MindStep: Structured Cognitive Prompting for Human Behavior Simulation

Anonymous ACL submission

Abstract

Large Language Models (LLMs) have shown great potential as simulators of human behavior in social sciences. However, existing approaches primarily rely on surface-level behavioral imitation and overlook the cognitive processes that give rise to human actions. This limits the fidelity and interpretability of LLM-based agents. In this paper, we propose **MindStep**, a three-stage cognitive prompting framework inspired by Theory of Mind (ToM) and Information Processing Theory. MindStep organizes the reasoning process into three sequential stages: Evocation, which anchors the model to a user-specific persona; Perception, which selectively extracts relevant contextual cues from the social environment; and Reaction, which infers actions conditioned on the internalized persona and perceived context. By explicitly modeling intermediate cognitive stages, MindStep provides a more interpretable and cognitively grounded framework for LLM-based social simulation. To comprehensively evaluate the effectiveness of MindStep in social simulation, we introduce **SocialAct**, a fine-grained benchmark for behavioral prediction across four real-world social events. Experiments on SocialAct show that MindStep consistently outperforms standard prompting baselines, achieving average gains of 14.0% to 18.6% across behavioral tasks. We further evaluate MindStep on the RecToM benchmark for mental-state reasoning, where it achieves average improvements of 9.21% over the 0-shot baseline on belief inference and up to 94.34% accuracy on desire inference. Our data and code is publicly available at: https://anonymous.4open.science/r/MindStep-9B30.

1 Introduction

Traditional social science research relies on empirical tools such as surveys, longitudinal studies, and fieldwork to study human behavior and social dynamics (Wang et al., 2017; Sigmund et al., 2010). Although effective, these methods are costly, difficult to scale, and often constrained by practical or ethical limitations. The emergence of Large Language Models (LLMs) has opened new opportunities for social simulation. By equipping LLM-based agents with diverse personas, researchers can simulate human behaviors ranging from survey responses (Cui et al., 2025) to opinion polarization (Mou et al., 2024) and group collaboration (Li et al., 2024).

Despite the growing use of LLMs in social simulation, existing approaches primarily rely on surface-level behavioral imitation. Recent studies have noted that most methods treat simulation as a simplified mapping from input profiles to output actions, without explicitly capturing the reasoning behind human behavior (Li et al., 2026a; Jiang et al., 2025). This behavior-driven paradigm has several limitations. First, the absence of explicit cognitive structure limits interpretability, making it difficult to explain why an agent produces a particular behavior. Second, in complex or long-context settings, standard prompting strategies such as Chain-of-Thought (CoT) frequently introduce noisy or unfaithful intermediate reasoning steps, thereby reducing simulation fidelity-the extent to which agents accurately reflect the responses of particular individuals or social groups in a given situation.(Xu et al., 2024; Shao et al., 2023). Third, LLM-based agents often converge toward overly homogenized behaviors, failing to capture the diversity and long-tail characteristics of real populations (Xie et al., 2026). Prior work further suggests that LLM-generated populations may not faithfully represent real human groups, potentially leading to biased or misleading conclusions (Bisbee et al., 2024).

To address these limitations, we propose MindStep, a cognitive prompting framework inspired by Theory of Mind (ToM) (Premack and Woodruff, 1978) and Information Processing Theory (Atkin-

son and Shiffrin, 1968). Drawing on findings that people anchor social reasoning on self-referential cognitive states (Todd and Tamir, 2024), and human responses emerge from multi-stage processing rather than direct input-output mapping (Baddeley and Hitch, 1974), MindStep organizes the simulation process into three sequential stages. It begins with *Evocation*, which anchors the model to a specific persona by activating background identity, ensuring that subsequent reasoning follows role-specific attributes rather than default model tendencies. Building on this foundation, *Perception* filters environmental information from the user's subjective perspective to extract situationally relevant cues. Finally, *Reaction* infers the behavioral outcomes conditioned on the internalized persona and perceived context. By structuring behavior generation as an explicit cognitive progression, MindStep provides interpretable reasoning traces while grounding responses in persona-specific attributes and contextual cues.

To comprehensively validate our framework, we evaluate MindStep on both behavioral prediction and mental-state inference tasks. For behavior prediction, we introduce SocialAct, a fine-grained benchmark based on real-world social events such as *BlackLivesMatter (BLM)*, *MeToo*, *COVID-19*, and *CaliforniaWildfire (CAL Fire)*, where models are required to infer a complete Action Triplet consisting of action type, action object, and action content, conditioned on user profiles, historical records, and current context. For mental-state inference, RecToM benchmark (Li et al., 2026b) is used to assess its generalized ToM capabilities beyond social media environments. Extensive evaluations demonstrate that MindStep consistently outperforms existing zero-shot and CoT baselines. Ablation studies and detailed qualitative analyses further reveal that persona anchoring and staged perception are critical for mitigating contextual noise. Our work shows that structuring LLM prompting into explicit cognitive stages is a promising direction for improving the fidelity and interpretability of social simulation without additional model training.

Our contributions are summarized as follows:

- We propose **MindStep**, a cognitive prompting framework that formulates social simulation as a cognitively-grounded multi-stage reasoning process.
- We introduce **SocialAct**, a fine-grained benchmark of 4,000 samples across four diverse real-world social events. Grounded in user profiles and temporally aligned social contexts, it formulates behavior prediction through structured Action Triplets, requiring models to infer behaviors along three explicit axes: action type, action object, and action content.
- Experimental results demonstrate that MindStep consistently outperforms standard 0-shot and CoT prompting across multiple LLMs and tasks, achieving average gains of 14.0% to 18.6% on SocialAct. Moreover, experiments on RecToM and additional qualitative analyses further demonstrate the generalization and interpretability of our proposed framework.

2 Related Work

LLMs in Social Simulation. Large Language Models (LLMs) have recently been adopted as simulators of human behavior in social science research. Prior work has used LLM-based agents to replicate psychological experiments (Cui et al., 2025), model social dynamics (Zhao et al., 2024), and predict responses to political or economic events (Tessler et al., 2024; Li et al., 2024). However, most existing approaches rely on direct profile-to-behavior mappings, emphasizing aggregate behavioral patterns or superficial persona consistency while overlooking the cognitive processes underlying human decisions. As a result, the internal mechanisms behind simulated behaviors remain difficult to interpret (Li et al., 2026a). Recent studies further suggest that LLM-generated populations often smooth out real-world heterogeneity and fail to capture important distributional characteristics of human groups (Chen et al., 2026). MindStep addresses these limitations by moving beyond output imitation toward structured cognitive prompting through staged reasoning.

Structured Reasoning and Theory of Mind. CoT prompting has shown strong performance in formal reasoning tasks, but its free-form reasoning process is often less reliable in behavioral simulation (Wei et al., 2022), where it may produce noisy or unfaithful reasoning traces(Shao et al., 2023). Existing structured reasoning methods, including hierarchical decomposition (Khot et al., 2023) and graph-based reasoning, improve traceability but are generally designed for broad reasoning tasks rather than social simulation. In parallel, ToM benchmarks such as Hi-ToM (Wu et al., 2023), OpenToM

Benchmark	Question Type	Quantity	Options	Answer Type	Input Context	Avg. Chars
SocialAct	BLM	1,000	4×3	Triplet	Profile, Memory, News, Twitter feed	3,796
	MeToo	1,000	4×3	Triplet	Profile, Memory, News, Twitter feed	3,807
	COVID-19	1,000	4×3	Triplet	Profile, Memory, News, Twitter feed	3,901
	CAL Fire	1,000	4×3	Triplet	Profile, Memory, News, Twitter feed	3,599
RecToM	Belief (Rec)	1,762	7	Single	Dialogue History	725
	Fine Intention (Rec)	2,205	10	Multi	Dialogue History	1,462
	Desire (Seek)	1,448	2	Single	Dialogue History	723
	Fine Intention (Seek)	2,205	16	Multi	Dialogue History	785

Table 1: Detailed statistical of SocialAct (ours) and RecToM subsets. Options denotes the choice space, where 4×3 representing the structured prediction of the Action Triplet. Chars count reflects the average length of the input context.

(Xu et al., 2024), and BigTOM (Gandhi et al., 2023) have been proposed to evaluate LLMs' abilities to reason about beliefs, intentions, and perspectives. However, most prior work treats ToM primarily as an evaluation target rather than a mechanism for improving simulation fidelity. MindStep builds on this line of work by integrating ToM into a structured cognitive prompting framework for social simulation.

3 Task Definition and Dataset Construction

3.1 Task Definition

We evaluate MindStep on two complementary tasks that together assess both behavioral simulation and mental-state inference. Table 1 summarizes the detailed statistics for each task.

Behavioral Prediction (SocialAct). We formulate behavior simulation as a user-conditioned structured prediction problem, evaluated on SocialAct, a fine-grained benchmark we newly introduce in this work (Section 3.2). Let $\mathcal{U}$ denote a user characterized by a profile $\mathcal{P}$ (including user's description and stance) and a memory $\mathcal{M}$ of k task-relevant historical tweets retrieved via BM25. At a given timestamp t, the user is exposed to a social context $\mathcal{C} = \{\mathcal{P}, \mathcal{M}, \mathcal{N}_t, \mathcal{F}_t, \mathcal{L}_t\}$, where $\mathcal{N}_t$ denotes timeline-aligned news content, $\mathcal{F}_t$ is the current Twitter feed providing three candidate tweets as options for A_{obj}, and $\mathcal{L}_t$ is the content list providing three candidate texts as options for A_{cont}.

The model is required to predict an *Action Triplet*:

$$(A_{\text{type}},\ A_{\text{obj}},\ A_{\text{cont}}) = f(\mathcal{C}), \tag{1}$$

where $A_{\text{type}} \in \{\textit{post}, \textit{retweet}, \textit{reply}, \textit{quote}\}$. Each dimension is formulated as a closed-set selection rather than open-ended generation.

