–881. http://dx.doi.org/10.1007/S10844-024-00840-5.

Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., & Mordatch, I. (2024). Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st international conference on machine learning* (pp. 11733–11763). URL: https://proceedings.mlr.press/v235/du24e.html.

Dua, D., Gupta, S., Singh, S., & Gardner, M. (2022). Successive prompting for decomposing complex questions. In *Proceedings of the 2022 conference on empirical methods in natural language processing* (pp. 1251–1265). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2022.emnlp-main.81, URL: https://aclanthology.org/2022.emnlp-main.81/.

Duan, N., Tang, D., & Zhou, M. (2020). Machine reasoning: Technology, dilemma and future. In *Proceedings of the 2020 conference on empirical methods in natural language processing: tutorial abstracts* (pp. 1–6). Association for Computational Linguistics, http://dx.doi.org/10.18653/V1/2020.EMNLP-TUTORIALS.1.

Dubey, M., Banerjee, D., Abdelkawi, A., & Lehmann, J. (2019). LC-quad 2.0: A large dataset for complex question answering over wikidata and dbpedia. In *Lecture notes in computer science, The semantic web - ISWC 2019 - 18th international semantic web conference, auckland, New zealand, October 26-30, 2019, proceedings, part II* (pp. 69–78). Springer, http://dx.doi.org/10.1007/978-3-030-30796-7_5.

Evans, J. S. B. (1984). Heuristic and analytic processes in reasoning. *British Journal of Psychology*, *75*(4), 451–468. http://dx.doi.org/10.1111/j.2044-8295.1984.tb01915.x, URL: https://bpspsychub.onlinelibrary.wiley.com/doi/abs/10.1111/j.2044-8295.1984.tb01915.x.

Evans, J. S. B. (2003). In two minds: dual-process accounts of reasoning. *Trends in Cognitive Sciences*, *7*(10), 454–459. http://dx.doi.org/10.1016/j.tics.2003.08.012, URL: https://www.sciencedirect.com/science/article/pii/S1364661303002250.

Evans, J. S. B. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology*, *59*(1), 255–278. http://dx.doi.org/10.1146/annurev.psych.59.103006.093629, URL: https://www.annualreviews.org/content/journals/10.1146/annurev.psych.59.103006.093629.

Fagin, R., Halpern, J. Y., Moses, Y., & Vardi, M. (2004). *Reasoning about knowledge*. MIT Press, URL: https://cse.buffalo.edu/~rapaport/676/F01/fagin.pdf.

Fang, Y., Sun, S., Gan, Z., Pillai, R., Wang, S., & Liu, J. (2020). Hierarchical graph network for multi-hop question answering. In *Proceedings of the 2020 conference on empirical methods in natural language processing* (pp. 8823–8838). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2020.emnlp-main.710, URL: https://aclanthology.org/2020.emnlp-main.710/.

Feng, S., Fang, G., Ma, X., & Wang, X. (2025). Efficient reasoning models: A survey. *Trans. Mach. Learn. Res.*, *2025*, URL: https://openreview.net/forum?id=sySqlxj8EB.

Ferguson, J., Gardner, M., Hajishirzi, H., Khot, T., & Dasigi, P. (2020). IIRC: A dataset of incomplete information reading comprehension questions. In *Proceedings of the 2020 conference on empirical methods in natural language processing* (pp. 1137–1147). Online: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2020.emnlp-main.86, URL: https://aclanthology.org/2020.emnlp-main.86.

Ferrada, S., Bustos, B., & Hogan, A. (2017). Imgpedia: A linked dataset with content-based analysis of wikimedia images. In *Lecture notes in computer science, The semantic web - ISWC 2017 - 16th international semantic web conference, vienna, Austria, October 21-25, 2017, proceedings, part II* (pp. 84–93). Springer, http://dx.doi.org/10.1007/978-3-319-68204-4_8.

Finn, C., Abbeel, P., & Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of machine learning research, Proceedings of the 34th international conference on machine learning, ICML 2017, sydney, NSW, Australia, 6-11 August 2017* (pp. 1126–1135). PMLR, URL: http://proceedings.mlr.press/v70/finn17a.html.

Fletcher, C. R. (1985). Understanding and solving arithmetic word problems: A computer simulation. *Behavior Research Methods, Instruments, & Computers*, *17*(5), 565–571. http://dx.doi.org/10.3758/BF03207654, URL: https://link.springer.com/article/10.3758/BF03207654.

Formal, T., Piwowarski, B., & Clinchant, S. (2021). SPLADE: Sparse lexical and expansion model for first stage ranking. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval* (pp. 2288–2292). ACM, http://dx.doi.org/10.1145/3404835.3463098.

Fortuna, P., & Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Computing Surveys*, *51*(4), 1–30. http://dx.doi.org/10.1145/3232676.

Frisoni, G., Cocchieri, A., Presepi, A., Moro, G., & Meng, Z. (2024). To generate or to retrieve? On the effectiveness of artificial contexts for medical open-domain question answering. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 9878–9919). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2024.acl-long.533, URL: https://aclanthology.org/2024.acl-long.533/.

Fu, T.-y., Lee, W.-C., & Lei, Z. (2017). Hin2vec: Explore meta-paths in heterogeneous information networks for representation learning. In *Proceedings of the 2017 ACM on conference on information and knowledge management* (pp. 1797–1806). http://dx.doi.org/10.1145/3132847.3132953.

Fu, Y., Peng, H., Sabharwal, A., Clark, P., & Khot, T. (2023). Complexity-based prompting for multi-step reasoning. In *The eleventh international conference on learning representations, ICLR 2023, kigali, rwanda, May 1-5, 2023*. OpenReview.net, URL: https://openreview.net/pdf?id=yf1icZHC-l9.

Galárraga, L. A., Teflioudi, C., Hose, K., & Suchanek, F. (2013). AMIE: association rule mining under incomplete evidence in ontological knowledge bases. In *Proceedings of the 22nd international conference on world wide web* (pp. 413–422). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/2488388.2488425.

Gan, E., Zhao, Y., Cheng, L., Yancan, M., Goyal, A., Kawaguchi, K., Kan, M.-Y., & Shieh, M. (2024). Reasoning robustness of llms to adversarial typographical errors. In *Proceedings of the 2024 conference on empirical methods in natural language processing* (pp. 10449–10459). http://dx.doi.org/10.18653/v1/2024.emnlp-main.584, URL: https://aclanthology.org/2024.emnlp-main.584/.

Gao, L., Madaan, A., Zhou, S., Alon, U., Liu, P., Yang, Y., Callan, J., & Neubig, G. (2023). PAL: Program-aided language models. In *Proceedings of machine learning research*: *202*, *Proceedings of the 40th international conference on machine learning* (pp. 10764–10799). PMLR, URL: https://proceedings.mlr.press/v202/gao23f.html.

Gao, Y., Qiao, L., Kan, Z., Wen, Z., He, Y., & Li, D. (2024). Two-stage generative question answering on temporal knowledge graph using large language models. In *Findings of the association for computational linguistics: ACL 2024* (pp. 6719–6734). Bangkok, Thailand: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2024.findings-acl.401, URL: https://aclanthology.org/2024.findings-acl.401.

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., & Wang, H. (2023). Retrieval-augmented generation for large language models: A survey. http://dx.doi.org/10.48550/arXiv.2312.10997, arXiv preprint arXiv:2312.10997.

Gao, F., Yang, Y., Gao, P., Gu, M., Zhao, S., Chen, Y., Yuan, H., Lan, M., Zhou, A., & He, L. (2024). Self-supervised BGP-graph reasoning enhanced complex KBQA via sparql generation. *Information Processing & Management*, *61*(5), Article 103802. http://dx.doi.org/10.1016/j.ipm.2024.103802, URL: https://www.sciencedirect.com/science/article/pii/S0306457324001614.

García-Durán, A., Dumančić, S., & Niepert, M. (2018). Learning sequence encoders for temporal knowledge graph completion. In *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 4816–4821). Brussels, Belgium: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/D18-1516, URL: https://aclanthology.org/D18-1516.

Geva, M., Khashabi, D., Segal, E., Khot, T., Roth, D., & Berant, J. (2021). Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, *9*, 346–361, URL: https://aclanthology.org/2021.tacl-1.21/.

Goel, R., Kazemi, S. M., Brubaker, M., & Poupart, P. (2020). Diachronic embedding for temporal knowledge graph completion. *34*, In *Proceedings of the AAAI conference on artificial intelligence* (04), (pp. 3988–3995). http://dx.doi.org/10.1609/aaai.v34i04.5815, URL: https://ojs.aaai.org/index.php/AAAI/article/view/5815.

Gou, Z., Shao, Z., Gong, Y., Shen, Y., Yang, Y., Duan, N., & Chen, W. (2024). CRITIC: large language models can self-correct with tool-interactive critiquing. In *The twelfth international conference on learning representations, ICLR 2024, vienna, Austria, May 7-11, 2024*. OpenReview.net, URL: https://openreview.net/forum?id=Sx038qxjek.

Groeneveld, D., Khot, T., Sabharwal, A., et al. (2020). A simple yet strong pipeline for hotpotqa. In *Proceedings of the 2020 conference on empirical methods in natural language processing* (pp. 8839–8845). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2020.emnlp-main.711, URL: https://aclanthology.org/2020.emnlp-main.711/.

Gu, Y., Deng, X., & Su, Y. (2023). Don't generate, discriminate: A proposal for grounding language models to real-world environments. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 4928–4949). Toronto, Canada: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2023.acl-long.270, URL: https://aclanthology.org/2023.acl-long.270.

Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, G., Cai, J., et al. (2018). Recent advances in convolutional neural networks. *Pattern Recognition*, *77*, 354–377. http://dx.doi.org/10.1016/J.PATCOG.2017.10.013.

Gu, Z., Ye, H., Chen, X., Zhou, Z., Feng, H., & Xiao, Y. (2025). StrucText-eval: Evaluating large language model's reasoning ability in structure-rich text. In *Proceedings of the 63rd annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 223–244). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2025.acl-long.11, URL: https://aclanthology.org/2025.acl-long.11/.

Guan, S., Wei, J., Jin, X., Guo, J., & Cheng, X. (2024). Look globally and reason: Two-stage path reasoning over sparse knowledge graphs. In *Proceedings of the 33rd ACM international conference on information and knowledge management* (pp. 695–705). Association for Computing Machinery, http://dx.doi.org/10.1145/3627673.3679845.

Guan, B., Zhu, X., & Yuan, S. (2024). A T5-based interpretable reading comprehension model with more accurate evidence training. *Information Processing & Management*, *61*(2), Article 103584. http://dx.doi.org/10.1016/j.ipm.2023.103584, URL: https://www.sciencedirect.com/science/article/pii/S0306457323003217.

Guo, D., Yang, D., Zhang, H., Song, J., Wang, P., Zhu, Q., Xu, R., Zhang, R., Ma, S., Bi, X., et al. (2025). Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, *645*(8081), 633–638. http://dx.doi.org/10.1038/S41586-025-09422-Z.

Gupta, S., Shrivastava, V., Deshpande, A., Kalyan, A., Clark, P., Sabharwal, A., & Khot, T. (2024). Bias runs deep: Implicit reasoning biases in persona-assigned LLMs. In *The twelfth international conference on learning representations*. OpenReview.net, URL: https://openreview.net/forum?id=kGteeZ18Ir.

Guu, K., Lee, K., Tung, Z., Pasupat, P., & Chang, M. (2020). Retrieval augmented language model pre-training. In *International conference on machine learning* (pp. 3929–3938). PMLR, URL: http://proceedings.mlr.press/v119/guu20a.html.

Haber, J., Vidgen, B., Chapman, M., Agarwal, V., Lee, R. K.-W., Yap, Y. K., & Röttger, P. (2023). Improving the detection of multilingual online attacks with rich social media data from Singapore. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 12705–12721). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2023.acl-long.711, URL: https://aclanthology.org/2023.acl-long.711/.

Hamilton, W. L., Bajaj, P., Zitnik, M., Jurafsky, D., & Leskovec, J. (2018). Embedding logical queries on knowledge graphs. In *Advances in neural information processing systems 31: annual conference on neural information processing systems 2018, neurIPS 2018, December 3-8, 2018, montréal, Canada* (pp. 2030–2041). URL: https://proceedings.neurips.cc/paper/2018/hash/ef50c335cca9f340bde656363ebd02fd-Abstract.html.

Han, X., Cao, S., Lv, X., Lin, Y., Liu, Z., Sun, M., & Li, J. (2018). Openke: An open toolkit for knowledge embedding. In *Proceedings of the 2018 conference on empirical methods in natural language processing: system demonstrations* (pp. 139–144). Association for Computational Linguistics, http://dx.doi.org/10.18653/V1/D18-2024.

Han, K., & Gardent, C. (2023). Generating and answering simple and complex questions from text and from knowledge graphs. In *The 13th international joint conference on natural language processing and the 3rd conference of the Asia-Pacific chapter of the association for computational linguistics (IJCNLP-AACL 2023)* (pp. 285–304). Association for Computational Linguistics, http://dx.doi.org/10.18653/V1/2023.IJCNLP-MAIN.19.

Hannan, D., Jain, A., & Bansal, M. (2020). Manymodalqa: Modality disambiguation and qa over diverse inputs. *34*, In *Proceedings of the AAAI conference on artificial intelligence* (05), (pp. 7879–7886). AAAI Press, http://dx.doi.org/10.1609/AAAI.V34I05.6294.

He, P., Liu, X., Gao, J., & Chen, W. (2021). Deberta: decoding-enhanced bert with disentangled attention. In *9th international conference on learning representations, ICLR 2021, virtual event, Austria, May 3-7, 2021*. OpenReview.net, URL: https://openreview.net/forum?id=XPZIaotutsD.

Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., Song, D., & Steinhardt, J. (2021). Measuring mathematical problem solving with the MATH dataset. In *Proceedings of the neural information processing systems track on datasets and benchmarks 1, neurIPS datasets and benchmarks 2021, December 2021, virtual*. URL: https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/be83ab3ecd0db773eb2dc1b0a17836a1-Abstract-round2.html.

Hermann, K. M., Kocisky, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., & Blunsom, P. (2015). Teaching machines to read and comprehend. *Advances in Neural Information Processing Systems*, *28*, 1693–1701, URL: https://proceedings.neurips.cc/paper/2015/hash/afdec7005cc9f14302cd0474fd0f3c96-Abstract.html.

Herzig, J., Nowak, P. K., Müller, T., Piccinno, F., & Eisenschlos, J. (2020). Tapas: Weakly supervised table parsing via pre-training. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 4320–4333). Online: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2020.acl-main.398, URL: https://aclanthology.org/2020.acl-main.398.

Hill, F., Bordes, A., Chopra, S., & Weston, J. (2016). The goldilocks principle: Reading children's books with explicit memory representations. In *4th international conference on learning representations, ICLR 2016*. URL: http://arxiv.org/abs/1511.02301.

Himmelstein, D. S., Lizee, A., Hessler, C., Brueggeman, L., Chen, S. L., Hadley, D., Green, A., Khankhanian, P., & Baranzini, S. E. (2017). Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *ELife*, *6*, Article e26726. http://dx.doi.org/10.7554/eLife.26726.

Hirschman, L., Light, M., Breck, E., & Burger, J. D. (1999). Deep read: A reading comprehension system. In *Proceedings of the 37th annual meeting of the association for computational linguistics* (pp. 325–332). Association for Computational Linguistics, http://dx.doi.org/10.3115/1034678.1034731, URL: https://aclanthology.org/P99-1042/.

Ho, X., Duong Nguyen, A.-K., Sugawara, S., & Aizawa, A. (2020). Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th international conference on computational linguistics* (pp. 6609–6625). Barcelona, Spain (Online): International Committee on Computational Linguistics, http://dx.doi.org/10.18653/v1/2020.coling-main.580, URL: https://aclanthology.org/2020.coling-main.580.

Hossain, M. I., Zamzmi, G., Mouton, P. R., Salekin, M. S., Sun, Y., & Goldgof, D. (2025). Explainable AI for medical data: current methods, limitations, and future directions. *ACM Computing Surveys*, *57*(6), 1–46. http://dx.doi.org/10.1145/3637487, URL: https://dl.acm.org/doi/10.1145/3637487.

Hosseini, M. J., Hajishirzi, H., Etzioni, O., & Kushman, N. (2014). Learning to solve arithmetic word problems with verb categorization. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 523–533). Doha, Qatar: Association for Computational Linguistics, http://dx.doi.org/10.3115/v1/D14-1058, URL: https://aclanthology.org/D14-1058/.

Hou, Z., Jin, X., Li, Z., & Bai, L. (2021). Rule-aware reinforcement learning for knowledge graph reasoning. In *Findings of the association for computational linguistics: ACL-IJCNLP 2021* (pp. 4687–4692). Online: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2021.findings-acl.412, URL: https://aclanthology.org/2021.findings-acl.412.

Hu, M., Dong, H., Luo, P., Han, S., & Zhang, D. (2024). KET-QA: A dataset for knowledge enhanced table question answering. In *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (LREC-cOLING 2024)* (pp. 9705–9719). Torino, Italia: ELRA and ICCL, URL: https://aclanthology.org/2024.lrec-main.848/.

Hu, Z., Gutierrez Basulto, V., Xiang, Z., Li, X., Li, R., & Z. Pan, J. (2022). Type-aware embeddings for multi-hop reasoning over knowledge graphs. In *Proceedings of the thirty-first international joint conference on artificial intelligence, IJCAI-22* (pp. 3078–3084). International Joint Conferences on Artificial Intelligence Organization, http://dx.doi.org/10.24963/ijcai.2022/427, Main Track.

Hu, Z., Liu, C., Feng, X., Zhao, Y., Ng, S.-K., Luu, A. T., He, J., Koh, P. W., & Hooi, B. (2024). Uncertainty of thoughts: Uncertainty-aware planning enhances information seeking in large language models. In *ICLR 2024 workshop on large language model (LLM) agents*. URL: https://openreview.net/forum?id=ZWyLjimciT.

Hu, M., Peng, Y., Huang, Z., & Li, D. (2019). A multi-type multi-span network for reading comprehension that requires discrete reasoning. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)* (pp. 1596–1606). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/D19-1170, URL: https://aclanthology.org/D19-1170/.

Hu, M., Wei, F., Peng, Y., Huang, Z., Yang, N., & Li, D. (2019). Read+ verify: Machine reading comprehension with unanswerable questions. *33*, In *Proceedings of the AAAI conference on artificial intelligence* (01), (pp. 6529–6537). AAAI Press, http://dx.doi.org/10.1609/AAAI.V33I01.33016529.

Hu, W., Zhang, W., Jiang, Y., Zhang, C. J., Wei, X., & Qing, L. (2025). Removal of hallucination on hallucination: Debate-augmented RAG. In *Proceedings of the 63rd annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 15839–15853). http://dx.doi.org/10.18653/v1/2025.acl-long.770, URL: https://aclanthology.org/2025.acl-long.770/.

Huang, J., & Chang, K. C.-C. (2023). Towards reasoning in large language models: A survey. In *Findings of the association for computational linguistics: ACL 2023* (pp. 1049–1065). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2023.findings-acl.67, URL: https://aclanthology.org/2023.findings-acl.67/.

Huang, J., Chen, X., Mishra, S., Zheng, H. S., Yu, A. W., Song, X., & Zhou, D. (2024). Large language models cannot self-correct reasoning yet. In *The twelfth international conference on learning representations, ICLR 2024, vienna, Austria, May 7-11, 2024*. OpenReview.net, URL: https://openreview.net/forum?id=IkmD3fKBPQ.

Huang, Y., Han, X., & Sun, M. (2024). FastFiD: Improve inference efficiency of open domain question answering via sentence selection. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 6262–6276). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2024.acl-long.340, URL: https://aclanthology.org/2024.acl-long.340/.

