Tracing Belief-Driven Thoughts with Theory-of-Mind Agents: An Opinion Analysis Framework

Jintao Wen
School of Computer Science,
Wuhan University
Wuhan, Hubei, China
23jtwen@whu.edu.cn

Yunfeng Ning
School of Computer Science,
Wuhan University
Wuhan, Hubei, China
ningyunfeng@whu.edu.cn

Hankun Kang
School of Computer Science,
Wuhan University
Wuhan, Hubei, China
kanghankun@whu.edu.cn

Xin Miao
School of Computer Science,
Wuhan University
Wuhan, Hubei, China
miaoxin@whu.edu.cn

Tieyun Qian[*]
School of Computer Science,
Wuhan University
Wuhan, Hubei, China
Zhongguancun Academy
Beijing, China
qty@whu.edu.cn

Abstract

Opinion Analysis (OA) is a critical task in Natural Language Processing (NLP), aimed at identifying stance, sentiment, or hate speech toward specific targets in social media text. With the advancement of Large Language Models (LLMs), researchers have begun exploring LLM-based agent frameworks for OA. However, existing LLM-based agent frameworks rely on heuristic simulations and lack the capacity for emotional resonance, which is achieved by understanding others' internal thoughts, intentions, and beliefs during interactions—a cognitive ability rooted in Theory of Mind (ToM). To bridge this gap, we introduce OpinionToM, a multi-agent framework that shifts from heuristic simulation to ToM-driven cognitive reasoning. This shift requires a formal process capable of capturing key aspects of ToM, particularly the uncertainty each agent has about the hidden mental states of others. To this end, we model multi-agent reasoning in OA as a Social Partially Observable Markov Decision Process (Social POMDP). The Social POMDP is designed to track agents' continuous belief states by dynamically generating beliefs and adjusting corresponding weights based on observations. The weight adjustment mechanism draws inspiration from Bayesian inverse planning, leveraging LLMs as a computational backend to perform approximate probabilistic inference over competing beliefs, conditioned on the agents' received perceptions. We evaluate OpinionToM on six benchmarks across three distinct opinion analysis tasks, demonstrating significant performance improvements compared to baselines. We release the code and dataset at https://github.com/wenjt/OpinionToM.

CCS Concepts

• **Computing methodologies** → **Natural language processing**; • **Information systems** → *Sentiment analysis*; *Web and social media search.*

Keywords

Opinion Analysis, Theory-of-Mind, Multi-agent Collaboration

ACM Reference Format:
Jintao Wen, Yunfeng Ning, Hankun Kang, Xin Miao, and Tieyun Qian. 2026. Tracing Belief-Driven Thoughts with Theory-of-Mind Agents: An Opinion Analysis Framework. In *Proceedings of the ACM Web Conference 2026 (WWW '26), April 13–17, 2026, Dubai, United Arab Emirates.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3774904.3792659

1 Introduction

Opinion Analysis (OA), which encompasses stance detection [16, 21], aspect-level sentiment analysis (ABSA) [5], and hate speech detection (HSD) [43], aims to elucidate the stance, sentiment, or intent that users convey towards the specific target in social media discourse. Conventional OA methods have typically treated text as a static object containing opinion attributes. However, with Large Language Models (LLMs) demonstrating impressive capabilities in natural language understanding and logical reasoning [3, 27], defining OA as a reasoning task has emerged as a mainstream trend. This shift is driven by the observation that users often reveal their opinions implicitly through communicative behaviors, rather than stating them explicitly [14].

While LLM-based OA frameworks have exhibited considerable promise, their methodological foundations predominantly rely on heuristic simulations and static, rule-based paradigms. These systems often emulate social interactions through fixed role-infused prompts [14], predefined collaboration strategies [28], or fine-tuning on preference-oriented datasets [47]. Consequently, the agents' understanding of the mental states of other agents or users during interactions remains limited, still make systematic errors in complex scenarios [18]. This limitation can be attributed to the following two inherent challenges.

[*] Corresponding author.

WWW '26, Dubai, United Arab Emirates

ACM ISBN 979-8-4007-2307-0/2026/04
https://doi.org/10.1145/3774904.3792659

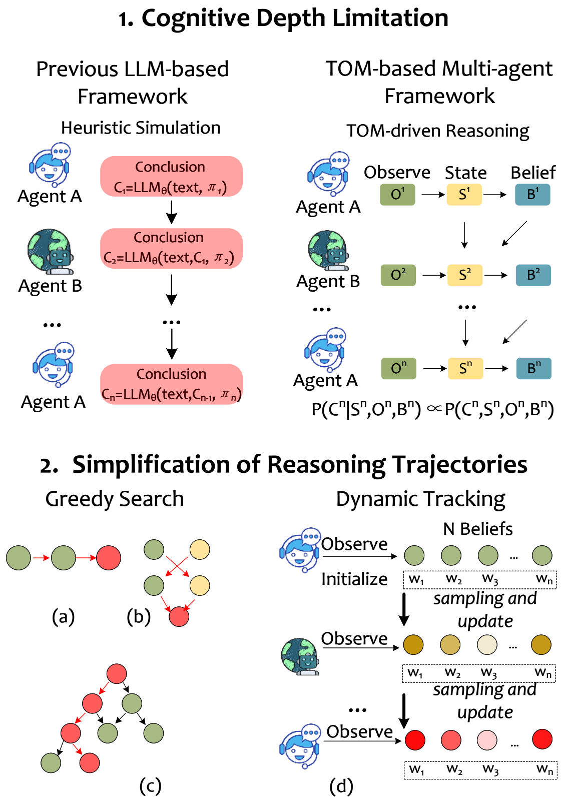

1. Cognitive Depth Limitation
Previous LLM-based Framework
Heuristic Simulation
Agent A
Conclusion C_1=LLM$_\theta$(text, π_1)
Agent B
Conclusion C_2=LLM$_\theta$(text,C_1, π_2)
...
Agent A
Conclusion C_n=LLM$_\theta$(text,C_{n-1}, π_n)
TOM-based Multi-agent Framework
TOM-driven Reasoning
Observe
State
Belief
Agent A
O^1 S^1 B^1
Agent B
O^2 S^2 B^2
...
Agent A
O^n S^n B^n
$P(C^n|S^n,O^n,B^n) \propto P(C^n,S^n,O^n,B^n)$
2. Simplification of Reasoning Trajectories
Greedy Search
Dynamic Tracking
(a)
(b)
(c)
N Beliefs
Observe
Initialize
w_1 w_2 w_3 ... w_n
sampling and update
Observe
sampling and update
...
Observe
(d)

Figure 1: Current LLM-based OA frameworks face two core challenges: 1. Cognitive Depth Limitation and 2. Simplification of Reasoning Trajectories. For cognitive depth limitation, we explicitly propose a multi-agent framework based on ToM-driven reasoning. For simplification of reasoning trajectories, we make an advance from existing (a) Sequential thinking, (b) Debating framework, and (c) Tree Search to (d) Sequential Monte Carlo reasoning.

(1) Cognitive Depth Limitation: Existing agent frameworks exhibit a fundamental cognitive gap: they reason rely solely on observed linguistic features or simplified sequential reasoning mechanisms, such as chain-of-thought (CoT) [39] or predefined rules [38], neglecting the hidden psychological dimensions (e.g., intentions, emotions, and beliefs) that underpin human interaction. This absence of explicit Theory of Mind (ToM) mechanisms [30, 35] prevents them from constructing internal models of others’ unspoken mental states. As a result, their reasoning tends to be superficial and brittle, especially when faced with ambiguous, conflicting, or strategically expressed viewpoints.

(2) Simplification of Reasoning Trajectories: Beyond the cognitive limitation, current collaborative reasoning paradigms (e.g., sequential thinking [15], debating frameworks [22, 28], or tree search [45]) suffer from simplifying opinion inference by focusing on the single most probable path or a final objective reward. Instead of dynamically tracking the full probability distribution of potential mental states, these greedy, single-path methods propagate limited information. This single-path propagation narrows the search space and hinders dynamic belief revision, making it difficult for agents to dynamically maintain multiple competing belief hypotheses—a capability essential for accurate mental state inference.

Figure 1 summarizes the two core challenges. To address them, we propose the first opinion analysis framework based on ToM. Theoretically, our framework is guided by the symbolic interactionism ToM [2] from social psychology. Technically, it enables a more flexible and comprehensive tracking of agents’ beliefs.