Mental-State Reasoning (RecToM). We additionally evaluate MindStep on RecToM (Li et al., 2026b), a benchmark for Theory-of-Mind (ToM) reasoning in conversational recommender systems. Given a dialogue history $\mathcal{D} = \{u_1, u_2, \ldots, u_T\}$ between a recommender and a seeker, the model is required to infer latent mental states:

$$\hat{s} = g(\mathcal{D}, q), \tag{2}$$

where q denotes a theory-of-mind question and $\hat{s}$ is the predicted mental state. We evaluate on three cognitive dimensions across both roles: *belief* (recommender), assessing inference of the seeker's attitude toward a recommended item; *desire* (seeker), evaluating the seeker's latent inclination to engage with an item; and *fine-grained intention* (recommender and seeker), requiring identification of the precise communicative purpose from a large candidate set.

3.2 SocialAct Benchmark

Existing social simulation benchmarks often focus on isolated scenarios, limited action spaces, or synthetic interactions, which restricts their ability to model user behavior in real-world social contexts. To address this limitation, we introduce SocialAct, a fine-grained benchmark for user behavior simulation during major societal events. SocialAct is designed to capture three aspects of social behavior: (1) contextual decision making, where behaviors are shaped by ongoing social discourse; (2) persona consistency, where actions are grounded in long-term user traits and stances; and (3) sensitivity to socially polarized topics, including race and gender issues (see in Figure 1).

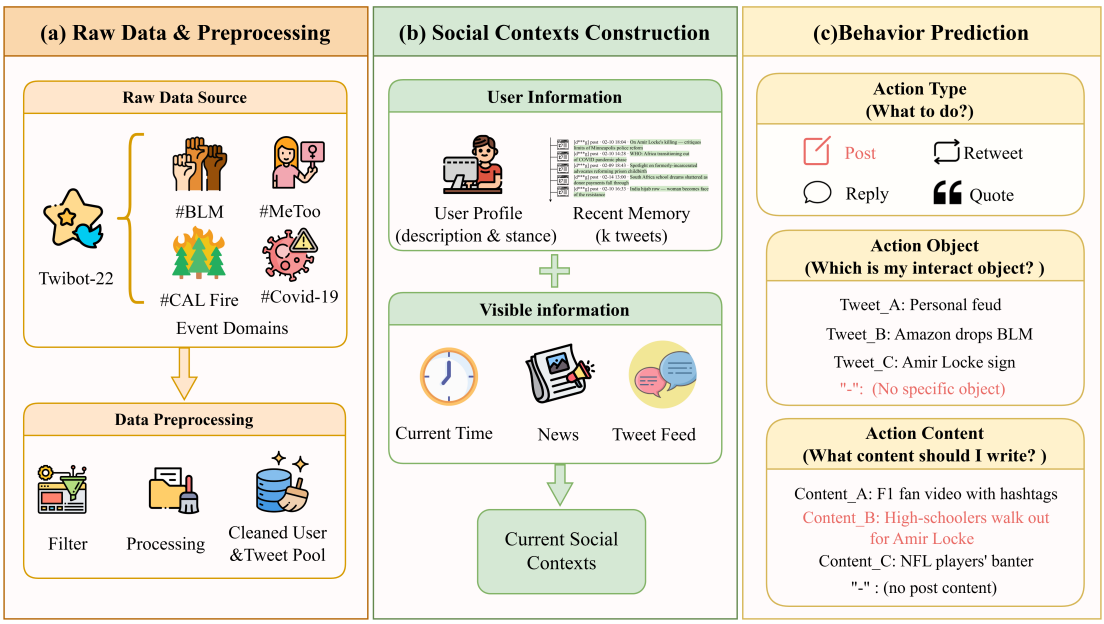

Figure 1: Overview of the SocialAct benchmark, illustrating the comprehensive process from dataset construction (a, b) to the structured behavior prediction task (c).

We construct SocialAct based on TwiBot-22 (Feng et al., 2022), a large-scale Twitter dataset. To identify event-related interactions, we filter the corpus using hashtags and keywords associated with four major societal events: #BlackLivesMatter (Giorgi et al., 2022), #MeToo (Maiorana et al., 2020), #COVID-19, and #CaliforniaWildfire (Jia et al., 2020). After removing bot accounts and users with insufficient activity history, the resulting corpus contains 51,273 action instances from 28,119 users. For each user, we retain profile attributes including user ID, follower and following counts, account creation time, location, interests, and historical tweets.

For each action instance in the preprocessed corpus, candidate options for A_{obj} and A_{cont} are sampled from the same event domain within a three-day temporal window to maintain contextual plausibility. To ensure balanced evaluation across domains, we then sample 1,000 action instances from each event, resulting in 4,000 test cases overall. Additional details on data processing and anonymization are provided in Appendix A.

4 MindStep: A Three-Stage Cognitive Prompting Framework

In this section, we introduce MindStep, a cognitively grounded prompting framework for social behavior simulation. MindStep organizes the generation process into three sequential stages: *Evocation*, *Perception*, and *Reaction* (see in Figure 2).

4.1 Evocation

The Evocation stage initializes the process by activating stable persona information from the profile description and historical behavior. At this stage, the language model only observes the user's profile and stance, and is prompted to produce step-by-step thoughts labeled as [Evocation-1], [Evocation-2], etc., summarizing the user's role, preferences, and expressive style before any situational context is introduced. This stage establishes a stable persona anchor and captures role-consistent habits, which helps reduce ambiguity and keep subsequent inference aligned with the user's profile.

4.2 Perception

The Perception stage mirrors selective attention in working memory. Given the current time, timeline-aligned news, and the visible feed, the model is prompted to generate thoughts labeled as [Perception-1], [Perception-2], etc., conditioned on the persona anchor from the previous stage. Rather than treating the entire context equally, it selects the facts, events, and interaction cues that are most relevant to the established role.

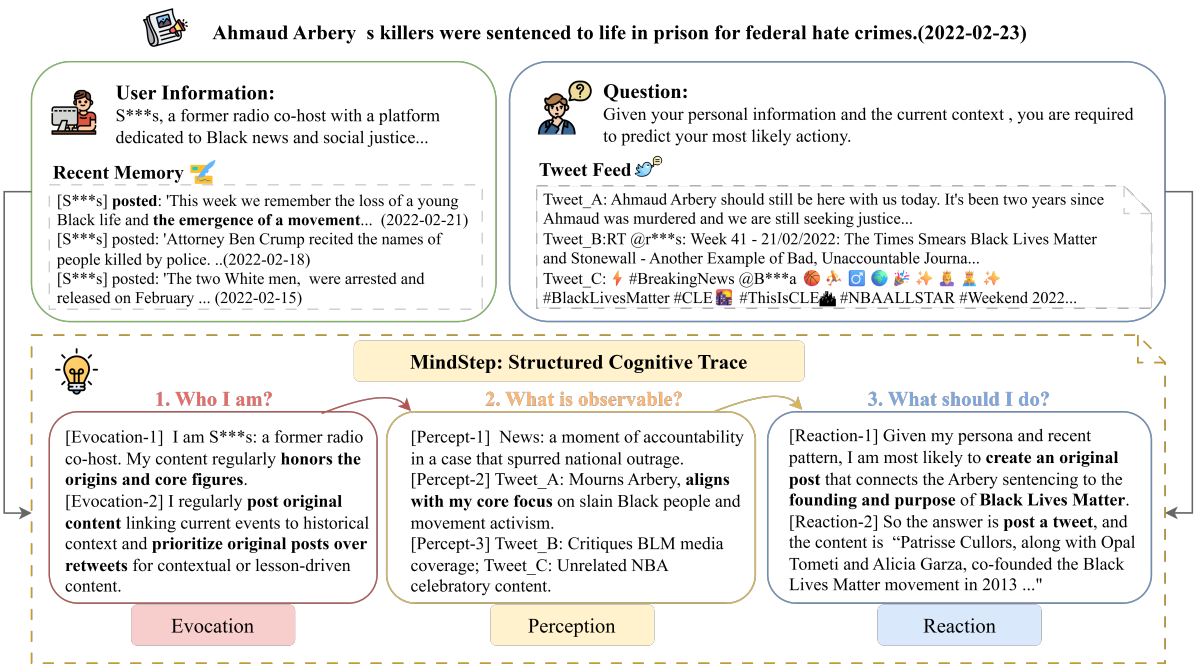

Figure 2: Overview of the MindStep framework. Given a user's profile, recent memory, and current social context, MindStep structures the reasoning process into three sequential stages: Evocation anchors the model to a user-specific persona, Perception selectively extracts situationally relevant cues, and Reaction infers the final action conditioned on the accumulated cognitive context.

4.3 Reaction

The Reaction stage combines the persona anchor with the selected percepts to infer the final action. The model generates thoughts labeled as [Reaction-1], [Reaction-2], etc., under instructions to prioritize persistent persona traits and behavioral habits over surface lexical matching in the environment. Following this trace, the model outputs the final answer in three target dimensions: $(A_{\text{type}}, A_{\text{obj}}, A_{\text{cont}})$.

5 Experiment

In this section, we conduct extensive experiments to evaluate MindStep. We first evaluate its primary performance on behavioral prediction with SocialAct, then examine its generalization ability on mental-state reasoning with RecToM. Finally, we analyze the contribution of each cognitive module through ablation studies.

5.1 Experimental Setup

Tasks and Metrics. We evaluate MindStep on two distinct benchmarks. Our primary task is behavioral prediction on the *SocialAct* benchmark, which requires models to predict a complete Action Triplet $(A_{\text{type}}, A_{\text{obj}}, A_{\text{cont}})$. Action-triplet accuracy is reported as the primary metric, where a prediction is considered correct only if the action type, action object, and action content are all correctly predicted. Macro-averaged accuracy, precision, recall, and F1-score are also calculated for each action element. Furthermore, we evaluate MindStep on *RecToM* for explicit mental-state reasoning across its belief, desire, and fine-grained intention subsets to examine the generalization capabilities of our prompting strategy. Since the benchmark contains both single-choice and multiple-choice questions, accuracy is reported as the primary metric.