Huang, X., Shen, J., Huang, S., Cheng, S., Wang, X., & Qu, Y. (2025). TARGA: Targeted synthetic data generation for practical reasoning over structured data. In *Proceedings of the 63rd annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 2704–2726). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2025.acl-long.137, URL: https://aclanthology.org/2025.acl-long.137/.

Huang, Z., Wang, Z., Xia, S., Li, X., Zou, H., Xu, R., Fan, R.-Z., Ye, L., Chern, E., Ye, Y., et al. (2024). Olympicarena: Benchmarking multi-discipline cognitive reasoning for superintelligent ai. *Advances in Neural Information Processing Systems*, *37*, 19209–19253, URL: http://papers.nips.cc/paper_files/paper/2024/hash/222d2eaf24cf8259a35d6c7130d31425-Abstract-Datasets_and_Benchmarks_Track.html.

Huang, R., Wei, W., Qu, X., Xie, W., Mao, X., & Chen, D. (2024). Joint multi-facts reasoning network for complex temporal question answering over knowledge graph. In *IEEE international conference on acoustics, speech and signal processing, ICASSP 2024, seoul, Republic of Korea, April 14-19, 2024* (pp. 10331–10335). IEEE, http://dx.doi.org/10.1109/ICASSP48485.2024.10447439.

Huang, Z., Zhou, J., Niu, C., & Cheng, G. (2023). Spans, not tokens: A span-centric model for multi-span reading comprehension. In *Proceedings of the 32nd ACM international conference on information and knowledge management* (pp. 874–884). ACM, http://dx.doi.org/10.1145/3583780.3615064.

Huth, M. (2004). Logic in computer science: Modelling and reasoning about systems. Cambridge University Press, URL: https://www.researchgate.net/profile/Michael-Huth-3/publication/220693544_Logic_in_computer_science_-_modelling_and_reasoning_about_systems_2_ed/links/543bce5e0cf204cab1db374e/Logic-in-computer-science-modelling-and-reasoning-about-systems-2-ed.pdf.

Izacard, G., & Grave, É. (2021a). Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th conference of the European chapter of the association for computational linguistics: main volume* (pp. 874–880). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2021.eacl-main.74, URL: https://aclanthology.org/2021.eacl-main.74/.

Izacard, G., & Grave, E. (2021b). Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th conference of the European chapter of the association for computational linguistics: main volume* (pp. 874–880). Online: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2021.eacl-main.74, URL: https://aclanthology.org/2021.eacl-main.74.

Jain, P., Rathi, S., Mausam, & Chakrabarti, S. (2020). Temporal knowledge base completion: New algorithms and evaluation protocols. In *Proceedings of the 2020 conference on empirical methods in natural language processing* (pp. 3733–3747). Online: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2020.emnlp-main.305, URL: https://aclanthology.org/2020.emnlp-main.305/.

Jamali, M., & Lakshmanan, L. (2013). Heteromf: recommendation in heterogeneous information networks using context dependent factor models. In *Proceedings of the 22nd international conference on world wide web* (pp. 643–654). http://dx.doi.org/10.1145/2488388.2488445.

Ji, G., He, S., Xu, L., Liu, K., & Zhao, J. (2015). Knowledge graph embedding via dynamic mapping matrix. In *Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (volume 1: long papers)* (pp. 687–696). Beijing, China: Association for Computational Linguistics, http://dx.doi.org/10.3115/v1/P15-1067, URL: https://aclanthology.org/P15-1067.

Ji, S., Pan, S., Cambria, E., Marttinen, P., & Yu, P. S. (2021). A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE Transactions on Neural Networks and Learning Systems*, *33*(2), 494–514. http://dx.doi.org/10.1109/TNNLS.2021.3070843.

Ji, Y., Wu, K., Li, J., Chen, W., Zhong, M., Jia, X., & Zhang, M. (2024). Retrieval and reasoning on KGs: Integrate knowledge graphs into large language models for complex question answering. In *Findings of the association for computational linguistics: EMNLP 2024* (pp. 7598–7610). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2024.findings-emnlp.446, URL: https://aclanthology.org/2024.findings-emnlp.446/.

Jia, Z., Abujabal, A., Saha Roy, R., Strötgen, J., & Weikum, G. (2018). TempQuestions: A benchmark for temporal question answering. In *Companion proceedings of the the web conference 2018* (pp. 1057–1062). Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee, http://dx.doi.org/10.1145/3184558.3191536.

Jia, Z., Pramanik, S., Saha Roy, R., & Weikum, G. (2021). Complex temporal question answering on knowledge graphs. In *Proceedings of the 30th ACM international conference on information & knowledge management* (pp. 792–802). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3459637.3482416.

Jiang, F., Drummond, T., & Cohn, T. (2024). Pre-training cross-lingual open domain question answering with large-scale synthetic supervision. In *Proceedings of the 2024 conference on empirical methods in natural language processing* (pp. 13906–13933). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2024.emnlp-main.770, URL: https://aclanthology.org/2024.emnlp-main.770/.

Jiang, T., Liu, T., Ge, T., Sha, L., Li, S., Chang, B., & Sui, Z. (2016). Encoding temporal information for time-aware link prediction. In *Proceedings of the 2016 conference on empirical methods in natural language processing* (pp. 2350–2354). Austin, Texas: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/D16-1260, URL: https://aclanthology.org/D16-1260.

Jiang, J., Zhou, K., Dong, Z., Ye, K., Zhao, X., & Wen, J.-R. (2023a). StructGPT: A general framework for large language model to reason over structured data. In *Proceedings of the 2023 conference on empirical methods in natural language processing* (pp. 9237–9251). Singapore: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2023.emnlp-main.574, URL: https://aclanthology.org/2023.emnlp-main.574.

Jiang, J., Zhou, K., Zhao, X., Song, Y., Zhu, C., Zhu, H., & Wen, J.-R. (2025). KG-agent: An efficient autonomous agent framework for complex reasoning over knowledge graph. In *Proceedings of the 63rd annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 9505–9523). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2025.acl-long.468, URL: https://aclanthology.org/2025.acl-long.468/.

Jiang, J., Zhou, K., Zhao, X., & Wen, J.-R. (2023b). UniKGQA: Unified retrieval and reasoning for solving multi-hop question answering over knowledge graph. In *The eleventh international conference on learning representations*. URL: https://openreview.net/forum?id=Z63RvyAZ2Vh.

Jiao, S., Zhu, Z., Wu, W., Zuo, Z., Qi, J., Wang, W., Zhang, G., & Liu, P. (2022). An improving reasoning network for complex question answering over temporal knowledge graphs. *Applied Intelligence: The International Journal of Artificial Intelligence, Neural Networks, and Complex Problem-Solving Technologies*, *53*(7), 8195–8208. http://dx.doi.org/10.1007/s10489-022-03913-6, URL: https://doi.org/10.1007/s10489-022-03913-6.

Jin, Z., Cao, P., Chen, Y., Liu, K., Jiang, X., Xu, J., Qiuxia, L., & Zhao, J. (2024). Tug-of-war between knowledge: Exploring and resolving knowledge conflicts in retrieval-augmented language models. In *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (LREC-cOLING 2024)* (pp. 16867–16878). URL: https://aclanthology.org/2024.lrec-main.1466/.

Jin, W., Qu, M., Jin, X., & Ren, X. (2020). Recurrent event network: Autoregressive structure inferenceover temporal knowledge graphs. In *Proceedings of the 2020 conference on empirical methods in natural language processing* (pp. 6669–6683). Online: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2020.emnlp-main.541, URL: https://aclanthology.org/2020.emnlp-main.541.

Jin, N., Siebert, J., Li, D., & Chen, Q. (2022). A survey on table question answering: recent advances. In *China conference on knowledge graph and semantic computing* (pp. 174–186). Springer, http://dx.doi.org/10.1007/978-981-19-7596-7_14, URL: https://doi.org/10.1007/978-981-19-7596-7_14.

Joshi, M., Choi, E., Weld, D. S., & Zettlemoyer, L. (2017). Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 1601–1611). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/P17-1147, URL: https://aclanthology.org/P17-1147/.

Kabra, S., Jha, A., & Reddy, C. K. (2025). Reasoning towards fairness: mitigating bias in language models through reasoning-guided fine-tuning. http://dx.doi.org/10.48550/arXiv.2504.05632, arXiv preprint arXiv:2504.05632.

Kadlec, R., Schmid, M., Bajgar, O., & Kleindienst, J. (2016). Text understanding with the attention sum reader network. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 908–918). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/P16-1086, URL: https://aclanthology.org/P16-1086/.

Kamoi, R., Zhang, Y., Zhang, N., Han, J., & Zhang, R. (2024). When can LLMs actually correct their own mistakes? A critical survey of self-correction of LLMs. *Transactions of the Association for Computational Linguistics*, *12*, 1417–1440. http://dx.doi.org/10.1162/tacl_a_00713, URL: https://aclanthology.org/2024.tacl-1.78/.

Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W. T. (2020). Dense passage retrieval for open-domain question answering. In *2020 conference on empirical methods in natural language processing, EMNLP 2020* (pp. 6769–6781). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2020.emnlp-main.550, URL: https://aclanthology.org/2020.emnlp-main.550/.

Kazemi, S. M., & Poole, D. (2018). Simple embedding for link prediction in knowledge graphs. In *Proceedings of the 32nd international conference on neural information processing systems* (pp. 4289–4300). Red Hook, NY, USA: Curran Associates Inc., URL: https://proceedings.neurips.cc/paper_files/paper/2018/file/b2ab001909a8a6f04b51920306046ce5-Paper.pdf.

Kebriaei, E., Homayouni, A., Faraji, R., Razavi, A., Shakery, A., Faili, H., & Yaghoobzadeh, Y. (2024). Persian offensive language detection. *Machine Learning*, *113*(7), 4359–4379. http://dx.doi.org/10.1007/S10994-023-06370-5.

Kemp, C., Tenenbaum, J. B., Griffiths, T. L., Yamada, T., & Ueda, N. (2006). Learning systems of concepts with an infinite relational model. In *Proceedings, the twenty-first national conference on artificial intelligence and the eighteenth innovative applications of artificial intelligence conference, July 16-20, 2006, boston, massachusetts, USA* (pp. 381–388). AAAI Press, URL: http://www.aaai.org/Library/AAAI/2006/aaai06-061.php.

Khalifa, M., Logeswaran, L., Lee, M., Lee, H., & Wang, L. (2023). GRACE: discriminator-guided chain-of-thought reasoning. In *Findings of the association for computational linguistics: EMNLP 2023, Singapore, December 6-10, 2023* (pp. 15299–15328). Association for Computational Linguistics, http://dx.doi.org/10.18653/V1/2023.FINDINGS-EMNLP.1022.

Khandelwal, A., Gupta, M., & Agrawal, P. (2025). Cocoa: Confidence-and context-aware adaptive decoding for resolving knowledge conflicts in large language models. In *Proceedings of the 2025 conference on empirical methods in natural language processing* (pp. 6846–6866). http://dx.doi.org/10.18653/v1/2025.emnlp-main.348, URL: https://aclanthology.org/2025.emnlp-main.348/.

Khot, T., Trivedi, H., Finlayson, M., Fu, Y., Richardson, K., Clark, P., & Sabharwal, A. (2023). Decomposed prompting: A modular approach for solving complex tasks. In *The eleventh international conference on learning representations*. URL: https://openreview.net/forum?id=_nGgzQjzaRy.

Kim, J., Kim, H., Cho, H., Kang, S., Chang, B., Yeo, J., & Lee, D. (2025). Driven personalized preference reasoning with large language models for recommendation. In *Proceedings of the 48th international ACM SIGIR conference on research and development in information retrieval* (pp. 1697–1706). http://dx.doi.org/10.1145/3726302.3730055.

Kim, J., Kim, T., Yoon, S., Kim, J., & Lee, D. (2026). RPM: Reasoning-level personalization for black-box large language models. URL: https://openreview.net/pdf?id=oKKVLHFzZ8.

Kim, K., & Lee, J.-Y. (2024). RE-RAG: Improving open-domain QA performance and interpretability with relevance estimator in retrieval-augmented generation. In *Proceedings of the 2024 conference on empirical methods in natural language processing* (pp. 22149–22161). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2024.emnlp-main.1236, URL: https://aclanthology.org/2024.emnlp-main.1236/.

Kim, M., Park, C., & Baek, S. (2024). Qpaug: Question and passage augmentation for open-domain question answering of LLMs. In *Findings of the association for computational linguistics: EMNLP 2024* (pp. 9024–9042). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2024.findings-emnlp.527, URL: https://aclanthology.org/2024.findings-emnlp.527/.

Kipf, T. N., & Welling, M. (2017). Semi-supervised classification with graph convolutional networks. In *5th international conference on learning representations, ICLR 2017, toulon, France, April 24-26, 2017, conference track proceedings*. OpenReview.net, URL: https://openreview.net/forum?id=SJU4ayYgl.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems*, *35*, 22199–22213, URL: http://papers.nips.cc/paper_files/paper/2022/hash/8bb0d291acd4acf06ef112099c16f326-Abstract-Conference.html.

Koncel-Kedziorski, R., Hajishirzi, H., Sabharwal, A., Etzioni, O., & Ang, S. D. (2015a). Parsing algebraic word problems into equations. *Transactions of the Association for Computational Linguistics*, *3*, 585–597. http://dx.doi.org/10.1162/TACL_A_00160.

Koncel-Kedziorski, R., Hajishirzi, H., Sabharwal, A., Etzioni, O., & Ang, S. D. (2015b). Parsing algebraic word problems into equations. *"Transactions of the Association for Computational Linguistics"*, *3*, 585–597. http://dx.doi.org/10.1162/tacl_a_00160, URL: https://aclanthology.org/Q15-1042/.

Koutra, D., Tong, H., & Lubensky, D. (2013). Big-align: Fast bipartite graph alignment. In *2013 IEEE 13th international conference on data mining* (pp. 389–398). IEEE, http://dx.doi.org/10.1109/ICDM.2013.152.

Krawczyk, D. C. (2012). The cognition and neuroscience of relational reasoning. *Brain Research*, *1428*, 13–23. http://dx.doi.org/10.1016/j.brainres.2010.11.080, URL: https://www.sciencedirect.com/science/article/pii/S000689931002593X.

Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Devlin, J., Lee, K., Toutanova, K., Jones, L., Kelcey, M., Chang, M.-W., Dai, A. M., Uszkoreit, J., Le, Q., & Petrov, S. (2019). Natural questions: A benchmark for question answering research. *"Transactions of the Association for Computational Linguistics"*, *7*, 452–466. http://dx.doi.org/10.1162/tacl_a_00276, URL: https://aclanthology.org/Q19-1026.

Lacroix, T., Obozinski, G., & Usunier, N. (2020). Tensor decompositions for temporal knowledge base completion. In *8th international conference on learning representations, ICLR 2020, addis ababa, ethiopia, April 26-30, 2020*. OpenReview.net, URL: https://openreview.net/forum?id=rke2P1BFwS.

Lai, G., Xie, Q., Liu, H., Yang, Y., & Hovy, E. (2017). RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 conference on empirical methods in natural language processing* (pp. 785–794). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/D17-1082, URL: https://aclanthology.org/D17-1082/.

Lan, Y., He, G., Jiang, J., Jiang, J., Zhao, W. X., & Wen, J.-R. (2021). A survey on complex knowledge base question answering: Methods, challenges and solutions. In *Proceedings of the thirtieth international joint conference on artificial intelligence* (pp. 4483–4491). http://dx.doi.org/10.24963/IJCAI.2021/611.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. http://dx.doi.org/10.1038/nature14539, URL: https://www.nature.com/articles/nature14539.

Lee, Y., Atreya, P., Ye, X., & Choi, E. (2024). Crafting in-context examples according to lms' parametric knowledge. In *2024 findings of the association for computational linguistics: NAACL 2024* (pp. 2069–2085). Association for Computational Linguistics.Association for Computational Linguistics (ACL), http://dx.doi.org/10.18653/v1/2024.findings-naacl.133, URL: https://aclanthology.org/2024.findings-naacl.133/.

Lee, J., & Hockenmaier, J. (2025). Evaluating step-by-step reasoning traces: A survey. In *Findings of the association for computational linguistics: EMNLP 2025* (pp. 1789–1814). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2025.findings-emnlp.94, URL: https://aclanthology.org/2025.findings-emnlp.94/.

Lee, E., Kim, M., Kwon, J., Lee, Y., Kim, J., Jang, S., & Kim, Y. (2026). HyCal: A training-free prototype calibration method for cross-discipline few-shot class-incremental learning. http://dx.doi.org/10.48550/arXiv.2604.15678, arXiv preprint arXiv:2604.15678.

Lee, N., Ping, W., Xu, P., Patwary, M., Fung, P. N., Shoeybi, M., & Catanzaro, B. (2022). Factuality enhanced language models for open-ended text generation. *Advances in Neural Information Processing Systems*, *35*, 34586–34599, URL: http://papers.nips.cc/paper_files/paper/2022/hash/df438caa36714f69277daa92d608dd63-Abstract-Conference.html.

Lee, J., Wang, Y., Li, J., & Zhang, M. (2024). Multimodal reasoning with multimodal knowledge graph. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 10767–10782). Association for Computational Linguistics, http://dx.doi.org/10.18653/V1/2024.ACL-LONG.579.

Lehmann, J., Bhandiwad, D., Gattogi, P., & Vahdati, S. (2024). Beyond boundaries: A human-like approach for question answering over structured and unstructured information sources. *Transactions of the Association for Computational Linguistics*, *12*, 786–802. http://dx.doi.org/10.1162/TACL_A_00671.

Lei, D., Jiang, G., Gu, X., Sun, K., Mao, Y., & Ren, X. (2020). Learning collaborative agents with rule guidance for knowledge graph reasoning. In *Proceedings of the 2020 conference on empirical methods in natural language processing* (pp. 8541–8547). Online: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2020.emnlp-main.688, URL: https://aclanthology.org/2020.emnlp-main.688.

Lei, F., Li, X., Wei, Y., He, S., Huang, Y., Zhao, J., & Liu, K. (2023). S3HQA: A three-stage approach for multi-hop text-table hybrid question answering. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 2: short papers)* (pp. 1731–1740). Toronto, Canada: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2023.acl-short.147, URL: https://aclanthology.org/2023.acl-short.147.

Lei, B., Zhang, Y., Zuo, S., Payani, A., & Ding, C. (2024). MACM: utilizing a multi-agent system for condition mining in solving complex mathematical problems. In *Advances in neural information processing systems 38: annual conference on neural information processing systems 2024, neurIPS 2024, vancouver, BC, Canada, December 10 - 15, 2024*. URL: http://papers.nips.cc/paper_files/paper/2024/hash/5fcedec09977357f32e8e0ec8957073b-Abstract-Conference.html.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2020). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (p. 7871). Association for Computational Linguistics, http://dx.doi.org/10.18653/V1/2020.ACL-MAIN.703.

Lewis, P., Oguz, B., Xiong, W., Petroni, F., Yih, S., & Riedel, S. (2022). Boosted dense retriever. In *Proceedings of the 2022 conference of the North American chapter of the association for computational linguistics: human language technologies* (pp. 3102–3117). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2022.naacl-main.226, URL: https://aclanthology.org/2022.naacl-main.226/.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., et al. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, *33*, 9459–9474, URL: https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

Lewkowycz, A., Andreassen, A., Dohan, D., Dyer, E., Michalewski, H., Ramasesh, V., Slone, A., Anil, C., Schlag, I., Gutman-Solo, T., et al. (2022). Solving quantitative reasoning problems with language models. *Advances in Neural Information Processing Systems*, *35*, 3843–3857, URL: http://papers.nips.cc/paper_files/paper/2022/hash/18abbeef8cfe9203fdf9053c9c4fe191-Abstract-Conference.html.

Ley, M. (2002). The DBLP computer science bibliography: Evolution, research issues, perspectives. In *International symposium on string processing and information retrieval* (pp. 1–10). Springer, http://dx.doi.org/10.1007/3-540-45735-6_1.

