# Format as a Prior:
# Quantifying and Analyzing Bias in LLMs for Heterogeneous Data

**Jiacheng Liu**[1, *], **Mayi Xu**[1, *], **Qiankun Pi**[1], **Wenli Li**[1], **Ming Zhong**[1], **Yuanyuan Zhu**[1], **Mengchi Liu**[1], **Tieyun Qian**[1, †]

[1]School of Computer Science, Wuhan University, China
{liu-jia-cheng, xumayi, qty}@whu.edu.cn

## Abstract

Large Language Models (LLMs) are increasingly employed in applications that require processing information from heterogeneous formats, including texts, tables, infoboxes, and knowledge graphs. However, systematic biases toward particular formats may undermine LLMs' ability to integrate heterogeneous data impartially, potentially resulting in reasoning errors and increased risks in downstream tasks. Yet it remains unclear *whether such biases are systematic*, *which data-level factors drive them*, and *what internal mechanisms underlie their emergence*.

In this paper, we present the first comprehensive study of format bias in LLMs through a three-stage empirical analysis. The first stage explores the presence and direction of bias across a diverse range of LLMs. The second stage examines how key data-level factors influence these biases. The third stage analyzes how format bias emerges within LLMs' attention patterns and evaluates a lightweight intervention to test its effectiveness. Our results show that format bias is consistent across model families, driven by information richness, structure quality, and representation type, and is closely associated with attention imbalance within the LLMs. Based on these investigations, we identify three future research directions to reduce format bias: enhancing data pre-processing through format repair and normalization, introducing inference-time interventions such as attention reweighting, and developing format-balanced training corpora. These directions will support the design of more robust and fair heterogeneous data processing systems.

**Code** — https://github.com/NLPGM/Format-as-a-prior

**Appendix** — https://github.com/NLPGM/Format-as-a-prior/appendix.pdf

## 1 Introduction

Large Language Models (LLMs) have demonstrated impressive capabilities across a wide range of natural language tasks (Brown et al. 2020). However, their practical deployment remains constrained by key limitations, including factual inaccuracies (commonly referred to as "hallucinations") (Ji et al. 2023) and incomplete or outdated knowledge (Petroni et al. 2019). One promising direction to address these issues is to incorporate external knowledge sources into the reasoning process—allowing models to ground their outputs in more accurate, up-to-date, and contextually relevant information (Lewis et al. 2020; Gao et al. 2023; Huang and Chang 2022).

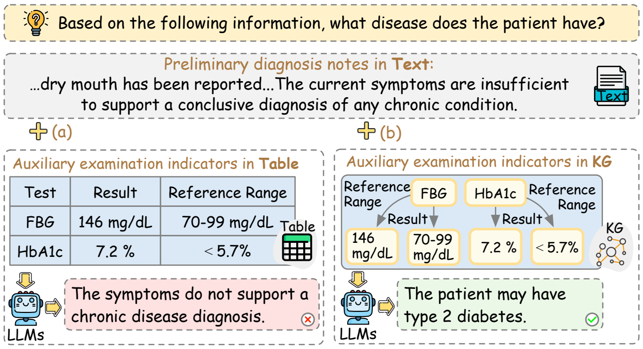

Based on the following information, what disease does the patient have?
Preliminary diagnosis notes in Text:
...dry mouth has been reported...The current symptoms are insufficient to support a conclusive diagnosis of any chronic condition.
Text
(a)
(b)
Auxiliary examination indicators in Table
Test
Result
Reference Range
FBG
146 mg/dL
70-99 mg/dL
HbA1c
7.2 %
< 5.7%
Table
LLMs
The symptoms do not support a chronic disease diagnosis.
Auxiliary examination indicators in KG
Reference Range
FBG
HbA1c
Reference Range
Result
Result
146 mg/dL
70-99 mg/dL
7.2 %
< 5.7%
KG
LLMs
The patient may have type 2 diabetes.

Figure 1: Format bias affects the LLM's decision.

In practice, external knowledge exists in diverse formats, ranging from unstructured texts to semi-structured infoboxes, as well as structured tables and Knowledge Graphs (KGs). These different sources of knowledge often complement one another, and the ability to effectively harness them together is crucial for real-world, knowledge-intensive applications.

The presence of different formats introduces a critical challenge: LLMs may not treat all formats equally when leveraging these heterogeneous data collaboratively. An LLM with strong format preferences interprets information through a distorted lens, giving undue weight to favored formats regardless of actual relevance. This can affect its ability to reason and synthesize effectively.

For example, in clinical decision support, an LLM (e.g., Qwen3-8b) given both textual notes and tabular examination data may overemphasize the text while overlooking key indicators in the table, leading to an incorrect diagnosis. In contrast, as shown in Figure 1, presenting the same information in a knowledge graph enables the model to identify abnormalities and reach the correct conclusion.

When homogeneous inputs are converted into heterogeneous ones with equivalent content, accuracy decreases

[*] Jiacheng Liu and Mayi Xu contribute equally to this work.
[†] Corresponding author

by 9% and 12% on HotpotQA (Yang et al. 2018) and MuSiQue (Trivedi et al. 2022) (200 samples each), confirming that format heterogeneity can directly impair reasoning performance. This phenomenon may widely arise in heterogeneous reasoning (Christmann, Saha Roy, and Weikum 2024), where key evidence is distributed across texts, tables, infoboxes, and knowledge graphs. LLMs may focus on information from their preferred formats, potentially overlooking crucial data in others. Such bias can result in incomplete or flawed conclusions and undermine the LLMs' role as impartial and effective synthesizers of heterogeneous inputs.