For cognitive depth limitation, we leverage the asymmetric nature of ToM—which involves inferring hidden mental states—to address reasoning uncertainties arising from insufficient cognitive depth. Inspired by Bayesian inverse planning, our framework, OpinionToM, explicitly instantiate ToM within a multi-agent architecture. Specifically, it employs an observation-belief collaboration mechanism grounded in the core tenets of ToM, as opposed to a predefined rule-based paradigm. Each agent assumes a distinct social role, observing, reasoning, and generating hypotheses at designated timesteps. This process allows them to progressively refine a shared situational understanding, thereby more accurately capturing evolving communicative intentions.

For simplification of reasoning trajectories, we formalize the OpinionToM reasoning process as a Social Partially Observable Markov Decision Process (Social POMDP) and model belief updates through Sequential Monte Carlo (SMC) inference [7]. In this formulation, the opinion inference unfolds dynamically: each agent continuously refines its belief distribution over possible mental states as new textual evidence emerges. Compared with conventional heuristic search or deterministic reasoning, SMC enables the exploration of richer mental-state hypotheses, thus better reflecting the uncertainty and diversity inherent in social opinion formation.

Comprehensive experiments on stance detection, ABSA, and hate speech detection tasks demonstrate that OpinionToM significantly outperforms existing LLM-based frameworks, achieving improved accuracy, interpretability, and cognitive plausibility.

Our contributions can be summarized as follows:

- We propose a novel multi-agent framework for social media opinion analysis that integrates Theory of Mind. During collaboration, our framework enables agents to dynamically adapt their own reasoning processes by inferring the beliefs of others.
- To simulate the ToM-based collaboration among agents, we formalize this process as a Social Partially Observable Markov Decision Process (Social POMDP). This formulation leverages Sequential Monte Carlo methods for more accurate modeling of mental state reasoning.
- Across six OA benchmarks, OpinionToM consistently outperforms existing baselines, demonstrating that ToM-driven cognitive reasoning can provide a powerful foundation for broader social and affective computing applications.

2 Related Work

Methods for Opinion Analysis. Previous studies on opinion analysis (OA) have primarily relied on textual features, such as keyword extraction [19], syntactic structures [5, 6, 46], or commonsense reasoning [34, 37]. Recently, LLM-based approaches employing specific prompting strategies or multi-agent frameworks have been developed to integrate extensive background knowledge, capture deeper textual features, and enhance reasoning robustness. For instance, KASD [16] retrieves topic-relevant Wikipedia documents to supplement background knowledge; COLA [14] leverages collaborative role-infused multi-agent framework to introduce multi-perspective knowledge; PREDICT [28] utilizes a multi-agent debating framework to achieve robust performance. FOL [40] leverages first-order logic to unify stance detection outputs from multiple LLMs, harnessing the logical rules derived from each.

Although existing LLM-based methods demonstrate impressive performance, they tend to exploit the LLMs' ToM solely implicitly. As a result, agents' understanding of other agents' or social media users' mental states remains unrobust, leading to suboptimal performance. To address this issue, we draw upon research on ToM and explicitly model agents' mental state representations to construct a novel multi-agent framework for OA tasks.

LLMs' Theory of Mind. ToM refers to the capacity to infer mental states that are not directly observable in others [13]. Recent studies have shown that LLMs can demonstrate a preliminary, human-like form of ToM by reasoning about beliefs, intentions, and perspectives [31]. To further investigate this capability, recent advances employ controlled role-playing scenarios [33] or multi-agent environments [18]. Beyond evaluation, recent frameworks explicitly enhance LLMs' ToM for social reasoning through various strategies—for example, symbolic prompting in SYMBOLICToM [32], multi-agent collaboration in MetaMind [42], and inverse planning in AutoToM [44]. ThoughtTracing [12] uses Monte Carlo simulation to trace the beliefs of an individual agent.

Despite their proficiency in ToM inference, these methods have largely neglected its potential for application in downstream tasks. To address this gap, we explore how ToM can enhance multi-agent collaboration to advance OA in social media scenarios.

3 Preliminary

3.1 Bayesian Inverse Planning for Theory of Mind

Bayesian Inverse Planning (BIP) serves as a computational foundation for model-based ToM inference. This approach posits that agents' observable behavior is generated by their latent mental states. Formally, BIP leverages a generative agent model (e.g., an LLM), which is defined as a Bayesian network that specifies the causal relationships from internal mental variables to observable actions. The task of ToM inference is then framed as an inverse problem: given observed behavior, BIP probabilistically infers the most likely latent mental states that produced it. This framework is formally defined by the following components:

- **Observable Variables (O_t)**: The set of states, actions, or utterances observable at time t, denoted as $O_t = \{o_i^t\}_{i \in N_O}$, where N_O denotes the index set of observable types, and o_i^t denotes the specific observed value of type i at time t. Their values are typically extracted from the problem context.
- **Latent Mental Variables (L_t)**: The set of unobservable mental states (e.g., goals, desires, and beliefs) at time t, denoted as $L_t = \{l_i^t\}_{i \in N_L}$. Here, N_L represents the index set of mental state types, and l_i^t represents the specific value of type i at time t.

The agent model is a Bayesian network defining the joint distribution $P(L_t, O_t)$. Given this model and the observations, BIP performs probabilistic inference to estimate the posterior distribution of the latent mental variables at step t:

$$P(L_t \mid O_{1:t}) \propto P(O_t \mid L_t) P(L_t) \propto P(O_t, L_t). \tag{1}$$

3.2 Sequential Monte Carlo

Sequential Monte Carlo (SMC) methods, also known as particle filters, comprise a class of algorithms designed for incremental inference on a sequence of posterior distributions. SMC methods are particularly well-suited for nonlinear, non-Gaussian state-space models and are capable of handling complex multimodal distributions, making them an ideal tool for inferring time-varying mental states. The core of SMC methods is a recursive "predict-update" cycle, which typically consists of the following three steps:

Sampling and Propagation. At timestep t, a new candidate state $\widetilde{L}_t^{(i)}$ is generated from each existing particle $L_{t-1}^{(i)}$ according to the system's state transition model:

$$\widetilde{L}_t^{(i)} \sim P(L_t \mid L_{t-1}^{(i)}) \quad \text{for } i = 1, \ldots, N \tag{2}$$

This forms a new set of candidate particles $\{\widetilde{L}_t^{(1)}, \ldots, \widetilde{L}_t^{(N)}\}$, representing the predictive prior belief before incorporating the new observation O_t.

Updating Weights. The importance of each candidate particle is evaluated based on the newly arrived observation O_t. The weight $w_t^{(i)}$ for each particle is computed via the observation model and is proportional to the likelihood of the observation data given the particle's state:

$$w_t^{(i)} \propto P(O_t \mid \widetilde{L}_t^{(i)}) \tag{3}$$

The weights are normalized such that $\sum_{i=1}^{N} w_t^{(i)} = 1$. The resulting weighted particle set $\{(\widetilde{L}_t^{(i)}, w_t^{(i)})\}_{i=1}^{N}$ constitutes a discrete approximation of the current posterior distribution $P(L_t \mid O_{1:t})$.

Resampling. To mitigate the particle degeneracy problem (only a few particles possess significant weights), the particles are resampled according to their weights. By drawing N times with replacement from the current weighted particle set, a new, unweighted particle set $\{L_t^{(1)}, \ldots, L_t^{(N)}\}$ is generated, where the probability of selecting each particle is equal to its normalized weight $w_t^{(i)}$. This allows the algorithm to discard low-weight trajectories and focus on high-likelihood regions of the state space. Finally, the resampled particle set can be used to approximate the desired marginal posterior distribution:

$$P(L_t \mid O_{1:t}) \approx \frac{1}{N} \sum_{i=1}^{N} \delta_{L_t^{(i)}}(L_t) \tag{4}$$

where δ is the Dirac delta function [9].