Li, C., Bi, B., Yan, M., Wang, W., & Huang, S. (2021). Addressing semantic drift in generative question answering with auxiliary extraction. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 2: short papers)* (pp. 942–947). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2021.acl-short.118, URL: https://aclanthology.org/2021.acl-short.118/.

Li, J., Cao, P., Chen, Y., Xu, J., Li, H., Jiang, X., Liu, K., & Zhao, J. (2025). Towards better chain-of-thought: A reflection on effectiveness and faithfulness. In *Findings of the association for computational linguistics: ACL 2025* (pp. 10747–10765). http://dx.doi.org/10.18653/v1/2025.findings-acl.560, URL: https://aclanthology.org/2025.findings-acl.560/.

Li, Z., Chen, L., Jian, Y., Wang, H., Zhao, Y., Zhang, M., Xiao, K., Zhang, Y., Deng, H., & Hou, X. (2025). Aggregation or separation? Adaptive embedding message passing for knowledge graph completion. *Information Sciences*, *691*, Article 121639. http://dx.doi.org/10.1016/J.INS.2024.121639.

Li, H., Chen, J., Yang, J., Ai, Q., Jia, W., Liu, Y., Lin, K., Wu, Y., Yuan, G., Hu, Y., et al. (2025). Legalagentbench: Evaluating llm agents in legal domain. In *Proceedings of the 63rd annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 2322–2344). http://dx.doi.org/10.18653/v1/2025.acl-long.116, URL: https://aclanthology.org/2025.acl-long.116/.

Li, Z., Fan, S., Gu, Y., Li, X., Duan, Z., Dong, B., Liu, N., & Wang, J. (2024). Flexkbqa: A flexible llm-powered framework for few-shot knowledge base question answering. *38*, In *Proceedings of the AAAI conference on artificial intelligence* (17), (pp. 18608–18616). AAAI Press, http://dx.doi.org/10.1609/AAAI.V38I17.29823.

Li, Z., Guan, S., Jin, X., Peng, W., Lyu, Y., Zhu, Y., Bai, L., Li, W., Guo, J., & Cheng, X. (2022). Complex evolutional pattern learning for temporal knowledge graph reasoning. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 2: short papers), ACL 2022, dublin, Ireland, May 22-27, 2022* (pp. 290–296). Association for Computational Linguistics, http://dx.doi.org/10.18653/V1/2022.ACL-SHORT.32.

Li, P., He, Y., Yashar, D., Cui, W., Ge, S., Zhang, H., Rifinski Fainman, D., Zhang, D., & Chaudhuri, S. (2024). Table-GPT: Table fine-tuned GPT for diverse table tasks. *Proc. ACM Manag. Data*, *2*(3), http://dx.doi.org/10.1145/3654979.

Li, X., Jin, J., Dong, G., Qian, H., Zhu, Y., Wu, Y., Wen, J., & Dou, Z. (2025). WebThinker: Empowering large reasoning models with deep research capability. http://dx.doi.org/10.48550/ARXIV.2504.21776, CoRR arXiv:2504.21776.

Li, Z., Jin, X., Guan, S., Li, W., Guo, J., Wang, Y., & Cheng, X. (2021). Search from history and reason for future: Two-stage reasoning on temporal knowledge graphs. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: long papers)* (pp. 4732–4743). Online: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2021.acl-long.365, URL: https://aclanthology.org/2021.acl-long.365.

Li, Z., Jin, X., Li, W., Guan, S., Guo, J., Shen, H., Wang, Y., & Cheng, X. (2021). Temporal knowledge graph reasoning based on evolutional representation learning. In *SIGIR '21: the 44th international ACM SIGIR conference on research and development in information retrieval, virtual event, Canada, July 11-15, 2021* (pp. 408–417). ACM, http://dx.doi.org/10.1145/3404835.3462963.

Li, X., Lei, W., & Yang, Y. (2023). From easy to hard: Two-stage selector and reader for multi-hop question answering. In *IEEE international conference on acoustics, speech and signal processing ICASSP 2023, rhodes island, Greece, June 4-10, 2023* (pp. 1–5). IEEE, http://dx.doi.org/10.1109/ICASSP49357.2023.10096119.

Li, Z., Li, X., Fan, S., & Wang, J. (2024). Optimization techniques for unsupervised complex table reasoning via self-training framework. *IEEE Transactions on Knowledge and Data Engineering*, *36*(12), 8996–9010. http://dx.doi.org/10.1109/TKDE.2024.3439405, URL: https://ieeexplore.ieee.org/document/10634137.

Li, Y., Lin, Z., Zhang, S., Fu, Q., Chen, B., Lou, J.-G., & Chen, W. (2023). Making language models better reasoners with step-aware verifier. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 5315–5333). Toronto, Canada: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2023.acl-long.291, URL: https://aclanthology.org/2023.acl-long.291/.

Li, X., Liu, Y., Ju, S., & Xie, Z. (2020). Dynamic reasoning network for multi-hop question answering. In *Natural language processing and Chinese computing: 9th cCF international conference, NLPCC 2020, zhengzhou, China, October 14–18, 2020, proceedings, part i 9* (pp. 29–40). Springer, http://dx.doi.org/10.1007/978-3-030-60450-9_3.

Li, T., Ma, X., Zhuang, A., Gu, Y., Su, Y., & Chen, W. (2023). Few-shot in-context learning on knowledge base question answering. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 6966–6980). Toronto, Canada: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2023.acl-long.385, URL: https://aclanthology.org/2023.acl-long.385.

Li, X., & Qiu, X. (2023). Mot: Memory-of-thought enables ChatGPT to self-improve. In *Proceedings of the 2023 conference on empirical methods in natural language processing* (pp. 6354–6374). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2023.emnlp-main.392, URL: https://aclanthology.org/2023.emnlp-main.392/.

Li, Y., Sun, S., & Zhao, J. (2022). Tirgn: Time-guided recurrent graph network with local-global historical patterns for temporal knowledge graph reasoning. In *Proceedings of the thirty-first international joint conference on artificial intelligence, IJCAI 2022, vienna, Austria, 23-29 July 2022* (pp. 2152–2158). ijcai.org, http://dx.doi.org/10.24963/IJCAI.2022/299.

Li, X., Taheri, A., Tu, L., & Gimpel, K. (2016). Commonsense knowledge base completion. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 1445–1455). Berlin, Germany: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/P16-1137, URL: https://aclanthology.org/P16-1137.

Li, D., Tan, Z., Qian, P., Li, Y., Chaudhary, K. S., Hu, L., & Shen, J. (2025). Smoa: Improving multi-agent large language models with sparse mixture-of-agents. In *Lecture notes in computer science*: *vol. 15872, Advances in knowledge discovery and data mining - 29th Pacific-Asia conference on knowledge discovery and data mining, PAKDD 2025, sydney, NSW, Australia, June 10-13, 2025, proceedings, part III* (pp. 54–65). Springer, http://dx.doi.org/10.1007/978-981-96-8180-8_5.

Li, M., Wang, W., Feng, F., Zhang, H., Wang, Q., & Chua, T.-S. (2023). Hypothetical training for robust machine reading comprehension of tabular context. In *Findings of the association for computational linguistics: ACL 2023* (pp. 1220–1236). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2023.findings-acl.79, URL: https://aclanthology.org/2023.findings-acl.79/.

Li, J., Wang, J., Zhang, Z., & Zhao, H. (2024). Self-prompting large language models for zero-shot open-domain QA. In *Proceedings of the 2024 conference of the North American chapter of the association for computational linguistics: human language technologies (volume 1: long papers)* (pp. 296–310). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2024.naacl-long.17, URL: https://aclanthology.org/2024.naacl-long.17/.

Li, Q., & Wu, G. (2025). Explainable reasoning over temporal knowledge graphs by pre-trained language model. *Information Processing & Management*, *62*(1), Article 103903. http://dx.doi.org/10.1016/j.ipm.2024.103903, URL: https://www.sciencedirect.com/science/article/pii/S0306457324002620.

Li, R., Xiao, Q., Yang, J., Zhang, L., & Chen, Y. (2024). MRC-PASCL: A few-shot machine reading comprehension approach via post-training and answer span-oriented contrastive learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *32*, 4838–4849. http://dx.doi.org/10.1109/TASLP.2024.3490373.

Li, S., Zhao, S., Zhang, Z., Fang, Z., Chen, W., & Wang, T. (2025). Basis is also explanation: Interpretable legal judgment reasoning prompted by multi-source knowledge. *Information Processing & Management*, *62*(3), Article 103996. http://dx.doi.org/10.1016/j.ipm.2024.103996, URL: https://www.sciencedirect.com/science/article/pii/S0306457324003558.

Liang, T., He, Z., Jiao, W., Wang, X., Wang, Y., Wang, R., Yang, Y., Shi, S., & Tu, Z. (2024). Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the 2024 conference on empirical methods in natural language processing* (pp. 17889–17904). Miami, Florida, USA: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2024.emnlp-main.992, URL: https://aclanthology.org/2024.emnlp-main.992.

Liang, K., Meng, L., Liu, M., Liu, Y., Tu, W., Wang, S., Zhou, S., Liu, X., Sun, F., & He, K. (2024). A survey of knowledge graph reasoning on graph types: Static, dynamic, and multi-modal. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *46*(12), 9456–9478. http://dx.doi.org/10.1109/TPAMI.2024.3417451.

Liang, K., Meng, L., Liu, Y., Liu, M., Wei, W., Liu, S., Tu, W., Wang, S., Zhou, S., & Liu, X. (2024). Simple yet effective: Structure guided pre-trained transformer for multi-modal knowledge graph reasoning. In *Proceedings of the 32nd ACM international conference on multimedia* (pp. 1554–1563). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3664647.3681112.

Liang, P. P., Wu, C., Morency, L.-P., & Salakhutdinov, R. (2021). Towards understanding and mitigating social biases in language models. In *Proceedings of the 38th international conference on machine learning* (pp. 6565–6576). PMLR, URL: http://proceedings.mlr.press/v139/liang21a.html.

Lightman, H., Kosaraju, V., Burda, Y., Edwards, H., Baker, B., Lee, T., Leike, J., Schulman, J., Sutskever, I., & Cobbe, K. (2024). Let's verify step by step. In *The twelfth international conference on learning representations, ICLR 2024, vienna, Austria, May 7-11, 2024*. OpenReview.net, URL: https://openreview.net/forum?id=v8L0pN6EOi.

Liguda, C., & Pfeiffer, T. (2012). Modeling math word problems with augmented semantic networks. In *International conference on application of natural language to information systems* (pp. 247–252). Springer, http://dx.doi.org/10.1007/978-3-642-31178-9_29.

Lin, Y., Liu, Z., Sun, M., Liu, Y., & Zhu, X. (2015). Learning entity and relation embeddings for knowledge graph completion. In *Proceedings of the twenty-ninth AAAI conference on artificial intelligence, January 25-30, 2015, austin, texas, USA* (pp. 2181–2187). AAAI Press, http://dx.doi.org/10.1609/AAAI.V29I1.9491.

Lin, L., & She, K. (2020). Tensor decomposition-based temporal knowledge graph embedding. In *32nd IEEE international conference on tools with artificial intelligence, ICTAI 2020, baltimore, MD, USA, November 9-11, 2020* (pp. 969–975). IEEE, http://dx.doi.org/10.1109/ICTAI50040.2020.00151.

Lin, X. V., Socher, R., & Xiong, C. (2018). Multi-hop knowledge graph reasoning with reward shaping. In *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 3243–3253). Brussels, Belgium: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/D18-1362, URL: https://aclanthology.org/D18-1362.

Lin, L., Wang, L., Guo, J., & Wong, K.-F. (2025). Investigating bias in llm-based bias detection: Disparities between llms and human perception. In *Proceedings of the 31st international conference on computational linguistics* (pp. 10634–10649). URL: https://aclanthology.org/2025.coling-main.709/.

Ling, W., Yogatama, D., Dyer, C., & Blunsom, P. (2017). Program induction by rationale generation: Learning to solve and explain algebraic word problems. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 158–167). Vancouver, Canada: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/P17-1015, URL: https://aclanthology.org/P17-1015/.

Liu, S., Fan, C., Cheng, K., Wang, Y., Cui, P., Sun, Y., & Liu, Z. (2024). Inductive meta-path learning for schema-complex heterogeneous information networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *46*(12), 10196–10209. http://dx.doi.org/10.1109/TPAMI.2024.3435055.

Liu, S., He, Y., Wang, Y., Zou, H., Cheng, H., Yang, W., Cui, P., & Liu, Z. (2025). Rule learning for knowledge graph reasoning under agnostic distribution shift. http://dx.doi.org/10.48550/ARXIV.2507.05110, arXiv preprint arXiv:2507.05110.

Liu, Y., Li, H., García-Durán, A., Niepert, M., Oñoro-Rubio, D., & Rosenblum, D. S. (2019). MMKG: multi-modal knowledge graphs. In *Lecture notes in computer science*, *The semantic web - 16th international conference, ESWC 2019, portorož, Slovenia, June 2-6, 2019, proceedings* (pp. 459–474). Springer, http://dx.doi.org/10.1007/978-3-030-21348-0_30.

Liu, Y., Liang, D., Fang, F., Wang, S., Wu, W., & Jiang, R. (2023). Time-aware multiway adaptive fusion network for temporal knowledge graph question answering. In *IEEE international conference on acoustics, speech and signal processing ICASSP 2023, rhodes island, Greece, June 4-10, 2023* (pp. 1–5). IEEE, http://dx.doi.org/10.1109/ICASSP49357.2023.10095395.

Liu, Y., Liang, D., Li, M., Giunchiglia, F., Li, X., Wang, S., Wu, W., Huang, L., Feng, X., & Guan, R. (2023). Local and global: Temporal question answering via information fusion. In *Proceedings of the thirty-second international joint conference on artificial intelligence, IJCAI-23* (pp. 5141–5149). International Joint Conferences on Artificial Intelligence Organization, http://dx.doi.org/10.24963/ijcai.2023/571, Main Track.

Liu, Y., Ma, Y., Hildebrandt, M., Joblin, M., & Tresp, V. (2022). Tlogic: Temporal logical rules for explainable link forecasting on temporal knowledge graphs. In *Thirty-sixth AAAI conference on artificial intelligence, AAAI 2022, thirty-fourth conference on innovative applications of artificial intelligence, IAAI 2022, the twelveth symposium on educational advances in artificial intelligence, EAAI 2022 virtual event, February 22 - March 1, 2022* (pp. 4120–4127). AAAI Press, http://dx.doi.org/10.1609/AAAI.V36I4.20330.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. http://dx.doi.org/10.48550/arXiv.1907.11692, arXiv preprint arXiv:1907.11692.

Liu, Z., Ping, W., Roy, R., Xu, P., Lee, C., Shoeybi, M., & Catanzaro, B. (2024). ChatQA: Surpassing GPT-4 on conversational QA and RAG. In *The thirty-eighth annual conference on neural information processing systems*. URL: https://openreview.net/forum?id=bkUvKPKafQ.

Liu, W., Qiang, B., Chen, R., Xie, Y., Chen, L., & Chen, Z. (2025). Linear self-attention with multi-relational graph for knowledge graph completion. *Applied Intelligence: The International Journal of Artificial Intelligence, Neural Networks, and Complex Problem-Solving Technologies*, *55*(10), 727. http://dx.doi.org/10.1007/S10489-025-06592-1.

Liu, S., Shang, Y.-M., & Zhang, X. (2026). Truthfulrag: Resolving factual-level conflicts in retrieval-augmented generation with knowledge graphs. *40*, In *Proceedings of the AAAI conference on artificial intelligence* (38), (pp. 32168–32176). http://dx.doi.org/10.1609/AAAI.V40I38.40489.

Liu, Y., Wu, J., He, Y., Gong, R., Xia, J., Li, L., Gao, H., Chen, H., Bi, B., Zhang, J., et al. (2025). Efficient inference for large reasoning models: A survey. http://dx.doi.org/10.48550/arXiv.2503.23077, arXiv preprint arXiv:2503.23077.

Liu, S., Xu, J., Tjangnaka, W., Semnani, S., Yu, C., & Lam, M. (2024). SUQL: Conversational search over structured and unstructured data with large language models. In *Findings of the association for computational linguistics: NAACL 2024* (pp. 4535–4555). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2024.findings-naacl.283, URL: https://aclanthology.org/2024.findings-naacl.283/.

Lu, P., Peng, B., Cheng, H., Galley, M., Chang, K., Wu, Y. N., Zhu, S., & Gao, J. (2023). Chameleon: Plug-and-play compositional reasoning with large language models. In *Advances in neural information processing systems 36: annual conference on neural information processing systems 2023, neurIPS 2023, new orleans, la, USA, December 10 - 16, 2023*. URL: http://papers.nips.cc/paper_files/paper/2023/hash/871ed095b734818cfba48db6aeb25a62-Abstract-Conference.html.

Lu, X., Wang, L., Jiang, Z., He, S., & Liu, S. (2022). MMKRL: A robust embedding approach for multi-modal knowledge graph representation learning. *Applied Intelligence: The International Journal of Artificial Intelligence, Neural Networks, and Complex Problem-Solving Technologies*, *52*(7), 7480–7497. http://dx.doi.org/10.1007/s10489-021-02693-9.

Lucas, M. M., Yang, J., Pomeroy, J. K., & Yang, C. C. (2024). Reasoning with large language models for medical question answering. *Journal of the American Medical Informatics Association*, *31*(9), 1964–1975. http://dx.doi.org/10.1093/JAMIA/OCAE131.

Luo, S., Deng, G., Xu, J., Zhang, X., Hou, H., & Song, L. (2025). Reasoning meets personalization: Unleashing the potential of large reasoning model for personalized generation. http://dx.doi.org/10.48550/arXiv.2505.17571, arXiv preprint arXiv:2505.17571.

Luo, H., E, H., Tang, Z., Peng, S., Guo, Y., Zhang, W., Ma, C., Dong, G., Song, M., Lin, W., Zhu, Y., & Luu, A. T. (2024). ChatKBQA: A generate-then-retrieve framework for knowledge base question answering with fine-tuned large language models. In *Findings of the association for computational linguistics: ACL 2024* (pp. 2039–2056). Bangkok, Thailand: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2024.findings-acl.122, URL: https://aclanthology.org/2024.findings-acl.122.

Luo, L., Li, Y.-F., Haf, R., & Pan, S. (2024). Reasoning on graphs: Faithful and interpretable large language model reasoning. In *ICLR*. URL: https://openreview.net/forum?id=ZGNWW7xZ6Q.

Lv, X., Gu, Y., Han, X., Hou, L., Li, J., & Liu, Z. (2019). Adapting meta knowledge graph information for multi-hop reasoning over few-shot relations. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)* (pp. 3376–3381). Hong Kong, China: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/D19-1334, URL: https://aclanthology.org/D19-1334.

Lv, S., Guo, D., Xu, J., Tang, D., Duan, N., Gong, M., Shou, L., Jiang, D., Cao, G., & Hu, S. (2020). Graph-based reasoning over heterogeneous external knowledge for commonsense question answering. *34*, In *Proceedings of the AAAI conference on artificial intelligence* (05), (pp. 8449–8456). AAAI Press, http://dx.doi.org/10.1609/AAAI.V34I05.6364.

Lv, X., Han, X., Hou, L., Li, J., Liu, Z., Zhang, W., Zhang, Y., Kong, H., & Wu, S. (2020). Dynamic anticipation and completion for multi-hop reasoning over sparse knowledge graph. In *Proceedings of the 2020 conference on empirical methods in natural language processing* (pp. 5694–5703). Online: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2020.emnlp-main.459, URL: https://aclanthology.org/2020.emnlp-main.459.

Lyu, Q., Apidianaki, M., & Callison-Burch, C. (2024). Towards faithful model explanation in nlp: A survey. *Computational Linguistics*, *50*(2), 657–723. http://dx.doi.org/10.1162/coli_a_00511, URL: https://aclanthology.org/2024.cl-2.6/.