Although there have been several studies exploring various types of bias in LLMs, such as bias between multi-modal data (Zhu et al. 2024; Zhang et al. 2025), there is a lack of systematic research on format bias. To address this gap, we present the first comprehensive investigation and analysis of format bias in LLMs. Our study centers on three critical questions: *whether such format biases are systematic*, *which data-level factors contribute to them*, and *what internal mechanisms in LLMs underlie their emergence*.

To systematically investigate the three questions, we conduct a three-stage empirical study by constructing a heterogeneous data conflict scenario for the exploration of bias. The first stage explores the presence and direction of bias across a diverse range of LLMs. The second stage aims to examine how key data-level factors, including information richness, structure quality, and format type, influence these biases. The third stage investigates the emergence of format bias within LLMs' attention mechanisms and evaluates a lightweight intervention strategy to assess its effectiveness in mitigating such bias.

Our results reveal that format bias is both systematic and consistent across models, driven primarily by differences in information richness, structure quality, and format type. We further show that such bias originates from imbalanced attention allocation during inference and can be partially mitigated through attention-based interventions.

Our key contributions are as follows:

1. To the best of our knowledge, we are the first to investigate the issue of format bias in LLMs and to present a comprehensive investigation of LLM biases toward different knowledge formats across a wide range of LLMs.
2. We conduct a three-stage empirical study to examine the presence and direction of bias, identify the data-level factors that give rise to the bias, and investigate the internal mechanisms in LLMs that contribute to their presence.
3. Based on the comprehensive investigation, we identify three future research directions that may reduce the format bias, which will contribute the development of a more effective heterogeneous data processing system.

## 2 Related Work

### 2.1 Heterogeneous reasoning

An important direction in AI research is to develop LLMs capable of reasoning over heterogeneous knowledge sources, including unstructured texts, tables, and KGs, especially when relevant evidence is dispersed across different formats. However, existing methods often struggle to integrate such fragmented information effectively for accurate inference.

To formalize this challenge, recent benchmarks such as *COMPMIX* (Christmann, Saha Roy, and Weikum 2024) and *CompMix-IR* (Min et al. 2024) require cross-format reasoning, stimulating the development of hybrid QA systems that combine structured and unstructured inputs.

Current approaches can be broadly categorized into two types: (1) *Unified retrieval frameworks*, which abstract away format heterogeneity using shared APIs or embedding spaces (Xia et al. 2025; Min et al. 2024); and (2) *LLM-centric pipelines*, which enhance downstream reasoning via evidence selection, re-ranking, or modular tool use (Christmann and Weikum 2024; Lehmann et al. 2024; Zhang et al. 2024; Biswal et al. 2024).

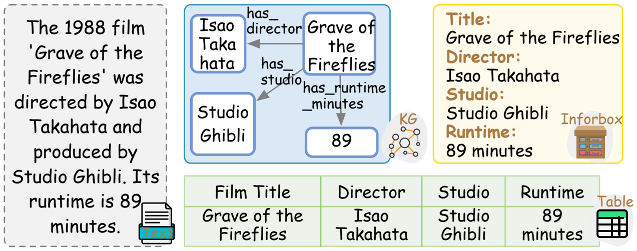
However, existing work often assumes that once evidence is retrieved, LLMs will evaluate it fairly based on content alone. Our study revisits this assumption by asking whether the format in which evidence is presented can influence the model's judgment, even when the underlying meaning remains the same.

### 2.2 LLM Behavior under Conflicting Evidence

Recent studies have uncovered a broad range of behavioral biases in LLMs, extending beyond social stereotypes to systematic patterns in reasoning and judgment. A key challenge lies in how LLMs handle conflicts—both between their *parametric knowledge* (internal beliefs) and *in-context evidence*, and among competing pieces of evidence (Xu et al. 2024). LLMs tend to favor information aligned with their pre-trained knowledge, even when contradicted by accurate inputs (Jin et al. 2024a; Xie et al. 2023). These LLM biases are shaped by factors such as entity popularity (Xie et al. 2023), event recency (Fang et al. 2023), and evidence frequency (Jin et al. 2024a). Input artifacts also affect LLM behavior, such as preferring self-generated content over retrieved passages (Tan et al. 2024). Related efforts have also examined knowledge conflicts in multi-modal (Zhang et al. 2025; Zhu et al. 2024) and multi-agent settings (Ju et al. 2025), where preference biases and inter-agent inconsistency further complicate reasoning.

Benchmarks like *ConflictBank* (Su et al. 2024), *WikiContradict* (Hou et al. 2024), and *WhoQA* (Pham et al. 2024) have been proposed to evaluate how LLMs handle factual or semantic inconsistencies, especially in ambiguous scenarios. In response, a range of strategies have emerged, including conflict-aware decoding (Yuan et al. 2024; Jin et al. 2024a), counterfactual data augmentation (Fang et al. 2023), internal intervention via attention pruning (Jin et al. 2024b), neuron reweighting (Shi et al. 2024), or prompting LLMs to generate multi-answer responses with source attribution (Shaier, Kobren, and Ogren 2024).

Our work extends prior research by identifying a previously overlooked source of bias: the format in which information is presented. In contrast to earlier studies that focus on content-level factors such as recency or frequency, we show that differences in format representation alone can systematically influence LLMs' behavior.



The 1988 film 'Grave of the Fireflies' was directed by Isao Takahata and produced by Studio Ghibli. Its runtime is 89 minutes.