Baselines and LLMs. Following established practices in the theory of mind literature (Wilf et al., 2024; Sabour et al., 2024), we compare MindStep against direct answering and standard CoT prompting in a zero-shot setting. Four LLMs are evaluated on SocialAct: Gemma-4-E4B-it, DeepSeek-v4-flash, GPT-5-mini, and Qwen3-max. GPT-4o-mini is additionally evaluated on RecToM. All prompt templates are provided in Appendix B.

Implementation Details. To ensure reproducibility, the generation temperature is set to 0.0 for SocialAct and 0.1 for RecToM. Open-source models are run on four RTX 5090 GPUs, while closed-source models are accessed through their respective APIs. The running time for one model on a single event benchmark is approximately three hours.

Method	BLM	MeToo	COVID-19	CAL Fire
0-shot				
Gemma-4-E4B-it	17.50	18.30	39.60	44.90
DeepSeek-v4-flash	40.80	43.30	63.70	57.80
GPT-5-mini	42.90	40.10	62.50	54.60
Qwen3-max	23.70	23.40	44.20	50.30
0-shot CoT				
Gemma-4-E4B-it	18.70	16.90	28.60	38.00
DeepSeek-v4-flash	26.80	30.40	48.50	45.00
GPT-5-mini	42.30	39.60	60.90	59.30
Qwen3-max	37.50	40.30	60.20	61.00
MindStep				
Gemma-4-E4B-it	24.90 (**+7.40, 6.20**)	22.70 (**+4.40, 5.80**)	41.00 (**+1.40, 12.40**)	49.50 (**+4.60, 11.50**)
DeepSeek-v4-flash	52.60 (**+11.80, 25.80**)	56.10 (**+12.80, 25.70**)	73.90 (**+10.20, 25.40**)	68.20 (**+10.40, 23.20**)
GPT-5-mini	55.30 (**+12.40, 13.00**)	52.70 (**+12.60, 13.10**)	70.90 (**+8.40, 10.00**)	67.10 (**+12.50, 7.80**)
Qwen3-max	64.80 (**+41.10, 27.30**)	68.00 (**+44.60, 27.70**)	80.10 (**+35.90, 19.90**)	80.10 (**+29.80, 19.10**)

Table 2: MindStep results on SocialAct across the BLM, MeToo, COVID-19, and CAL Fire events. All results are reported as action-triplet accuracy (%). Numbers in parentheses denote the absolute accuracy improvement of MindStep over the corresponding baselines (0-shot and 0-shot CoT).

5.2 Results

5.2.1 Overall Performance on SocialAct.

Table 2 reports model performance across four SocialAct event benchmarks. Overall, simulating human behavior remains challenging for standard methods, while our framework provides consistent improvements across models of different scales.

Performance Across Model Scales. Our approach improves accuracy across models of varying capacities. Qwen3-max achieves the highest performance, reaching 80.1% on both the COVID-19 and CAL Fire benchmarks. On more complex topics, such as BLM, Qwen3-max shows absolute gains of 41.1% over the 0-shot baseline and 27.3% over CoT, with similar improvements of 44.6% and 27.7% observed on MeToo. Smaller models, such as Gemma-4-E4B-it, also benefit from our method (e.g., +7.4% on BLM), though their absolute performance remains lower than that of larger models. GPT-5-mini demonstrates steady gains across all events, including +12.4% on BLM and +12.6% on MeToo.

Effects of Chain-of-Thought Prompting. We also observe that zero-shot Chain-of-Thought (0-shot CoT) prompting often reduces performance relative to direct zero-shot prompting. For example, applying 0-shot CoT to DeepSeek-v4-flash decreases accuracy from 40.8% to 26.8% on BLM, and from 43.3% to 30.4% on MeToo. This suggests that unconstrained reasoning chains may introduce noise in nuanced social simulation tasks. A qualitative analysis of this failure mode is provided in Appendix C.

Analysis by Event Type. Performance also varies substantially by event type. Baseline models consistently score lower on sociopolitical controversies (BLM, MeToo) than on public crises (COVID-19, CAL Fire), underscoring the inherent difficulty of simulating these complex events. Our framework produces the largest gains on sociopolitical controversies, with average improvements of 18.2% on BLM and 18.6% on MeToo, compared to 14.0% on COVID-19 and 14.3% on CAL Fire.

5.2.2 Action Element Analysis on SocialAct

Table 3 illustrates the average Accuracy and F1 scores for the three action elements, aggregated over four benchmark events.

Across models, action content consistently achieves the highest scores, with Accuracy and F1 closely aligned. In contrast, action type and action object exhibit lower F1 values compared to their accuracy, reflecting the challenge posed by less frequent behaviors in real-world social interactions.

Consistent with the overall results, MindStep improves performance on all three elements across different models. Under our method, Qwen3-max achieves the highest scores, reaching 83.0% accuracy and 52.8% F1 on action type, 81.0% accuracy and 64.7% F1 on action object, and 84.8% accuracy and 84.4% F1 on action content. DeepSeek-v4-flash achieves the second-highest results, and similar positive trends are observed for GPT-5-mini and Gemma-4-E4B-it. These results demonstrate that MindStep effectively enhances the prediction of each action element.

Method	Action Type				Action Object				Action Content			
	Acc.	Pre.	Rec.	F1	Acc.	Pre.	Rec.	F1	Acc.	Pre.	Rec.	F1
0-shot												
Gemma-4-E4B-it	42.63	38.91	37.51	29.73	36.48	36.81	43.92	30.78	44.80	57.97	47.39	45.20
DeepSeek-v4-flash	66.15	44.74	55.10	44.01	63.95	48.80	58.45	50.87	72.78	74.30	73.16	72.49
GPT-5-mini	57.33	42.37	61.68	40.71	55.50	45.20	60.11	46.45	71.20	72.42	71.41	70.89
Qwen3-max	42.80	41.37	47.43	33.63	40.33	41.18	55.44	38.70	57.65	68.48	59.86	58.50
0-shot CoT												
Gemma-4-E4B-it	36.35	37.68	35.57	27.82	35.58	37.94	50.12	32.41	50.25	55.13	51.15	50.51
DeepSeek-v4-flash	44.23	38.87	51.19	32.95	43.65	42.81	59.76	39.95	63.15	70.46	64.40	63.83
GPT-5-mini	58.78	42.81	55.61	41.52	56.25	45.49	60.20	46.95	71.45	73.05	72.03	71.25
Qwen3-max	57.90	46.26	51.99	41.92	57.43	45.73	60.21	45.75	70.98	72.09	71.29	70.68
MindStep												
Gemma-4-E4B-it	47.53	37.16	36.27	32.47	48.43	41.12	53.43	40.27	55.03	60.33	56.50	55.14
DeepSeek-v4-flash	72.23	48.15	57.65	48.11	70.48	54.04	**65.94**	57.64	78.70	79.14	79.17	78.35
GPT-5-mini	70.40	47.09	**62.62**	47.07	68.00	50.69	60.84	53.69	75.85	75.81	75.70	75.29
Qwen3-max	**82.95**	**60.78**	58.47	**52.76**	**80.98**	**65.09**	64.56	**64.71**	**84.78**	**84.37**	**84.51**	**84.36**

Table 3: Accuracy, Precision, Recall and F1 scores on action type, action object, and action content across models under 0-shot, 0-shot CoT, and MindStep, averaged across four benchmark events. Bold indicates the best result and underline indicates the second-best result.

5.2.3 Performance on RecToM

Table 4 reports the results on RecToM across the recommender and seeker roles, with random guess and human results included as reference baselines following the original benchmark setting.

Overall Performance. The empirical results demonstrate that MindStep yields consistent performance gains across all evaluated LLMs and conversational roles, with average improvements of 1.57–9.21% over the 0-shot baseline and 3.30–8.87% over the 0-shot CoT baseline across tasks. These improvements suggest that structured cognitive prompting may better support the integration of dialogue history and contextual social cues during mental-state reasoning.

Belief and Desire Inference. The largest improvements are observed in the belief and desire inference subtasks. For belief reasoning, MindStep yields consistent gains across all evaluated models, achieving average improvements of 9.21% over the 0-shot baseline and 4.78% over the 0-shot CoT baseline on recommender belief tasks. For desire inference, the improvements translate into strong absolute performance, with DeepSeek-v4-flash achieving 94.34% accuracy on the seeker desire subset and reducing the gap to the human baseline to within 4%.

Challenges in Fine-Grained Intention Inference. MindStep also improves performance on fine-grained intention tasks. DeepSeek-v4-flash achieves the highest accuracy of 35.83% and 33.15% for the recommender and seeker roles, respectively. However, absolute performance on these subsets remains relatively low across models. One possible explanation is that fine-grained intention inference requires distinguishing among multiple plausible and context-dependent alternatives, leading to a more challenging reasoning setting than belief or desire inference. These results suggest that accurately modeling nuanced communicative intentions remains difficult for current LLMs in Theory-of-Mind reasoning settings.

Method	Recommender		Seeker	
	Intention	Belief	Intention	Desire
Random Guess	0.10	14.29	0.00	50.00
Human	64.32	96.84	59.92	98.25
0-shot				
GPT-4o-mini	29.12	56.24	26.71	88.05
DeepSeek-v4-flash	31.52	66.74	32.11	83.70
GPT-5-mini	29.71	69.30	27.57	85.08
Qwen3-max	30.93	81.73	28.84	87.57
0-shot CoT				
GPT-4o-mini	25.71	62.49	29.66	88.33
DeepSeek-v4-flash	28.84	75.82	27.53	76.38
GPT-5-mini	29.07	68.05	24.54	87.02
Qwen3-max	28.75	**85.36**	28.03	83.29
MinStep				
GPT-4o-mini	28.30	70.66	28.28	93.92
DeepSeek-v4-flash	**35.83**	81.61	**33.15**	**94.34**
GPT-5-mini	33.99	73.44	28.45	90.26
Qwen3-max	29.43	85.13	33.06	91.99
Avg. $\Delta_{\text{0-shot}}$	1.57	9.21	1.93	6.53
Avg. $\Delta_{\text{0-shot CoT}}$	3.80	4.78	3.30	8.87

Table 4: Results on the RecToM benchmark (accuracy in %). Bold and underlined numbers denote the best and second-best results among prompting methods, respectively. Avg. $\Delta_{\text{0-shot}}$ and Avg. $\Delta_{\text{0-shot CoT}}$ represent the average performance gains of MindStep over the corresponding baselines across evaluated models.

