Beyond Inherent Cognition Biases in LLM-Based Event Forecasting: A Multi-Cognition Agentic Framework

Anonymous ACL submission

Abstract

Large Language Models (LLMs) exhibit strong reasoning capabilities and are widely applied in event forecasting. However, studies have demonstrated that LLMs exhibit human-like cognitive biases, systematic patterns of deviation from rationality in decision-making. To explore the cognitive biases in event forecasting, we introduce CogForecast, a human-curated dataset comprising six topics. Experimental results on three LLMs reveal significant cognitive biases in LLM-based event forecasting methods. To address this issue, we propose MCA, a Multi-Cognition Agentic framework. Specifically, MCA leverages LLMs to act as multi-cognition event participants, performing perspective-taking based on the cognitive patterns of event participants to alleviate the inherent cognitive biases in LLMs and offer diverse analytical perspectives. Then, MCA clusters agents according to their predictions and derives a final answer through a group-level reliability scoring method. Experimental results on a dataset including eight event categories demonstrate the effectiveness of MCA. Using Llama-3.1-70B, MCA achieves an accuracy of 82.3% (79.5% for the human crowd). Additionally, we demonstrate that MCA can alleviate the cognitive biases in LLMs and investigate three influencing factors. Our code and dataset will be publicly released.

1 Introduction

Recently, large language models (LLMs, Zhao et al., 2023) such as ChatGPT have shown remarkable reasoning capabilities across various applications, including event forecasting. Event forecasting (Granroth-Wilding and Clark, 2016; Zhao, 2022; Zhou et al., 2022) is a challenging task that aims to predict future developments based on the analysis of background information. Basically, LLM-based forecasting methods can be categorized into prompt engineering (Shi et al., 2023; Schoenegger et al., 2024), Retrieval-Augmented Generation (Liao et al., 2024; Luo et al., 2024), instruction tuning (Tao et al., 2024a; Yuan et al., 2024), and LLM-based agent methods (Ye et al., 2024; Cheng and Chin, 2024). These studies treat LLMs as objective analysts and contribute significantly to the progression of event forecasting.

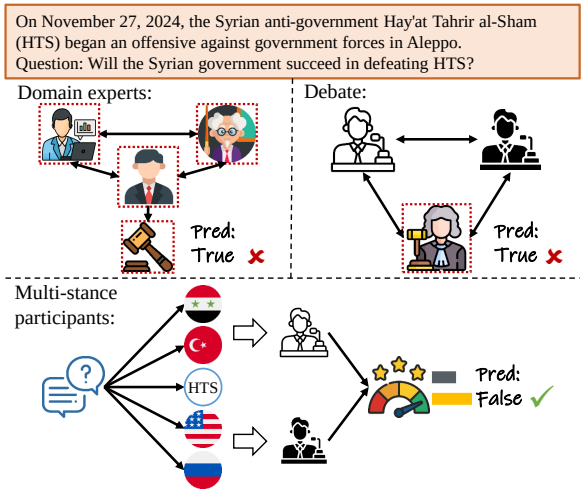

Figure 1: Comparison of different multi-agent methods. Agents with red dashed boxes inherit cognitive biases from LLMs, as demonstrated in Section 2.

However, as demonstrated in Talboy and Fuller (2023) and Echterhoff et al. (2024), LLMs inherit human-like cognitive biases from human-created data. The cognitive biases are systematic patterns of deviation from norm or rationality in decision-making, thus rendering LLM-based methods insufficient for objective decision-making. To investigate the cognitive biases in event forecasting, we introduce CogForecast, a human-curated dataset comprising six topics (each with a pair of entities). Using a cognitive preference score as the metric, three LLMs show significant cognitive biases. Furthermore, cognitive biases are also observed in agents using domain experts, such as political scholars and analysts, and the final judge in multi-

agent debate systems (MAD, Du et al., 2024; Liang et al., 2024), as depicted in Figure 1.