Title:
Grave of the Fireflies
Director:
Isao Takahata
Studio:
Studio Ghibli
Runtime:
89 minutes

| Film Title | Director | Studio | Runtime |
|---|---|---|---|
| Grave of the Fireflies | Isao Takahata | Studio Ghibli | 89 minutes |

Figure 2: Examples of the four data formats used in our experiments: texts, tables, infoboxes, and KGs.

## 3 Investigation Framework

This section establishes a framework for analyzing format bias in LLMs, encompassing dataset construction, confounding factor exclusion, and automated response evaluation. The dataset construction process is detailed in Appendix A, with evaluation details in Appendix B.

### 3.1 Dataset and Format Construction

We construct our dataset based on *ConflictBank* (Su et al. 2024), a public corpus designed to evaluate LLM behavior under factual conflicts. We randomly sample 4,000 entries, each containing one factual claim and three counterclaims with corresponding supporting evidence. This results in 12,000 contradiction pairs, as each claim is paired individually with each of the three counterclaims. Each pair consists of two different claims about the same subject and relation, each supported by its own piece of evidence.

While the original evidence in *ConflictBank* is presented in plain text, our goal is to examine how LLM behavior is influenced by presenting supporting evidence in different data formats. To construct each heterogeneous contradiction pair, we randomly convert the two pieces of evidence, each of which supports a different claim about the same subject, into different formats. If the selected format is text, no conversion is applied.

We use GPT-4o-mini as a transformation engine to convert selected texts into one of the following Wikipedia-inspired formats.

- KGs: A set of (Subject, Predicate, Object) triples capturing core semantic relations.
- Infobox: A structured key-value format modeled after Wikipedia infobox templates, summarizing factual information.
- Table: A tabular format styled after Wikipedia tables, presenting comparable facts in labeled rows and columns.

Figure 2 provides illustrative examples of these four formats, showing how semantically equivalent information can be presented in structurally distinct ways. These formats reflect the most commonly used representations in prior work on heterogeneous reasoning. Manual inspection of 5% samples confirms 98.7% factual and 99.3% syntactic accuracy, reflecting strong data fidelity.

### 3.2 Confounding Factor Control

To ensure that our evaluation isolates the effect of evidence format itself, we implement controls to eliminate two major confounding factors: internal knowledge bias and evidence presentation order.

- Filtering Internal Knowledge: To ensure that LLM responses are based on external evidence rather than parametric memory, we adopt a filtering procedure consistent with prior work (Gekhman et al. 2024). Each factual claim is tested 16 times by directly querying a given LLM with the corresponding question in a zero-shot setting, and only those samples for which the model fails to reproduce the factual claim in all trials are retained.
- Randomizing Evidence Order: To eliminate the known bias introduced by input order (Xie et al. 2023), we randomize the sequence of all evidence segments for each input.

### 3.3 Evaluated LLMs

This evaluation covers ten LLMs across six major series: GPT-4o-mini (Achiam et al. 2023), LLaMA-3.1 (8B), Mistral (7B), Qwen3 (8B, 14B, 30B-A3B, 32B) (Team 2025), Gemma-2 (9B, 27B) (Team et al. 2024), and GLM-4 (9B) (GLM et al. 2024). These LLMs span a range of sizes and architectures, enabling cross-family comparison of format-driven biases.

To ensure reproducibility and eliminate stochasticity, all evaluations were conducted in deterministic inference mode (temperature = 0, without sampling randomness).

### 3.4 Evaluation Protocol and Metrics

Given the dataset's scale, we adopt an automated evaluation pipeline using LLMs as adjudicators. Specifically, GPT-4o-mini, GLM-4.5-Air (Zeng et al. 2025), and Qwen-plus are employed to judge which of the two conflicting claims (Source A or Source B) each target response supports.

Each model independently evaluates every sample three times for stability, with the final label per model determined by majority vote. We then compute FPR and DCR for each model and report their averages across the three evaluators. This multi-model setup improves consistency and mitigates variance in individual LLM judgments. Manual checks on a randomly sampled 5% subset show 99.8% agreement between human and averaged LLM judgments, confirming high reliability.

Each LLM response is classified into one of three mutually exclusive categories:

- **Pref-A:** The response predominantly or exclusively supports the claim from Source A.
- **Pref-B:** The response predominantly or exclusively supports the claim from Source B.
- **Both:** The response acknowledges the contradiction and presents both perspectives in a comparative or side-by-side manner.

All responses in our experiments fall unambiguously into one of these three categories, and no additional response types are observed. This categorization enables two quantitative bias metrics used throughout our analysis.:

- **Dual Coverage Rate (DCR)**: Measures the proportion of responses that acknowledge both claims, indicating an LLM's capacity to represent multiple perspectives:

$$\text{DCR} = \frac{\text{Both}}{\text{Pref-A} + \text{Pref-B} + \text{Both}} \tag{1}$$

- **Format Preference Ratio (FPR)**: Captures asymmetric bias between conflicting claims when an LLM gives a single-sided response. For the A vs. B experiments, the FPR is calculated as:

$$\text{FPR} = \frac{\text{Pref-A}}{\text{Pref-A} + \text{Pref-B}} \tag{2}$$

These two metrics correspond to the two bias types introduced earlier: **DCR** captures the presence of bias, while **FPR** captures the direction of bias.

## 4 Experimental Results and Analysis

In this section, we present the empirical results of our three-stage investigation. Complete experimental details and raw results are provided in Appendix C.

### 4.1 Establishing the Existence of Format Bias

**Objective and Setup** The primary aim of this initial experiment is to test the null hypothesis that LLMs process information in a format-agnostic manner. To this end, we conduct a large-scale evaluation using the experimental framework outlined in Section 3. We assess ten state-of-the-art LLMs from various families and parameter scales. Each LLM is evaluated on all six possible pairs among the four target data formats: texts, tables, infoboxes, and KGs. This design yields 60 unique experimental conditions (10 LLMs × 6 format pairs), enabling a comprehensive assessment of systematic format biases.