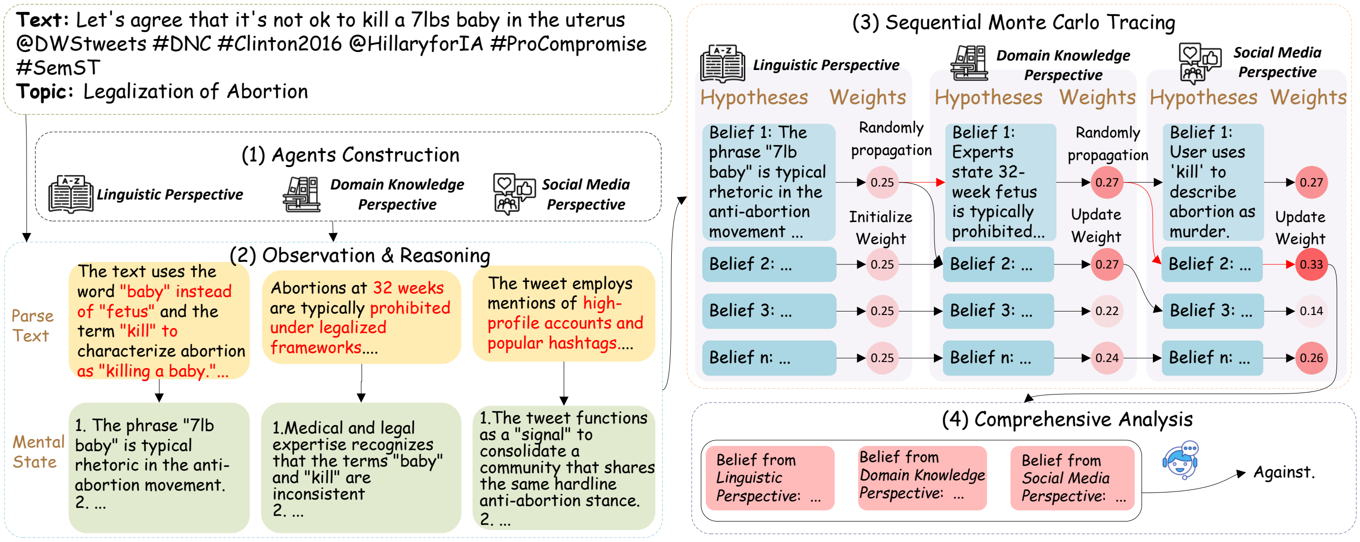

Text: Let's agree that it's not ok to kill a 7lbs baby in the uterus @DWStweets #DNC #Clinton2016 @HillaryforIA #ProCompromise #SemST
Topic: Legalization of Abortion
(1) Agents Construction
Linguistic Perspective
Domain Knowledge Perspective
Social Media Perspective
(2) Observation & Reasoning
Parse Text
The text uses the word "baby" instead of "fetus" and the term "kill" to characterize abortion as "killing a baby."...
Abortions at 32 weeks are typically prohibited under legalized frameworks....
The tweet employs mentions of high-profile accounts and popular hashtags....
Mental State
1. The phrase "7lb baby" is typical rhetoric in the anti-abortion movement. 2. ...
1.Medical and legal expertise recognizes that the terms "baby" and "kill" are inconsistent 2. ...
1.The tweet functions as a "signal" to consolidate a community that shares the same hardline anti-abortion stance. 2. ...
(3) Sequential Monte Carlo Tracing
Linguistic Perspective
Domain Knowledge Perspective
Social Media Perspective
Hypotheses Weights
Belief 1: The phrase "7lb baby" is typical rhetoric in the anti-abortion movement ...
Belief 2: ...
Belief 3: ...
Belief n: ...
Randomly propagation
Initialize Weight
0.25
0.25
0.25
0.25
Belief 1: Experts state 32-week fetus is typically prohibited...
Update Weight
0.27
0.27
0.22
0.24
Belief 1: User uses 'kill' to describe abortion as murder.
0.27
0.33
0.14
0.26
(4) Comprehensive Analysis
Belief from Linguistic Perspective: ...
Belief from Domain Knowledge Perspective: ...
Belief from Social Media Perspective: ...
Against.

Figure 2: An overview of OpinionToM framework, which consists of four crucial steps: Agents Construction, Observation & Reasoning, Sequential Monte Carlo Tracing, and Comprehensive Analysis.

Table 1: Opinion Analysis Task Definition: The overall task is decomposed into independent subtasks. For a given input text, each subtask is designed to address only its corresponding objective. These include Stance Detection (SD), Aspect-Based Sentiment Analysis (ABSA), and Hate Speech Detection (HSD).

OA	Input	Output	Evaluation Metrics
SD	Text x and Topic T	Support/Oppose/Neutral	Accuracy and macro-F1
ABSA	Text x and Target T	Positive/Negative/Neutral	Accuracy and macro-F1
HSD	Text x and Target T	Hate/Offensive/Normal	Accuracy and weighted-F1

4 Task Definition and an Overview of Our Framework

4.1 Task definition

Opinion Analysis (OA) aims at identifying the stance, sentiment, or intent toward specific targets t expressed in social media text x. Table 1 provides the formal task definition, including input/output specifications and evaluation metrics. We propose a novel multi-agent framework that enhances social reasoning by incorporating ToM into the agent interaction. Specifically, at each timestep, the agent first parses the text a^T and generates a mental state s^T, while simultaneously incorporating belief hypotheses from the previous timestep as prior conditions $\mathcal{P}^{T-1}$ to infer new belief hypotheses $\mathcal{P}^T_{(i)} \sim p_\theta\left(\mathcal{P}^T \mid \mathcal{P}^{T-1}_{(i)}, \{(s_\tau, a_\tau)\}_{\tau=1}^{T}, x\right)$. Through multiple iterations, we aim to obtain a plausible hypothesis trajectories that approximates the evolution of the user's expressed opinion o.

To achieve this, we first develop three agents f_1, f_2, and f_3 that can construct parsed text a^T, mental state s^T, and hypothesis $\mathcal{P}^T$ from an input x at each timestep. Next, we develop a function f_θ that propagates p^T and updates it for the next timestep. Finally, we correct the generated hypothesis trajectories $\mathcal{P}^{Final} = \{\mathcal{P}^1, \mathcal{P}^2, \dots\}$, and utilize a function $f_{analysis}$ to perform the opinion analysis.

In the subsequent section on the "Tracing Belief with Theory-of-Mind Agents Framework", the modules for *Agents Construction* include the functions f_1, f_2 and f_3 respectively. The modules for *Sequential Monte Carlo Tracing* implement the function f_θ.

4.2 The Overall Structure of Our Framework

Based on the task definition, we propose a framework for tracing belief-driven thoughts with Theory-of-Mind agents, named **OpinionToM**. We frame opinion analysis as a multi-timestep reasoning task, where at each step, an agent refines its beliefs by observing text and reasoning about the others' beliefs. As shown in Figure 2, it consists of the following steps.

Agents Construction. Opinion analysis requires synthesizing multiple perspectives. To this end, we construct three expert agents, each specialized in a distinct domain.

Observation and Reasoning. At each timestep, the expert parses the input from their specialized perspective and generates a parsed text a^t and mental state s^t. This output serves as the prerequisite for generating belief hypotheses.

Sequential Monte Carlo Tracing. Hypotheses from each timestep propagate subsequently to the next agent. To reflect the inherent uncertainty in social reasoning, we formalize this propagation as a Social Partially Observable Markov Decision Process (Social POMDP), which uses Sequential Monte Carlo methods to simulate the evolution of the hypothesis set.

Comprehensive Analysis. The analysis LLM processes the generated hypothetical trajectories alongside the original text. This synthesis produces a final analysis of the stance toward a specific topic or target.

5 Tracing Belief with Theory-of-Mind Agents Framework

5.1 Agents Construction

In our framework, we introduce three agents—a Linguist, a Domain Expert, and a Sociologist—to achieve a comprehensive analysis of the user's viewpoints.

The details are shown in Appendix A.

Linguist. The Linguist's primary role is to deconstruct the form and function of the text, revealing how linguistic choices subtly shape and convey viewpoints.

Domain Expert. The Domain Expert's primary role is to provide factual grounding. It excavates the underlying, implicit knowledge networks and social contexts.

Sociologist. The Sociologist's primary role is to analyze how content functions as a social token that is produced, disseminated, and interpreted within digital spaces.

5.2 Observation and Reasoning

Let f denote the set of agents. We require it to be capable of the following steps: at any time, given a text input x and a target t, the agent obtains a description $\{(s_t, a_t)\}_{t=1}^{T}$ representing the **parsed text** and the **mental state**. Here, $\{(s_t, a_t)\}_{t=1}^{T}$ records what the agent objectively observes and subjectively perceives from a certain perspective at that moment.