To mitigate the cognitive biases of LLMs, we propose MCA, a Multi-Cognition Agentic framework for complex event forecasting. The method is motivated by perspective-taking in cognitive theory, which is widely applied in international relations analysis. As illustrated in Figure 1, MCA profiles agents as multi-cognition event participants for perspective-taking, facilitating LLMs in shedding inherent cognitive biases and offering a comprehensive perspective. Specifically, MCA includes two stages: agent construction and forecasting. In the agent construction stage, MCA proposes an automatic agent construction method that clusters historical events and extracts multi-cognition participants, resulting in a large-scale agent collection. In the forecasting stage, given a question, MCA dynamically retrieves relevant multi-cognition agents from the agent collection. Subsequently, a retrieval assistant collects multilingual, multi-cognition information from news websites and YouTube to alleviate information cocoons. Based on retrieved information, agents engage in perspective-taking from the viewpoint of event participants, facilitating comprehensive analysis from diverse perspectives. Finally, to support objective collective decision-making (CDM), MCA clusters agents according to their predictions and derives a final answer using a group-level reliability scoring method.

In experiments, we evaluate MCA on a challenging forecasting benchmark, including eight categories. MCA demonstrates its superiority across four LLMs, yielding an average accuracy improvement of 4.7% (especially in the "Security" category). Notably, using Llama-3.1-70B as the LLM, MCA yields an accuracy of 82.3%, surpassing the human crowd's 79.5%. Additionally, we demonstrate that MCA can alleviate cognitive biases and explore three factors influencing the cognitive biases and prediction performance of LLMs, including agent profiling, information source, and cognitive certainty. Regarding the CDM, we compare various CDM mechanisms, highlighting the sensitivities of dictatorial and debating methods, and demonstrating the effectiveness of our method. Our contributions are as follows:

- We introduce a dataset, CogForecast, revealing the cognitive biases of LLM-based methods.
- We introduce MCA to alleviate the cognitive biases and achieve superior performance.
- We investigate three factors influencing the cognitive biases of LLMs, providing insights for future research.

2 The Cognitive Biases of LLMs in Event Forecasting

Dataset To address the dataset gap in assessing cognitive biases of event forecasting, we introduce CogForecast, a human-curated dataset comprising 6 topics (6 pairs of entities $\{T_i = [e_i^1, e_i^2]\}_{i=1}^{6}$, 218 samples). These entity pairs exhibit significant cognitive discrepancies, including *"US-China", "US-Iran", "Ukraine-Russia", "Palestine-Israel", "South Korea-North Korea", and "Syrian-HTS"*. Each sample contains a question and three options, such as: *"Question: In 2024, the Syrian opposition HTS succeeded in overthrowing the Assad government. Will Syria gain more freedom and democracy? Options: (A) Cannot answer; (B) Yes; (C) No"*. Given the significant cognitive divergence between e_i^1 and e_i^2, correctness evaluation, which annotates a correct answer for each question, results in serious inconsistencies among annotators. Therefore, for question q_i^j, we propose annotating the cognitive preferences p_j^b of option (B) and p_j^c of (C) from $\{e_i^1, e_i^2\}$. For the example above, p_j^b for option "(B) Yes" is e_i^2 "HTS", as this option aligns with the cognition of HTS. This annotation method demonstrates substantial agreement between two annotators, with a Fleiss' Kappa score of 96.7%. The dataset construction details can be found in Appendix A.2.

Metrics For topic T_i, prediction on question q_i^j is mapped to preference p_i^j. Then, for e_i^1, e_i^2, and neutral option "Cannot answer", we calculate their cognitive preference scores as:

$$P_e = \frac{\sum_{j=1}^{count(q_i^j)} 1_{(p_i^j = e)}}{count(q_i^j)}, e \in \{e_i^1, e_i^2, neutral\} \quad (1)$$

Results As depicted in Figure 2, when employing CoT (first row), three LLMs consistently exhibit a pronounced cognitive preference for e_i^1 (blue bar) over e_i^2 (gray bar) across all topics. Additionally, different LLMs show varying degrees of cognitive biases, with Llama-3-8B exhibiting the most significant biases. Using the similar prompt template as CoT, we evaluate three kinds of agents: ExpertPrompting (*You are an international relations*

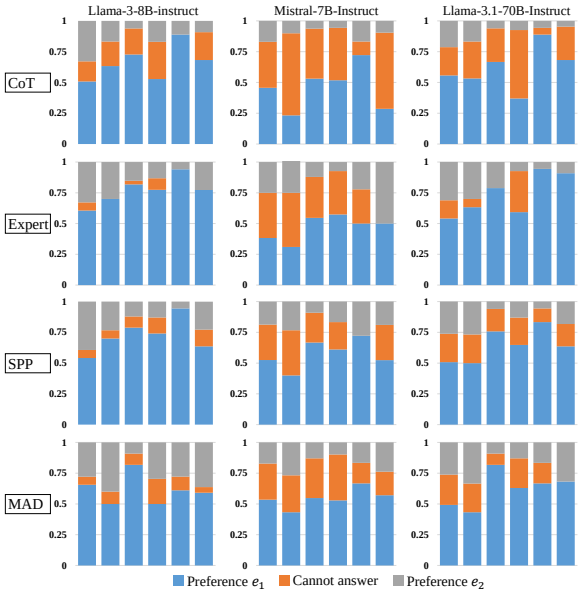

Figure 2: The cognitive biases of three LLMs using CoT, ExpertPrompting, SPP, and MAD.