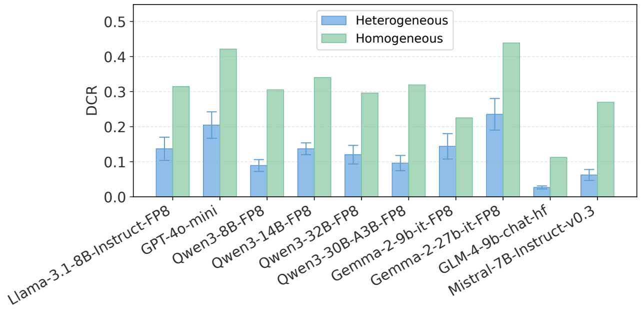


Figure 3: Average DCR across models under heterogeneous and homogeneous formats. Error bars show standard error.

**Top-Level Findings** Our results provide strong evidence against the null hypothesis of format impartiality, revealing instead a consistent and multifaceted pattern of bias in both its presence and direction.

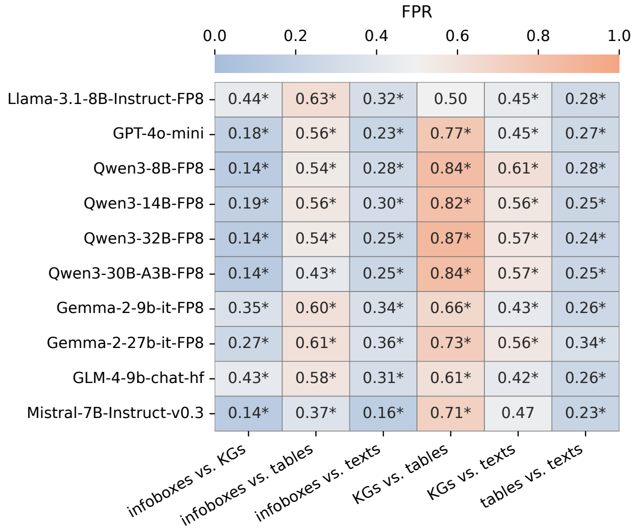


Figure 4: Heatmap of FPR between format pairs across LLMs. Asterisks (*) indicate statistical significance under a two-sided binomial test with null hypothesis FPR $= 0.5$.

The first pattern, the presence of bias, is pervasive under heterogeneous format conditions. This is reflected in the uniformly low Dual Coverage Rate (DCR), ranging from 3.01% to 24.02%, indicating that LLMs often fail to acknowledge conflicting information across different formats, typically exhibiting a preference for one input while disregarding the other.

To isolate the role of format, we introduce a control condition where both inputs are presented in plain text. As shown in Figure 3, DCR increases markedly under this homogeneous setting. This contrast highlights a broader pattern: format heterogeneity alone can independently and substantially impair a model's ability to jointly consider multiple inputs, even when the content is semantically equivalent.

This limitation persists across models of varying size, as no clear scaling trend is observed. For example, within the Qwen3 series, larger models do not exhibit improved performance in this regard, suggesting that increased parameter size alone does not resolve this form of processing asymmetry.

The second, and more decisive, pattern is a strong directional bias that emerges when an LLM commits to a single-sided response. Despite considerable variation in architecture and scale, LLMs demonstrate a surprisingly consistent pattern of format preferences. As shown in Figure 4, our cross-model analysis reveals a clear preference hierarchy: semantically rich formats such as texts and KGs are consistently favored over visually structured ones like infoboxes and tables. Furthermore, when we group the data by topical domain, we find that these biases persist across domains, suggesting that the observed biases are robust and generalizable. See Appendix C.7 for detailed domain-level results.

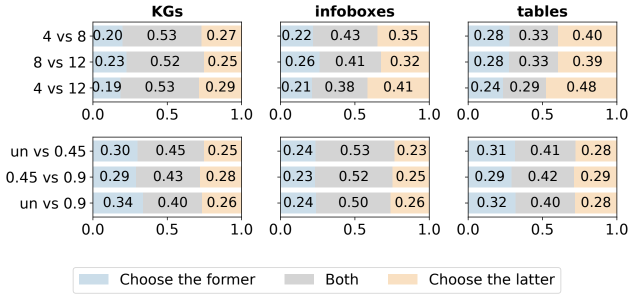


Figure 5: LLM biases across conditions (averaged over ten LLMs). Top: Information Richness; Bottom: Structure Quality. Bars show the proportion of responses favoring the former input, the latter, or both.

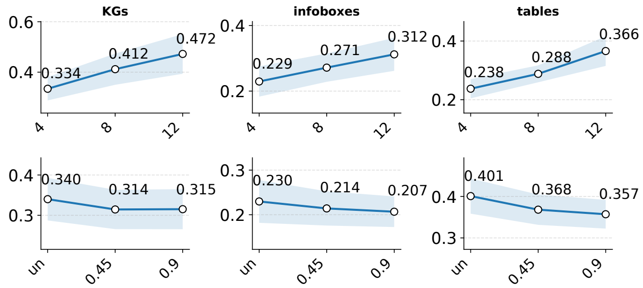


Figure 6: Average FPR for structured data vs. texts across ten LLMs. Top: Information Richness; Bottom: Structure Quality. Shaded areas indicate mean ± one standard deviation.