5.3 Ablation Study

To examine the contribution of each component in MindStep, we conduct ablation experiments on SocialAct using Qwen3-max. Three variants are evaluated:

- **MindStep w/o Evocation:** The Evocation module is removed, so responses are generated without explicit memory-based persona activation.
- **MindStep w/o Perception:** The Perception module is removed, evaluating performance without explicit parsing of external context.
- **MindStep-Reversed:** The cognitive sequence is reversed, processing contextual perception before persona evocation (Perception → Evocation → Reaction), to assess the effect of persona anchoring.

Table 5 reports the ablation results of MindStep on SocialAct using Qwen3-max.

Impact of the Evocation Module. Ablating the Evocation module leads to the most substantial performance drop, with accuracy decreasing by 12.2% and 16.6% on BLM and MeToo, respectively, and by 8.5% and 8.3% on COVID-19 and CAL Fire. This indicates that anchoring the agent's persona is particularly crucial for sociopolitical events, where stable attitudes and social stances strongly influence behavior.

Role of the Perception Module. Removing the Perception module results in moderate declines, ranging from 5.8% to 6.3% on BLM and MeToo, and from 4.8% to 4.9% on COVID-19 and CAL Fire. This suggests that context filtering provides essential complementary evidence for action prediction, functioning effectively alongside the persona anchor.

Significance of the Cognitive Sequence. Furthermore, reversing the cognitive sequence causes accuracy decreases of 4.5% and 3.4% on sociopolitical controversies (BLM and MeToo), compared to drops of less than 2% on public crises (COVID-19 and CAL Fire). This demonstrates that in events heavily shaped by social attitudes, persona anchoring and self-referential biases play a primary role, whereas contextual cues dominate in public crises, rendering the stage ordering less critical.

Overall, these results empirically validate our proposed three-stage architecture: Evocation establishes a stable persona anchor, Perception extracts relevant contextual cues, and Reaction integrates both to generate accurate action predictions.

Method	BLM	MeToo	COVID-19	CAL Fire
0-shot	23.70	23.40	44.20	50.30
0-shot CoT	37.50	40.30	60.20	61.00
MindStep	64.80	68.00	80.10	80.10
w/o Evocation	52.60	51.40	71.60	71.80
w/o Perception	59.00	59.00	76.20	76.30
Reversed Order	60.30	64.60	78.40	79.40

Table 5: Ablation experiment results on SocialAct behavioral prediction. All results are reported as action-triplet accuracy (%). "w/o Evocation" and "w/o Perception" denote removing the Evocation and Perception stages, respectively. "Reversed Order" denotes the reversed cognitive sequence (Perception → Evocation → Reaction).

6 Conclusion

In this work, we propose MindStep, a three-stage cognitive prompting framework for LLM-based social simulation inspired by Theory of Mind and Information Processing Theory. Rather than relying on surface-level behavioral imitation, MindStep decomposes behavior generation into the sequential stages of Evocation, Perception, and Reaction, enabling more transparent reasoning traces while grounding simulated behaviors in persona-specific attributes rather than generic model priors. We further introduce SocialAct, a fine-grained benchmark grounded in real-world social events for evaluating contextualized behavior prediction. Results demonstrate that MindStep consistently outperforms standard prompting methods on both behavioral prediction and mental-state reasoning tasks, improving action inference in complex contexts and strengthening belief and desire reasoning on RecToM. These findings suggest that structured cognitive prompting is a promising direction for improving the fidelity and interpretability of LLM-based social simulation without additional training. In future work, we plan to extend MindStep from constrained selection tasks to open-ended generation environments, and further investigate how structured cognitive prompting can support robust multi-agent interactions and long-context social reasoning in dynamic environments.

Limitations

Although MindStep demonstrates consistent improvements across behavioral prediction and

mental-state reasoning tasks, several limitations remain.

First, our evaluation primarily focuses on constrained selection and action discrimination tasks within SocialAct and RecToM. While this closed-set design ensures baseline reproducibility and avoids the output variance inherent in open-ended text generation, it does not fully capture the unconstrained, generative nature of human social communication. Future work is needed to extend our structured cognitive pathways to open-ended generation environments.

Second, the data configuration of SocialAct is grounded in temporally aligned real-world interactions. Candidate actions and distractors are sampled from the same event domain around the timestamp of the target action, creating a contextually coherent and behaviorally challenging setting for fine-grained behavioral prediction. However, this setup does not model how human behavior and user personas evolve over longer time horizons or shift across different topic domains. Incorporating dynamic memory mechanisms to support long-term behavioral adaptation remains an open challenge.

Finally, our evaluation is limited to social media and conversational benchmarks covering a relatively small set of social events. Model predictions may still be affected by demographic bias, political alignment, and safety constraints in current LLMs, which could limit generalization across broader social settings. Future work should evaluate structured cognitive prompting under more diverse populations and interaction scenarios.

Ethical Considerations

SocialAct is constructed from publicly available Twitter data in TwiBot-22 and is intended solely for research use. We follow the original dataset terms and apply extensive anonymization during preprocessing. User-identifying information, including usernames, mentions, and URLs, is obfuscated or removed to reduce identity leakage while preserving the conversational structure of the data. To minimize harm, we filter content that may violate platform policies or social norms, including hate speech, violence, self-harm, extremist content, sexual content, illegal activities, misinformation, and other unsafe material. However, because the benchmark is derived from real social media data, residual bias and representation imbalance may still persist.

For mental-state reasoning, we evaluate on the public RecToM benchmark using its original task setup and reported random-guess and human reference results. Since no new human-subject data or annotations are collected in our experiments, the evaluation introduces no additional privacy risk. Although MindStep is a structured prediction framework rather than an open-ended generation system, its outputs may still reflect biases inherited from the underlying LLMs. Accordingly, we emphasize that the benchmark and results should be used only for research and analysis.

References

Richard C Atkinson and Richard M Shiffrin. 1968. Human memory: A proposed system and its control processes. In *Psychology of learning and motivation*, volume 2, pages 89–195. Elsevier.

Alan D. Baddeley and Graham Hitch. 1974. Working memory. volume 8 of *Psychology of Learning and Motivation*, pages 47–89. Academic Press.

James Bisbee, Joshua D Clinton, Cassy Dorff, Brenton Kenkel, and Jennifer M Larson. 2024. Synthetic replacements for human survey data? the perils of large language models. *Political Analysis*, 32(4):401–416.

Jiawei Chen, Ruoxi Xu, Boxi Cao, Ruotong Pan, Yunfei Zhang, Yifei Hu, Yong Du, Tingting Gao, Yaojie Lu, Yingfei Sun, and 1 others. 2026. Towards real-world human behavior simulation: Benchmarking large language models on long-horizon, cross-scenario, heterogeneous behavior traces. *arXiv preprint arXiv:2604.08362*.

Ziyan Cui, Ning Li, and Huaikang Zhou. 2025. A large-scale replication of scenario-based experiments in psychology and management using large language models. *Nature Computational Science*, 5(8):627–634.

Shangbin Feng, Zhaoxuan Tan, Herun Wan, Ningnan Wang, Zilong Chen, Binchi Zhang, Qinghua Zheng, Wenqian Zhang, Zhenyu Lei, Shujie Yang, and 1 others. 2022. Twibot-22: Towards graph-based twitter bot detection. *Advances in Neural Information Processing Systems*, 35:35254–35269.

Kanishk Gandhi, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah Goodman. 2023. Understanding social reasoning in language models with language models. *Advances in Neural Information Processing Systems*, 36:13518–13529.

Salvatore Giorgi, Sharath Chandra Guntuku, McKenzie Himelein-Wachowiak, Amy Kwarteng, Sy Hwang, Muhammad Rahman, and Brenda Curtis. 2022. Twitter corpus of the# blacklivesmatter movement and counter protests: 2013 to 2021. In *Proceedings of the*

International AAAI Conference on Web and Social Media, volume 16, pages 1228–1235.

Shenyue Jia, Seung Hee Kim, Son V Nghiem, Paul Doherty, and Menas C Kafatos. 2020. Patterns of population displacement during mega-fires in california detected using facebook disaster maps. *Environmental Research Letters*, 15(7):074029.

Dongzhi Jiang, Renrui Zhang, Ziyu Guo, Yanwei Li, Yu Qi, Xinyan Chen, Liuhui Wang, Jianhan Jin, Claire Guo, Shen Yan, Bo Zhang, Chaoyou Fu, Peng Gao, and Hongsheng Li. 2025. MME-cot: Benchmarking chain-of-thought in large multimodal models for reasoning quality, robustness, and efficiency. In *Forty-second International Conference on Machine Learning*.

Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. 2023. Decomposed prompting: A modular approach for solving complex tasks. In *The Eleventh International Conference on Learning Representations*.

Chance Jiajie Li, Jiayi Wu, Zhenze Mo, Ao Qu, Yuhan Tang, Kaiya Zhao, Yulu Gan, Jie Fan, Jiangbo Yu, Jinhua Zhao, and 1 others. 2026a. Simulating society requires simulating thought. *Advances in Neural Information Processing Systems*, 38.

Mengfan Li, Xuanhua Shi, and Yang Deng. 2026b. Rectom: A benchmark for evaluating machine theory of mind in llm-based conversational recommender systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 31636–31644.

Nian Li, Chen Gao, Mingyu Li, Yong Li, and Qingmin Liao. 2024. Econagent: large language model-empowered agents for simulating macroeconomic activities. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15523–15536.