Lyu, X., Min, S., Beltagy, I., Zettlemoyer, L., & Hajishirzi, H. (2023). Z-ICL: Zero-shot in-context learning with pseudo-demonstrations. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 2304–2317). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2023.acl-long.129, URL: https://aclanthology.org/2023.acl-long.129/.

Ma, K., Cheng, H., Liu, X., Nyberg, E., & Gao, J. (2022a). Open-domain question answering via chain of reasoning over heterogeneous knowledge. In *Findings of the association for computational linguistics: EMNLP 2022* (pp. 5360–5374). Association for Computational Linguistics, http://dx.doi.org/10.18653/V1/2022.FINDINGS-EMNLP.392.

Ma, K., Cheng, H., Liu, X., Nyberg, E., & Gao, J. (2022b). Open domain question answering with a unified knowledge interface. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 1605–1620). Association for Computational Linguistics, http://dx.doi.org/10.18653/V1/2022.ACL-LONG.113.

Ma, J., Li, K., Zhang, F., Wang, Y., Luo, X., Li, C., & Qiao, Y. (2024). Taret: Temporal knowledge graph reasoning based on topology-aware dynamic relation graph and temporal fusion. *Information Processing & Management*, *61*(6), Article 103848. http://dx.doi.org/10.1016/j.ipm.2024.103848, URL: https://www.sciencedirect.com/science/article/pii/S0306457324002073.

Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegreffe, S., Alon, U., Dziri, N., Prabhumoye, S., Yang, Y., Gupta, S., Majumder, B. P., Hermann, K., Welleck, S., Yazdanbakhsh, A., & Clark, P. (2023). Self-refine: Iterative refinement with self-feedback. In *Advances in neural information processing systems 36: annual conference on neural information processing systems 2023, neurIPS 2023, new orleans, la, USA, December 10 - 16, 2023*. URL: http://papers.nips.cc/paper_files/paper/2023/hash/91edff07232fb1b55a505a9e9f6c0ff3-Abstract-Conference.html.

Maharana, A., & Bansal, M. (2022). GRADA: Graph generative data augmentation for commonsense reasoning. In *Proceedings of the 29th international conference on computational linguistics* (pp. 4499–4516). International Committee on Computational Linguistics, URL: https://aclanthology.org/2022.coling-1.397/.

Mai, S., Sun, Y., Xiong, A., Zeng, Y., & Hu, H. (2024). Multimodal boosting: Addressing noisy modalities and identifying modality contribution. *IEEE Transactions on Multimedia*, *26*, 3018–3033. http://dx.doi.org/10.1109/TMM.2023.3306489.

Mavromatis, C., Subramanyam, P. L., Ioannidis, V. N., Adeshina, A., Howard, P. R., Grinberg, T., Hakim, N., & Karypis, G. (2022). Tempoqr: temporal question reasoning over knowledge graphs. *36*, In *Proceedings of the AAAI conference on artificial intelligence* (5), (pp. 5825–5833). AAAI Press, http://dx.doi.org/10.1609/AAAI.V36I5.20526.

Meilicke, C., Chekol, M. W., Betz, P., Fink, M., & Stuckenschmidt, H. (2024). Anytime bottom-up rule learning for large-scale knowledge graph completion. *The VLDB Journal*, *33*(1), 131–161. http://dx.doi.org/10.1007/S00778-023-00800-5.

Meilicke, C., Chekol, M. W., Ruffinelli, D., & Stuckenschmidt, H. (2019). Anytime bottom-up rule learning for knowledge graph completion. In *Proceedings of the twenty-eighth international joint conference on artificial intelligence, IJCAI 2019, macao, China, August 10-16, 2019* (pp. 3137–3143). ijcai.org, http://dx.doi.org/10.24963/IJCAI.2019/435.

Meilicke, C., Fink, M., Wang, Y., Ruffinelli, D., Gemulla, R., & Stuckenschmidt, H. (2018). Fine-grained evaluation of rule- and embedding-based systems for knowledge graph completion. In *The semantic web – ISWC 2018: 17th international semantic web conference, Monterey, CA, USA, October 8–12, 2018, proceedings, part i* (pp. 3–20). Berlin, Heidelberg: Springer-Verlag, http://dx.doi.org/10.1007/978-3-030-00671-6_1.

Meng, X., Bai, L., Hu, J., & Zhu, L. (2024). Multi-hop path reasoning over sparse temporal knowledge graphs based on path completion and reward shaping. *Information Processing & Management*, *61*(2), Article 103605. http://dx.doi.org/10.1016/j.ipm.2023.103605, URL: https://www.sciencedirect.com/science/article/pii/S0306457323003424.

Meng, S., Zhou, J., Chen, X., Liu, Y., Lu, F., & Huang, X. (2024). Structure-information-based reasoning over the knowledge graph: A survey of methods and applications. *ACM Transactions on Knowledge Discovery from Data*, *18*(8), 1–42. http://dx.doi.org/10.1145/3671148.

Merenda, F., Gomez-Perez, J. M., & Rigau, G. (2026). Can LLMs reason like doctors? Exploring the limits of large language models in complex medical reasoning. In *Findings of the association for computational linguistics: EACL 2026* (pp. 2432–2452). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2026.findings-eacl.127, URL: https://aclanthology.org/2026.findings-eacl.127/.

Messner, J., Abboud, R., & Ceylan, İ. İ. (2022). Temporal knowledge graph completion using box embeddings. In *Thirty-sixth AAAI conference on artificial intelligence, AAAI 2022, thirty-fourth conference on innovative applications of artificial intelligence, IAAI 2022, the twelveth symposium on educational advances in artificial intelligence, EAAI 2022 virtual event, February 22 - March 1, 2022* (pp. 7779–7787). AAAI Press, http://dx.doi.org/10.1609/AAAI.V36I7.20746.

Miller, A., Fisch, A., Dodge, J., Karimi, A.-H., Bordes, A., & Weston, J. (2016). Key-value memory networks for directly reading documents. In *Proceedings of the 2016 conference on empirical methods in natural language processing* (pp. 1400–1409). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/D16-1147, URL: https://aclanthology.org/D16-1147/.

Min, S., Zhong, V., Zettlemoyer, L., & Hajishirzi, H. (2019). Multi-hop reading comprehension through question decomposition and rescoring. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 6097–6109). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/P19-1613, URL: https://aclanthology.org/P19-1613/.

Minsky, M. (1988). *Society of mind*. Simon and Schuster, URL: http://wearcam.org/sensularity.pdf.

Mirzadeh, I., Alizadeh, K., Shahrokhi, H., Tuzel, O., Bengio, S., & Farajtabar, M. (2025). GSM-symbolic: Understanding the limitations of mathematical reasoning in large language models. In *The thirteenth international conference on learning representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, URL: https://openreview.net/forum?id=AjXkRZIvjB.

Mohammadi, A., Ramezani, R., & Baraani, A. (2023). Topic-aware multi-hop machine reading comprehension using weighted graphs. *Expert Systems with Applications*, *224*, Article 119873. http://dx.doi.org/10.1016/J.ESWA.2023.119873.

Montgomery Jr, E. B. (2018). *Medical reasoning: the nature and use of medical knowledge*. Oxford University Press, http://dx.doi.org/10.1093/med/9780190912925.001.0001, URL: https://academic.oup.com/book/24755.

Moosavi, N. S., Rücklé, A., Roth, D., & Gurevych, I. (2021). SciGen: a dataset for reasoning-aware text generation from scientific tables. In *Thirty-fifth conference on neural information processing systems datasets and benchmarks track (round 2)*. URL: https://openreview.net/forum?id=Jul-uX7EV_I.

Mousselly-Sergieh, H., Botschen, T., Gurevych, I., & Roth, S. (2018). A multimodal translation-based approach for knowledge graph representation learning. In *Proceedings of the seventh joint conference on lexical and computational semantics* (pp. 225–234). New Orleans, Louisiana: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/S18-2027, URL: https://aclanthology.org/S18-2027.

Mozafari, J., Abdallah, A., Piryani, B., & Jatowt, A. (2024). Exploring hint generation approaches for open-domain question answering. In *Findings of the association for computational linguistics: EMNLP 2024* (pp. 9327–9352). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2024.findings-emnlp.546, URL: https://aclanthology.org/2024.findings-emnlp.546/.

Mueller, T., Piccinno, F., Shaw, P., Nicosia, M., & Altun, Y. (2019). Answering conversational questions on structured data without logical forms. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)* (pp. 5902–5910). Hong Kong, China: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/D19-1603, URL: https://aclanthology.org/D19-1603.

Naik, R., Chandrasekaran, V., Yuksekgonul, M., Palangi, H., & Nushi, B. (2024). Diversity of thought improves reasoning abilities of LLMs. http://dx.doi.org/10.48550/arXiv.2310.07088, arXiv:2310.07088.

Namburi, S. S. S., Sreedhar, M., Srinivasan, S., & Sala, F. (2023). The cost of compression: Investigating the impact of compression on parametric knowledge in language models. In *Findings of the association for computational linguistics: EMNLP 2023* (pp. 5255–5273). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2023.findings-emnlp.349, URL: https://aclanthology.org/2023.findings-emnlp.349/.

Nan, L., Hsieh, C., Mao, Z., Lin, X. V., Verma, N., Zhang, R., Kryściński, W., Schoelkopf, H., Kong, R., Tang, X., Mutuma, M., Rosand, B., Trindade, I., Bandaru, R., Cunningham, J., Xiong, C., Radev, D., & Radev, D. (2022). Fetaqa: Free-form table question answering. *Transactions of the Association for Computational Linguistics*, *10*, 35–49. http://dx.doi.org/10.1162/tacl_a_00446, URL: https://aclanthology.org/2022.tacl-1.3/.

Nathani, D., Wang, D., Pan, L., & Wang, W. Y. (2023). MAF: multi-aspect feedback for improving reasoning in large language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing, EMNLP 2023, Singapore, December 6-10, 2023* (pp. 6591–6616). Association for Computational Linguistics, http://dx.doi.org/10.18653/V1/2023.EMNLP-MAIN.407.

Ng, M. K.-P., Li, X., & Ye, Y. (2011). Multirank: co-ranking for objects and relations in multi-relational data. In *Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1217–1225). http://dx.doi.org/10.1145/2020408.2020594.

Nguyen, C. D., French, T., Stewart, M., Hodkiewicz, M., & Liu, W. (2025). Representation learning in complex logical query answering on knowledge graphs: A survey. *ACM Computing Surveys*, *58*(5), 1–36. http://dx.doi.org/10.1145/3771692.

Nguyen, T., Rosenberg, M., Song, X., Gao, J., Tiwary, S., Majumder, R., & Deng, L. (2016). MS MARCO: a human generated machine reading comprehension dataset. In *CEUR workshop proceedings*, *Proceedings of the workshop on cognitive computation: integrating neural and symbolic approaches 2016 co-located with the 30th annual conference on neural information processing systems (NIPS 2016), Barcelona, Spain, December 9, 2016*. CEUR-WS.org, URL: https://ceur-ws.org/Vol-1773/CoCoNIPS_2016_paper9.pdf.

Ni, A., Iyer, S., Radev, D., Stoyanov, V., Yih, W.-T., Wang, S., & Lin, X. V. (2023). LEVER: Learning to verify language-to-code generation with execution. In *Proceedings of machine learning research*: *202*, *Proceedings of the 40th international conference on machine learning* (pp. 26106–26128). PMLR, URL: https://proceedings.mlr.press/v202/ni23b.html.

Ni, R., Ma, Z., Yu, K., & Xu, X. (2020). Specific time embedding for temporal knowledge graph completion. In *19th IEEE international conference on cognitive informatics & cognitive computing, iCCI*cC 2020, Beijing, China, September 26-28, 2020* (pp. 105–110). IEEE, http://dx.doi.org/10.1109/ICCICC50026.2020.9450214, URL: https://doi.org/10.1109/ICCICC50026.2020.9450214.

Nickel, M., Tresp, V., & Kriegel, H. (2011). A three-way model for collective learning on multi-relational data. In *Proceedings of the 28th international conference on machine learning, ICML 2011, bellevue, washington, USA, June 28 - July 2, 2011* (pp. 809–816). Omni Press, URL: https://icml.cc/2011/papers/438_icmlpaper.pdf.

Ning, X., Lin, Z., Zhou, Z., Yang, H., & Wang, Y. (2023). Skeleton-of-thought: Large language models can do parallel decoding. http://dx.doi.org/10.48550/ARXIV.2307.15337, CoRR. arXiv:2307.15337.

Ning, Y., Xu, M., Wen, J., Pi, Q., Zhu, Y., Zhong, M., Jiang, J., & Qian, T. (2026). Privacy-protected retrieval-augmented generation for knowledge graph question answering. *40*, In *Proceedings of the AAAI conference on artificial intelligence* (38), (pp. 32573–32581). http://dx.doi.org/10.1609/AAAI.V40I38.40534.

Niu, G., & Li, B. (2023). Logic and commonsense-guided temporal knowledge graph completion. In *Thirty-seventh AAAI conference on artificial intelligence, AAAI 2023, thirty-fifth conference on innovative applications of artificial intelligence, IAAI 2023, thirteenth symposium on educational advances in artificial intelligence, EAAI 2023, washington, DC, USA, February 7-14, 2023* (pp. 4569–4577). AAAI Press, http://dx.doi.org/10.1609/AAAI.V37I4.25579.

Nye, M. I., Andreassen, A. J., Gur-Ari, G., Michalewski, H., Austin, J., Bieber, D., Dohan, D., Lewkowycz, A., Bosma, M., Luan, D., Sutton, C., & Odena, A. (2021). Show your work: Scratchpads for intermediate computation with language models. CoRR. arXiv:2112.00114.

Oguz, B., Chen, X., Karpukhin, V., Peshterliev, S., Okhonko, D., Schlichtkrull, M., Gupta, S., Mehdad, Y., & Yih, S. (2022). Unik-QA: Unified representations of structured and unstructured knowledge for open-domain question answering. In *Findings of the association for computational linguistics: NAACL 2022* (pp. 1535–1546). Association for Computational Linguistics, http://dx.doi.org/10.18653/V1/2022.FINDINGS-NAACL.115.

Ohsugi, Y., Saito, I., Nishida, K., Asano, H., & Tomita, J. (2019). A simple but effective method to incorporate multi-turn context with BERT for conversational machine comprehension. In *Proceedings of the first workshop on NLP for conversational AI* (pp. 11–17). Florence, Italy: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/W19-4102, URL: https://aclanthology.org/W19-4102/.

Opedal, A., Shirakami, H., Schölkopf, B., Saparov, A., & Sachan, M. (2025). MathGAP: Out-of-distribution evaluation on problems with arbitrarily complex proofs. In *The thirteenth international conference on learning representations*. URL: https://openreview.net/forum?id=5ck9PIrTpH.

OpenAI (2024). Openai o1 system card. http://dx.doi.org/10.48550/ARXIV.2412.16720, CoRR. arXiv:2412.16720.

Ouyang, S., Zhang, Z., & Zhao, H. (2024). Fact-driven logical reasoning for machine reading comprehension. In *Proceedings of the AAAI conference on artificial intelligence*: *vol. 38*, (17), (pp. 18851–18859). AAAI Press, http://dx.doi.org/10.1609/AAAI.V38I17.29850.

Pal, V., Kanoulas, E., Yates, A., & de Rijke, M. (2024). Table question answering for low-resourced indic languages. In *Proceedings of the 2024 conference on empirical methods in natural language processing* (pp. 75–92). Miami, Florida, USA: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2024.emnlp-main.5, URL: https://aclanthology.org/2024.emnlp-main.5.

Pan, L., Albalak, A., Wang, X., & Wang, W. Y. (2023). Logic-LM: Empowering large language models with symbolic solvers for faithful logical reasoning. In *Findings of the association for computational linguistics: EMNLP 2023, Singapore, December 6-10, 2023* (pp. 3806–3824). Association for Computational Linguistics, http://dx.doi.org/10.18653/V1/2023.FINDINGS-EMNLP.248.

Pan, P., Lei, J., Wang, J., Ouyang, D., Qu, J., & Li, Z. (2025). Concept-aware embedding for logical query reasoning over knowledge graphs. *Information Processing & Management*, *62*(2), Article 103971. http://dx.doi.org/10.1016/j.ipm.2024.103971, URL: https://www.sciencedirect.com/science/article/pii/S0306457324003303.

Pan, S., Luo, L., Wang, Y., Chen, C., Wang, J., & Wu, X. (2024). Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*, *36*(7), 3580–3599. http://dx.doi.org/10.1109/TKDE.2024.3352100.

Pan, L., Saxon, M., Xu, W., Nathani, D., Wang, X., & Wang, W. Y. (2024). Automatically correcting large language models: *Surveying the Landscape of Diverse Automated Correction Strategies*. *Trans. Assoc. Comput. Linguistics*, *12*, 484–506. http://dx.doi.org/10.1162/TACL_A_00660.

Passmore, J. A. (1961). Philosophical reasoning. URL: https://philarchive.org/archive/PASPR-2.

Patel, A., Bhattamishra, S., & Goyal, N. (2021). Are NLP models really able to solve simple math word problems? In *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: human language technologies* (pp. 2080–2094). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2021.naacl-main.168, URL: https://aclanthology.org/2021.naacl-main.168/.

Paul, D., Ismayilzada, M., Peyrard, M., Borges, B., Bosselut, A., West, R., & Faltings, B. (2024). REFINER: reasoning feedback on intermediate representations. In *Proceedings of the 18th conference of the European chapter of the association for computational linguistics, EACL 2024 - volume 1: long papers, st. julian's, malta, March 17-22, 2024* (pp. 1100–1126). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2024.eacl-long.67, URL: https://aclanthology.org/2024.eacl-long.67/.

Petruzzellis, F., Cornelio, C., & Lio, P. (2025). Hierarchical planning for complex tasks with knowledge graph-RAG and symbolic verification. In *Forty-second international conference on machine learning, ICML 2025*. URL: https://proceedings.mlr.press/v267/petruzzellis25a.html.

Pezeshkpour, P., Chen, L., & Singh, S. (2018). Embedding multimodal relational data for knowledge base completion. In *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 3208–3218). Brussels, Belgium: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/D18-1359, URL: https://aclanthology.org/D18-1359.

Pirrò, G. (2020). Relatedness and tbox-driven rule learning in large knowledge bases. In *Proceedings of the AAAI conference on artificial intelligence*: *vol. 34*, (03), (pp. 2975–2982). AAAI Press, http://dx.doi.org/10.1609/AAAI.V34I03.5690.

Plaat, A., Wong, A., Verberne, S., Broekens, J., van Stein, N., & Back, T. (2024). Reasoning with large language models, a survey. http://dx.doi.org/10.48550/arXiv.2407.11511, arXiv preprint arXiv:2407.11511.

Pramanik, S., Alabi, J., Roy, R. S., & Weikum, G. (2024). UNIQORN: unified question answering over RDF knowledge graphs and natural language text. *Journal of Web Semantics*, *83*, Article 100833. http://dx.doi.org/10.1016/J.WEBSEM.2024.100833.

Press, O., Zhang, M., Min, S., Schmidt, L., Smith, N., & Lewis, M. (2023). Measuring and narrowing the compositionality gap in language models. In *Findings of the association for computational linguistics: EMNLP 2023* (pp. 5687–5711). Singapore: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2023.findings-emnlp.378, URL: https://aclanthology.org/2023.findings-emnlp.378/.

Purohit, K., V, V., Devalla, R., Yerragorla, K. M., Bhattacharya, S., & Anand, A. (2024). EXPLORA: Efficient exemplar subset selection for complex reasoning. In *Proceedings of the 2024 conference on empirical methods in natural language processing* (pp. 5367–5388). Miami, Florida, USA: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2024.emnlp-main.307, URL: https://aclanthology.org/2024.emnlp-main.307/.