## 4.2 Identifying the Factors Behind Format Bias

Having established the widespread presence of format bias, we now turn to investigating its data-level factors. While prior work has extensively explored factors influencing LLM biases in textual inputs, our focus is on structured data and the properties that may shape LLM behavior in this context. To move from identifying whether such bias exists to understanding why it arises, we decompose the abstract notion of "format" into three representative and controllable dimensions: the structure itself, and the content, which we further divide into quantity and quality.

Building on this decomposition, we hypothesize that three key dimensions (the amount of information conveyed, referred to as quantity; the structural quality of that information, or quality; and the mode of presentation, or format type) play a central role in shaping how LLMs evaluate evidence.

To assess the influence of these factors, we design a unified experimental framework that applies to two of the three factors (with the exception of the format type). Each of these two factors is examined under two complementary conditions:

- **Homogeneous setting**: Two evidence sources of the same format are compared, differing only in the factor under investigation. This controlled design isolates the variable to reveal an LLM's intrinsic sensitivity to that property.
- **Heterogeneous setting**: A structured evidence source is paired with unstructured plain texts. This setup assesses the relative strength of the structured format and how its properties influence an LLM's preference when compared against a universal baseline.

**Factor 1: Information Richness** This factor concerns the volume of factual detail within an external knowledge source used during reasoning. In real-world applications, the external context provided to LLMs often varies widely in the amount of information it contains. It serves as a proxy for the completeness of factual detail. To examine whether LLMs apply a "more is better" heuristic, interpreting quantity as indicative of evidentiary strength, we systematically vary the number of entries in structured formats (e.g., table rows or knowledge graph triples).

In the homogeneous setting, we conduct experiments within each format type (tables, KGs, and infoboxes), comparing three levels of information richness across three pairwise conditions: 4 vs. 8 entries, 8 vs. 12 entries, and 4 vs. 12 entries.Results consistently indicate that LLMs favor the richer variant in each pair, irrespective of format. This suggests that even when structure is held constant, LLMs exhibit a systematic bias for inputs with more factual content (see Figure 6).

In the heterogeneous setting, we assess whether this preference generalizes when structured inputs are compared against unstructured texts. For each format, we construct three pairs: texts vs. 4-entry structure, texts vs. 8-entry structure, and texts vs. 12-entry structure. All structured inputs are generated from the same source texts as the corresponding text versions, ensuring content consistency. Across all formats, LLMs' bias for the structured inputs increases with the number of entries (see Figure 6).

These findings suggest that LLMs tend to associate greater volumes of factual content with higher evidentiary value, both within individual formats and when comparing structured and unstructured inputs.

**Factor 2: Structure Quality** This factor examines whether LLMs are sensitive to the structural integrity of external knowledge inputs. In practice, structured formats like tables or knowledge graphs may contain noise or malformed syntax. To simulate this, we introduce controlled corruption into structure-defining tokens (e.g., brackets, colons, separators), randomly replacing them with other characters or blanks at fixed probabilities (0.45 and 0.9), while preserving the underlying factual content.

In the homogeneous setting, we compare clean and corrupted versions within each format type. LLMs consistently favor the well-formed input, confirming that structure quality serves as a reliability signal. Notably, the preference saturates beyond moderate corruption (e.g., 0.45), indicating that LLMs tend to treat inputs as either structurally valid or invalid (see Figure 6).

In the heterogeneous setting, we pair each corrupted version with its corresponding clean texts. Each corruption level is applied to the same intact structured data instance, guaranteeing that the underlying information remains identical across differently corrupted versions. LLMs' bias for the structured inputs declines sharply as corruption increases, despite identical semantics. This suggests that structural degradation alone can undermine the perceived credibility of otherwise accurate structured inputs (see Figure 6).

**Factor 3: Format Type** This factor investigates the impact of the representational structure and layout used to represent logically equivalent information. The choice between plain text, a relational graph (KGs), or a visual grid (tables, infoboxes) embodies core differences in format semantics. Holding the content constant, we ask: do LLMs possess intrinsic preferences for certain data structures?

| Metric | Infobox | Table | KGs |
|---|---|---|---|
| FPR | 0.235 | 0.398 | 0.336 |

Table 1: Average FPR between structured data and texts (mean across ten LLMs).

We construct matched pairs with identical factual entries in both texts and structured variants. As shown in Table 1, the results reveal a consistent hierarchy: tables are most competitive, followed by KGs, with infoboxes least preferred. All three structured variants contain nearly identical content, differing only in their representational layout. This suggests that the format itself, rather than the informational content, modulates LLMs' bias.

**Cross-Factor Insight** Consistent with the findings in Section 4.1, we observe that format homogeneity substantially reduces bias in presence. This trend holds not only for unstructured inputs but also extends to structured formats such as tables and knowledge graphs, where DCR increases significantly (28%–53%) when both inputs adopt the same format. These results suggest that LLMs are more capable of jointly considering multiple sources of information when they are presented in a uniform structure. In contrast, even when the content is semantically equivalent, presenting information in heterogeneous formats tends to impair integration and leads to partial or selective processing.

### 4.3 Mechanism Behind Format Bias

In this analysis, we move beyond identifying data-level factors to analyzing how format bias manifests within the internal processing of LLMs. Our aim is to examine whether differences in attention allocation across input formats are associated with the presence and direction of bias identified earlier. We focus our analysis on three representative LLMs from different model families: Qwen3-8B, Mistral-7B-Instruct-v0.3, and Llama-3.1-8B-Instruct.

**The Relation Between Attention Allocation and Presence of Bias** We begin by analyzing how attention is distributed between conflicting evidence inputs during inference. For each input pair, we compute the mean attention mass assigned to each segment and then calculate the absolute difference between the two values to quantify the degree of imbalance.