Algorithm 1: Sequential Monte Carlo Tracing

```
Input: Social media text $x_i$, target $t_i$, LLM for final analysis
       $f_{analysis}$, LLM-based agents set $f$, number of
       hypotheses $n_h$
Output: Opinion Label $\hat{y}_i$ and Hypotheses set $\mathcal{P}^{Final}$
while $0 < T \leq Size(f)$ do
    if $T = 0$ then
        $\mathcal{P}^T \leftarrow$ InitializeHypotheses$(M_T, x_i, t_i, n_h)$;
        $\mathcal{P}^T \leftarrow$ InitializeWeights$(\mathcal{P}^T)$;
        $\mathcal{P}^{Final} = \{\mathcal{P}^T\}$;
        continue;
    end
    $T \leftarrow T + 1$;
    $a_T \leftarrow$ ParseText$(f_T, x_i)$;
    $s_T \leftarrow$ GenerateMentalState$(f_T, a_T, x_i)$;
    $\mathcal{P}^T \leftarrow$ Propagate$(f_T, \mathcal{P}^{T-1}, a_T, s_T)$;
    $\mathcal{P}^T \leftarrow$ UpdateWeights$(\mathcal{P}^T, s_T)$;
    $\mathcal{P}^{Final} \leftarrow \mathcal{P}^{Final} \cup \{\mathcal{P}^T\}$
end
$\hat{y}_i \leftarrow$ AggregateFinalStance$(f_{analysis}, \mathcal{P}^{Final}, x_i, t_i)$;
return $(\hat{y}_i, \mathcal{P}^{Final})$;
```

5.3 Sequential Monte Carlo Tracing

If an agent has only partial observability of the state, the model becomes a Partially Observable Markov Decision Process (POMDP) [11]. Inspired by the POMDP, the agent receives partial observations o_t (i.e., the parsed text a_t and the mental state s_t), maintaining belief hypotheses $\mathcal{P}^t$ over possible states. Consequently, We formalize propagating and updating belief hypotheses as a **Social POMDP** and simulate it using Sequential Monte Carlo Tracing. Sequential Monte Carlo Tracing employs Sequential Monte Carlo principles to maintain a set of belief hypotheses about other agents' belief. These hypotheses are dynamically updated, and their weights are updated based on new observations derived from parsed text and mental states. The updates integrate the agent's current parse text, states and prior beliefs, as detailed in Algorithm 1.

Initial Hypothesis Generation. Sequential Monte Carlo Tracing begins by generating a set of N weighted hypotheses based on the agent's text parsing content and mental state at the first timestep. This hypothesis set reflects the agent's initial belief states. Formally, these hypotheses are drawn from an initial distribution: $p(\mathcal{P}^1 \mid s^1, a^1, x, t)$ where $\mathcal{P}^1$ represents the hypothesis, a^1 is the parsing text, s^1 denotes the initial state, x represents the input text, and t indicates the target. In practice, we first analyze the linguistic features using an LLM-based linguist agent. The agent takes (s^1, a^1, x, t) as input and generates N hypotheses, with each hypothesis assigned a uniform initial weight: $w_1^{(i)} = \frac{1}{N} \quad \text{for } i = 1, \ldots, N$ This uniform initialization ensures equal consideration of all hypotheses at the beginning of the reasoning process.

Propagate. At each timestep t, every hypothesis $\mathcal{P}_{(i)}^{t-1}$ from the previous step is propagated forward to generate an updated hypothesis $\mathcal{P}_{(i)}^{t}$. This propagation is conditioned on the full trajectory up to time t:

$$\mathcal{P}_{(i)}^{t} \sim p_\theta\left(\mathcal{P}^t \mid \mathcal{P}^{t-1}, \{(s_\tau, a_\tau)\}_{\tau=1}^{t}\right). \tag{5}$$

Here, θ denotes the parameters of the LLM, and $(s_\tau, a_\tau)_{\tau=1}^{t}$ represents the sequence of parsing-state pairs. The LLM serves as a computational backend, generating plausible mental state hypotheses conditioned on this sequence and the preceding belief hypotheses.

During propagation, hypotheses from $\mathcal{P}^{t-1}$ are randomly sampled according to their weights. The new hypotheses in $\mathcal{P}^t$ inherit weights from their corresponding parents.

UpdateWeights. Following propagation, each hypothesis $h_t^{(i)}$ undergoes a weight update based on its ability to explain the observed parsing text a^t or mental state s^t. Using inverse Bayesian reasoning, the LLM evaluates the likelihood of s_t under each hypothesis.

$$w_t^{(i)} := p_\theta\left(s^t \mid \mathcal{P}_{(i)}^{t}, \{(s_\tau, a_\tau)\}_{\tau=1}^{t-1}, a^t\right). \tag{6}$$

Empirically, we find that having the LLM select from five calibrated likelihood options—from "very likely (80%)" to "very unlikely (<20%)"—yields more robust performance and great stability. The updated weights are normalized to satisfy $\sum_{i=1}^{N} w_t^{(i)} = 1$.

5.4 Comprehensive Analysis

After traversing the entire trajectory, we aggregate all hypothesis trajectories while preserving their corresponding weights as hypothesis trajectories. Our final output retains all belief hypotheses $\mathcal{P}^{Final}$ with their associated weights, and synthesizes the textual content to form a unified stance judgment.

Table 2: The main results for in-target stance detection in Sem16 and P-stance. The best and second best results are in bold and underlined. * denotes the statistically significant improvements of the best results over the second best ones (t-test with a p-value < 0.01).

Method	P-stance				Sem16						
	Biden	Sanders	Trump	Avg	DT	HC	MF	LA	A	CC	Avg
JointCL	-	-	-	-	50.50	54.80	53.80	49.50	54.50	39.70	50.47
GPT-4o	62.90	80.00	71.50	71.47	62.50	68.70	44.70	51.50	9.10	31.10	44.60
GPT-4o$_{COT}$	84.08	80.12	82.24	82.15	64.16	78.69	73.22	**71.48**	65.15	34.00	47.10
KASD-BERT	79.04	75.09	70.84	74.99	54.74	64.78	57.13	51.63	55.97	40.11	54.06
KASD-GPT4o	83.60	79.66	84.31	82.52	64.23	80.32	70.41	62.71	63.95	55.83	66.24
COLA	86.60	**84.00**	79.70	83.43	71.20	75.90	69.10	71.00	62.30	64.00	68.92
FACTUAL	85.14	81.05	85.08	83.76	71.66	79.32	75.76	67.77	64.56	70.08	72.52
GPT-EDDA	-	-	-	-	69.50	80.10	69.20	62.70	67.20	68.50	69.50
Ours	**86.63***	80.67	**87.61***	**84.31***	**74.62***	**82.91***	**77.58***	71.35	**70.78***	**73.98***	**75.58***

Table 3: The main results for zero-shot and few-shot stance detection in VAST. The markers are the same as those in Table 2.

Model	VAST		
	Zero-Shot	Few-shot	Overall
GPT-4o	65.2	64.8	64.8
GPT-4o$_{COT}$	66.9	66.0	66.4
COLA	73.4	-	-
KASD-BERT	76.8	-	-
CKI	81.9	79.6	80.7
EDDA-LLaMA	76.3	-	-
Ours	**84.6***	**81.7***	**82.6***

6 Experiments

In this section, we describe the datasets, baselines, and experimental setup, and then present the results.

6.1 Experimental Setup

Datasets. Our experimental evaluation spans six benchmarks across three opinion analysis tasks: Stance Detection (SD), Hate Speech Detection (HSD), and Aspect-Based Sentiment Analysis (ABSA). The data partitions for these six datasets, detailing training and test set sizes, are summarized in Appendix B.

Implementation Details. For all agents in OpinionToM, we employ GPT-4o as the computational backend. The agent prompts are provided in Appendix C. For other agents that rely on the prompt design utilized in our framework, we also adopt GPT-4o [1]. To guarantee reproducibility, we set the temperature of the LLM to 0. The reported results represent the average over five independent runs, ensuring statistical reliability.

Baselines. We select the following classic and state-of-the-art baselines for each task.

We use two kinds of baselines, including fine-tuning methods based on BERT-like models, i.e., KASD-BERT [16], JointCL [21] and KEprompt [10], and LLM-based methods, i.e., GPT-4o, GPT-4o$_{CoT}$ [39], KASD-GPT-4o [16], COLA [14], GPT-EDDA [8], and FACTUAL [17].

Hate Speech Detection (HSD): We use two hate speech detection tools, i.e., PerspectiveAPI and PaiBERT [43]. We also use a fine-tuning method, i.e., TKE$_{BERT/RoBERTa}$ [23], which is based on BERT-like models. Besides, we include two LLM-based baselines, including GPT-4o and GPT-4o$_{CoT}$ [39] for HSD.

Aspect-based Sentiment Analysis (ABSA): For the fine-tuning methods based on BERT-like models, we use R-GAT [36], APARN [24] and CEIB [4]. For LLM-based baselines, we use GPT-4o and GPT-4o$_{CoT}$ [39] as before.

Metrics. For SD and ABSA, we calculate the average *macro-F1* score and accuracy (*Acc.*). For HSD, we calculate the average *weighted-F1* score [41] and accuracy (*Acc.*).

6.2 Main Results

The experimental results for Stance Detection (SD), Hate Speech Detection (HSD), and Aspect-Based Sentiment Analysis (ABSA) are summarized in Tables 2~5, respectively. Overall, the results indicate that our framework consistently surpasses all baselines across each opinion analysis task, demonstrating strong generalizability and task adaptability. A detailed analysis of the results for each subtask is provided below.