Puterman, M. L. (1990). Markov decision processes. *Handbooks in Operations Research and Management Science*, *2*, 331–434. http://dx.doi.org/10.1016/S0927-0507(05)80172-0, URL: https://www.sciencedirect.com/science/chapter/handbook/abs/pii/S0927050705801720.

Qian, X., Zhang, Y., Zhao, Y., Zhou, B., Sui, X., Zhang, L., & Song, K. (2024). Timer[4] : Time-aware retrieval-augmented large language models for temporal knowledge graph question answering. In *Proceedings of the 2024 conference on empirical methods in natural language processing* (pp. 6942–6952). Miami, Florida, USA: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2024.emnlp-main.394, URL: https://aclanthology.org/2024.emnlp-main.394.

QianyiHu, Q., Tu, X., Cong, G., & Zhang, S. (2025). Time-aware ReAct agent for temporal knowledge graph question answering. In *Findings of the association for computational linguistics: NAACL 2025* (pp. 6013–6024). http://dx.doi.org/10.18653/v1/2025.findings-naacl.334, URL: https://aclanthology.org/2025.findings-naacl.334/.

Qiao, S., Ou, Y., Zhang, N., Chen, X., Yao, Y., Deng, S., Tan, C., Huang, F., & Chen, H. (2023). Reasoning with language model prompting: A survey. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 5368–5393). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2023.acl-long.294, URL: https://aclanthology.org/2023.acl-long.294/.

Qiu, P., Wu, C., Liu, S., Fan, Y., Zhao, W., Chen, Z., Gu, H., Peng, C., Zhang, Y., Wang, Y., et al. (2025). Quantifying the reasoning abilities of LLMs on clinical cases. *Nature Communications*, *16*(1), 9799. http://dx.doi.org/10.1038/s41467-025-64769-1, URL: https://www.nature.com/articles/s41467-025-64769-1#citeas.

Qiu, L., Xiao, Y., Qu, Y., Zhou, H., Li, L., Zhang, W., & Yu, Y. (2019). Dynamically fused graph network for multi-hop reasoning. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 6140–6150). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/P19-1617, URL: https://aclanthology.org/P19-1617/.

Qu, M., Chen, J., Xhonneux, L.-P., Bengio, Y., & Tang, J. (2020). Rnnlogic: Learning logic rules for reasoning on knowledge graphs. In *International conference on learning representations*. OpenReview.net, URL: https://openreview.net/forum?id=tGZu6DlbreV.

Qu, Y., Ding, Y., Liu, J., Liu, K., Ren, R., Zhao, W. X., Dong, D., Wu, H., & Wang, H. (2021). Rocketqa: An optimized training approach to dense passage retrieval for open-domain question answering. In *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: human language technologies* (pp. 5835–5847). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2021.naacl-main.466, URL: https://aclanthology.org/2021.naacl-main.466/.

Qu, X., Li, Y., Su, Z.-C., Sun, W., Yan, J., Liu, D., Cui, G., Liu, D., Liang, S., He, J., et al. (2025). A survey of efficient reasoning for large reasoning models: Language, multimodality, and beyond. http://dx.doi.org/10.48550/ARXIV.2503.21614, arXiv preprint arXiv:2503.21614.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. URL: https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, *21*(140), 1–67, URL: https://jmlr.org/papers/v21/20-074.html.

Rajani, N. F., McCann, B., Xiong, C., & Socher, R. (2019). Explain yourself! leveraging language models for commonsense reasoning. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 4932–4942). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/P19-1487, URL: https://aclanthology.org/P19-1487/.

Rajpurkar, P., Jia, R., & Liang, P. (2018). Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 2: short papers)* (pp. 784–789). Melbourne, Australia: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/P18-2124, URL: https://aclanthology.org/P18-2124.

Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 conference on empirical methods in natural language processing* (pp. 2383–2392). Austin, Texas: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/D16-1264, URL: https://aclanthology.org/D16-1264.

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)* (pp. 3982–3992). Hong Kong, China: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/D19-1410, URL: https://aclanthology.org/D19-1410.

Ren, H., Hu, W., & Leskovec, J. (2020). Query2box: Reasoning over knowledge graphs in vector space using box embeddings. In *8th international conference on learning representations*. OpenReview.net, URL: https://openreview.net/forum?id=BJgr4kSFDS.

Ren, H., & Leskovec, J. (2020). Beta embeddings for multi-hop logical reasoning in knowledge graphs. *Advances in Neural Information Processing Systems*, *33*, 19716–19726, URL: https://proceedings.neurips.cc/paper/2020/hash/e43739bba7cdb577e9e3e4e42447f5a5-Abstract.html.

Rescher, N. (2001). *Philosophical reasoning: A study in the methodology of philosophizing*. Malden, Mass.: Wiley-Blackwell.

Riloff, E., & Thelen, M. (2000). A rule-based question answering system for reading comprehension tests. In *ANLP/NAACL-readingComp '00*, *Proceedings of the 2000 ANLP/NAACL workshop on reading comprehension tests as evaluation for computer-based language understanding sytems - volume 6* (pp. 13–19). USA: Association for Computational Linguistics, http://dx.doi.org/10.3115/1117595.1117598.

Robertson, S. E., & Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends® in Information Retrieval*, *3*(4), 333–389. http://dx.doi.org/10.1561/1500000019.

Roy, S., & Roth, D. (2015). Solving general arithmetic word problems. In *Proceedings of the 2015 conference on empirical methods in natural language processing* (pp. 1743–1752). Association for Computational Linguistics. Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/D15-1202, URL: https://aclanthology.org/D15-1202/.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533–536. http://dx.doi.org/10.1038/323533a0, URL: https://www.nature.com/articles/323533a0.

Sachan, M., & Xing, E. (2016). Machine comprehension using rich semantic representations. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: short papers)* (pp. 486–492). The Association for Computer Linguistics, http://dx.doi.org/10.18653/V1/P16-2079.

Sachdeva, R., Tutek, M., & Gurevych, I. (2024). Catfood: Counterfactual augmented training for improving out-of-domain performance and calibration. In *Proceedings of the 18th conference of the European chapter of the association for computational linguistics (volume 1: long papers)* (pp. 1876–1898). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2024.eacl-long.113, URL: https://aclanthology.org/2024.eacl-long.113/.

Sadeghian, A., Armandpour, M., Colas, A. M., & Wang, D. Z. (2021). Chronor: Rotation based temporal knowledge graph embedding. In *Thirty-fifth AAAI conference on artificial intelligence, AAAI 2021, thirty-third conference on innovative applications of artificial intelligence, IAAI 2021, the eleventh symposium on educational advances in artificial intelligence, EAAI 2021, virtual event, February 2-9, 2021* (pp. 6471–6479). AAAI Press, http://dx.doi.org/10.1609/AAAI.V35I7.16802.

Sadeghian, A., Armandpour, M., Ding, P., & Wang, D. Z. (2019). Drum: End-to-end differentiable rule mining on knowledge graphs. *Advances in Neural Information Processing Systems*, *32*, 15321–15331, URL: https://proceedings.neurips.cc/paper/2019/hash/0c72cb7ee1512f800abe27823a792d03-Abstract.html.

Safavi, T., & Koutra, D. (2020). CoDEx: A comprehensive knowledge graph completion benchmark. In *Proceedings of the 2020 conference on empirical methods in natural language processing* (pp. 8328–8350). Online: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2020.emnlp-main.669, URL: https://aclanthology.org/2020.emnlp-main.669/.

Saito, K., Lee, C.-Y., Sohn, K., & Ushiku, Y. (2025). Where is the answer? An empirical study of positional bias for parametric knowledge extraction in language model. In *Proceedings of the 2025 conference of the nations of the americas chapter of the association for computational linguistics: human language technologies (volume 1: long papers)* (pp. 1252–1269). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2025.naacl-long.58, URL: https://aclanthology.org/2025.naacl-long.58/.

Salton, G., Fox, E. A., & Wu, H. (1983). Extended boolean information retrieval. *Communications of the ACM*, *26*(11), 1022–1036. http://dx.doi.org/10.1145/182.358466.

Sap, M., Le Bras, R., Allaway, E., Bhagavatula, C., Lourie, N., Rashkin, H., Roof, B., Smith, N. A., & Choi, Y. (2019). ATOMIC: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI conference on artificial intelligence*: *vol. 33*, (01), (pp. 3027–3035). AAAI Press, http://dx.doi.org/10.1609/AAAI.V33I01.33013027.

Saxena, A., Chakrabarti, S., & Talukdar, P. (2021). Question answering over temporal knowledge graphs. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: long papers)* (pp. 6663–6676). Online: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2021.acl-long.520, URL: https://aclanthology.org/2021.acl-long.520.

Saxena, Y., & Gaur, M. (2026). Neurosymbolic retrievers for retrieval-augmented generation. *IEEE Intelligent Systems*, *41*(1), 96–104. http://dx.doi.org/10.1109/MIS.2025.3642666, URL: https://www.computer.org/csdl/magazine/ex/2026/01/11373694/2dZJoCmTppC.

Saxena, A., Kochsiek, A., & Gemulla, R. (2022). Sequence-to-sequence knowledge graph completion and question answering. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 2814–2828). Dublin, Ireland: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2022.acl-long.201, URL: https://aclanthology.org/2022.acl-long.201.

Saxena, A., Tripathi, A., & Talukdar, P. (2020). Improving multi-hop question answering over knowledge graphs using knowledge base embeddings. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 4498–4507). Online: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2020.acl-main.412, URL: https://aclanthology.org/2020.acl-main.412.

Schlichtkrull, M., Kipf, T. N., Bloem, P., van den Berg, R., Titov, I., & Welling, M. (2018). Modeling relational data with graph convolutional networks. (pp. 593–607). Berlin, Heidelberg: Springer-Verlag, http://dx.doi.org/10.1007/978-3-319-93417-4_38.

Segal, E., Efrat, A., Shoham, M., Globerson, A., & Berant, J. (2020). A simple and effective model for answering multi-span questions. In *Proceedings of the 2020 conference on empirical methods in natural language processing* (pp. 3074–3080). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2020.emnlp-main.248, URL: https://aclanthology.org/2020.emnlp-main.248/.

Seo, M. J., Kembhavi, A., Farhadi, A., & Hajishirzi, H. (2017). Bidirectional attention flow for machine comprehension. In *5th international conference on learning representations, ICLR 2017, toulon, France, April 24-26, 2017, conference track proceedings*. OpenReview.net, URL: https://openreview.net/forum?id=HJ0UKP9ge.

Shaikh, O., Zhang, H., Held, W., Bernstein, M., & Yang, D. (2023). On second thought, let's not think step by step! bias and toxicity in zero-shot reasoning. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 4454–4470). http://dx.doi.org/10.18653/v1/2023.acl-long.244, URL: https://aclanthology.org/2023.acl-long.244/.

Shang, C., Wang, G., Qi, P., & Huang, J. (2022). Improving time sensitivity for question answering over temporal knowledge graphs. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 8017–8026). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2022.acl-long.552, URL: https://aclanthology.org/2022.acl-long.552/.

Shao, N., Cui, Y., Liu, T., Wang, S., & Hu, G. (2020). Is graph structure necessary for multi-hop question answering? In *Proceedings of the 2020 conference on empirical methods in natural language processing* (pp. 7187–7192). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2020.emnlp-main.583, URL: https://aclanthology.org/2020.emnlp-main.583/.

Sharma, A., Saxena, A., Gupta, C., Kazemi, M., Talukdar, P., & Chakrabarti, S. (2023). TwiRGCN: Temporally weighted graph convolution for question answering over temporal knowledge graphs. In *Proceedings of the 17th conference of the European chapter of the association for computational linguistics* (pp. 2049–2060). Dubrovnik, Croatia: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2023.eacl-main.150, URL: https://aclanthology.org/2023.eacl-main.150.

Shen, X., Barlacchi, G., Del Tredici, M., Cheng, W., Byrne, B., & Gispert, A. (2022a). Product answer generation from heterogeneous sources: A new benchmark and best practices. In *Proceedings of the fifth workshop on e-commerce and NLP (ECNLP 5)* (pp. 99–110). Dublin, Ireland: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2022.ecnlp-1.13, URL: https://aclanthology.org/2022.ecnlp-1.13.

Shen, X., Vakulenko, S., Del Tredici, M., Barlacchi, G., Byrne, B., & de Gispert, A. (2022b). Low-resource dense retrieval for open-domain question answering: A comprehensive survey. http://dx.doi.org/10.48550/ARXIV.2208.03197, arXiv preprint arXiv:2208.03197.

Shen, X., Wang, S., Tan, Z., Yao, L., Zhao, X., Xu, K., Wang, X., & Chen, T. (2025). FaithCoT-bench: Benchmarking instance-level faithfulness of chain-of-thought reasoning. http://dx.doi.org/10.48550/arXiv.2510.04040, arXiv preprint arXiv:2510.04040.

Shi, C., Ding, J., Cao, X., Hu, L., Wu, B., & Li, X. (2021). Entity set expansion in knowledge graph: a heterogeneous information network perspective. *Frontiers of Computer Science*, *15*(1), Article 151307. http://dx.doi.org/10.1007/S11704-020-9240-8.

Shi, C., Hu, B., Zhao, W. X., & Yu, P. S. (2018). Heterogeneous information network embedding for recommendation. *IEEE Transactions on Knowledge and Data Engineering*, *31*(2), 357–370. http://dx.doi.org/10.1109/TKDE.2018.2833443.

Shi, F., Li, D., Wang, X., Li, B., & Wu, X. (2024). Tgformer: A graph transformer framework for knowledge graph embedding. *IEEE Transactions on Knowledge and Data Engineering*, *37*(1), 526–541. http://dx.doi.org/10.1109/TKDE.2024.3486747.

Shi, C., Li, Y., Zhang, J., Sun, Y., & Yu, P. S. (2016). A survey of heterogeneous information network analysis. *IEEE Transactions on Knowledge and Data Engineering*, *29*(1), 17–37. http://dx.doi.org/10.1109/TKDE.2016.2598561.

Shi, S., Wang, Y., Lin, C.-Y., Liu, X., & Rui, Y. (2015). Automatically solving number word problems by semantic parsing and reasoning. In *Proceedings of the 2015 conference on empirical methods in natural language processing* (pp. 1132–1142). http://dx.doi.org/10.18653/v1/D15-1135, URL: https://aclanthology.org/D15-1135.pdf.

Shi, B., & Weninger, T. (2018). Open-world knowledge graph completion. In *Proceedings of the thirty-second AAAI conference on artificial intelligence, (AAAI-18)* (pp. 1957–1964). AAAI Press, http://dx.doi.org/10.1609/AAAI.V32I1.11535.

Shi, W., Xu, R., Zhuang, Y., Yu, Y., Sun, H., Wu, H., Yang, C., & Wang, M. D. (2024). Medadapter: Efficient test-time adaptation of large language models towards medical reasoning. *2024*, In *Proceedings of the conference on empirical methods in natural language processing. conference on empirical methods in natural language processing* (p. 22294). Association for Computational Linguistics, http://dx.doi.org/10.18653/V1/2024.EMNLP-MAIN.1244.

Shinn, N., Cassano, F., Gopinath, A., Narasimhan, K., & Yao, S. (2023). Reflexion: language agents with verbal reinforcement learning. In *Advances in neural information processing systems 36: annual conference on neural information processing systems 2023, neurIPS 2023, new orleans, la, USA, December 10 - 16, 2023*. URL: http://papers.nips.cc/paper_files/paper/2023/hash/1b44b878bb782e6954cd888628510e90-Abstract-Conference.html.

Shrestha, I., & Srinivasan, P. (2025). LLM bias detection and mitigation through the lens of desired distributions. In *Proceedings of the 2025 conference on empirical methods in natural language processing* (pp. 1464–1480). http://dx.doi.org/10.18653/v1/2025.emnlp-main.76, URL: https://aclanthology.org/2025.emnlp-main.76/.

Simon, H. A., & Newell, A. (1976). Computer science as empirical inquiry: symbols and search. *Communications of the ACM*, *19*(3), 11–126. http://dx.doi.org/10.1145/360018.360022.

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, *119*(1), 3, URL: https://psycnet.apa.org/buy/1996-01401-001.

Son, G., Jung, H., Hahm, M., Na, K., & Jin, S. (2023). Beyond classification: Financial reasoning in state-of-the-art language models. In *Proceedings of the fifth workshop on financial technology and natural language processing and the second multimodal AI for financial forecasting* (pp. 34–44). URL: https://aclanthology.org/2023.finnlp-1.3/.

Speer, R., Chin, J., & Havasi, C. (2017). ConceptNet 5.5: An open multilingual graph of general knowledge. *Proceedings of the AAAI Conference on Artificial Intelligence*, *31*(1), 4444–4451. http://dx.doi.org/10.1609/AAAI.V31I1.11164.

Su, M., Li, Z., Chen, Z., Bai, L., Jin, X., & Guo, J. (2024). Temporal knowledge graph question answering: A survey. http://dx.doi.org/10.48550/ARXIV.2406.14191, arXiv preprint arXiv:2406.14191.

Su, D., Li, X., Zhang, J., Shang, L., Jiang, X., Liu, Q., & Fung, P. (2022). Read before generate! faithful long form question answering with machine reading. In *60th annual meeting of the association-for-computational-linguistics* (pp. 744–756). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2022.findings-acl.61, URL: https://aclanthology.org/2022.findings-acl.61/.

Suadaa, L. H., Kamigaito, H., Funakoshi, K., Okumura, M., & Takamura, H. (2021). Towards table-to-text generation with numerical reasoning. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: long papers)* (pp. 1451–1465). Online: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2021.acl-long.115, URL: https://aclanthology.org/2021.acl-long.115.

Sui, Y., Zhou, M., Zhou, M., Han, S., & Zhang, D. (2024). Table meets LLM: can large language models understand structured table data? A benchmark and empirical study. In *Proceedings of the 17th ACM international conference on web search and data mining, WSDM 2024, merida, Mexico, March 4-8, 2024* (pp. 645–654). ACM, http://dx.doi.org/10.1145/3616855.3635752.

Sun, H., Arnold, A. O., Bedrax-Weiss, T., Pereira, F., & Cohen, W. W. (2020). Faithful embeddings for knowledge base queries. In *Proceedings of the 34th international conference on neural information processing systems*. Red Hook, NY, USA: Curran Associates Inc., URL: https://proceedings.neurips.cc/paper/2020/hash/fe74074593f21197b7b7be3c08678616-Abstract.html.

Sun, H., Bi, X., Tu, Z., Zhao, B., Zhang, K., Chu, D., & Xu, X. (2026). Enhancing privacy-preserving knowledge graph embeddings with federated learning for iot services. *ACM Transactions on Internet Technology*, *26*(1), 1–24. http://dx.doi.org/10.1145/3747351.

Sun, Z., Deng, Z., Nie, J., & Tang, J. (2019). Rotate: Knowledge graph embedding by relational rotation in complex space. In *7th international conference on learning representations, ICLR 2019, new orleans, la, USA, May 6-9, 2019*. OpenReview.net, URL: https://openreview.net/forum?id=HkgEQnRqYQ.

Sun, Y., & Han, J. (2013). Mining heterogeneous information networks: a structural analysis approach. *ACM SIGKDD Explorations Newsletter*, *14*(2), 20–28. http://dx.doi.org/10.1145/2481244.2481248.

Sun, Y., Han, J., Yan, X., Yu, P. S., & Wu, T. (2011). Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. *Proceedings of the VLDB Endowment*, *4*(11), 992–1003, URL: http://www.vldb.org/pvldb/vol4/p992-sun.pdf.

Sun, Y., Han, J., Yan, X., Yu, P. S., & Wu, T. (2022). Heterogeneous information networks: the past, the present, and the future. *Proceedings of the VLDB Endowment*, *15*(12), http://dx.doi.org/10.14778/3554821.3554901, URL: https://www.vldb.org/pvldb/vol15/p3807-sun.pdf.