To quantify the negative correlation between attention gap and DCR, we employ Spearman's rank correlation coefficient as a measurement. The results show coefficients of –0.31, –0.37, and –0.54 across the three LLMs, indicating a weak to moderate negative correlation. This suggests that greater imbalance in attention distribution is associated with a lower likelihood of the model recognizing both sources of information.

This implies that format bias arises, at least in part, from skewed attention allocation early in processing, which increases the chance of one source being overlooked.

**The Relation Between Attention Allocation and Direction of Bias** We further examine whether attention allocation can explain which input LLMs tend to prefer when selecting only one side. Interestingly, in 82.35 percent of such cases, they favor the source that received less attention.

This observation implies that while attention imbalance affects whether both sources are represented in the output, it does not consistently determine the direction of preference. In other words, a segment receiving more attention does not necessarily have a higher chance of being selected as the final answer.

These results indicate that attention allocation is related to both types of bias, though the nature of this relationship may differ.

**Attention-Guided Intervention** To move from correlation to causation, we design an intervention that directly modifies the internal representations produced by the attention mechanism.

Specifically, we apply a normalization-based reweighting to the attention distribution at each generation step. Let $A \in \mathbb{R}^{H \times L_q \times L_k}$ denote the attention tensor, where $H$ is the number of heads, $L_q$ the query sequence length, and $L_k$ the key sequence length. At the current decoding step, we denote the attention distribution over keys as $a \in \mathbb{R}^{L_k}$, where $a_j$ represents the attention weight assigned to the $j$-th key token.

To ensure that the two pieces of evidence receive equal total attention, we define the corresponding token index sets $I_1$ and $I_2$, and compute the total attention mass on each:

$$m_1 = \sum_{j \in I_1} a_j, \quad m_2 = \sum_{j \in I_2} a_j \tag{3}$$

We then compute their average:

$$\bar{m} = \frac{m_1 + m_2 + \varepsilon}{2} \tag{4}$$

where $\varepsilon$ is a small constant for numerical stability. The attention weights are reweighted accordingly:

$$a'_j = \begin{cases} \frac{\bar{m}}{m_1+\varepsilon} \cdot a_j, & j \in I_1 \\ \frac{\bar{m}}{m_2+\varepsilon} \cdot a_j, & j \in I_2 \\ a_j, & \text{otherwise} \end{cases} \tag{5}$$

This procedure enforces equal total attention mass across the two evidence segments, while preserving the original intra-segment distribution.

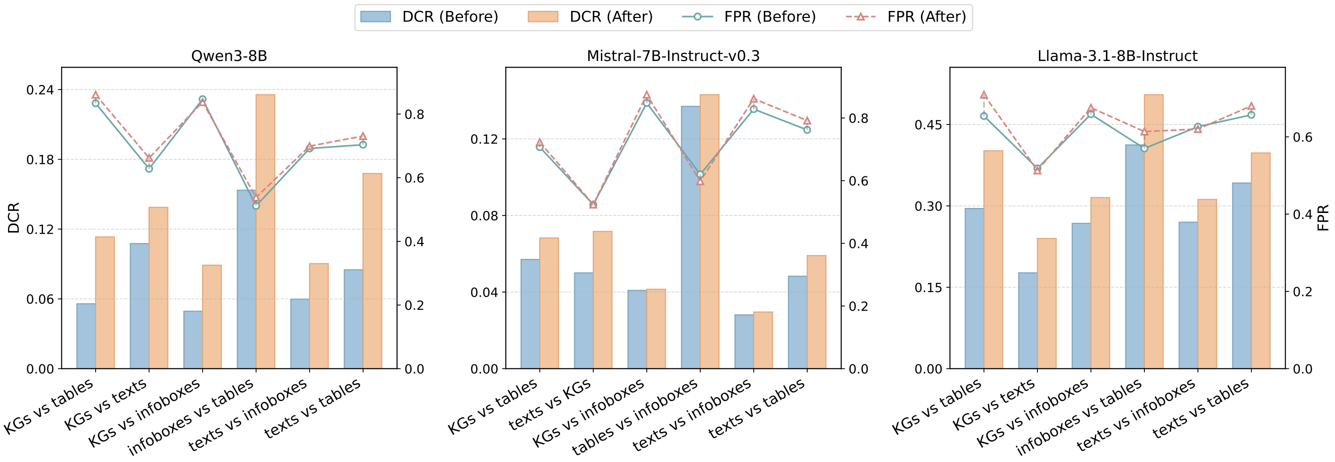


Figure 7: Effects of intervention methods in terms of DCR and FPR.

**Experiment Results** We apply the attention-balancing intervention to three representative LLMs and compare their behavior before and after modification.

The results reveal a consistent improvement in the models' ability to attend to both conflicting sources: DCR significantly increases across all format pairs and all three models (see Figure 7). This suggests that enforcing a more balanced attention distribution during inference effectively encourages the model to integrate information from both input segments, rather than overlooking one entirely. When applied to heterogeneous-input reasoning tasks, the attention-balancing method also improves performance on downstream RAG-style QA datasets, including HotpotQA and MuSiQue, where accuracy increases by 6.5% and 9.5%, respectively.

In contrast, FPR remains largely stable after the intervention, and the changes are not statistically significant across the evaluated models (see Figure 7). This implies that although the LLMs process both sources more evenly, the intervention has limited effect on their final output preferences once a directional bias has emerged.

The different effects of attention interventions on DCR and FPR suggest that presence of biases are at least partially controllable at inference time, whereas direction of biases are more resistant to modification. These more stable preferences may reflect deeper inductive biases acquired during pretraining, such as stronger alignment with textual inputs.