Zachary Maiorana, Pablo Morales Henry, and Jennifer Weintraub. 2020. Metoo digital media collection-twitter dataset.

Xinyi Mou, Zhongyu Wei, and Xuanjing Huang. 2024. Unveiling the truth and facilitating change: Towards agent-based large-scale social movement simulation. In *ACL (Findings)*, pages 4789–4809.

David Premack and Guy Woodruff. 1978. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, 1(4):515–526.

Sahand Sabour, Siyang Liu, Zheyuan Zhang, June Liu, Jinfeng Zhou, Alvionna Sunaryo, Tatia Lee, Rada Mihalcea, and Minlie Huang. 2024. Emobench: Evaluating the emotional intelligence of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5986–6004.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-llm: A trainable agent for role-playing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13153–13187.

Karl Sigmund, Hannelore De Silva, Arne Traulsen, and Christoph Hauert. 2010. Social learning promotes institutions for governing the commons. *Nature*, 466(7308):861–863.

Michael Henry Tessler, Michiel A Bakker, Daniel Jarrett, Hannah Sheahan, Martin J Chadwick, Raphael Koster, Georgina Evans, Lucy Campbell-Gillingham, Tantum Collins, David C Parkes, and 1 others. 2024. Ai can help humans find common ground in democratic deliberation. *Science*, 386(6719):eadq2852.

Andrew R Todd and Diana I Tamir. 2024. Factors that amplify and attenuate egocentric mentalizing. *Nature Reviews Psychology*, 3(3):164–180.

Mo Wang, Daniel J Beal, David Chan, Daniel A Newman, Jeffrey B Vancouver, and Robert J Vandenberg. 2017. Longitudinal research: A panel discussion on conceptual issues, research design, and statistical techniques. *Work, Aging and Retirement*, 3(1):1–24.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Alex Wilf, Sihyun Lee, Paul Pu Liang, and Louis-Philippe Morency. 2024. Think twice: Perspective-taking improves large language models' theory-of-mind capabilities. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8292–8308.

Yufan Wu, Yinghui He, Yilin Jia, Rada Mihalcea, Yulong Chen, and Naihao Deng. 2023. Hi-tom: A benchmark for evaluating higher-order theory of mind reasoning in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10691–10706.

Yueqi Xie, Lemeng Liang, Shuzhen Li, Yifu Lu, Zhiwen Xiao, Mengdi Shi, Junming Huang, Mengdi Wang, and Yu Xie. 2026. Evaluating the statistical realism of llm-generated social science data. *Proceedings of the National Academy of Sciences*, 123(19):e2538145123.

Hainiu Xu, Runcong Zhao, Lixing Zhu, Jinhua Du, and Yulan He. 2024. Opentom: A comprehensive benchmark for evaluating theory-of-mind reasoning capabilities of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8593–8623.

Xiutian Zhao, Ke Wang, and Wei Peng. 2024. An electoral approach to diversify llm-based multi-agent collective decision-making. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2712–2727.

A Construction of the SocialAct Benchmark

We utilize the TwiBot-22 dataset (Feng et al., 2022), which is publicly available under the MIT License, as our foundation. All derivative data generated in this study is used exclusively for academic research, strictly following the original access terms.

From this source, we construct the SocialAct dataset through a multi-stage pipeline for contextualized behavior prediction on Twitter. During this process, the data is carefully filtered to protect privacy, preserve conversational context, and ensure plausible social actions.

Ultimately, SocialAct serves as a benchmark for modeling user behavior across real-world events by incorporating both historical tweets and temporally aligned news. The final dataset samples 1,000 actions per event, and Table 6 details the statistics of the full, unsampled dataset.

Event	Users	Actions	Avg. Hist. Tweets/User
BLM	5,184	8,358	252.41
MeToo	2,008	3,143	219.03
Covid-19	21,337	38,586	248.82
CAL Fire	884	1,186	306.8

Table 6: Statistics of the original, unsampled SocialAct data derived from TwiBot-22.

A.1 Data Preprocessing

We derive SocialAct from the TwiBot-22 dataset. To ensure data quality and user privacy, we first preprocess the raw corpus and then construct the final benchmark from the processed data.

A.2 Sample Filtering

We filter the raw data according to three criteria. First, we exclude users with fewer than thirty historical tweets, as insufficient history makes it difficult to model stable user traits. Second, for actions involving a target tweet, we require the target content to contain at least three effective words after removing URLs, hashtags, and mentions. Third, we remove samples containing violent, explicit, or otherwise safety-sensitive content, and further conduct an LLM-based audit to reduce harmful or inappropriate instances.

A.3 Privacy Preservation and Text Normalization

To mitigate privacy risks while preserving conversational fidelity, we partially obfuscate user-identifying information. Specifically, usernames and mentions are masked by retaining only partial character information (e.g., "@a***z"), while URLs are truncated and anonymized. This design reduces the risk of identity leakage while maintaining the natural format of social media interactions. Furthermore, we normalize textual content using NFKC normalization, removing zero-width characters and redundant whitespace.

A.4 Task and Context Construction

After preprocessing, we formulate the contextualized behavior prediction task by constructing the input social context and the corresponding action candidates. All candidate pools are built from the full preprocessed corpus within each event domain.

A.4.1 Social Context Construction

For each instance, we construct the input social context $\mathcal{C}$ from three sources. First, we use DeepSeek-V4-Flash to generate a concise summary of the user's profile and infer the user's stance toward the event based on the user's profile information, account metadata, interest lists, and 20 historical tweets. The prompt template used for profile and stance generation is shown in Figure 3. Second, we retrieve 5 task-relevant historical tweets using BM25 to provide a social memory of prior behavior. Third, we retrieve timeline-aligned news content from an event-specific news timeline. For each action, we select the most recent news preceding the action timestamp, ensuring the model is conditioned on the relevant temporal background rather than the direct future cause of the interaction.

A.4.2 Action Triplet and Option Construction

We formulate the task as predicting an *Action Triplet* $(A_{\text{type}}, A_{\text{obj}}, A_{\text{cont}})$, where $A_{\text{type}} \in \{\textit{post}, \textit{retweet}, \textit{reply}, \textit{quote}\}$ denotes the action type. To evaluate the prediction of A_{obj} and A_{cont}, we construct multiple-choice candidate sets. To ensure contextual plausibility, all distractors are sampled from the same event domain within a three-day temporal window ($[T-2, T]$), covering the day of the action and the two preceding days.

For A_{obj}, we construct a visible Twitter feed containing three candidate tweets. If the ground-truth action is *retweet*, *reply*, or *quote*, the feed contains

Prompt for Profile and Stance Generation

Given the following observation about an individual {Name}, please summarize the relevant details from the profile and deduce their stance on {target}. His or her profile information is as follows:.
Name: {Name}
Location: {Location}
Description: {Description}
Stats: {Followers Count} followers, {Following Count} following, {Tweet Count} tweets
Account Created: {Account Created}
List of interest: {Related List}
Sample of posted tweets: {Sample of Previous Posts}
Task:
1. Summary: You can deduce the preferences and personality from profile, lists, and tweets, but please avoid repeating the observation in the summary.
2. Stance: Deduce the stance on {target}. Choose from [Support, Neutral, Oppose, Unknown].
Constraint: Only output the following format. No extra words.
Summary: [You are {Name}. followed by the narrative description]
Stance: [Your choice]

Figure 3: Prompt used to generate the user profile summary and stance.

one gold target tweet and two distractors. If the action is *post*, the feed contains three distractors, and A_{obj} is set to the placeholder "-".

For A_{cont}, we construct a candidate response pool for actions that require textual content (*post*, *reply*, and *quote*). The pool contains one gold response and two distractors. To make the task non-trivial, we exclude candidates authored by the current user and remove texts that are identical to the gold response or the target object tweet via fuzzy matching. If the action is *retweet*, A_{cont} is set to "-".

A.5 Benchmark Sampling

After constructing the candidate pools and contextual inputs from the full preprocessed corpus, we sample the final evaluation benchmark in a balanced manner across domains. Specifically, we sample 1,000 action instances from each event domain, resulting in 4,000 test cases in total. The prompt template used for SocialAct question generation is shown in Figure 4.

Prompt for SocialAct Question Generation

You are a Twitter user {Name}. Given your personal information and recent memory, along with the current context and options, you are required to predict your most likely action:
(1) Your description: {role_description};
(2) Your stance on {target}: {stance};
(3) Current time is {current_time};
(4) The news page you got is {trigger_news}
(5) Your recent memory is {past_event}
(6) The twitter page you can see is {tweet_page}
Determine your next action by selecting one option from each dimension:
(7) Action Type: select from [post, retweet, reply, quote]
(8) Action Object: "-" ONLY if type is 'post'; otherwise, you must select the specific tweet ID from tweet_page that you are interacting with. (tweet_A or tweet_B or tweet_C)
(9) Action Content: "-" ONLY if type is 'retweet'; otherwise, you must select one item from {tweet_content}. (Do not generate your own or copy from elsewhere)

Figure 4: Prompt used to generate SocialAct question.

B Prompt Templates for MindStep

In this section, we provides the complete prompt templates used in our experiments, including the MindStep prompts, ablation variants, and baseline prompts.

Although the core structure of MindStep remains consistent across tasks, the prompts are adapted to different domains and input formats. Specifically, SocialAct involves social media environments with persona descriptions, stance information, recent memory, and dynamic Twitter interactions, while RecToM focuses on movie recommendation dialogues that only contain conversational context without explicit personal profiles or social histories.

Accordingly, the prompts are domain-adapted while preserving the same underlying cognitive reasoning framework. The SocialAct prompts emphasize the evocation of behavioral tendencies and social posting habits from persona-related information, whereas the RecToM prompts adapt the evocation process to infer preferences, conversational attitudes, and interpersonal understanding from dialogue context alone. Despite these domain-specific adaptations, all prompts follow the same three-stage reasoning process: Evocation, Perception, and Reaction.