Sun, Y., Han, J., Zhao, P., Yin, Z., Cheng, H., & Wu, T. (2009). Rankclus: integrating clustering with ranking for heterogeneous information network analysis. In *Proceedings of the 12th international conference on extending database technology: advances in database technology* (pp. 565–576). http://dx.doi.org/10.1145/1516360.1516426.

Sun, J., Xu, C., Tang, L., Wang, S., Lin, C., Gong, Y., Ni, L., Shum, H.-Y., & Guo, J. (2024). Think-on-graph: Deep and responsible reasoning of large language model on knowledge graph. In *The twelfth international conference on learning representations*. URL: https://openreview.net/forum?id=nnVO1PvbTv.

Sun, J., Zheng, C., Xie, E., Liu, Z., Chu, R., Qiu, J., Xu, J., Ding, M., Li, H., Geng, M., et al. (2025). A survey of reasoning with foundation models: Concepts, methodologies, and outlook. *ACM Computing Surveys*, *57*(11), 1–43. http://dx.doi.org/10.1145/3729218.

Sun, H., Zhong, J., Ma, Y., Han, Z., & He, K. (2021). TimeTraveler: Reinforcement learning for temporal knowledge graph forecasting. In *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 8306–8319). Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2021.emnlp-main.655, URL: https://aclanthology.org/2021.emnlp-main.655.

Talmor, A., Herzig, J., Lourie, N., & Berant, J. (2019). CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)* (pp. 4149–4158). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/N19-1421, URL: https://aclanthology.org/N19-1421/.

Talmor, A., Tafjord, O., Clark, P., Goldberg, Y., & Berant, J. (2020a). Leap-of-thought: Teaching pre-trained models to systematically reason over implicit knowledge. *Neural Information Processing Systems,Neural Information Processing Systems*, URL: https://proceedings.neurips.cc/paper/2020/hash/e992111e4ab9985366e806733383bd8c-Abstract.html.

Talmor, A., Tafjord, O., Clark, P., Goldberg, Y., & Berant, J. (2020b). Leap-of-thought: Teaching pre-trained models to systematically reason over implicit knowledge. *Advances in Neural Information Processing Systems*, *33*, 20227–20237, URL: https://proceedings.neurips.cc/paper/2020/hash/e992111e4ab9985366e806733383bd8c-Abstract.html.

Talmor, A., Yoran, O., Catav, A., Lahav, D., Wang, Y., Asai, A., Ilharco, G., Hajishirzi, H., & Berant, J. (2021). MultiModalQA: complex question answering over text, tables and images. In *9th international conference on learning representations, ICLR 2021, virtual event, Austria, May 3-7, 2021*. OpenReview.net, URL: https://openreview.net/forum?id=ee6W5UgQLa.

Tang, J., Hu, S., Chen, Z., Xu, H., & Tan, Z. (2022). Incorporating phrases in latent query reformulation for multi-hop question answering. *Mathematics*, *10*(4), 646. http://dx.doi.org/10.3390/math10040646, URL: https://www.mdpi.com/2227-7390/10/4/646.

Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., & Su, Z. (2008). Arnetminer: extraction and mining of academic social networks. In *Proceedings of the 14th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 990–998). http://dx.doi.org/10.1145/1401890.1402008.

Tang, X., Zhu, S.-C., Liang, Y., & Zhang, M. (2024). Rule: Knowledge graph reasoning with rule embedding. In *Findings of the association for computational linguistics: ACL 2024* (pp. 4316–4335). Association for Computational Linguistics, http://dx.doi.org/10.18653/V1/2024.FINDINGS-ACL.256.

Tao, Y., Li, Y., & Wu, Z. (2021). Temporal link prediction via reinforcement learning. In *IEEE international conference on acoustics, speech and signal processing, ICASSP 2021, toronto, on, Canada, June 6-11, 2021* (pp. 3470–3474). IEEE, http://dx.doi.org/10.1109/ICASSP39728.2021.9413413.

Teru, K., Denis, E., & Hamilton, W. (2020). Inductive relation prediction by subgraph reasoning. In *International conference on machine learning* (pp. 9448–9457). PMLR, URL: http://proceedings.mlr.press/v119/teru20a.html.

Toroghi, A., Guo, W., Abdollah Pour, M. M., & Sanner, S. (2024). Right for right reasons: Large language models for verifiable commonsense knowledge graph question answering. In *Proceedings of the 2024 conference on empirical methods in natural language processing* (pp. 6601–6633). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2024.emnlp-main.378, URL: https://aclanthology.org/2024.emnlp-main.378/.

Toutanova, K., Chen, D., Pantel, P., Poon, H., Choudhury, P., & Gamon, M. (2015). Observed versus latent features for knowledge base and text inference. In *Proceedings of the 3rd workshop on continuous vector space models and their compositionality* (pp. 57–66). Association for Computational Linguistics, http://dx.doi.org/10.18653/V1/W15-4007.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. (2023). Llama 2: Open foundation and fine-tuned chat models. http://dx.doi.org/10.48550/ARXIV.2307.09288, arXiv preprint arXiv:2307.09288.

Trischler, A., Wang, T., Yuan, X., Harris, J., Sordoni, A., Bachman, P., & Suleman, K. (2017). NewsQA: A machine comprehension dataset. In *Proceedings of the 2nd workshop on representation learning for NLP* (pp. 191–200). Vancouver, Canada: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/W17-2623, URL: https://aclanthology.org/W17-2623.

Trivedi, R. S., Dai, H., Wang, Y., & Song, L. (2017). Know-evolve: Deep temporal reasoning for dynamic knowledge graphs. In *Proceedings of machine learning research*, *Proceedings of the 34th international conference on machine learning, ICML 2017, sydney, NSW, Australia, 6-11 August 2017* (pp. 3462–3471). PMLR, URL: http://proceedings.mlr.press/v70/trivedi17a.html.

Trivedi, P., Maheshwari, G., Dubey, M., & Lehmann, J. (2017). LC-quad: A corpus for complex question answering over knowledge graphs. In *The semantic web – ISWC 2017: 16th international semantic web conference, vienna, Austria, October 21-25, 2017, proceedings, part II* (pp. 210–218). Berlin, Heidelberg: Springer-Verlag, http://dx.doi.org/10.1007/978-3-319-68204-4_22.

Trouillon, T., Welbl, J., Riedel, S., Gaussier, E., & Bouchard, G. (2016). Complex embeddings for simple link prediction. In *Proceedings of machine learning research*: *48*, *Proceedings of the 33rd international conference on machine learning* (pp. 2071–2080). New York, New York, USA: PMLR, URL: https://proceedings.mlr.press/v48/trouillon16.html.

Tu, M., Huang, K., Wang, G., Huang, J., He, X., & Zhou, B. (2020). Select, answer and explain: Interpretable multi-hop reading comprehension over multiple documents. In *Proceedings of the AAAI conference on artificial intelligence*: *vol. 34*, (05), (pp. 9073–9080). AAAI Press, http://dx.doi.org/10.1609/AAAI.V34I05.6441.

Tu, M., Wang, G., Huang, J., Tang, Y., He, X., & Zhou, B. (2019). Multi-hop reading comprehension across multiple documents by reasoning over heterogeneous graphs. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 2704–2713). Association for Computational Linguistics. Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/P19-1260, URL: https://aclanthology.org/P19-1260/.

Tutek, M., Chaleshtori, F. H., Marasović, A., & Belinkov, Y. (2025). Measuring chain of thought faithfulness by unlearning reasoning steps. In *Proceedings of the 2025 conference on empirical methods in natural language processing* (pp. 9946–9971). http://dx.doi.org/10.18653/v1/2025.emnlp-main.504, URL: https://aclanthology.org/2025.emnlp-main.504/.

Umeyama, S. (2002). An eigendecomposition approach to weighted graph matching problems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *10*(5), 695–703. http://dx.doi.org/10.1109/34.6778.

Upadhyay, S., Chang, M.-W., Chang, K.-W., & Yih, W.-t. (2016). Learning from explicit and implicit supervision jointly for algebra word problems. In *Proceedings of the 2016 conference on empirical methods in natural language processing* (pp. 297–306). http://dx.doi.org/10.18653/V1/D16-1029.

Usbeck, R., Yan, X., Perevalov, A., Jiang, L., Schulz, J., Kraft, A., Möller, C., Huang, J., Reineke, J., Ngomo, A. N., Saleem, M., & Both, A. (2024). QALD-10 - the 10th challenge on question answering over linked data: Shifting from dbpedia to wikidata as a KG for KGQA. *Semantic Web*, *15*(6), 2193–2207. http://dx.doi.org/10.3233/SW-233471.

Vanoirbeek, C. (1992). Formatting structured tables. In *EP92 (proceedings of electronic publishing, 1992)* (pp. 291–309). Cambridge University Press UK, URL: https://books.google.co.jp/books?hl=zh-CN&lr=&id=YN1sLgZtHC8C&oi=fnd&pg=PA291&dq=related:22XtCHodsPIJ:scholar.google.com/&ots=6giPymVgE9&sig=nyrwOrtJF2d4DeUbkvKTvCwom0I&redir_esc=y#v=onepage&q&f=false.

Vashishth, S., Sanyal, S., Nitin, V., Agrawal, N., & Talukdar, P. (2020). Interacte: Improving convolution-based knowledge graph embeddings by increasing feature interactions. *34*, In *Proceedings of the AAAI conference on artificial intelligence* (03), (pp. 3009–3016). AAAI Press, http://dx.doi.org/10.1609/AAAI.V34I03.5694.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*, URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Wan, G., & Du, B. (2021). GaussianPath: A Bayesian multi-hop reasoning framework for knowledge graph reasoning. In *Thirty-fifth AAAI conference on artificial intelligence, AAAI 2021, thirty-third conference on innovative applications of artificial intelligence, IAAI 2021, the eleventh symposium on educational advances in artificial intelligence, EAAI 2021, virtual event, February 2-9, 2021* (pp. 4393–4401). AAAI Press, http://dx.doi.org/10.1609/AAAI.V35I5.16565.

Wan, G., Pan, S., Gong, C., Zhou, C., & Haffari, G. (2020). Reasoning like human: Hierarchical reinforcement learning for knowledge graph reasoning. In *Proceedings of the twenty-ninth international joint conference on artificial intelligence, IJCAI 2020* (pp. 1926–1932). ijcai.org, http://dx.doi.org/10.24963/IJCAI.2020/267.

Wan, X., Sun, R., Dai, H., Arik, S., & Pfister, T. (2023). Better zero-shot reasoning with self-adaptive prompting. In *Findings of the association for computational linguistics: ACL 2023* (pp. 3493–3514). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2023.findings-acl.216, URL: https://aclanthology.org/2023.findings-acl.216/.

Wang, D., Chen, Y., & Cuenca Grau, B. (2023). Efficient embeddings of logical variables for query answering over incomplete knowledge graphs. *Proceedings of the AAAI Conference on Artificial Intelligence*, *37*(4), 4652–4659. http://dx.doi.org/10.1609/aaai.v37i4.25588, URL: https://ojs.aaai.org/index.php/AAAI/article/view/25588.

Wang, J., Chen, H., & Zhang, W. (2025). Structure-aware transformer for hyper-relational knowledge graph completion. *Expert Systems with Applications*, *277*, Article 126992. http://dx.doi.org/10.1016/J.ESWA.2025.126992.

Wang, X., Cucala, D. J. T., Grau, B. C., & Horrocks, I. (2024). Faithful rule extraction for differentiable rule learning models. In *The twelfth international conference on learning representations, ICLR 2024, vienna, Austria, May 7-11, 2024*. OpenReview.net, URL: https://openreview.net/forum?id=kBTzlxM2J1.

Wang, D., & Li, L. (2023). Learning from mistakes via cooperative study assistant for large language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing, EMNLP 2023, Singapore, December 6-10, 2023* (pp. 10667–10685). Association for Computational Linguistics, http://dx.doi.org/10.18653/V1/2023.EMNLP-MAIN.659.

Wang, Z., Li, L., Li, Q., & Zeng, D. (2019). Multimodal data enhanced representation learning for knowledge graphs. In *International joint conference on neural networks, IJCNN 2019 budapest, Hungary, July 14-19, 2019* (pp. 1–8). IEEE, http://dx.doi.org/10.1109/IJCNN.2019.8852079.

Wang, M., Li, Z., Wang, J., Zou, W., Zhou, J., & Gan, J. (2024). Trackge: Transformer with relation-pattern adaptive contrastive learning for knowledge graph embedding. *Knowledge-Based Systems*, *301*, Article 112218. http://dx.doi.org/10.1016/J.KNOSYS.2024.112218.

Wang, Y., Lipka, N., Rossi, R. A., Siu, A. F., Zhang, R., & Derr, T. (2024). Knowledge graph prompting for multi-document question answering. In *Thirty-eighth AAAI conference on artificial intelligence, AAAI 2024* (pp. 19206–19214). AAAI Press, http://dx.doi.org/10.1609/AAAI.V38I17.29889.

Wang, C., Liu, Y., Bi, B., Zhang, D., Li, Z.-Z., Ma, Y., He, Y., Yu, S., Li, X., Fang, J., Zhang, J., & Hooi, B. (2025). Safety in large reasoning models: A survey. In *Findings of the association for computational linguistics: EMNLP 2025* (pp. 3468–3482). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2025.findings-emnlp.185, URL: https://aclanthology.org/2025.findings-emnlp.185/.

Wang, S., Liu, Z., Zhong, W., Zhou, M., Wei, Z., Chen, Z., & Duan, N. (2022). From lsat: The progress and challenges of complex reasoning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *30*, 2201–2216. http://dx.doi.org/10.1109/TASLP.2022.3164218.

Wang, L., Lu, Z., Zhu, Y., Hu, K., Yin, Z., Tang, S., Wang, Z., Ouyang, W., & Ma, X. (2026). Charting empirical laws for LLM fine-tuning in scientific multi-discipline learning. http://dx.doi.org/10.48550/arXiv.2602.11215, arXiv preprint arXiv:2602.11215.

Wang, Z., Ma, S., Wang, K., & Zhuang, Z. (2025). Rule-guided graph neural networks for explainable knowledge graph reasoning. *39*, In *Proceedings of the AAAI conference on artificial intelligence* (12), (pp. 12784–12791). http://dx.doi.org/10.1609/AAAI.V39I12.33394.

Wang, Y., Ren, R., Li, J., Zhao, X., Liu, J., & Wen, J.-R. (2024). REAR: A relevance-aware retrieval-augmented framework for open-domain question answering. In *Proceedings of the 2024 conference on empirical methods in natural language processing* (pp. 5613–5626). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2024.emnlp-main.321, URL: https://aclanthology.org/2024.emnlp-main.321/.

Wang, Z., Su, M., Yang, Y., Zeng, C., & Ye, L. (2024). Cross-disciplinary cognitive diagnosis leveraging deep transfer learning for smart education. In *2024 international conference on intelligent education and intelligent research* (pp. 1–8). http://dx.doi.org/10.1109/IEIR62538.2024.10959920, URL: https://ieeexplore.ieee.org/document/10959920.

Wang, M., Wang, H., Qi, G., & Zheng, Q. (2020). Richpedia: A large-scale, comprehensive multi-modal knowledge graph. *Big Data Research*, *22*, Article 100159. http://dx.doi.org/10.1016/j.bdr.2020.100159, URL: https://www.sciencedirect.com/science/article/pii/S2214579620300277.

Wang, Q., Wang, Z., Su, Y., Tong, H., & Song, Y. (2024). Rethinking the bounds of LLM reasoning: Are multi-agent discussions the key? In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 6106–6131). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2024.acl-long.331, URL: https://aclanthology.org/2024.acl-long.331/.

Wang, M., Wang, S., Yang, H., Zhang, Z., Chen, X., & Qi, G. (2021). Is visual context really helpful for knowledge graph? A representation learning perspective. In *Proceedings of the 29th ACM international conference on multimedia* (pp. 2735–2743). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3474085.3475470.

Wang, X., Wei, J., Schuurmans, D., Le, Q. V., Chi, E. H., Narang, S., Chowdhery, A., & Zhou, D. (2023). Self-consistency improves chain of thought reasoning in language models. In *The eleventh international conference on learning representations, ICLR 2023, kigali, rwanda, May 1-5, 2023*. OpenReview.net, URL: https://openreview.net/pdf?id=1PL1NIMMrw.

Wang, L., Xu, W., Lan, Y., Hu, Z., Lan, Y., Lee, R. K., & Lim, E. (2023a). Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers), ACL 2023, toronto, Canada, July 9-14, 2023* (pp. 2609–2634). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2023.acl-long.147.

Wang, L., Yang, N., Huang, X., Jiao, B., Yang, L., Jiang, D., Majumder, R., & Wei, F. (2023b). SimLM: Pre-training with representation bottleneck for dense passage retrieval. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 2244–2258). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2023.acl-long.125, URL: https://aclanthology.org/2023.acl-long.125/.

Wang, M., Yao, Y., Xu, Z., Qiao, S., Deng, S., Wang, P., Chen, X., Gu, J.-C., Jiang, Y., Xie, P., et al. (2024). Knowledge mechanisms in large language models: A survey and perspective. In *Findings of the association for computational linguistics: EMNLP 2024* (pp. 7097–7135). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2024.findings-emnlp.416, URL: https://aclanthology.org/2024.findings-emnlp.416/.

Wang, Z., Zhang, J., Feng, J., & Chen, Z. (2014). Knowledge graph embedding by translating on hyperplanes. In *Proceedings of the twenty-eighth AAAI conference on artificial intelligence* (pp. 1112–1119). AAAI Press, http://dx.doi.org/10.1609/aaai.v28i1.8870, URL: https://ojs.aaai.org/index.php/AAAI/article/view/8870.

Wason, P. C. (1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology*, *20*(3), 273–281. http://dx.doi.org/10.1080/14640746808400161, URL: https://journals.sagepub.com/doi/abs/10.1080/14640746808400161.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. In *Advances in neural information processing systems 35: annual conference on neural information processing systems 2022, neurIPS 2022, new orleans, la, USA, November 28 - December 9, 2022*. URL: http://papers.nips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.

Weng, Y., Zhu, M., Xia, F., Li, B., He, S., Liu, S., Sun, B., Liu, K., & Zhao, J. (2023). Large language models are better reasoners with self-verification. In *Findings of the association for computational linguistics: EMNLP 2023, Singapore, December 6-10, 2023* (pp. 2550–2575). Association for Computational Linguistics, http://dx.doi.org/10.18653/V1/2023.FINDINGS-EMNLP.167.

Wenzel, F., Dittadi, A., Gehler, P., Simon-Gabriel, C.-J., Horn, M., Zietlow, D., Kernert, D., Russell, C., Brox, T., Schiele, B., et al. (2022). Assaying out-of-distribution generalization in transfer learning. *Advances in Neural Information Processing Systems*, *35*, 7181–7198, URL: http://papers.nips.cc/paper_files/paper/2022/hash/2f5acc925919209370a3af4eac5cad4a-Abstract-Conference.html.

Wiseman, S., Shieber, S., & Rush, A. (2017). Challenges in data-to-document generation. In *Proceedings of the 2017 conference on empirical methods in natural language processing* (pp. 2253–2263). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/D17-1239, URL: https://aclanthology.org/D17-1239/.

Wu, Z., & Feng, Y. (2024). Protrix: Building models for planning and reasoning over tables with sentence context. In *Findings of the association for computational linguistics: EMNLP 2024* (pp. 4378–4406). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2024.findings-emnlp.253, URL: https://aclanthology.org/2024.findings-emnlp.253/.

Wu, S., Fung, Y. R., Qian, C., Kim, J., Hakkani-Tur, D., & Ji, H. (2025). Aligning llms with individual preferences via interaction. In *Proceedings of the 31st international conference on computational linguistics* (pp. 7648–7662). URL: https://aclanthology.org/2025.coling-main.511/.