## 5 Discussion

Our findings show that format bias in LLMs is systematic and has practical consequences for systems processing heterogeneous data. To mitigate it, interventions can be applied at three levels: pre-processing, inference, and model development.

**Data Pre-processing** Pre-processing is a cost-effective way to reduce bias. Since LLMs tend to trust well-structured inputs, automatically repairing corrupted format in tables or KGs can prevent valuable content from being dismissed. Additionally, using a consistent input format can help reduce format-induced bias, as LLMs perform better when differences in input format is minimized. During the transformation process, it is important to preserve all essential information while avoiding the introduction of noise or unintended alterations.

**Inference-Time Intervention** Re-balancing attention across inputs during inference can help LLMs more effectively integrate information from multiple sources, reducing the tendency to overlook less-preferred formats. By encouraging the model to distribute attention more evenly across heterogeneous data, it enhances the model's ability to incorporate information from all inputs and supports more robust and balanced reasoning. Although such adjustments may not fully shift the model's final output bias, they improve its intermediate processing. Future work may explore more fine-grained or deeper intervention strategies to better align model attention with content relevance and reduce structural bias during inference.

**Model Development and Fine-tuning** Format preferences likely originate from pretraining data imbalance. Mitigation may involve training on format-balanced corpora, using contrastive objectives, or incorporating format-aware modules. Though costlier than inference-time methods, such approaches offer a more fundamental solution.

## 6 Conclusion

This work presents the first in-depth study of format bias in LLMs, identifying it as a consistent effect driven by information richness, structure quality, and format type. The bias manifests in two types: the presence of bias that can be mitigated, and the direction of bias likely rooted in pretraining. These findings underscore the importance of treating data format as a core factor in LLM design and evaluation. We also outline practical mitigation strategies, including data preprocessing, inference-time attention adjustment, and format-aware training, which together offer clear paths for reducing format bias in heterogeneous reasoning.

## Acknowledgments

We sincerely thank the anonymous reviewers for their insightful comments and helpful suggestions, which helped improve this work.

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

## References

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. *arXiv preprint arXiv:2303.08774*.

Biswal, A.; Patel, L.; Jha, S.; Kamsetty, A.; Liu, S.; Gonzalez, J. E.; Guestrin, C.; and Zaharia, M. 2024. Text2sql is not enough: Unifying ai and databases with tag. *arXiv preprint arXiv:2408.14717*.

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.

Christmann, P.; Saha Roy, R.; and Weikum, G. 2024. Compmix: A benchmark for heterogeneous question answering. In *Companion Proceedings of the ACM Web Conference 2024*, 1091–1094.

Christmann, P.; and Weikum, G. 2024. Rag-based question answering over heterogeneous data and text. *arXiv preprint arXiv:2412.07420*.

Fang, T.; Wang, Z.; Zhou, W.; Zhang, H.; Song, Y.; and Chen, M. 2023. Getting sick after seeing a doctor? diagnosing and mitigating knowledge conflicts in event temporal reasoning. *arXiv preprint arXiv:2305.14970*.

Gao, Y.; Xiong, Y.; Gao, X.; Jia, K.; Pan, J.; Bi, Y.; Dai, Y.; Sun, J.; Wang, H.; and Wang, H. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2(1).

Gekhman, Z.; Yona, G.; Aharoni, R.; Eyal, M.; Feder, A.; Reichart, R.; and Herzig, J. 2024. Does fine-tuning LLMs on new knowledge encourage hallucinations? *arXiv preprint arXiv:2405.05904*.

GLM, T.; :; Zeng, A.; Xu, B.; Wang, B.; Zhang, C.; Yin, D.; Zhang, D.; Rojas, D.; Feng, G.; Zhao, H.; Lai, H.; Yu, H.; Wang, H.; Sun, J.; Zhang, J.; Cheng, J.; Gui, J.; Tang, J.; Zhang, J.; Sun, J.; Li, J.; Zhao, L.; Wu, L.; Zhong, L.; Liu, M.; Huang, M.; Zhang, P.; Zheng, Q.; Lu, R.; Duan, S.; Zhang, S.; Cao, S.; Yang, S.; Tam, W. L.; Zhao, W.; Liu, X.; Xia, X.; Zhang, X.; Gu, X.; Lv, X.; Liu, X.; Liu, X.; Yang, X.; Song, X.; Zhang, X.; An, Y.; Xu, Y.; Niu, Y.; Yang, Y.; Li, Y.; Bai, Y.; Dong, Y.; Qi, Z.; Wang, Z.; Yang, Z.; Du, Z.; Hou, Z.; and Wang, Z. 2024. ChatGLM: A Family of Large Language Models from GLM-130B to GLM-4 All Tools. arXiv:2406.12793.

Hou, Y.; Pascale, A.; Carnerero-Cano, J.; Tchrakian, T.; Marinescu, R.; Daly, E.; Padhi, I.; and Sattigeri, P. 2024. Wikicontradict: A benchmark for evaluating llms on real-world knowledge conflicts from wikipedia. *Advances in Neural Information Processing Systems*, 37: 109701–109747.

Huang, J.; and Chang, K. C.-C. 2022. Towards reasoning in large language models: A survey. *arXiv preprint arXiv:2212.10403*.

Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y. J.; Madotto, A.; and Fung, P. 2023. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12): 1–38.

Jin, Z.; Cao, P.; Chen, Y.; Liu, K.; Jiang, X.; Xu, J.; Li, Q.; and Zhao, J. 2024a. Tug-of-war between knowledge: Exploring and resolving knowledge conflicts in retrieval-augmented language models. *arXiv preprint arXiv:2402.14409*.