B.1 Prompts for SocialAct

Prompt for Behavior Prediction in SocialAct

To predict your most likely action, you will simulate your cognitive processing step by step. Please follow this format strictly, No extra words:
My cognitive trace:
[Evocation-1] [Step-by-step summarize your persona and inherent behavioral patterns purely from your description, stance, and recent memory. Focus on your behavioral habits, identity, expression, and inclination towards posting original content or interacting with others.]
[Evocation-2] ...
...
[Percept-1] [Guided by your evoked persona and purpose, step-by-step summarize the key events from the current twitter page and news page that resonate with your persona, including both potential topics for independent posting and existing tweets for interaction.]
[Percept-2] ...
...
[Reaction-1] [Synthesizing your evocations and percepts, predict the specific action you would most likely take by prioritizing your persona and inherent behavioral habits over literal topic matching or merely reacting to visible tweets.]
[Reaction-...] ...
...
Therefore my action is:
Action type:
Action object:
Action content:

Prompt for Ablation: MindStep w/o Evocation

To predict your most likely action, you will simulate your cognitive processing step by step. Please follow this format strictly, No extra words:
My cognitive trace:
[Percept-1] [Step-by-step objectively summarize the key events from the current twitter page and news page, including both potential topics for independent posting and existing tweets for interaction.]
[Percept-2] ...
...
[Reaction-1] [Synthesizing your percepts, predict the specific action you would most likely take, avoiding mechanically matching topics literally or merely reacting to visible tweets.]
[Reaction-...] ...
...
Therefore my action is:
Action type:
Action object:
Action content:

Prompt for Ablation: MindStep w/o Perception

To predict your most likely action, you will simulate your cognitive processing step by step. Please follow this format strictly, No extra words:
My cognitive trace:
[Evocation-1] [Step-by-step summarize your persona and inherent behavioral patterns purely from your description, stance, and recent memory. Focus on your behavioral habits, identity, expression, and inclination towards posting original content or interacting with others.]
[Evocation-2] ...
...
[Reaction-1] [Synthesizing your evocations, predict the specific action you would most likely take, avoiding mechanically matching topics literally or merely reacting to visible tweets.]
[Reaction-...] ...
...
Therefore my action is:
Action type:
Action object:
Action content:

Prompt for Ablation: MindStep-Reversed

To predict your most likely action, you will simulate your cognitive processing step by step. Please follow this format strictly, No extra words:
My cognitive trace:
[Evocation-1] [Step-by-step summarize your persona and inherent behavioral patterns purely from your description, stance, and recent memory. Focus on your behavioral habits, identity, expression, and inclination towards posting original content or interacting with others.]
[Evocation-2] ...
...
[Percept-1] [Guided by your evoked persona and purpose, step-by-step summarize the key events from the current twitter page and news page that resonate with your persona, including both potential topics for independent posting and existing tweets for interaction.]
[Percept-2] ...
...
[Reaction-1] [Synthesizing your evocations and percepts, predict the specific action you would most likely take by prioritizing your persona and inherent behavioral habits over literal topic matching or merely reacting to visible tweets.]
[Reaction-...] ...
...
Therefore my action is:
Action type:
Action object:
Action content:

B.1.1 Baseline Prompts

0-shot

Only output the following format, No extra words:
Action type:
Action object:
Action content:

0-shot CoT

To predict your most likely action, you need to think step by step. Please follow this format strictly, No extra words:
My thinking process:
the thinking process for choosing the action
Therefore my action is:
Action type:
Action object:
Action content:

B.2 Prompts for RecToM

Prompt for Belief Inference

To infer the recommender's belief, simulate the cognitive processing step by step. Please follow this format strictly, No extra words:
[Evocation-1] Step into the recommender's shoes. Recall your interaction and get a feel for the seeker's general taste.
[Percept-2] Spot the target movie. Notice who brought it up. Did the seeker watch it? Mentioning plot details or saying "seen it" means yes.
[Reaction-3] From the recommender's view, sense the seeker's attitude: are they positive (accepting, eager, or fond if seen) or resistant (declining or critical if seen)?
Therefore the answer is: X (where X is one of A-G)

Prompt for Desire Inference

To predict the seeker's most likely action, simulate the cognitive processing step by step. Please follow this format strictly, No extra words:
[Evocation-1] Summarize seeker's preferred genres and explicitly rejected types.
[Percept-2] Locate the target movie. Extract seeker's exact reaction (accepted, liked, seen-it, or rejected).
[Reaction-3] Determine likelihood: "Likely to watch" includes wanting to watch, OR having watched and liked it. Only choose "no" if explicitly rejected or disliked.
Therefore the answer is: X (where X is A or B)

Prompt for Recommender Intention Inference

To infer the recommender's intention, simulate the cognitive processing step by step. Please follow this format strictly, No extra words:
[Evocation-1] Step into the recommender's shoes. Summarize the seeker's current state and implicit needs based on the prior dialogue context.
[Percept-2] Focus strictly on the seeker's latest input. Identify the exact query, feedback, or emotional cue that serves as the direct trigger for your next response.
[Reaction-3] Analyze the target utterance. Pinpoint its primary conversational goal (e.g., asking for preferences, recommending, chitchatting, greeting, or giving feedback,...) and map your motivation directly to the provided choices.
Therefore the answer is: X (where X is a combination of letters, e.g., 'A', 'BC', 'ADE')

MindStep Prompt for Seeker Intention Inference

To infer the seeker's intention, simulate the cognitive processing step by step. Please follow this format strictly, No extra words:
[Evocation-1] Step into the seeker's shoes. Walk through the dialogue history to capture the subtle context and feel how the conversational vibe has evolved.
[Percept-2] Spot the target utterance. Catch your immediate reaction, reading between the lines to grasp its literal meaning, hidden implications, and shifts in intent.
[Reaction-3] From your view, sense the drive: hunting for recs, reacting to a pick, or just chatting? If reacting, are you accepting, rejecting, or critiquing?
Therefore the answer is: 'XX' (where X is a combination of letters, e.g., 'A', 'BC', 'ADE')

C Case Study

To further analyze behavioral differences among prompting paradigms, we examine a representative example from the BLM benchmark (Table 7). Figure 5 presents the reasoning traces produced by different prompting paradigms on this example. The user is an anti-carceral feminist activist and academic whose historical posts are characterized by original, movement-oriented commentary on structural violence. The current scenario presents a news article about the Kimberly Potter conviction and three candidate tweets: an unrelated domestic dispute (Tweet_A), a report on Amazon suspending BLM from its charity platform (Tweet_B), and a personal anecdote referencing an Amir Locke solidarity moment (Tweet_C). The target Action Triplet is (post, -, content_A), corresponding to post a news article about student-organized walkouts in support of Amir Locke.

MindStep produces the correct prediction. In the Evocation stage, the model activates the user's identity as an analytically oriented activist whose prior behaviors primarily involve posting original content rather than interacting with others. In the Perception stage, Tweet_B is identified as topically salient to the user's documented interest in racial justice, while Tweet_C overlaps with the user's prior event-specific posting history but is framed as a personal anecdote, making it less salient as a candidate interaction object. Crucially, none of the candidate content options in the content list directly responds to either Tweet_B or Tweet_C. In the Reaction stage, the model combines persona traits with the selected percepts and correctly infers that the user's behavioral pattern favors generating an original post rather than quoting or retweeting one of the candidate tweets.

0-Shot CoT predicts quote Tweet_C. The model correctly identifies the overlap between Tweet_C and the user's prior posts on Amir Locke, but appears to over-rely on surface lexical matching and selects quoting as the most likely action. This suggests that unconstrained reasoning chains may over-rely on surface lexical matching without sufficiently grounding the prediction in persistent persona traits and behavioral habits.

0-Shot predicts quote Tweet_B. Although Tweet_B is contextually relevant to the user's interests, the predicted action type remains incorrect. Without intermediate reasoning, the model appears to rely on generic interaction priors rather than persona-specific behavioral patterns.

Overall, both baselines identify socially relevant candidate tweets but fail to correctly infer the action type. MindStep's staged decomposition makes the decision process more traceable: Evocation anchors the user's posting behavior and persona traits, Perception filters contextually relevant information, and Reaction infers the behavioral outcome based on these processed signals. This structured progression helps mitigate the model's reliance on surface lexical matching and generic interaction patterns, leading to the correct prediction.

D Detailed Results and Ablation Analysis for Behavioral Prediction

D.1 Detailed Metrics of Action Elements

In this section, Table 8 - Table 11 reports the full Accuracy, Precision, Recall, and F1 scores of action object, action type, and action content on the four SocialAct events. Overall, action content is consistently easier to predict than action object and action type, while the latter two remain more challenging due to their finer-grained behavioral semantics. Across all events, MindStep consistently outperforms both 0-shot and 0-shot CoT prompting, with the largest gains observed on BLM and MeToo.

On BLM, Qwen3-max under MindStep reaches 55.54% F1 on action object, 50.08% F1 on action type, and 78.85% F1 on action content, compared with 28.44%, 30.30%, and 52.80% under 0-shot, respectively. This corresponds to absolute improvements of +27.10, +19.78, and +26.05 percentage points, or relative gains of 95.28%, 65.31%, and 49.34%, respectively. Compared with 0-shot CoT, MindStep still yields substantial improvements of +16.41, +13.00, and +15.75 points on the three elements. Similar trends are observed on MeToo, where the F1 scores increase from 26.93%, 28.17%, and 52.17% under 0-shot to 57.61%, 49.08%, and 83.28% under MindStep, yielding absolute gains of +30.68, +20.91, and +31.11 points, respectively. Even on COVID-19 and CAL Fire, where the baseline performance is relatively higher, MindStep remains consistently effective, further confirming its ability to strengthen the prediction of each action element across different event types.