Wu, Y., Huang, Y., Hu, N., Hua, Y., Qi, G., Chen, J., & Pan, J. Z. (2024). CoTKR: Chain-of-thought enhanced knowledge rewriting for complex knowledge graph question answering. In *Proceedings of the 2024 conference on empirical methods in natural language processing* (pp. 3501–3520). Miami, Florida, USA: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2024.emnlp-main.205, URL: https://aclanthology.org/2024.emnlp-main.205.

Wu, T., Khan, A., Yong, M., Qi, G., & Wang, M. (2022). Efficiently embedding dynamic knowledge graphs. *Knowledge-Based Systems*, *250*, Article 109124. http://dx.doi.org/10.1016/j.knosys.2022.109124, URL: https://www.sciencedirect.com/science/article/pii/S0950705122005548.

Wu, Y., Minervini, P., Stenetorp, P., & Riedel, S. (2020). Don't read too much into it: Adaptive computation for open-domain question answering. In *Proceedings of sustaiNLP: workshop on simple and efficient natural language processing* (pp. 63–72). Association for Computational Linguistics, http://dx.doi.org/10.18653/V1/2020.SUSTAINLP-1.9.

Wu, X., Nian, J., Wei, T.-R., Tao, Z., Wu, H.-T., & Fang, Y. (2025a). Does reasoning introduce bias? A study of social bias evaluation and mitigation in LLM reasoning. In *Findings of the association for computational linguistics: EMNLP 2025* (pp. 18534–18555). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2025.findings-emnlp.1006, URL: https://aclanthology.org/2025.findings-emnlp.1006/.

Wu, X., Yang, J., Chai, L., Zhang, G., Liu, J., Du, X., Liang, D., Shu, D., Cheng, X., Sun, T., et al. (2025b). Tablebench: A comprehensive and complex benchmark for table question answering. *39*, In *Proceedings of the AAAI conference on artificial intelligence* (24), (pp. 25497–25506). AAAI Press, http://dx.doi.org/10.1609/AAAI.V39I24.34739.

Wu, Z., Zeng, Q., Zhang, Z., Tan, Z., Shen, C., & Jiang, M. (2025). Enhancing mathematical reasoning in LLMs by stepwise correction. In *Proceedings of the 63rd annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 21602–21623). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2025.acl-long.1048, URL: https://aclanthology.org/2025.acl-long.1048/.

Xia, S., Li, X., Liu, Y., Wu, T., & Liu, P. (2025). Evaluating mathematical reasoning beyond accuracy. In *Proceedings of the AAAI conference on artificial intelligence*: *vol. 39*, (26), (pp. 27723–27730). http://dx.doi.org/10.1609/AAAI.V39I26.34987.

Xia, Y., Luo, J., Zhou, G., Lan, M., Chen, X., & Chen, J. (2024). DT4kgr: Decision transformer for fast and effective multi-hop reasoning over knowledge graphs. *Information Processing & Management*, *61*(3), Article 103648. http://dx.doi.org/10.1016/j.ipm.2024.103648, URL: https://www.sciencedirect.com/science/article/pii/S0306457324000086.

Xiao, Y., Zhou, G., & Liu, J. (2022). Modeling temporal-sensitive information for complex question answering over knowledge graphs. In *Natural language processing and Chinese computing* (pp. 418–430). Cham: Springer International Publishing, http://dx.doi.org/10.1007/978-3-031-17120-8_33.

Xiao, Y., Zhou, G., Xie, Z., Liu, J., & Huang, J. X. (2024). Learning dual disentangled representation with self-supervision for temporal knowledge graph reasoning. *Information Processing & Management*, *61*(3), Article 103618. http://dx.doi.org/10.1016/j.ipm.2023.103618, URL: https://www.sciencedirect.com/science/article/pii/S0306457323003552.

Xie, R., Liu, Z., Luan, H., & Sun, M. (2017). Image-embodied knowledge representation learning. In *Proceedings of the twenty-sixth international joint conference on artificial intelligence, IJCAI 2017, melbourne, Australia, August 19-25, 2017* (pp. 3140–3146). ijcai.org, http://dx.doi.org/10.24963/IJCAI.2017/438.

Xin, A., Liu, J., Yao, Z., Lee, Z., Cao, S., Hou, L., & Li, J. (2025). Atomr: Atomic operator-empowered large language models for heterogeneous knowledge reasoning. In *Proceedings of the 31st ACM SIGKDD conference on knowledge discovery and data mining v. 2* (pp. 3344–3355). ACM, http://dx.doi.org/10.1145/3711896.3736849.

Xin, C., Lu, Y., Lin, H., Zhou, S., Zhu, H., Wang, W., Liu, Z., Han, X., & Sun, L. (2024). Chain-of-rewrite: Aligning question and documents for open-domain question answering. In *Findings of the association for computational linguistics: EMNLP 2024* (pp. 1884–1896). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2024.findings-emnlp.104, URL: https://aclanthology.org/2024.findings-emnlp.104/.

Xiong, K., Ding, X., Cao, Y., Liu, T., & Qin, B. (2023). Examining inter-consistency of large language models collaboration: An in-depth analysis via debate. In *Findings of the association for computational linguistics: EMNLP 2023* (pp. 7572–7590). Singapore: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2023.findings-emnlp.508, URL: https://aclanthology.org/2023.findings-emnlp.508/.

Xiong, W., Hoang, T., & Wang, W. Y. (2017). DeepPath: A reinforcement learning method for knowledge graph reasoning. In *Proceedings of the 2017 conference on empirical methods in natural language processing* (pp. 564–573). Copenhagen, Denmark: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/D17-1060, URL: https://aclanthology.org/D17-1060.

Xiong, W., Li, X. L., Iyer, S., Du, J., Lewis, P. S. H., Wang, W. Y., Mehdad, Y., Yih, S., Riedel, S., Kiela, D., & Oguz, B. (2021). Answering complex open-domain questions with multi-hop dense retrieval. In *9th international conference on learning representations, ICLR 2021, virtual event, Austria, May 3-7, 2021*. URL: https://openreview.net/forum?id=EMHoBG0avc1.

Xiong, L., Xiong, C., Li, Y., Tang, K., Liu, J., Bennett, P. N., Ahmed, J., & Overwijk, A. (2021). Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *9th international conference on learning representations, ICLR 2021, virtual event, Austria, May 3-7, 2021*. OpenReview.net, URL: https://openreview.net/forum?id=zeFrfgyZln.

Xu, K., Chen, M., Feng, Y., & Dong, Z. (2025). Advancing rule learning in knowledge graphs with structure-aware graph transformer. *Information Processing & Management*, *62*(2), Article 103976. http://dx.doi.org/10.1016/j.ipm.2024.103976, URL: https://www.sciencedirect.com/science/article/pii/S0306457324003352.

Xu, M., Li, Y., Sun, K., & Qian, T. (2024). Adaption-of-thought: Learning question difficulty improves large language models for reasoning. In *Proceedings of the 2024 conference on empirical methods in natural language processing* (pp. 5468–5495). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2024.emnlp-main.313, URL: https://aclanthology.org/2024.emnlp-main.313/.

Xu, F., Lin, Q., Han, J., Zhao, T., Liu, J., & Cambria, E. (2025). Are large language models really good logical reasoners? A comprehensive evaluation and beyond. *IEEE Trans. Knowl. Data Eng.*, *37*(4), 1620–1634. http://dx.doi.org/10.1109/TKDE.2025.3536008.

Xu, W., Liu, B., Peng, M., Jiang, Z., Jia, X., Liu, K., Liu, L., & Peng, M. (2025). Historical facts learning from long-short terms with language model for temporal knowledge graph reasoning. *Information Processing & Management*, *62*(3), Article 104047. http://dx.doi.org/10.1016/j.ipm.2024.104047, URL: https://www.sciencedirect.com/science/article/pii/S0306457324004060.

Xu, X., Ma, R., Zhou, B., Yan, L., & Ma, Z. (2025). Spatial and temporal twin-guided pattern recurrent graph network for implementing reasoning of spatiotemporal knowledge graph. *Information Processing & Management*, *62*(1), Article 103942. http://dx.doi.org/10.1016/j.ipm.2024.103942, URL: https://www.sciencedirect.com/science/article/pii/S0306457324003017.

Xu, C., Nayyeri, M., Alkhoury, F., Yazdi, H. S., & Lehmann, J. (2020). Tero: A time-aware knowledge graph embedding via temporal rotation. In *Proceedings of the 28th international conference on computational linguistics, COLING 2020, Barcelona, Spain (online), December 8-13, 2020* (pp. 1583–1593). International Committee on Computational Linguistics, http://dx.doi.org/10.18653/V1/2020.COLING-MAIN.139.

Xu, Y., Ou, J., Xu, H., & Fu, L. (2023). Temporal knowledge graph reasoning with historical contrastive learning. In *Thirty-seventh AAAI conference on artificial intelligence, AAAI 2023, thirty-fifth conference on innovative applications of artificial intelligence, IAAI 2023, thirteenth symposium on educational advances in artificial intelligence, EAAI 2023, washington, DC, USA, February 7-14, 2023* (pp. 4765–4773). AAAI Press, http://dx.doi.org/10.1609/AAAI.V37I4.25601.

Xu, R., Qi, Z., Guo, Z., Wang, C., Wang, H., Zhang, Y., & Xu, W. (2024). Knowledge conflicts for llms: A survey. In *Proceedings of the 2024 conference on empirical methods in natural language processing* (pp. 8541–8565). http://dx.doi.org/10.18653/v1/2024.emnlp-main.486, URL: https://aclanthology.org/2024.emnlp-main.486/.

Xu, M., Sun, K., Li, Y., & Qian, T. (2023). Cold-start multi-hop reasoning by hierarchical guidance and self-verification. In *Machine learning and knowledge discovery in databases: research track: European conference, ECML pKDD 2023, turin, Italy, September 18–22, 2023, proceedings, part II* (pp. 577–592). Berlin, Heidelberg: Springer-Verlag, http://dx.doi.org/10.1007/978-3-031-43415-0_34.

Xu, D., Xu, T., Wu, S., Zhou, J., & Chen, E. (2022). Relation-enhanced negative sampling for multimodal knowledge graph completion. In *Proceedings of the 30th ACM international conference on multimedia* (pp. 3857–3866). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3503161.3548388.

Xu, Z., Zhang, S., Xia, Y., Xiong, L., & Tong, H. (2020). Ranking on network of heterogeneous information networks. In *2020 IEEE international conference on big data (big data)* (pp. 848–857). IEEE, http://dx.doi.org/10.1109/BIGDATA50022.2020.9378121.

Xue, C., Liang, D., Wang, P., & Zhang, J. (2024). Question calibration and multi-hop modeling for temporal question answering. In *Thirty-eighth AAAI conference on artificial intelligence, AAAI 2024, thirty-sixth conference on innovative applications of artificial intelligence, IAAI 2024, fourteenth symposium on educational advances in artificial intelligence, EAAI 2014, February 20-27, 2024, vancouver, Canada* (pp. 19332–19340). AAAI Press, http://dx.doi.org/10.1609/AAAI.V38I17.29903.

Yamada, I., Asai, A., & Hajishirzi, H. (2021). Efficient passage retrieval with hashing for open-domain question answering. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 2: short papers)* (pp. 979–986). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2021.acl-short.123, URL: https://aclanthology.org/2021.acl-short.123/.

Yan, C., Zhao, F., & Jin, H. (2022). ExKGR: Explainable multi-hop reasoning for evolving knowledge graph. In *Lecture notes in computer science: vol. 13245, Database systems for advanced applications - 27th international conference, DASFAA 2022, virtual event, April 11-14, 2022, proceedings, part i* (pp. 153–161). Springer, http://dx.doi.org/10.1007/978-3-031-00123-9_11.

Yang, S.-X., Mao, X.-L., Shang, Y.-M., & Huang, H. (2025). Toward balanced denoising: Building a structural and textual denoiser for table understanding. *IEEE Transactions on Knowledge and Data Engineering*, *37*(12), 7414–7425. http://dx.doi.org/10.1109/TKDE.2025.3612217, URL: https://ieeexplore.ieee.org/abstract/document/11174007.

Yang, Z., Qi, P., Zhang, S., Bengio, Y., Cohen, W., Salakhutdinov, R., & Manning, C. D. (2018a). HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*. Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/D18-1259, URL: https://aclanthology.org/D18-1259/.

Yang, Z., Qi, P., Zhang, S., Bengio, Y., Cohen, W., Salakhutdinov, R., & Manning, C. D. (2018b). HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 2369–2380). Brussels, Belgium: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/D18-1259, URL: https://aclanthology.org/D18-1259.

Yang, D., Qing, P., Li, Y., Lu, H., & Lin, X. (2022). Gammae: Gamma embeddings for logical queries on knowledge graphs. In *Proceedings of the 2022 conference on empirical methods in natural language processing* (pp. 745–760). Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2022.emnlp-main.47, URL: https://aclanthology.org/2022.emnlp-main.47.

Yang, Y., & Song, L. (2019). Learn to explain efficiently via neural logic inductive learning. In *International conference on learning representations*. OpenReview.net, URL: https://openreview.net/forum?id=SJlh8CEYDB.

Yang, F., Yang, Z., & Cohen, W. W. (2017). Differentiable learning of logical rules for knowledge base reasoning. *Advances in Neural Information Processing Systems*, *30*, 2319–2328, URL: https://proceedings.neurips.cc/paper/2017/hash/0e55666a4ad822e0e34299df3591d979-Abstract.html.

Yang, B., Yih, W., He, X., Gao, J., & Deng, L. (2015). Embedding entities and relations for learning and inference in knowledge bases. In *3rd international conference on learning representations*. URL: http://arxiv.org/abs/1412.6575.

Yang, Y., Yih, W.-t., & Meek, C. (2015). WikiQA: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 conference on empirical methods in natural language processing* (pp. 2013–2018). Lisbon, Portugal: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/D15-1237, URL: https://aclanthology.org/D15-1237/.

Yang, J., Zhang, Z., & Zhao, H. (2021). Multi-span style extraction for generative reading comprehension. In *CEUR workshop proceedings*, *Proceedings of the workshop on scientific document understanding co-located with 35th AAAI conference on artificial inteligence, sDU@aAAI 2021, virtual event, February 9, 2021*. CEUR-WS.org, URL: https://ceur-ws.org/Vol-2831/paper7.pdf.

Yang, Z., Zhu, Z., & Zhu, J. (2025). CuriousLLM: Elevating multi-document question answering with LLM-enhanced knowledge graph reasoning. In *Proceedings of the 2025 conference of the nations of the americas chapter of the association for computational linguistics: human language technologies (volume 3: industry track)* (pp. 274–286). Albuquerque, New Mexico: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2025.naacl-industry.23, URL: https://aclanthology.org/2025.naacl-industry.23/.

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T., Cao, Y., & Narasimhan, K. (2024). Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, *36*, URL: http://papers.nips.cc/paper_files/paper/2023/hash/271db9922b8d1f4dd7aaef84ed5ac703-Abstract-Conference.html.

Ye, X., Yavuz, S., Hashimoto, K., Zhou, Y., & Xiong, C. (2022). RNG-KBQA: Generation augmented iterative ranking for knowledge base question answering. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 6032–6043). Dublin, Ireland: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2022.acl-long.417, URL: https://aclanthology.org/2022.acl-long.417.

Yih, W.-t., Richardson, M., Meek, C., Chang, M.-W., & Suh, J. (2016). The value of semantic parse labeling for knowledge base question answering. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: short papers)* (pp. 201–206). Berlin, Germany: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/P16-2033, URL: https://aclanthology.org/P16-2033.

Yin, Z., & Wang, S. (2025). Enhancing scientific table understanding with type-guided chain-of-thought. *Information Processing & Management*, *62*(4), Article 104159. http://dx.doi.org/10.1016/j.ipm.2025.104159, URL: https://www.sciencedirect.com/science/article/pii/S0306457325001001.

Yoran, O., Wolfson, T., Bogin, B., Katz, U., Deutch, D., & Berant, J. (2023). Answering questions by meta-reasoning over multiple chains of thought. In *The 2023 conference on empirical methods in natural language processing* (pp. 5942–5966). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2023.emnlp-main.364, URL: https://aclanthology.org/2023.emnlp-main.364/.

Yoran, O., Wolfson, T., Ram, O., & Berant, J. (2024). Making retrieval-augmented language models robust to irrelevant context. In *The twelfth international conference on learning representations, ICLR 2024, vienna, Austria, May 7-11, 2024*. OpenReview.net, URL: https://openreview.net/forum?id=ZS4m74kZpH.

Yu, H., Atanasova, P., & Augenstein, I. (2024). Revealing the parametric knowledge of language models: A unified framework for attribution methods. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 8173–8186). Association for Computational Linguistics, http://dx.doi.org/10.18653/V1/2024.ACL-LONG.444.

Yu, P., Chen, G., & Wang, J. (2025). Table-critic: A multi-agent framework for collaborative criticism and refinement in table reasoning. In *Proceedings of the 63rd annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 17432–17451). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2025.acl-long.853, URL: https://aclanthology.org/2025.acl-long.853/.

Yu, W., Iter, D., Wang, S., Xu, Y., Ju, M., Sanyal, S., Zhu, C., Zeng, M., & Jiang, M. (2023). Generate rather than retrieve: Large language models are strong context generators. In *The eleventh international conference on learning representations, ICLR 2023, kigali, rwanda, May 1-5, 2023*. OpenReview.net, URL: https://openreview.net/forum?id=fB0hRu9GZUS.

Yu, T., Jing, Y., Zhang, X., Jiang, W., Wu, W., Wang, Y., Hu, W., Du, B., & Tao, D. (2025). Benchmarking reasoning robustness in large language models. http://dx.doi.org/10.48550/arXiv.2503.04550, arXiv preprint arXiv:2503.04550.

Yu, X., Ren, X., Sun, Y., Sturt, B., Khandelwal, U., Gu, Q., Norick, B., & Han, J. (2013). Recommendation in heterogeneous information networks with implicit user feedback. In *Proceedings of the 7th ACM conference on recommender systems* (pp. 347–350). http://dx.doi.org/10.1145/2507157.2507230.

Yu, T., Wu, C., Lin, X. V., Wang, B., Tan, Y. C., Yang, X., Radev, D. R., Socher, R., & Xiong, C. (2021). GraPPa: Grammar-augmented pre-training for table semantic parsing. In *9th international conference on learning representations, ICLR 2021, virtual event, Austria, May 3-7, 2021*. OpenReview.net, URL: https://openreview.net/forum?id=kyaIeYj4zZ.

Yu, D., Yang, B., Liu, D., Wang, H., & Pan, S. (2023). A survey on neural-symbolic learning systems. *Neural Networks*, *166*, 105–126. http://dx.doi.org/10.1016/J.NEUNET.2023.06.028.

Yu, D., Zhang, S., Ng, P., Zhu, H., Li, A. H., Wang, J., Hu, Y., Wang, W. Y., Wang, Z., & Xiang, B. (2023). DecAF: Joint decoding of answers and logical forms for question answering over knowledge bases. In *The eleventh international conference on learning representations*. URL: https://openreview.net/forum?id=XHc5zRPxqV9.

Yu, F., Zhang, H., Tiwari, P., & Wang, B. (2024). Natural language reasoning, a survey. *ACM Computing Surveys*, *56*(12), 1–39. http://dx.doi.org/10.1145/3664194.

Yuan, Z., Wang, K., Zhu, S., Yuan, Y., Zhou, J., Zhu, Y., & Wei, W. (2025). Finllms: A framework for financial reasoning dataset generation with large language models. *IEEE Transactions on Big Data*, *11*(5), 2264–2277. http://dx.doi.org/10.1109/TBDATA.2024.3524083.

Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., et al. (2024). Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 9556–9567). http://dx.doi.org/10.1109/CVPR52733.2024.00913.