Jin, Z.; Cao, P.; Yuan, H.; Chen, Y.; Xu, J.; Li, H.; Jiang, X.; Liu, K.; and Zhao, J. 2024b. Cutting off the head ends the conflict: A mechanism for interpreting and mitigating knowledge conflicts in language models. *arXiv preprint arXiv:2402.18154*.

Ju, T.; Wang, B.; Fei, H.; Lee, M.-L.; Hsu, W.; Li, Y.; Wang, Q.; Cheng, P.; Wu, Z.; Zhang, Z.; et al. 2025. Investigating the Adaptive Robustness with Knowledge Conflicts in LLM-based Multi-Agent Systems. *arXiv preprint arXiv:2502.15153*.

Lehmann, J.; Bhandiwad, D.; Gattogi, P.; and 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.

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.

Min, D.; Xu, Z.; Qi, G.; Huang, L.; and You, C. 2024. UniHGKR: Unified Instruction-aware Heterogeneous Knowledge Retrievers. *arXiv preprint arXiv:2410.20163*.

Petroni, F.; Rocktäschel, T.; Lewis, P.; Bakhtin, A.; Wu, Y.; Miller, A. H.; and Riedel, S. 2019. Language models as knowledge bases? *arXiv preprint arXiv:1909.01066*.

Pham, Q. H.; Ngo, H.; Luu, A. T.; and Nguyen, D. Q. 2024. Who's Who: Large Language Models Meet Knowledge Conflicts in Practice. *arXiv preprint arXiv:2410.15737*.

Shaier, S.; Kobren, A.; and Ogren, P. 2024. Adaptive question answering: Enhancing language model proficiency for addressing knowledge conflicts with source citations. *arXiv preprint arXiv:2410.04241*.

Shi, D.; Jin, R.; Shen, T.; Dong, W.; Wu, X.; and Xiong, D. 2024. Ircan: Mitigating knowledge conflicts in llm generation via identifying and reweighting context-aware neurons. *Advances in Neural Information Processing Systems*, 37: 4997–5024.

Su, Z.; Zhang, J.; Qu, X.; Zhu, T.; Li, Y.; Sun, J.; Li, J.; Zhang, M.; and Cheng, Y. 2024. Conflictbank: A benchmark for evaluating the influence of knowledge conflicts in llm. *arXiv preprint arXiv:2408.12076*.

Tan, H.; Sun, F.; Yang, W.; Wang, Y.; Cao, Q.; and Cheng, X. 2024. Blinded by generated contexts: How language models merge generated and retrieved contexts when knowledge conflicts? *arXiv preprint arXiv:2401.11911*.

Team, G.; Riviere, M.; Pathak, S.; Sessa, P. G.; Hardin, C.; Bhupatiraju, S.; Hussenot, L.; Mesnard, T.; Shahriari, B.; Ramé, A.; et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Team, Q. 2025. Qwen3 Technical Report. arXiv:2505.09388.

Trivedi, H.; Balasubramanian, N.; Khot, T.; and Sabharwal, A. 2022. MuSiQue: Multihop Questions via Single-hop Question Composition. *Transactions of the Association for Computational Linguistics*, 10: 539–554.

Xia, Y.; Chen, J.; Zhan, Y.; Zhao, S.; Jiang, W.; Zhang, C.; Han, W.; Bai, B.; and Gao, J. 2025. ER-RAG: Enhance RAG with ER-Based Unified Modeling of Heterogeneous Data Sources. *arXiv preprint arXiv:2504.06271*.

Xie, J.; Zhang, K.; Chen, J.; Lou, R.; and Su, Y. 2023. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *The Twelfth International Conference on Learning Representations*.

Xu, R.; Qi, Z.; Guo, Z.; Wang, C.; Wang, H.; Zhang, Y.; and Xu, W. 2024. Knowledge conflicts for llms: A survey. *arXiv preprint arXiv:2403.08319*.

Yang, Z.; Qi, P.; Zhang, S.; Bengio, Y.; Cohen, W.; Salakhutdinov, R.; and Manning, C. D. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, 2369–2380.

Yuan, X.; Yang, Z.; Wang, Y.; Liu, S.; Zhao, J.; and Liu, K. 2024. Discerning and resolving knowledge conflicts through adaptive decoding with contextual information-entropy constraint. *arXiv preprint arXiv:2402.11893*.

Zeng, A.; Lv, X.; Zheng, Q.; Hou, Z.; Chen, B.; Xie, C.; Wang, C.; Yin, D.; Zeng, H.; Zhang, J.; et al. 2025. Glm-4.5: Agentic, reasoning, and coding (arc) foundation models. *arXiv preprint arXiv:2508.06471*.

Zhang, H. C.; Semnani, S. J.; Ghassemi, F.; Xu, J.; Liu, S.; and Lam, M. S. 2024. Spaghetti: Open-domain question answering from heterogeneous data sources with retrieval and semantic parsing. *arXiv preprint arXiv:2406.00562*.

Zhang, Y.; Ma, J.; Hou, Y.; Bai, X.; Chen, K.; Xiang, Y.; Yu, J.; and Zhang, M. 2025. Evaluating and Steering Modality Preferences in Multimodal Large Language Model. *arXiv preprint arXiv:2505.20977*.

Zhu, T.; Liu, Q.; Wang, F.; Tu, Z.; and Chen, M. 2024. Unraveling cross-modality knowledge conflicts in large vision-language models. *arXiv preprint arXiv:2410.03659*.