D.2 Detailed Ablation Metrics of MindStep

Table 12 presents the detailed ablation results of MindStep on Qwen3-max across the four event domains. Removing the Evocation module leads to the most substantial and consistent performance drop, indicating that persona anchoring is the key factor for Action Triplet prediction. On BLM, ablating Evocation reduces F1 from 55.54%, 50.08%, and 78.85% to 51.11%, 43.20%, and 68.98% on action object, action type, and action content, respectively. On MeToo, the corresponding drops are even larger, decreasing from 57.61%, 49.08%, and 83.28% to 46.01%, 42.19%, and 70.76%, respectively. These results suggest that anchoring the agent in a stable persona is particularly important for sociopolitical events, where social stance and personal attitude play a dominant role.

Removing the Perception module also leads to moderate but consistent declines across all events, showing that contextual parsing provides essential complementary evidence for behavioral prediction. Reversing the cognitive order further harms performance on all four events, although the degradation is generally smaller than that caused by ablating Evocation. Overall, these ablation results empirically validate the proposed three-stage architecture: Evocation establishes a stable persona anchor, Perception extracts event-specific contextual cues, and Reaction integrates both sources to generate accurate Action Triplets.

Question: You are a Twitter user d***g. Given your personal information and recent memory, along with the current context and options, you are required to predict your most likely action
Description: You are d***g, a long-standing feminist activist and academic focused on global gender justice, with a geographic and professional nexus between Washington D.C. and Cape Town. You operate the forum Women In and Beyond the Global, centering your analysis on systemic oppression within households, prisons, and cities. Your prolific tweeting reveals a deep, intersectional commitment to **anti-racism**, prison abolition, housing justice, and labor rights, consistently highlighting the specific struggles of Black women globally. Your content **actively opposes conservative backlash against racial equity education and critical race theor**y, while **amplifying movements from Minneapolis to South Africa**.
Stance on BlackLivesMatter: Support
Current Time: 2022-02-20 14:30:59+00:00
News: Kimberly Potter was convicted of manslaughter in the death of Daunte Wright. (2021-12-23)
Recent Memory: • [d***g] **posted**: School's dream dashed as donor fails to pay up #SouthAfrica #EasternCape Time: 2022-02-14 13:00:12+00:00) • [d***g] posted: In response to **Amir Locke**'s killing, Minneapolis mayor gestures to limits of police reform #AmirLocke #BlackLivesMatter (2022-02-10 18:04:26+00:00) • [d***g] **posted**: Hijab row: The India woman who is the face of the fight to wear headscarf #India (2022-02-10 16:33:25+00:00) • [d***g] **posted**: Africa transitioning out of pandemic phase of Covid, WHO says (2022-02-10 14:28:45+00:00) • [d***g] **posted**: Meet the Formerly Incarcerated Advocates Revolutionizing Giving Birth in Prison (2022-02-09 18:43:33+00:00)
Twitter Feed: **tweet_A**: again, #BlackLivesMatter, eh? Reed beat his ex-wife, & was driving OTHER women's cars around... **tweet_B**: RT @D***l: Amazon SUSPENDS Black Lives Matter from its charity platform **tweet_C**: I was at a swim meet with my son in Warrenton, VA & happened upon group of people carrying Black Lives Matter and justice for **Amir Locke** signs...
Content List: **Content_A**: @Ms @Sr foguete do tipo nasa – uma fancam do #W13 #FIA #F1 #FaustaoNaBand #JN #SirPATRON44 #TEAMLH #WeStandWithLewisHamilton #WSL #mercedesf1 #TotoWolff #MercedesW13 #MercedesAMGF1 #OpatrãoTãON #F1naBand #BlackLivesMatter #NFL #wslnosportv #WSLBrasil **Content_B**: Meet the High School Students Who Organized Thousands to Walk Out for **Amir Locke** @ig #AmirLocke #BlackLivesMatter **Content_C**: @Wp @Te @MD Nobody cheated more than us! Me: "Hey Warren, what are you doing here in the D-Line meeting, you're on offense!?" Sapp: "I've come to take your job" Me: "Shiiit, you ain't taking my job!" * cut to * me watching every game from the sidelines #HackAndGash4LIFE
Determine your next action by selecting one option from each dimension: **Action Options:** Type: select from [post, retweet, reply, quote] Object: "-" ONLY if type is 'post'; otherwise, select the specific tweet ID. Content: "-" ONLY if type is 'retweet'; otherwise, select one item from the provided content list.

Table 7: An Example Question in SocialAct: BLM Event.

MindStep

My cognitive trace:
Evocation-1] As d***g, I am a feminist activist and academic. My posts consistently engage with issues of police violence, carceral systems, housing, labor, and anti-racist education. I frequently highlight specific cases of state violence (e.g., Amir Locke) and connect them to broader structural critiques, especially prison abolition and racial capitalism. My recent memory shows a pattern of posting original content—often linking to in-depth reporting or movement work—rather than retweeting or replying casually.
[Evocation-2] My stance is firmly pro–Black Lives Matter, but my expression of support is analytical and grounded in abolitionist praxis, not performative solidarity. My content avoids superficial endorsements and instead offers contextualized critique or resource-sharing.
[Percept-1] The news page is a case I have likely addressed before. tweet_A discusses domestic violence involving a man named Reed, which lacks clear connection to systemic analysis or movement contexts I prioritize; tweet_B reports Amazon suspending BLM from its charity platform—an institutional action relevant to defunding of racial justice, which aligns with my interests; tweet_C describes a personal encounter with BLM/Amir Locke solidarity in Virginia, which resonates with my prior post about Amir Locke but is anecdotal rather than analytical.
[Reaction-1] Among these, tweet_B touches on institutional backlash against BLM—a theme I actively oppose, per my description ("opposes conservative backlash against racial equity education and critical race theory"). However, my behavioral pattern favors original posting over retweeting or quoting others' observations, especially when I can offer deeper context or alternative framing. Moreover, none of the provided action content options directly respond to tweet_B or tweet_C in a way that matches my voice.
[Reaction-2] Given my habit of posting original, movement-centered content and my recent focus on Amir Locke (Feb 10), the most aligned action content option is "Meet the High School Students Who Organized Thousands to Walk Out for Amir Locke @i***g https://***pI #AmirLocke #BlackLivesMatter" —it continues my engagement with the Amir Locke case through youth-led resistance, fitting my intersectional, anti-carceral lens.

Therefore my action is:
Action type: post
Action object: -
Action content: Meet the High School Students Who Organized Thousands to Walk Out for Amir Locke @i***g https://***pI #AmirLocke #BlackLivesMatter"

0-shot Cot

My thinking process:\nThe news about Kimberly Potter's conviction for Daunte Wright's death aligns with my consistent focus on police violence and systemic injustice, especially concerning Black lives. My recent memory shows active engagement with the killing of Amir Locke and broader #BlackLivesMatter issues, including critiques of police reform and solidarity actions. Among the visible tweets, tweet_C reflects grassroots solidarity with #BlackLivesMatter and specifically references Amir Locke—a case I've already tweeted about—making it highly relevant to my ongoing advocacy. Retweeting would amplify this moment of public solidarity but given my pattern of adding analytical or mobilizing context (e.g., linking student organizing or incarcerated women's struggles), a quote tweet that elevates youth-led action aligns best The only non-irrelevant, on-topic content option is "Meet the High School Students Who Organized Thousands to Walk Out for Amir Locke," which directly connects to my prior post about Amir Locke and my commitment to highlighting collective resistance, particularly by marginalized groups. Thus, I would quote tweet tweet_C with that content to bridge spontaneous public solidarity with organized youth activism.\n\n

Therefore my action is:
Action type: quote
Action object: tweet_C
Action content: Meet the High School Students Who Organized Thousands to Walk Out for Amir Locke @i***g https://***pI #AmirLocke #BlackLivesMatter"

0-shot

Action type: quote
Action object: tweet_B
Action content: Meet the High School Students Who Organized Thousands to Walk Out for Amir Locke @i***g https://***pI #AmirLocke #BlackLivesMatter

Figure 5: Overview of the MindStep framework. Given a user's profile, recent memory, and current social context, MindStep structures the reasoning process into three sequential stages: Evocation anchors the model to a user-specific persona, Perception selectively extracts situationally relevant cues, and Reaction infers the final action conditioned on the accumulated cognitive context.

Method	Action Type				Action Object				Action Content			
	Acc.	**Pre.**	**Rec.**	**F1**	**Acc.**	**Pre.**	**Rec.**	**F1**	**Acc.**	**Pre.**	**Rec.**	**F1**
0-shot												
Gemma-4-E4B-it	31.80	39.67	34.36	25.16	20.90	32.32	32.01	19.96	34.80	49.81	37.18	33.62
Deepseek-v4-flash	58.70	41.97	**71.33**	41.16	54.40	43.24	53.23	43.93	65.20	68.38	66.10	65.44
Gpt-5-mini	53.30	40.34	67.42	38.10	49.20	41.29	52.01	40.79	65.90	68.32	66.29	66.08
Qwen3-max	33.90	42.01	60.79	30.30	29.10	37.66	44.91	28.44	52.20	61.73	53.86	52.80
0-shot CoT												
Gemma-4-E4B-it	32.20	37.50	29.02	25.25	28.00	33.24	38.62	25.83	44.70	49.35	45.63	45.08
Deepseek-v4-flash	35.70	37.88	60.25	27.89	33.40	39.76	50.44	32.20	51.70	63.39	53.16	52.86
Gpt-5-mini	53.90	40.73	68.12	38.59	48.60	40.70	50.71	40.00	65.70	68.29	66.40	65.88
Qwen3-max	48.90	43.35	60.10	37.08	47.70	39.90	49.68	39.13	63.20	64.16	63.26	63.10
our method												
Gemma-4-E4B-it	41.30	35.52	30.55	29.35	39.10	35.14	41.53	31.85	45.10	50.04	46.32	45.09
Deepseek-v4-flash	65.90	46.31	70.15	45.15	62.20	46.60	**56.81**	48.95	72.00	73.25	72.58	71.97
Gpt-5-mini	67.90	44.32	70.79	44.72	63.50	46.46	55.25	48.73	72.00	72.55	72.16	71.89
Qwen3-max	**80.10**	**58.65**	70.88	**50.08**	**75.90**	**55.94**	55.20	**55.54**	**79.10**	**79.04**	**78.91**	**78.85**

Table 8: Detailed evaluation metrics on the BLM event benchmark across action type, action object, and action content prediction. Results include Accuracy, Precision, Recall, and F1. Bold indicates the best result in each column, and underline indicates the second-best result.