analyst specializing in the analysis of e_i^1-e_i^2 relations), SPP (a multi-agent system that simulates collaboration among domain experts), and MAD (a two-round multi-agent debate system that simulates debates between affirmative and negative sides). However, compared to CoT, ExpertPrompting exacerbates the cognitive biases of LLMs. After incorporating additional experts, SPP exhibits minor fluctuations in cognitive bias across dialog rounds (average shift in preference e_i^1 is 4.3% across three LLMs). For MAD, while the affirmative and negative sides display different cognitive preferences (averaged preference difference is 11.3% across three LLMs), the final judge remains the cognitive preferences of LLMs.

3 Method

3.1 Task Definition and Framework Overview

Following the task definition of binary event forecasting in (Halawi et al., 2024), the objective is to predict answers (True/False) of binary forecasting questions and to assign occurring probabilities. Each data d consists of a question q, a background description, a resolution criterion, and four kinds of timestamps: a begin date $date_{begin}$ when the question is published, a close date $date_{close}$ when no further forecasts can be submitted on forecasting platform, a resolve date $date_{resolve}$ when the outcome is determined, and 1-5 retrieval dates $date_{retrieval}$ when the model can retrieve additional information up to this date. The retrieval dates are sampled between the $date_{begin}$ and $date_{close}$, as well as before $date_{resolve}$, to prevent knowledge leakage.

As illustrated in Figure 3, MCA consists of two stages: the multi-cognition agent construction stage and the forecasting stage. In the first stage, MCA constructs a large-scale collection of agents from the trainset. In the second stage, MCA retrieves relevant multi-cognition agents and leverages their collective intelligence for forecasting.

3.2 Multi-cognition Agent Construction

Unlike existing agent profiling methods, such as domain experts and debating roles, MCA profiles agents as event participants and encourages LLMs to "step into the other person's shoes". However, complex events often involve potential participants not explicitly referenced in a single question, and similar events may share participants. Consequently, we propose an automatic agent construction method to extract agents. MCA first utilizes a text embedding model, *bge-large-en-v1.5*, to extract embeddings for all questions from the training set. Subsequently, following BERTopic (Grootendorst, 2022), we apply UMAP to reduce the embedding dimension to 100 and HDBSCAN (with $min\_samples$ and $min\_cluster\_size$ set to 3 and 7) to cluster questions into 237 topics. For each topic cluster, we concatenate questions and background information as textual input and prompt LLM to identify relevant multi-cognition entities (agents). Additionally, for each agent, we generate four attributes: (1) type (e.g., country, organization, individual); (2) a brief description; (3) professional field (e.g., Politics & Governance, Security & Defense); and (4) official languages. Finally, agents sharing the same name, type, and professional field are aggregated, resulting in 2,496 distinct agents.

3.3 Multi-cognition Event Forecasting

Step 1: Multi-cognition Agent Retrieving. Given a question q_i and its background, MCA first prompts LLM to identify relevant agents and to generate their names, types, and professional fields. The multi-cognition agents include three types: (1) the affirmative side, which argues that the event is more likely to occur and may benefit from it; (2) the negative side, which argues that the event is less likely to occur and may be adversely affected; (3) the neutral side, such as neutral international organizations and domain experts. Sub-