Stance Detection (SD). Our framework demonstrates superior performance over state-of-the-art methods on both in-target datasets and zero/few-shot datasets. For instance, on the SemEval-2016 dataset, our approach achieves significant improvements across diverse topics—from real-world subjects like "Hillary Clinton (HC)"

[1] To ensure a fair comparison, we use the same `gpt4o-2024-08-06` for all baselines and our method.

to abstract themes such as "Atheism (A)" and "Climate Change is a Real Concern (CC)". Compared to the second-best method, our framework elevates macro-F1 scores by approximately 6%, 2%, and 3% on these topics, respectively. These results substantiate that our architecture, through its collaborative reasoning empowered by Theory of Mind, enables deeper comprehension of social issues. Collectively, the experimental evidence demonstrates its overall effectiveness.

Table 4: The main results for hate speech detection. The markers are the same as those in Table 2.

Method	HateXPlain_2		HateXPlain_3		Avg.	
	F1	Acc	F1	Acc	F1	Acc
PerspectiveAPI	67.50	67.40	-	-	-	-
PaiBERT	64.00	63.90	-	-	-	-
TKE$_{\text{BERT}}$	78.20	78.80	64.00	64.00	71.10	71.40
TKE$_{\text{RoBERTa}}$	78.60	78.90	63.60	62.80	71.10	70.85
GPT-4o	73.40	73.40	39.30	40.00	56.35	56.70
GPT-4o$_{\text{COT}}$	77.50	77.60	43.80	45.00	60.65	61.30
Ours	**80.10***	**81.30***	**65.00***	**65.30***	**72.56***	**73.29***

Hate Speech Detection (HSD). For the HSD task, we conduct two subtasks: detecting whether the speech is toxic and classifying the types of toxicity, which is a much more challenging task. Our framework achieves the best results in both subtasks. Compared with the baselines directly using GPT-4o, our proposed framework improves the weighted-F1 score by approximately 2% in the binary classification task and 22% in the ternary classification task, demonstrating a substantial improvement that clearly shows the effectiveness of our framework in the HSD task.

Table 5: The main results for ABSA. The markers are the same as those in Table 2.

Method	Rest16		laptop14		Avg.	
	F1	Acc	F1	Acc	F1	Acc
R-GAT	76.62	89.71	74.07	78.21	75.35	83.96
APARN	82.44	87.76	79.10	81.96	78.96	80.91
CEIB	81.08	92.86	79.50	82.92	80.29	87.89
GPT-4o	68.75	87.18	74.70	77.37	71.73	82.28
GPT-4o$_{\text{COT}}$	69.12	88.96	77.37	79.70	73.25	84.33
Ours	**87.95***	**93.99***	**80.63***	**83.78***	**84.18***	**88.96***

Aspect-based Sentiment Analysis (ABSA). Our framework delivers strong gains on ABSA as well. On the Rest16 and Lap14 benchmarks, it surpasses SOTA baselines by 6% and 1% in macro-F1, respectively. These improvements on a well-studied task indicate that OpinionToM remains beneficial even when prior methods already operate near their performance ceiling.

Generalizability. Taken together, the experiments show that our framework not only achieves *superior performance* but also exhibits strong *cross-task generalization.*

(1) Most of the aforementioned baselines are designed for specific tasks. In contrast, our ToM-based opinion analysis framework can be readily deployed across different tasks and achieves SOTA performance.

(2) Although GPT-4o is powerful and can be applied to the three tasks above when used with simple prompt engineering, its performance remains relatively poor. This highlights the importance of incorporating ToM into agent collaboration for opinion analysis.

6.3 Ablation Study

To evaluate the individual contributions of the Theory of Mind (ToM) and Sequential Monte Carlo (SMC) modules, we perform ablation studies. As shown in Table 6, we systematically remove each module to assess its impact on overall performance.

Ablation on ToM. First, we simultaneously remove the entire ToM reasoning process and the SMC tracing, which reduces the entire model to the most naive collaborative approach. As shown in Table 6 under "*ToM* × *SMC* ×", we observe that after removing the user profile component, the model's performance drops to a level comparable to GPT-4o. When only ablating ToM while retaining SMC, the model's performance shows a significant improvement, but it still falls short of the full OpinionToM. This empirical finding strongly validates the indispensable role of ToM reasoning in opinion analysis, indicating that neglecting user perspectives significantly impairs the model's ability to analyze opinions.

Ablation on SMC. We remove the SMC process corresponding to "*ToM* ✓ *SMC* ×" in Table 6. After eliminating SMC, the performance of our framework decreases. By comparing "*ToM* ✓ *SMC* ×" with "our framework", we can observe that SMC can effectively enhance ToM-based belief propagation. Moreover, even without ToM, SMC still brings noticeable improvements to naive agent collaboration.

6.4 Deep Analysis

Hyperparameter Analysis. (1) We analyze the impact of timesteps and the number of belief hypotheses on the performance, as shown in Figure 3. It can be observed that when the number of hypotheses n=1, the Monte Carlo tracing degenerates into chain-of-thought reasoning. As the number of hypotheses increases, performance improves but eventually plateaus. (2) Our framework also demonstrates consistent performance across varying context window sizes, exhibiting strong robustness in handling texts of different lengths.

Performance Across Different Models. We evaluated the performance of our method on four LLMs with different architectures and parameters: Qwen2.5-7B, LLaMA2-7B, LLaMA2-14B, and LLaMA3-8B. The Table 7 demonstrates that our method achieves excellent performance across various base LLMs.

Computational Cost. We observe that from Figure 4: (1) Theory of Mind (ToM) reasoning elicit more extensive output representations compared to direct questions. This indicates that processing social information imposes higher computational demands on these models. (2) When answering ToM questions incorrectly, the models still generate a comparable number of output tokens (and sometimes even more in certain models). (3) Our approach consumes fewer tokens compared to current rule-based agent.

Table 6: The results for ablation study in opinion analysis.

Methods		Stance Detection						Hate Speech Detection				ABSA			
		Sem16		P-stance		VAST		HateXplain_2		HateXplain_3		Rest16		Lap14	
ToM	SMC	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc
×	×	52.70	51.74	72.26	71.29	66.33	62.96	74.30	73.91	45.22	45.13	68.98	87.20	74.98	77.10
×	✓	63.05	63.56	81.33	81.82	77.22	74.12	74.71	75.85	55.90	56.38	79.75	88.53	75.10	79.10
✓	×	72.31	71.55	82.96	81.89	79.26	79.88	75.81	76.32	63.20	62.92	86.13	90.12	78.01	79.73
✓	✓	75.58	76.42	84.31	82.98	82.60	83.23	80.10	81.30	65.00	65.30	87.95	93.99	80.63	83.78

Table 7: The results for Performance Across Different Models.

LLM	Stance Detection						Hate Speech Detection				ABSA			
	Sem16		P-stance		VAST		HateXplain_2		HateXplain_3		Rest16		Lap14	
	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc
LLaMa2-7B	71.80	70.50	79.60	78.40	77.10	76.80	73.60	76.50	61.20	61.00	82.70	88.30	75.70	78.70
LLaMa2-14B	73.50	72.10	81.50	80.20	78.80	78.50	75.20	78.30	62.80	62.50	84.50	90.20	77.50	80.50
LLaMa3-8B	75.92	74.72	84.63	83.52	82.06	81.93	77.90	81.20	65.00	65.20	87.88	93.99	80.43	83.82
Qwen2.5-7B	75.32	74.22	84.13	83.82	81.56	82.43	78.40	80.70	64.50	65.70	87.38	94.49	80.93	83.32

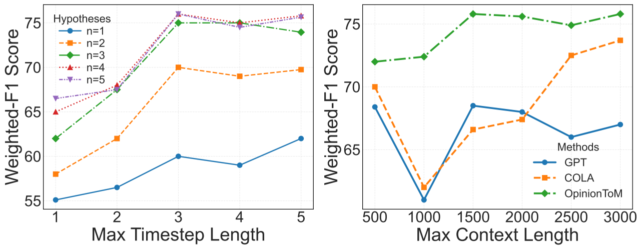


Figure 3: Hyperparameter analysis. Performance (%) of OpinionToM with different hypothesis number, max timestep length, and max context length.

Ordering of Agents. The order of agent collaboration follows the "from shallow to deep" principle (text⟶domain knowledge⟶social). We conduct experiments with different sequences, as shown in Appendix D. The results indicate that the reasoning order has only a minor impact on performance, and that different reasoning orders consistently outperform the direct aggregation and debate-based methods.