Method	Action Type				Action Object				Action Content			
	Acc.	**Pre.**	**Rec.**	**F1**	**Acc.**	**Pre.**	**Rec.**	**F1**	**Acc.**	**Pre.**	**Rec.**	**F1**
0-shot												
Gemma-4-E4B-it	51.10	40.24	38.76	33.46	47.80	42.33	56.04	41.06	50.40	66.57	52.61	51.73
Deepseek-v4-flash	77.90	49.91	50.61	48.55	78.60	62.08	67.96	64.66	78.20	79.74	78.32	78.12
Gpt-5-mini	68.00	45.00	72.50	44.89	67.60	53.08	71.12	57.43	76.50	77.97	77.07	76.54
Qwen3-max	48.70	40.05	39.40	33.90	49.20	45.43	65.92	44.84	58.90	74.38	61.03	60.52
0-shot CoT												
Gemma-4-E4B-it	37.00	36.53	31.32	26.21	38.60	41.09	58.60	36.64	47.40	57.73	49.04	48.49
Deepseek-v4-flash	52.40	40.34	44.97	37.07	53.60	47.77	70.80	49.11	70.70	76.30	71.52	71.50
Gpt-5-mini	67.00	43.59	47.15	43.41	66.30	52.37	70.22	56.57	74.60	77.15	75.36	74.82
Qwen3-max	66.90	48.69	44.91	44.05	66.60	52.40	70.57	56.85	73.90	77.64	74.89	74.28
our method												
Gemma-4-E4B-it	54.80	35.72	35.45	33.20	57.40	46.18	62.87	48.54	55.80	64.94	57.70	56.47
Deepseek-v4-flash	81.10	51.76	51.18	51.32	80.80	64.92	**75.11**	68.94	84.60	84.57	84.71	84.39
Gpt-5-mini	77.70	51.37	**74.59**	50.65	76.50	59.11	69.26	63.06	79.90	80.16	79.97	79.68
Qwen3-max	**88.40**	**61.82**	50.17	**52.58**	**87.40**	**76.73**	72.77	**74.63**	**87.50**	**87.25**	**87.40**	**87.30**

Table 9: Detailed evaluation metrics on the Covid-19 event benchmark across action type, action object, and action content prediction. Results include Accuracy, Precision, Recall, and F1. Bold indicates the best result in each column, and underline indicates the second-best result.

Method	Action Type				Action Object				Action Content			
	Acc.	**Pre.**	**Rec.**	**F1**	**Acc.**	**Pre.**	**Rec.**	**F1**	**Acc.**	**Pre.**	**Rec.**	**F1**
0-shot												
Gemma-4-E4B-it	31.20	37.05	34.31	24.78	24.20	33.41	36.05	22.98	37.30	47.75	39.96	37.85
Deepseek-v4-flash	58.30	41.60	44.95	40.68	53.90	41.40	50.64	42.46	69.80	72.04	70.66	69.79
Gpt-5-mini	47.00	39.95	41.68	35.82	44.60	41.11	53.40	39.34	66.10	68.31	66.74	66.10
Qwen3-max	32.30	39.31	36.91	28.17	27.30	36.46	43.72	26.93	51.30	63.19	54.08	52.17
0-shot CoT												
Gemma-4-E4B-it	28.00	37.68	30.27	24.34	26.40	36.05	42.94	26.16	46.70	49.85	47.62	46.95
Deepseek-v4-flash	38.00	37.75	37.81	29.98	36.60	40.12	52.73	34.85	60.30	68.07	61.70	61.23
Gpt-5-mini	48.10	40.45	41.56	36.60	45.10	40.76	52.53	39.24	67.20	69.00	68.07	67.16
Qwen3-max	47.80	43.30	35.73	37.05	47.80	40.66	51.63	40.16	66.10	66.75	66.16	65.70
our method												
Gemma-4-E4B-it	33.60	36.68	31.23	27.78	34.10	37.41	46.85	31.67	49.40	55.83	51.34	49.89
Deepseek-v4-flash	66.30	47.09	**48.48**	46.54	63.40	49.07	**61.14**	52.01	76.50	77.44	77.30	76.32
Gpt-5-mini	61.30	44.03	45.61	43.20	58.40	45.35	56.41	47.29	72.60	73.23	72.93	72.29
Qwen3-max	**75.80**	**56.16**	46.80	**49.08**	**73.80**	**55.68**	60.12	**57.61**	**83.80**	**83.22**	**83.70**	**83.28**

Table 10: Detailed evaluation metrics on the MeToo event benchmark across action type, action object, and action content prediction. Results include Accuracy, Precision, Recall, and F1. Bold indicates the best result in each column, and underline indicates the second-best result.

Method	Action Type				Action Object				Action Content			
	Acc.	**Pre.**	**Rec.**	**F1**	**Acc.**	**Pre.**	**Rec.**	**F1**	**Acc.**	**Pre.**	**Rec.**	**F1**
0-shot												
Gemma-4-E4B-it	56.40	38.67	42.62	35.50	53.00	39.19	51.56	39.09	56.70	67.75	59.81	57.61
Deepseek-v4-flash	69.70	45.46	53.50	45.67	68.90	48.48	61.96	52.41	77.90	77.06	77.55	76.62
Gpt-5-mini	61.00	44.19	65.11	44.04	60.60	45.34	63.90	48.25	76.30	75.08	75.54	74.85
Qwen3-max	56.30	44.11	52.64	42.14	55.70	45.16	67.21	46.83	68.20	74.63	70.45	68.51
0-shot CoT												
Gemma-4-E4B-it	48.20	39.01	51.68	35.50	49.30	41.38	60.33	41.00	62.20	63.61	62.30	61.50
Deepseek-v4-flash	50.80	39.50	61.72	36.84	51.00	43.59	65.08	43.64	69.90	74.11	71.20	69.75
Gpt-5-mini	66.10	46.49	65.60	47.49	65.00	48.14	67.33	51.97	78.30	77.75	78.29	77.13
Qwen3-max	68.00	49.71	**67.21**	49.48	67.60	49.94	68.95	54.58	80.70	79.79	80.84	79.64
our method												
Gemma-4-E4B-it	60.40	40.72	47.84	39.55	63.10	45.76	62.45	49.01	69.80	70.49	70.65	69.11
Deepseek-v4-flash	75.60	47.41	60.80	49.43	75.50	55.56	**70.68**	60.65	81.70	81.30	82.11	80.71
Gpt-5-mini	74.70	48.62	59.49	49.72	73.60	51.84	62.43	55.66	78.90	77.27	77.75	77.29
Qwen3-max	**87.50**	**66.50**	66.02	**59.28**	**86.80**	**72.01**	70.15	**71.05**	**88.70**	**87.98**	**88.02**	**88.00**

Table 11: Detailed evaluation metrics on the CAL Fire event benchmark across action type, action object, and action content prediction. Results include Accuracy, Precision, Recall, and F1. Bold indicates the best result in each column, and underline indicates the second-best result.

Event	Method	Action Type				Action Object				Action Content			
		Acc.	**Pre.**	**Rec.**	**F1**	**Acc.**	**Pre.**	**Rec.**	**F1**	**Acc.**	**Pre.**	**Rec.**	**F1**
Covid	MindStep	88.40	61.82	50.17	52.58	87.40	76.73	72.77	74.63	87.50	87.25	87.40	87.30
	w/o Evocation	79.80	54.97	47.25	50.03	81.70	66.24	74.62	69.81	80.90	80.53	80.69	80.56
	w/o Perception	84.30	59.14	49.40	51.11	84.40	70.95	76.70	73.46	84.90	84.82	85.13	84.78
	Reversed Order	86.40	62.43	72.87	51.95	86.20	75.28	72.40	73.69	86.90	86.62	86.73	86.67
BLM	MindStep	80.10	58.65	70.88	50.08	75.90	55.94	55.20	55.54	79.10	79.04	78.91	78.85
	w/o Evocation	65.10	48.64	63.97	43.20	65.60	48.38	56.55	51.11	69.60	69.04	68.99	68.98
	w/o Perception	72.80	53.98	69.40	48.56	70.40	51.33	56.61	53.44	74.60	74.37	74.21	74.23
	Reversed Order	74.70	53.83	67.70	46.68	72.50	52.79	55.00	53.80	74.60	74.32	74.29	74.18
MeToo	MindStep	75.80	56.16	46.80	49.08	73.80	55.68	60.12	57.61	83.80	83.22	83.70	83.28
	w/o Evocation	60.00	47.05	39.32	42.19	59.50	44.10	52.81	46.01	71.50	70.78	71.00	70.76
	w/o Perception	68.20	51.32	42.22	45.39	67.10	49.12	56.04	51.52	77.80	77.23	77.45	77.26
	Reversed Order	73.60	53.05	44.52	47.46	72.80	54.63	59.16	56.53	80.80	79.99	80.10	80.01
CAL Fire	MindStep	87.50	66.50	66.02	59.28	86.80	72.01	70.15	71.05	88.70	87.98	88.02	88.00
	w/o Evocation	78.90	58.17	58.57	53.19	79.00	58.72	64.43	61.08	82.80	81.02	80.72	80.86
	w/o Perception	82.80	60.85	65.99	56.85	81.80	62.99	69.81	65.90	87.90	86.66	87.39	86.96
	Reversed Order	87.20	67.71	63.98	58.27	87.10	73.96	68.25	70.85	87.60	87.16	85.99	86.47

Table 12: Detailed ablation results of MindStep on the four SocialAct event benchmarks. Results are reported across action type, action object, and action content prediction.