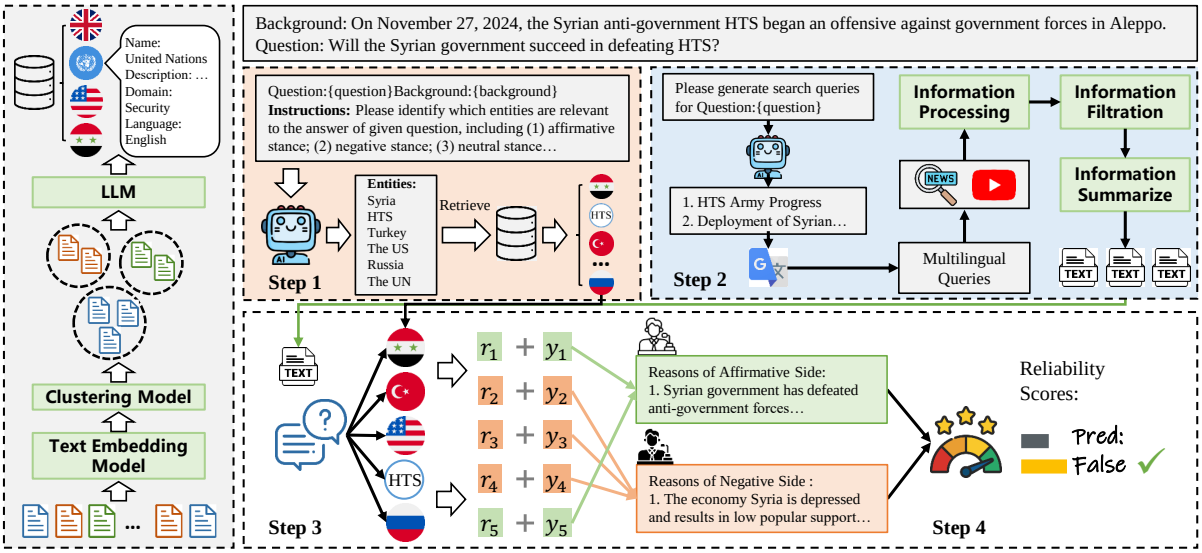

Figure 3: Illustration of the agent construction and forecasting pipeline for MCA. The forecasting stage includes four steps: (1) multi-cognition agent retrieving; (2) multilingual information retrieving; (3) multi-cognition reasoning, and (4) group-level cognition aggregating.

sequently, we employ text matching (name, type, and professional field) to retrieve agents $A_i = \left\{a_i^1, a_i^2 \dots a_i^j\right\}$ from the agent collection. Those unmatched agents will be created and added to the agent collection.

Step 2: Multilingual Information Retrieving. As highlighted in Yang (2024), information cocoons may exacerbate cognitive biases in both humans and LLMs. Therefore, unlike Halawi et al. (2024); Guan et al. (2024), which retrieve monolingual data from news websites, a retrieval assistant retrieves multilingual, multi-cognition information from news websites and YouTube. There are five steps: (1) Search Query Generation, (2) Information Retrieval, (3) Information Processing, (4) Information Filtration and (5) Information Summarizing. Details are provided in Appendix A.3.

Step 3: Multi-cognition Reasoning. As illustrated in Figure 3, j retrieved agents exhibit diverse identities and cognition, thereby facilitating multi-cognition reasoning and diverse predictions. Using prompting method, we convert each agent profile into textual prompt $p_{profile}$ (*You are an AI agent who specializes in event forecasting, and here is your profile. Name: {name} Type: {type} Description: {description} Professional field: {domain} Please answer the following question from your perspective and objectively.*) Subsequently, MCA obtains the forecasting prompt by concatenating $p_{profile}$ with the question prompt $p_{question}$ of data d_i and the chain-of-thought (CoT) prompt $p_{instruct}$.

$$P_{reasoning} = p_{profile} \oplus p_{question} \oplus p_{instruct} \quad (2)$$

For a fair comparison, we select the best $p_{question}$ and $p_{instruct}$ from Halawi et al. (2024) based on accuracy and apply this prompt template across all methods. Then, j agents leverage LLM M to perform CoT reasoning and obtain their forecasting results Y_i of data d_i.

$$Y_i = \left\{[a_i^1, r_i^1, y_i^1], [a_i^2, r_i^2, y_i^2] \dots [a_i^j, r_j^j, y_i^j]\right\} \quad (3)$$

where r_i^j and y_i^j denote the textual reasoning and prediction probability for agent a_i^j.