Yuhui, M., Ying, Z., Guangzuo, C., Yun, R., & Ronghuai, H. (2010). Frame-based calculus of solving arithmetic multi-step addition and subtraction word problems. In *2010 second international workshop on education technology and computer science*: *vol. 2*, (pp. 476–479). IEEE, URL: https://ieeexplore.ieee.org/abstract/document/5458590.

Zack, T., Dhaliwal, G., Geha, R., Margaretten, M., Murray, S., & Hong, J. C. (2023). A clinical reasoning-encoded case library developed through natural language processing. *Journal of General Internal Medicine*, *38*(1), 5–11. http://dx.doi.org/10.1007/s11606-022-07758-0, URL: https://link.springer.com/article/10.1007/s11606-022-07758-0.

Zha, Z., Qi, P., Bao, X., Tian, M., & Qin, B. (2024). MÎTQA: Multi-view, multi-hop and multi-stage reasoning for temporal question answering. In *IEEE international conference on acoustics, speech and signal processing, ICASSP 2024, seoul, Republic of Korea, April 14-19, 2024* (pp. 10086–10090). IEEE, http://dx.doi.org/10.1109/ICASSP48485.2024.10448071.

Zhang, Z., Cai, J., Zhang, Y., & Wang, J. (2020). Learning hierarchy-aware knowledge graph embeddings for link prediction. In *The thirty-fourth AAAI conference on artificial intelligence, AAAI 2020, the thirty-second innovative applications of artificial intelligence conference, IAAI 2020, the tenth AAAI symposium on educational advances in artificial intelligence, EAAI 2020, new york, NY, USA, February 7-12, 2020* (pp. 3065–3072). AAAI Press, http://dx.doi.org/10.1609/AAAI.V34I03.5701.

Zhang, Z., Cao, Y., & Liao, L. (2025). XFinbench: Benchmarking LLMs in complex financial problem solving and reasoning. In *Findings of the association for computational linguistics: ACL 2025* (pp. 8715–8758). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2025.findings-acl.457, URL: https://aclanthology.org/2025.findings-acl.457/.

Zhang, Q., Chen, S., Xu, D., Cao, Q., Chen, X., Cohn, T., & Fang, M. (2023). A survey for efficient open domain question answering. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 14447–14465). http://dx.doi.org/10.18653/v1/2023.acl-long.808, URL: https://aclanthology.org/2023.acl-long.808/.

Zhang, Y., Dai, H., Kozareva, Z., Smola, A., & Song, L. (2018). Variational reasoning for question answering with knowledge graph. In *Proceedings of the AAAI conference on artificial intelligence*: *vol. 32*, (1), (pp. 6069–6076). AAAI Press, http://dx.doi.org/10.1609/AAAI.V32I1.12057.

Zhang, Z., Fang, M., & Chen, L. (2024). RetrievalQA: Assessing adaptive retrieval-augmented generation for short-form open-domain question answering. In *Findings of the association for computational linguistics: ACL 2024* (pp. 6963–6975). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2024.findings-acl.415, URL: https://aclanthology.org/2024.findings-acl.415/.

Zhang, B., Gong, M., Huang, J., & Ma, X. (2021). Clustering heterogeneous information network by joint graph embedding and nonnegative matrix factorization. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, *15*(4), 1–25. http://dx.doi.org/10.1145/3441449.

Zhang, D., Jiang, H., Li, X., Li, G., Ning, B., & Chen, H. (2025). Pair-wise or high-order? A self-adaptive graph framework for knowledge graph embedding. *Neural Networks*, *188*, Article 107494. http://dx.doi.org/10.1016/J.NEUNET.2025.107494.

Zhang, W., Jin, L., Zhu, Y., Chen, J., Huang, Z., Wang, J., Hua, Y., Liang, L., & Chen, H. (2025). TrustUQA: A trustful framework for unified structured data question answering. In *Thirty-ninth AAAI conference on artificial intelligence, thirty-seventh conference on innovative applications of artificial intelligence, fifteenth symposium on educational advances in artificial intelligence, AAAI 2025, philadelphia, PA, USA, February 25 - March 4, 2025* (pp. 25931–25939). AAAI Press, http://dx.doi.org/10.1609/AAAI.V39I24.34787.

Zhang, J., Kong, X., & Philip, S. Y. (2013). Predicting social links for new users across aligned heterogeneous social networks. In *2013 IEEE 13th international conference on data mining* (pp. 1289–1294). IEEE, http://dx.doi.org/10.1109/ICDM.2013.134.

Zhang, Y., Kong, X., Ye, K., Shen, G., & Zheng, S. (2025). Tackling sparse facts for temporal knowledge graph completion. In *Proceedings of the ACM on web conference 2025* (pp. 3561–3570). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3696410.3714839.

Zhang, C., Lai, Y., Feng, Y., & Zhao, D. (2021). Extract, integrate, compete: Towards verification style reading comprehension. In *Findings of the association for computational linguistics: EMNLP 2021* (pp. 2976–2986). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2021.findings-emnlp.255, URL: https://aclanthology.org/2021.findings-emnlp.255/.

Zhang, D., Li, W., Qiu, T., & Li, G. (2025). Co-occurrence graph convolutional networks with approximate entailment for knowledge graph embedding. *Applied Soft Computing, 170*, Article 112666. http://dx.doi.org/10.1016/J.ASOC.2024.112666.

Zhang, G., Liu, J., Zhou, G., Zhao, K., Xie, Z., & Huang, B. (2024). Question-directed reasoning with relation-aware graph attention network for complex question answering over knowledge graph. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 32*, 1915–1927. http://dx.doi.org/10.1109/TASLP.2024.3375631.

Zhang, H., Semnani, S., Ghassemi, F., Xu, J., Liu, S., & Lam, M. (2024a). SPAGHETTI: Open-domain question answering from heterogeneous data sources with retrieval and semantic parsing. In *Findings of the association for computational linguistics: ACL 2024* (pp. 1663–1678). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2024.findings-acl.96, URL: https://aclanthology.org/2024.findings-acl.96/.

Zhang, H., Si, S., Zhao, Y., Xie, L., Xu, Z., Chen, L., Nan, L., Wang, P., Tang, X., & Cohan, A. (2024b). OpenT2T: An open-source toolkit for table-to-text generation. In *Proceedings of the 2024 conference on empirical methods in natural language processing: system demonstrations* (pp. 259–269). Miami, Florida, USA: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2024.emnlp-demo.27, URL: https://aclanthology.org/2024.emnlp-demo.27/.

Zhang, Y., Tang, J., Yang, Z., Pei, J., & Yu, P. S. (2015). Cosnet: Connecting heterogeneous social networks with local and global consistency. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1485–1494). http://dx.doi.org/10.1145/2783258.2783268.

Zhang, S., Tay, Y., Yao, L., & Liu, Q. (2019). Quaternion knowledge graph embeddings. In *Proceedings of the 33rd international conference on neural information processing systems*. Red Hook, NY, USA: Curran Associates Inc., URL: https://proceedings.neurips.cc/paper_files/paper/2019/file/d961e9f236177d65d21100592edb0769-Paper.pdf.

Zhang, Z., Wang, J., Chen, J., Ji, S., & Wu, F. (2024). Cone: cone embeddings for multi-hop reasoning over knowledge graphs. In *Proceedings of the 35th international conference on neural information processing systems*. Red Hook, NY, USA: Curran Associates Inc., URL: https://proceedings.neurips.cc/paper/2021/hash/a0160709701140704575d499c997b6ca-Abstract.html.

Zhang, Y., Wang, T., Chen, S., Wang, K., Zeng, X., Lin, H., Han, X., Sun, L., & Lu, C. (2025). ARise: Towards knowledge-augmented reasoning via risk-adaptive search. In *Proceedings of the 63rd annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 10978–10995). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2025.acl-long.538, URL: https://aclanthology.org/2025.acl-long.538/.

Zhang, Z., Wen, L., & Zhao, W. (2024). A GAIL fine-tuned LLM enhanced framework for low-resource knowledge graph question answering. In *Proceedings of the 33rd ACM international conference on information and knowledge management* (pp. 3300–3309). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3627673.3679753.

Zhang, Z., Wu, Y., Zhou, J., Duan, S., Zhao, H., & Wang, R. (2020). SG-net: Syntax-guided machine reading comprehension. In *Proceedings of the AAAI conference on artificial intelligence: vol. 34*, (05), (pp. 9636–9643). AAAI Press, http://dx.doi.org/10.1609/AAAI.V34I05.6511.

Zhang, M., Xia, Y., Liu, Q., Wu, S., & Wang, L. (2023). Learning long- and short-term representations for temporal knowledge graph reasoning. In *Proceedings of the ACM web conference 2023, WWW 2023, austin, TX, USA, 30 April 2023 - 4 May 2023* (pp. 2412–2422). ACM, http://dx.doi.org/10.1145/3543507.3583242.

Zhang, J., Xu, X., Zhang, N., Liu, R., Hooi, B., & Deng, S. (2024). Exploring collaboration mechanisms for LLM agents: A social psychology view. In *ICLR 2024 workshop on large language model (LLM) agents*. URL: https://openreview.net/forum?id=7hjIA8xAOD.

Zhang, Z., Yang, J., & Zhao, H. (2021). Retrospective reader for machine reading comprehension. *35*, In *Proceedings of the AAAI conference on artificial intelligence* (16), (pp. 14506–14514). AAAI Press, http://dx.doi.org/10.1609/AAAI.V35I16.17705.

Zhang, C., Yu, L., Saebi, M., Jiang, M., & Chawla, N. (2020). Few-shot multi-hop relation reasoning over knowledge bases. In *Findings of the association for computational linguistics: EMNLP 2020* (pp. 580–585). Online: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2020.findings-emnlp.51, URL: https://aclanthology.org/2020.findings-emnlp.51.

Zhang, T., Yue, X., Li, Y., & Sun, H. (2024). TableLlama: Towards open large generalist models for tables. In *Proceedings of the 2024 conference of the North American chapter of the association for computational linguistics: human language technologies (volume 1: long papers)* (pp. 6024–6044). Mexico City, Mexico: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2024.naacl-long.335, URL: https://aclanthology.org/2024.naacl-long.335.

Zhang, F., Zhang, Z., Ao, X., Zhuang, F., Xu, Y., & He, Q. (2022). Along the time: Timeline-traced embedding for temporal knowledge graph completion. In *Proceedings of the 31st ACM international conference on information & knowledge management* (pp. 2529–2538). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3511808.3557233.

Zhang, H., Zhang, Y., Li, X., Shi, W., Xu, H., Liu, H., Wang, Y., Shang, L., Liu, Q., Liu, Y., et al. (2024). Evaluating the external and parametric knowledge fusion of large language models. http://dx.doi.org/10.48550/ARXIV.2405.19010, arXiv preprint arXiv:2405.19010.

Zhang, Z., Zhang, A., Li, M., & Smola, A. (2023). Automatic chain of thought prompting in large language models. In *The eleventh international conference on learning representations, ICLR 2023, kigali, rwanda, May 1-5, 2023*. OpenReview.net, URL: https://openreview.net/pdf?id=5NTt8GFjUHkr.

Zhang, J., Zhang, X., Yu, J., Tang, J., Tang, J., Li, C., & Chen, H. (2022). Subgraph retrieval enhanced model for multi-hop knowledge base question answering. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 5773–5784). Dublin, Ireland: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2022.acl-long.396, URL: https://aclanthology.org/2022.acl-long.396.

Zhang, Z., Zhang, Y., & Zhao, H. (2021). Syntax-aware multi-spans generation for reading comprehension. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 30*, 260–268. http://dx.doi.org/10.1109/TASLP.2021.3138679.

Zhang, K., Zuo, Y., He, B., Sun, Y., Liu, R., Jiang, C., Fan, Y., Tian, K., Jia, G., Li, P., Fu, Y., Lv, X., Zhang, Y., Zeng, S., Qu, S., Li, H., Wang, S., Wang, Y., Long, X., ... Zhou, B. (2025). A survey of reinforcement learning for large reasoning models. http://dx.doi.org/10.48550/ARXIV.2509.08827, CoRR. arXiv:2509.08827.

Zhangyue, Y., Yuxin, W., Xiannian, H., Yiguang, W., Hang, Y., Xinyu, Z., Zhao, C., Xuanjing, H., & Xipeng, Q. (2023). Rethinking label smoothing on multi-hop question answering. In *Proceedings of the 22nd Chinese national conference on computational linguistics* (pp. 611–623). Springer, http://dx.doi.org/10.1007/978-981-99-6207-5_5.

Zhao, W., Liu, Y., Niu, T., Wan, Y., Yu, P., Joty, S., Zhou, Y., & Yavuz, S. (2024). DIVKNOWQA: Assessing the reasoning ability of LLMs via open-domain question answering over knowledge base and text. In *Findings of the association for computational linguistics: NAACL 2024* (pp. 51–68). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2024.findings-naacl.5, URL: https://aclanthology.org/2024.findings-naacl.5/.

Zhao, Y., Long, Y., Liu, H., Kamoi, R., Nan, L., Chen, L., Liu, Y., Tang, X., Zhang, R., & Cohan, A. (2024). Docmath-eval: Evaluating math reasoning capabilities of llms in understanding financial documents. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 16103–16120). http://dx.doi.org/10.18653/v1/2024.acl-long.852, URL: https://aclanthology.org/2024.acl-long.852/.

Zheng, S., Chen, W., Zhao, P., Liu, A., Fang, J., & Zhao, L. (2021). When hardness makes a difference: Multi-hop knowledge graph reasoning over few-shot relations. In *Proceedings of the 30th ACM international conference on information & knowledge management* (pp. 2688–2697). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3459637.3482402.

Zheng, M., Hao, Y., Jiang, W., Lin, Z., Lyu, Y., She, Q., & Wang, W. (2023). IM-TQA: A Chinese table question answering dataset with implicit and multi-type table structures. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 5074–5094). Toronto, Canada: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2023.acl-long.278, URL: https://aclanthology.org/2023.acl-long.278.

Zheng, C., Liu, Z., Xie, E., Li, Z., & Li, Y. (2024). Progressive-hint prompting improves reasoning in large language models. In *AI for math workshop @ ICML 2024*. URL: https://openreview.net/forum?id=UkFEs3ciz8.

Zheng, S., Yin, H., Chen, T., Nguyen, Q. V. H., Chen, W., & Zhao, L. (2023). DREAM: adaptive reinforcement learning based on attention mechanism for temporal knowledge graph reasoning. In *Proceedings of the 46th international ACM SIGIR conference on research and development in information retrieval, SIGIR 2023, taipei, Taiwan, July 23-27, 2023* (pp. 1578–1588). ACM, http://dx.doi.org/10.1145/3539618.3591671.

Zhong, E., Fan, W., Zhu, Y., & Yang, Q. (2013). Modeling the dynamics of composite social networks. In *Proceedings of the 19th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 937–945). http://dx.doi.org/10.1145/2487575.2487652.

Zhong, W., Huang, J., Liu, Q., Zhou, M., Wang, J., Yin, J., & Duan, N. (2022). Reasoning over hybrid chain for table-and-text open domain question answering. In *Proceedings of the thirty-first international joint conference on artificial intelligence, IJCAI-22* (pp. 4531–4537). International Joint Conferences on Artificial Intelligence Organization, http://dx.doi.org/10.24963/ijcai.2022/629, Main Track.

Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., Wang, L., Li, C., & Sun, M. (2020). Graph neural networks: A review of methods and applications. *AI Open*, *1*, 57–81. http://dx.doi.org/10.1016/j.aiopen.2021.01.001, URL: https://www.sciencedirect.com/science/article/pii/S2666651021000012.

Zhou, Y., Huang, J., Sun, H., Sun, Y., Qiao, S., & Wambura, S. (2019). Recurrent meta-structure for robust similarity measure in heterogeneous information networks. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, *13*(6), 1–33. http://dx.doi.org/10.1145/3364226.

Zhou, K., Liu, C., Zhao, X., Jangam, S., Srinivasa, J., Liu, G., Song, D., & Wang, X. E. (2025). The hidden risks of large reasoning models: A safety assessment of r1. In *Proceedings of the 14th international joint conference on natural language processing and the 4th conference of the Asia-Pacific chapter of the association for computational linguistics* (pp. 3250–3265). The Asian Federation of Natural Language Processing and The Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2025.ijcnlp-long.173, URL: https://aclanthology.org/2025.ijcnlp-long.173/.

Zhou, P., Peng, X., Zhang, F., Xu, Z., Ai, J., Qiu, Y., Zhao, W., Song, J., Li, C., Tang, W., et al. (2026). Mdk12-bench: a multi-discipline benchmark for evaluating reasoning in multimodal large language models. In *Proceedings of the AAAI conference on artificial intelligence*: *vol. 40*, (34), (pp. 28982–28990). AAAI Press, http://dx.doi.org/10.1609/AAAI.V40I34.40134.

Zhou, D., Schärli, N., Hou, L., Wei, J., Scales, N., Wang, X., Schuurmans, D., Cui, C., Bousquet, O., Le, Q. V., & Chi, E. H. (2023). Least-to-most prompting enables complex reasoning in large language models. In *The eleventh international conference on learning representations, ICLR 2023, kigali, rwanda, May 1-5, 2023*. OpenReview.net, URL: https://openreview.net/pdf?id=WZH7099tgfM.

Zhou, J., Zhong, W., Wang, Y., & Wang, J. (2025). Adaptive-solver framework for dynamic strategy selection in large language model reasoning. *Information Processing & Management*, *62*(3), Article 104052. http://dx.doi.org/10.1016/j.ipm.2024.104052, URL: https://www.sciencedirect.com/science/article/pii/S0306457324004114.

Zhu, C., Chen, M., Fan, C., Cheng, G., & Zhang, Y. (2021). Learning from history: Modeling temporal knowledge graphs with sequential copy-generation networks. In *Thirty-fifth AAAI conference on artificial intelligence, AAAI 2021, thirty-third conference on innovative applications of artificial intelligence, IAAI 2021, the eleventh symposium on educational advances in artificial intelligence, EAAI 2021, virtual event, February 2-9, 2021* (pp. 4732–4740). AAAI Press, http://dx.doi.org/10.1609/AAAI.V35I5.16604.

Zhu, Z., Galkin, M., Zhang, Z., & Tang, J. (2022). Neural-symbolic models for logical queries on knowledge graphs. In *Proceedings of machine learning research*: *vol. 162*, *Proceedings of the 39th international conference on machine learning* (pp. 27454–27478). PMLR, URL: https://proceedings.mlr.press/v162/zhu22c.html.

Zhu, A., Hwang, A., Dugan, L., & Callison-Burch, C. (2024). FanOutQA: A multi-hop, multi-document question answering benchmark for large language models. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 2: short papers)* (pp. 18–37). Bangkok, Thailand: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2024.acl-short.2, URL: https://aclanthology.org/2024.acl-short.2.

Zhu, F., Lei, W., Huang, Y., Wang, C., Zhang, S., Lv, J., Feng, F., & Chua, T.-S. (2021a). TAT-QA: A question answering benchmark on a hybrid of tabular and textual content in finance. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: long papers)* (pp. 3277–3287). Online: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2021.acl-long.254, URL: https://aclanthology.org/2021.acl-long.254.

Zhu, F., Lei, W., Wang, C., Zheng, J., Poria, S., & Chua, T.-S. (2021b). Retrieving and reading: A comprehensive survey on open-domain question answering. http://dx.doi.org/10.48550/arXiv.2101.00774, arXiv preprint arXiv:2101.00774.

Zong, C., Yan, Y., Lu, W., Shao, J., Huang, Y., Chang, H., & Zhuang, Y. (2024). Triad: A framework leveraging a multi-role LLM-based agent to solve knowledge base question answering. In *Proceedings of the 2024 conference on empirical methods in natural language processing* (pp. 1698–1710). Miami, Florida, USA: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2024.emnlp-main.101, URL: https://aclanthology.org/2024.emnlp-main.101.