7 Conclusion

In this work, we identified two critical challenges in existing LLM-based agent frameworks for Opinion Analysis (OA): cognitive depth limitation and simplification of reasoning trajectories. To address these gaps, we proposed OpinionToM, a novel multi-agent framework that shifts from heuristic simulation to Theory of Mind (ToM)-driven cognitive reasoning. By formally modeling multi-agent interaction as a Social Partially Observable Markov Decision Process (Social POMDP), our framework dynamically tracks and updates belief states about agents' hidden mental states, with weight adjustments inspired by Bayesian inverse planning. Extensive experiments across six benchmarks and three opinion tasks demonstrate that OpinionToM achieves significant performance improvements over strong baselines, validating the effectiveness of integrating ToM principles with probabilistic reasoning for enhanced social understanding in NLP systems.

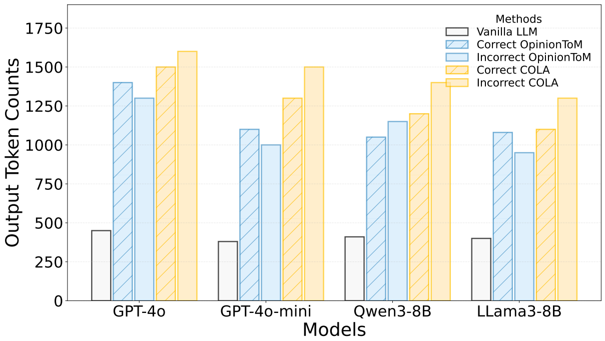


Figure 4: The average token count in reasoning trajectories of Vanilla LLM, OpinionToM, and COLA across different LLMs.

Acknowledgments

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

References

[1] Emily Allaway and Kathleen McKeown. 2020. Zero-shot stance detection: A dataset and model using generalized topic representations. *arXiv preprint arXiv:2010.03640* (2020).

[2] Ian Apperly. 2010. *Mindreaders: the cognitive basis of" theory of mind"*. Psychology Press.

[3] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems* (Vancouver, BC, Canada) *(NIPS '20)*. Curran Associates Inc., Red Hook, NY, USA, Article 159, 25 pages.

[4] Mingshan Chang, Min Yang, Qingshan Jiang, and Ruifeng Xu. 2024. Counterfactual-Enhanced Information Bottleneck for Aspect-Based Sentiment Analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 17736–17744.

[5] Bingfeng Chen, Qihan Ouyang, Yongqi Luo, Boyan Xu, Ruichu Cai, and Zhifeng Hao. 2024. S^{2}GSL: Incorporating Segment to Syntactic Enhanced Graph Structure Learning for Aspect-based Sentiment Analysis. *arXiv preprint arXiv:2406.02902* (2024).

[6] Hao Chen, Zepeng Zhai, Fangxiang Feng, Ruifan Li, and Xiaojie Wang. 2022. Enhanced multi-channel graph convolutional network for aspect sentiment triplet extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2974–2985.

[7] Pierre Del Moral, Arnaud Doucet, and Ajay Jasra. 2006. Sequential Monte Carlo Samplers. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 68, 3 (05 2006), 411–436. doi:10.1111/j.1467-9868.2006.00553.x

[8] Daijun Ding, Li Dong, Zhichao Huang, Guangning Xu, Xu Huang, Bo Liu, Liwen Jing, and Bowen Zhang. 2024. EDDA: An Encoder-Decoder Data Augmentation Framework for Zero-Shot Stance Detection. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. 5484–5494.

[9] Sadri Hassani. 2000. *Dirac Delta Function*. Springer New York, New York, NY, 289–319. doi:10.1007/978-0-387-21562-4_7

[10] Hu Huang, Bowen Zhang, Yangyang Li, Baoquan Zhang, Yuxi Sun, Chuyao Luo, and Cheng Peng. 2023. Knowledge-enhanced prompt-tuning for stance detection. *ACM Transactions on Asian and Low-Resource Language Information Processing* 22, 6 (2023), 1–20.

[11] Leslie Pack Kaelbling, Michael L Littman, and Anthony R Cassandra. 1998. Planning and acting in partially observable stochastic domains. *Artificial intelligence* 101, 1-2 (1998), 99–134.

[12] Hyunwoo Kim, Melanie Sclar, Tan Zhi-Xuan, Lance Ying, Sydney Levine, Yang Liu, Joshua B Tenenbaum, and Yejin Choi. 2025. Hypothesis-driven theory-of-mind reasoning for large language models. *arXiv preprint arXiv:2502.11881* (2025).

[13] Michal Kosinski. 2024. Evaluating Large Language Models in Theory of Mind Tasks. arXiv:2302.02083 [cs.CL] doi:10.1073/pnas.2405460121

[14] Xiaochong Lan, Chen Gao, Depeng Jin, and Yong Li. 2024. Stance detection with collaborative role-infused llm-based agents. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 18. 891–903.

[15] Bin Lei, Yi Zhang, Shan Zuo, Ali Payani, and Caiwen Ding. 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*, Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (Eds.). http://papers.nips.cc/paper_files/paper/2024/hash/5fcedec09977357f32e8e0ec8957073b-Abstract-Conference.html

[16] Ang Li, Bin Liang, Jingqian Zhao, Bowen Zhang, Min Yang, and Ruifeng Xu. 2023. Stance detection on social media with background knowledge. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 15703–15717.

[17] Ang Li, Jingqian Zhao, Bin Liang, Lin Gui, Hui Wang, Xi Zeng, Kam-Fai Wong, and Ruifeng Xu. 2024. Mitigating Biases of Large Language Models in Stance Detection with Calibration. *arXiv preprint arXiv:2402.14296* (2024).

[18] Huao Li, Yu Chong, Simon Stepputtis, Joseph Campbell, Dana Hughes, Charles Lewis, and Katia Sycara. 2023. Theory of Mind for Multi-Agent Collaboration via Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 180–192. doi:10.18653/v1/2023.emnlp-main.13

[19] Yingjie Li, Krishna Garg, and Cornelia Caragea. 2023. A new direction in stance detection: Target-stance extraction in the wild. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 10071–10085.

[20] Yingjie Li, Tiberiu Sosea, Aditya Sawant, Ajith Jayaraman Nair, Diana Inkpen, and Cornelia Caragea. 2021. P-stance: A large dataset for stance detection in political domain. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. 2355–2365.

[21] Bin Liang, Qinglin Zhu, Xiang Li, Min Yang, Lin Gui, Yulan He, and Ruifeng Xu. 2022. JointCL: A Joint Contrastive Learning Framework for Zero-Shot Stance Detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, Dublin, Ireland, 81–91. doi:10.18653/v1/2022.acl-long.7

[22] Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 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*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 17889–17904. doi:10.18653/v1/2024.emnlp-main.992

[23] Junyu Lu, Bo Xu, Xiaokun Zhang, Changrong Min, Liang Yang, and Hongfei Lin. 2023. Facilitating Fine-grained Detection of Chinese Toxic Language: Hierarchical Taxonomy, Resources, and Benchmarks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 16235–16250.

[24] Fukun Ma, Xuming Hu, Aiwei Liu, Yawen Yang, Shuang Li, Philip S. Yu, and Lijie Wen. 2023. AMR-based Network for Aspect-based Sentiment Analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 322–337. doi:10.18653/v1/2023.acl-long.19

[25] Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 35. 14867–14875.

[26] Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*. 31–41.

[27] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative Agents: Interactive Simulacra of Human Behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (San Francisco, CA, USA) *(UIST '23)*. Association for Computing Machinery, New York, NY, USA, Article 2, 22 pages. doi:10.1145/3586183.3606763

[28] Someen Park, Jaehoon Kim, Seungwan Jin, Sohyun Park, and Kyungsik Han. 2024. PREDICT: Multi-Agent-based Debate Simulation for Generalized Hate Speech Detection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 20963–20987.

[29] Maria Pontiki, Dimitrios Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad Al-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, et al. 2016. Semeval-2016 task 5: Aspect based sentiment analysis. In *International workshop on semantic evaluation*. 19–30.

[30] David Premack and Guy Woodruff. 1978. Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences* 1, 4 (1978), 515–526. doi:10.1017/S0140525X00076512

[31] Maarten Sap, Ronan Le Bras, Daniel Fried, and Yejin Choi. 2022. Neural Theory-of-Mind? On the Limits of Social Intelligence in Large LMs. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 3762–3780. doi:10.18653/v1/2022.emnlp-main.248

[32] Melanie Sclar, Sachin Kumar, Peter West, Alane Suhr, Yejin Choi, and Yulia Tsvetkov. 2023. Minding Language Models' (Lack of) Theory of Mind: A Plug-and-Play Multi-Character Belief Tracker. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 13960–13980. doi:10.18653/v1/2023.acl-long.780

[33] Kazutoshi Shinoda, Nobukatsu Hojo, Kyosuke Nishida, Saki Mizuno, Keita Suzuki, Ryo Masumura, Hiroaki Sugiyama, and Kuniko Saito. 2025. ToMATO: Verbalizing the Mental States of Role-Playing LLMs for Benchmarking Theory of Mind. *Proceedings of the AAAI Conference on Artificial Intelligence* 39, 2 (Apr. 2025), 1520–1528. doi:10.1609/aaai.v39i2.32143

[34] Xinjie Sun, Kai Zhang, Qi Liu, Meikai Bao, and Yanjiang Chen. 2024. Harnessing domain insights: A prompt knowledge tuning method for aspect-based sentiment analysis. *Knowledge-Based Systems* 298 (2024), 111975.