Step 4: Group-level Cognition Aggregating. As demonstrated in Zhao et al. (2024), dictatorial methods, which designate a special agent to determine the final decision, are fragile due to their complete reliance on a single agent. In this work, we introduce a reliability scoring agent A_{CDM} to leverage the collective intelligence of multi-cognition agents. Despite the diversity of cognition, certain agents may share overlapping viewpoints. Therefore, A_{CDM} first divides Y_i into groups.

$$G_y = \left\{[a_i^k, r_k^k, y_i^k] \mid y_i^k = y\right\} \quad (4)$$

where G_y denotes the group with y_i^k equal to y. Taking the binary forecasting task as an example,

Methods	Security		Politics		Economics		Sports		Technology		All	
	Brier ↓	Acc ↑	Brier ↓	Acc ↑	Brier ↓	Acc ↑	Brier ↓	Acc ↑	Brier ↓	Acc ↑	Brier ↓	Acc ↑
Human Crowd	**0.129**	**78.4**	**0.145**	**78.2**	**0.147**	**78.3**	**0.171**	**73.1**	**0.114**	**84.3**	**0.149**	**77.0**
Claude-2.1	/	/	/	/	/	/	/	/	/	/	0.215	/
GPT-4-1106	0.188	69.6	0.184	71.8	0.213	64.9	0.181	71.1	0.152	80.2	0.190	69.6
+3CoT	0.180	70.8	0.181	70.6	0.209	65.7	0.178	72.1	0.151	79.7	0.186	70.2
+3SFT+3CoT	0.174	71.0	0.172	72.6	0.198	68.8	0.175	73.0	0.143	71.5	0.179	71.5
Llama-3-8B	0.236	60.5	0.205	68.7	0.222	61.5	0.190	72.5	0.149	78.9	0.204	68.1
+ExpertPrompt	0.24	59.7	0.206	69.2	0.233	62.2	0.196	69.5	0.176	75.1	0.210	67.4
+Self Consistency	0.227	62.7	0.196	71.6	0.211	67.0	0.193	70.7	0.157	78.0	0.201	69.9
+SPP	0.245	57.8	0.253	60.9	0.217	65.2	0.229	63.5	0.205	69.6	0.239	61.0
+MAD	0.296	42.4	0.297	43.3	0.285	43.9	0.271	50.0	0.287	49.1	0.285	45.8
+MCA	**0.204**	**74.6**	**0.187**	**75.9**	**0.202**	**74.4**	**0.182**	**73.6**	**0.141**	**86.0**	**0.194**	**74.3**
Δ	-0.023	+11.9	-0.009	+3.3	-0.009	+7.4	-0.008	+1.1	-0.008	+7.1	-0.007	+4.4
Human Crowd*	**0.103**	**84.1**	**0.112**	**81.3**	**0.143**	**79.7**	**0.176**	**71.9**	**0.066**	**94.9**	**0.133**	**79.5**
Llama-3.1-70B	0.189	68.3	0.134	79.5	0.150	71.9	0.170	74.5	0.070	91.8	0.162	74.2
+Self Consistency	0.172	74.3	**0.123**	82.3	0.145	73.1	0.161	78.6	0.060	92.9	0.152	77.8
+MCA	**0.122**	**93.4**	0.129	**85.0**	**0.133**	**76.0**	**0.155**	**79.0**	**0.052**	**95.3**	**0.145**	**82.3**
Δ	-0.050	+19.1	+0.006	+2.7	-0.012	+2.9	-0.006	+0.4	-0.008	+2.4	-0.007	+4.5

Table 1: Comparison between our MCA and other methods. The lower part presents the results on the test subset.

we divide predictions into two groups:

$$Y_i = G_{true} \cup G_{false} \tag{5}$$

$$G_i^{true} = \left\{ [a_i^k, r_k^k, y_i^k] \mid y_i^k > 0.5 \right\} \tag{6}$$

$$G_i^{false} = \left\{ [a_i^k, r_k^k, y_i^k] \mid y_i^k \leq 0.5 \right\} \tag{7}$$

where G_{true} and G_{false} denote the agent groups that predict event as more likely or less likely to occur, respectively. Subsequently, using an aggregation prompt, A_{CDM} aggregates their textual reasoning to provide comprehensive reasoning. Then, A_{CDM} evaluates their reliability scores S_i^{true} and S_i^{false} (0.0-1.0, with 0.7 indicating unchanged reliability) based on their reasoning rationality. The final prediction is derived as the weighted average of all predictions to avoid cognitive bias in dictatorial judgment:

$$y_{final} = \frac{1}{j} \sum_{k=1}^{j} \frac{y_i^k}{0.7} \cdot \left(S_i^{true} \cdot 1_{(y_i^k > 0.5)} + S_i^{false} \cdot 1_{(y_i^k \leq 0.5)} \right) \tag{8}$$

3.4 Collective Experience Acquisition

Capability acquisition is a critical process in agents, enabling dynamic learning and evolution. Drawing inspiration from trial-and-error learning, we integrate an experience memory into each cognitive agent and A_{CDM}. After collective prediction on a training sample, we check the correctness of cognitive agents in multi-cognition reasoning (predictions vs label) and A_{CDM} (whether the aggregated score y_{final} is better than averaging). For agents with mistakes, they are prompted to revise, add, or delete their memory.