[35] Hedwig Teglasi, Maryke H. Caputo, and Arianna L. Scott. 2022. Explicit and implicit theory of mind and social competence: A social information processing framework. *New Ideas in Psychology* 64 (2022), 100915. doi:10.1016/j.

newideapsych.2021.100915

[36] Kai Wang, Weizhou Shen, Yunyi Yang, Xiaojun Quan, and Rui Wang. 2020. Relational graph attention network for aspect-based sentiment analysis. *arXiv preprint arXiv:2004.12362* (2020).

[37] Zhunheng Wang, Xiaoyi Liu, Mengting Hu, Rui Ying, Ming Jiang, Jianfeng Wu, Yalan Xie, Hang Gao, and Renhong Cheng. 2024. ECoK: Emotional Commonsense Knowledge Graph for Mining Emotional Gold. In *Findings of the Association for Computational Linguistics ACL 2024*. 8055–8074.

[38] Bowen Zhang, Daijun Ding, Liwen Jing, and Hu Huang. 2025. A Logically Consistent Chain-of-Thought Approach for Stance Detection. arXiv:2312.16054 [cs.CL] https://arxiv.org/abs/2312.16054

[39] Bowen Zhang, Xianghua Fu, Daijun Ding, Hu Huang, Genan Dai, Nan Yin, Yangyang Li, and Liwen Jing. 2023. Investigating chain-of-thought with chatgpt for stance detection on social media. *arXiv preprint arXiv:2304.03087* (2023).

[40] Bowen Zhang, Jun Ma, Xianghua Fu, and Genan Dai. 2025. Logic Augmented Multi-Decision Fusion framework for stance detection on social media. *Information Fusion* (2025), 103214.

[41] Tianhua Zhang, Hongyin Luo, Yung-Sung Chuang, Wei Fang, Luc Gaitskell, Thomas Hartvigsen, Xixin Wu, Danny Fox, Helen Meng, and James Glass. 2023. Interpretable unified language checking. *arXiv preprint arXiv:2304.03728* (2023).

[42] Xuanming Zhang, Yuxuan Chen, Min-Hsuan Yeh, and Yixuan Li. 2025. MetaMind: Modeling Human Social Thoughts with Metacognitive Multi-Agent Systems. arXiv:2505.18943 [cs.CL] https://arxiv.org/abs/2505.18943

[43] Zhehao Zhang, Jiaao Chen, and Diyi Yang. 2023. Mitigating Biases in Hate Speech Detection from A Causal Perspective. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. 6610–6625.

[44] Zhining Zhang, Chuanyang Jin, Mung Yao Jia, and Tianmin Shu. 2025. AutoToM: Automated Bayesian Inverse Planning and Model Discovery for Open-ended Theory of Mind. In *Workshop on Reasoning and Planning for Large Language Models*. https://openreview.net/forum?id=EwdEdgCewj

[45] Andy Zhou, Kai Yan, Michal Shlapentokh-Rothman, Haohan Wang, and Yu-Xiong Wang. 2024. Language agent tree search unifies reasoning, acting, and planning in language models. In *Proceedings of the 41st International Conference on Machine Learning* (Vienna, Austria) *(ICML'24)*. JMLR.org, Article 2572, 23 pages.

[46] Shen Zhou and Tieyun Qian. 2023. On the strength of sequence labeling and generative models for aspect sentiment triplet extraction. In *Findings of the Association for Computational Linguistics: ACL 2023*. 12038–12050.

[47] İlker Gül, Rémi Lebret, and Karl Aberer. 2024. Stance Detection on Social Media with Fine-Tuned Large Language Models. arXiv:2404.12171 [cs.CL] https://arxiv.org/abs/2404.12171

Appendix

A Details of Agents Construction

In our framework OpinionToM, we introduce three agents—a Linguist, a Domain Expert, and a Sociologist—to analyze text from linguistic, domain-specific, and sociological dimensions, thereby achieving a comprehensive analysis of the user's viewpoints. The details are as follows:

Linguist. The Linguist's primary role is to deconstruct the form and function of the text, revealing how linguistic choices subtly shape and convey viewpoints.

1. Syntactic Structure and Information Flow: Analyzing how grammatical constructions (e.g., active/passive voice, clause embedding) guide information focus, establish causality, or obscure agency.

2. Discourse Coherence and Logical Relations: Examining how discourse markers (e.g., 'but', 'therefore') structure arguments, and signal contrast, concession, or reinforcement.

3. Rhetorical Strategies and Persuasive Mechanisms: Identifying rhetorical devices (e.g., metaphor, irony, parallelism) and explaining how they evoke emotion.

Domain Expert. The Domain Expert's primary role is to provide factual grounding. It excavates the underlying, implicit knowledge networks and social contexts, including:

1. Entity Relation and Network Construction: Identifying key entities and uncovering the historical, political, or social relationships among them.

2. Event Contextualization and Timeline Positioning: Situating mentioned events within a broader historical or political timeline and assessing their authenticity and significance.

3. Interpretation of Ideologies and Belief Systems: Parsing the religious, political, or cultural belief systems referenced in the text and understanding how they shape the viewpoint.

Sociologist. The Sociologist's primary role is to analyze how content functions as a social token that is produced, disseminated, and interpreted within digital spaces:

1. Community Engagement and Agenda-Setting: Parsing the evolution, convergence, and competition of hashtags.

2. Cultural Capital and Identity Signaling: Interpreting internet slang, inside jokes, and memes.

3. Emotional Contagion and Public Sentiment: Assessing the emotional tone of the text and its potential evolution through sharing chains.

B Datasets

Table 8: Statistics of six datasets in three subtasks.

Dataset	Training Set	Test Set
Sem16	4987	1867
P-stance	6846	2158
VAST	13477	3006
HateXplain	15383	1924
Lap14	2313	638
Rest16	1748	616

For the stance detection task, we employed three datasets. Sem16 [26] and P-stance [20] are typically used for in-target stance detection, while VAST [1] supports zero-shot stance detection. The Sem16 dataset contains six topics: Atheism (A), Climate Change (CC), Feminist Movement (FM), Hillary Clinton (HC), Donald Trump (DT), and Legalization of Abortion (LA). The P-stance dataset includes three topics: Joe Biden, Bernie Sanders, and Donald Trump. VAST comprises 4,986 distinct topics.

In hate speech detection, HateXplain [25] structures the test task as hate detection (binary classification: hate speech/normal) and hate type detection (three-way classification: hate speech/offensive/normal).

For ABSA, we selected Rest16 [29] and Lap14 [29], which are sentiment polarity datasets focusing on restaurant reviews and laptop reviews, respectively.

C Prompt

C.1 Hypotheses Initialization

```
"""
```

```
From a linguistic perspective, analyze the provided tweet, its parses, and state assessments to generate a numbered list of {n_hypotheses_str} hypotheses about the author's internal reasoning that led to the tweet's stance (favor/against/none) on the {self.target}
"""
```

C.2 Text Parsing

```
% The prompt used for Linguist Agent
"""
As a linguist, analyze the provided text by examining its linguistic features and their contribution to meaning. Address elements such as grammatical structure, tense and inflection, speech acts, rhetorical devices, and lexical choices. Do nothing else.
"""
```

```
% The prompt used for Domain Expert Agent
"""
Analyze the provided text by examining its key elements contained in the quote, such as characters, events, parties, religions, etc. Also explain their relationship with {target} (if exist). Do nothing else.
"""
```

```
% The prompt used for Sociologist agent
"""
Analyze the provided text by examining its user features and their contribution to meaning. Focus on the content, hashtags, Internet slang and colloquialisms, emotional tone, implied meaning, and so on. Do nothing else.
"""
```

C.3 Mental State Generation

```
% The prompt used for Linguist Agent
"""
As a linguist, describe the author's state of mind as evidenced by the language used in this tweet.
"""
```

```
% The prompt used for Domain Expert Agent
"""
As a social media veteran, describe the author's state of mind as evidenced by the language used in this tweet.
"""
```

```
% The prompt used for Sociologist agent
"""
As a {self.role}, describe the author's state of mind as evidenced by the language used in this tweet.
"""
```

C.4 Hypotheses Update

```
"""
<current_parse>
({agent_role})
{current_parse['parse']}
</current_parse>
<current_thought_state>
{current_thought['state']}
</current_thought_state>
Task: Generate a new hypothesis about the author's stance on {self.target} based on the above.
Only use information from the previous hypothesis, current parse, and current thought state.
Do nothing else.
"""
```

C.5 Weight Update (Inverse Bayesian)

```
"""
You distinguish probabilities for different hypotheses.
If hypothesis H directly supports observed state D, P(D|H) = 80-100%.
If H contradicts D, P(D|H) = 0-20%.
If H is neutral to D, P(D|H) = 40-60%.
Output ONLY a percentage (no extra text) and brief reasoning.

<hypothesis H>
{hyp.hypothesis}
</hypothesis H>
<observed state D>
{observed_state}
</observed state D>


Question: What is the probability P(D|H) that state D would be observed if hypothesis H is true?
Output a percentage (0%-100%) and explain your reasoning
"""
```

D Ordering of Agents

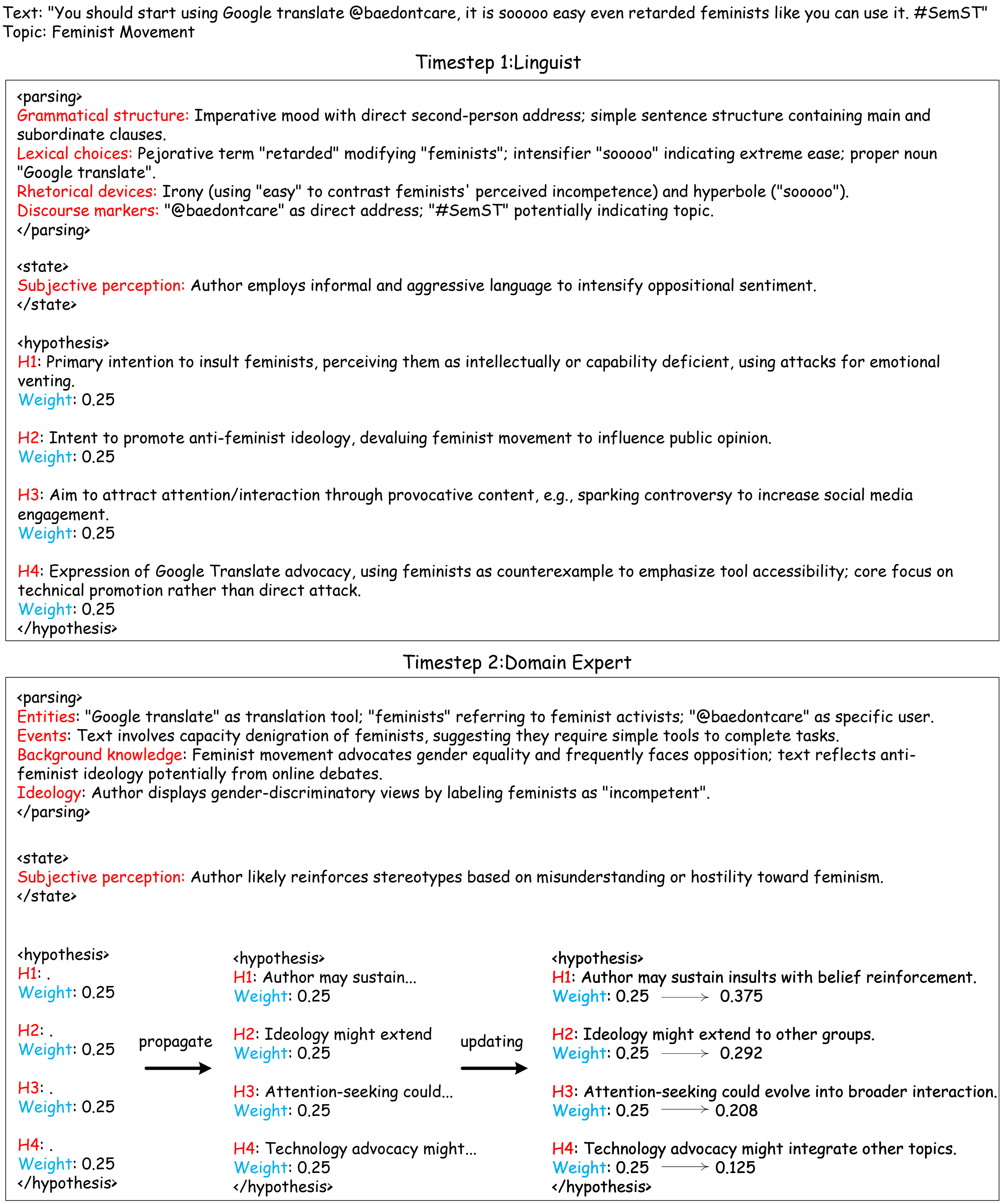
Table 9: Effect of reasoning order on performance.

Reasoning order	Stance	Sentiment	Hate Speech
Linguistic → Knowledge → Social Media	75.58	88.96	73.29
Linguistic → Social Media → Knowledge	74.95	88.28	72.15
Knowledge → Linguistic → Social Media	75.05	87.26	72.50
Knowledge → Social Media → Linguistic	73.83	87.77	71.30
Social Media → Knowledge → Linguistic	73.91	86.99	72.01
Social Media → Linguistic → Knowledge	73.77	87.29	71.85
Baselines			
Aggregation	62.44	64.71	64.00
Debate	68.92	72.56	66.25

We evaluate the impact of different reasoning orders on the results across three tasks.

E Sequential Monte Carlo Tracing Example

We demonstrate an example of Sequential Monte Carlo tracing within a single timestep.



Text: "You should start using Google translate @baedontcare, it is sooooo easy even retarded feminists like you can use it. #SemST"
Topic: Feminist Movement

Timestep 1:Linguist

<parsing>
Grammatical structure: Imperative mood with direct second-person address; simple sentence structure containing main and subordinate clauses.
Lexical choices: Pejorative term "retarded" modifying "feminists"; intensifier "sooooo" indicating extreme ease; proper noun "Google translate".
Rhetorical devices: Irony (using "easy" to contrast feminists' perceived incompetence) and hyperbole ("sooooo").
Discourse markers: "@baedontcare" as direct address; "#SemST" potentially indicating topic.
</parsing>

<state>
Subjective perception: Author employs informal and aggressive language to intensify oppositional sentiment.
</state>

<hypothesis>
H1: Primary intention to insult feminists, perceiving them as intellectually or capability deficient, using attacks for emotional venting.
Weight: 0.25

H2: Intent to promote anti-feminist ideology, devaluing feminist movement to influence public opinion.
Weight: 0.25

H3: Aim to attract attention/interaction through provocative content, e.g., sparking controversy to increase social media engagement.
Weight: 0.25

H4: Expression of Google Translate advocacy, using feminists as counterexample to emphasize tool accessibility; core focus on technical promotion rather than direct attack.
Weight: 0.25
</hypothesis>

Timestep 2:Domain Expert

<parsing>
Entities: "Google translate" as translation tool; "feminists" referring to feminist activists; "@baedontcare" as specific user.
Events: Text involves capacity denigration of feminists, suggesting they require simple tools to complete tasks.
Background knowledge: Feminist movement advocates gender equality and frequently faces opposition; text reflects anti-feminist ideology potentially from online debates.
Ideology: Author displays gender-discriminatory views by labeling feminists as "incompetent".
</parsing>

<state>
Subjective perception: Author likely reinforces stereotypes based on misunderstanding or hostility toward feminism.
</state>

<hypothesis>
H1: .
Weight: 0.25

H2: .
Weight: 0.25

H3: .
Weight: 0.25

H4: .
Weight: 0.25
</hypothesis>

propagate

<hypothesis>
H1: Author may sustain...
Weight: 0.25

H2: Ideology might extend
Weight: 0.25

H3: Attention-seeking could...
Weight: 0.25

H4: Technology advocacy might...
Weight: 0.25
</hypothesis>

updating

<hypothesis>
H1: Author may sustain insults with belief reinforcement.
Weight: 0.25 → 0.375

H2: Ideology might extend to other groups.
Weight: 0.25 → 0.292

H3: Attention-seeking could evolve into broader interaction.
Weight: 0.25 → 0.208

H4: Technology advocacy might integrate other topics.
Weight: 0.25 → 0.125
</hypothesis>

Figure 5: An example of Text Parsing, State Generation, and Sequential Monte Carlo Hypothesis Tracing.