4 Experiments

4.1 Experimental Setup

Datasets such as ICEWS (García-Durán et al., 2018) and SCTc-TE (Ma et al., 2023) are widely adopted. However, the most recent knowledge cutoff of these datasets is 2022, resulting in knowledge leakage for LLMs. Therefore, we employ the dataset released by Halawi et al. (2024), which contains 5,516 binary forecasting questions, including 3,762 questions for training, 840 for validation, and 914 for testing (published after June 1, 2023). The dataset is curated from platforms such as Metaculus, including 8 categories such as "Security" and "Politics". These platforms aggregate predictions from individual forecasters, providing a strong benchmark: the *Human Crowd*.

Models. To thoroughly assess the performance of MCA, we employ four LLMs for comparison: Llama-3-8B-Instruct, Mistral-7B-Instruct-v0.2, Llama-3.1-8B-Instruct, and Llama-3.1-70B-Instruct. To avoid knowledge leakage for the latter three LLMs, we create a test subset comprising instances with resolve dates after December 2023. Furthermore, we select a variety of competitive methods for comparison: (1) Human Crowd, the collective intelligence of human forecasters; (2) GPT-4 and its variants from Halawi et al. (2024); (3) CoT, which elicits step-by-step reasoning of LLMs; (4) Self-Consistency, which samples multiple (n=10) reasoning paths and uses the averaged prediction as final answer; (5) ExpertPrompting (Xu et al., 2023), which dynamically generates

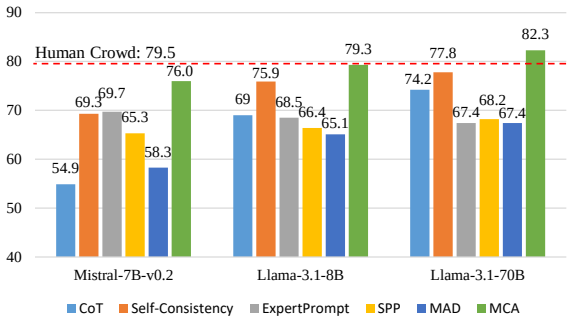

Figure 4: Comparison of Accuracy between MCA and other methods on three LLMs.

MA	MR	EM	GA	Llama-3-8B Brier	Llama-3-8B Acc	Mistral-7B Brier	Mistral-7B Acc
✓	✓	✓	✓	**0.194**	**74.3**	0.181	**76.0**
✗	✓	✓	✓	0.204	71.6	0.187	73.5
✓	✗	✓	✓	0.204	72.1	0.189	75.2
✓	✓	✗	✓	0.194	73.4	0.187	74.3
✓	✓	✓	✗	0.200	71.9	0.180	73.7
✗	✗	✗	✗	0.205	67.7	0.192	65.3
✓	✗	✗	✗	0.206	70.0	0.188	69.9
✓	✓	✗	✗	0.198	71.2	0.185	72.1
✓	✓	✓	✗	0.198	71.1	**0.180**	73.7

Table 2: Ablation results of MCA on two LLMs. MA, MR, EM, and GA denote the multi-cognition agents, multilingual retrieval assistant, experience memory, and group-level aggregating, respectively.

a domain expert to facilitate LLMs to answer as distinguished experts; (6) MAD (Liang et al., 2024), which employs a two-round debate, moderated by a judge; (7) SPP (Wang et al., 2024b), which engages in multi-turn collaboration with diverse domain experts. To ensure fairness, all methods utilize uniform prompt templates ($p_{question}$ and $p_{instruct}$) and multilingual information retriever, except for the necessary descriptive prefixes for methods. Implementation details are provided in Appendix A.1.

Metrics. We employ accuracy and Brier score as the metrics. Denoting $f_i \in [0,1]$ as the i-th probabilistic prediction and $o_i \in \{0,1\}$ as the gold answer, the accuracy is defined as $\frac{1}{n}\sum_{i=1}^{n} 1\{1\{f_i > 0.5\} = o_i\}$, while the Brier score is computed as $\frac{1}{n}\sum_{i=1}^{n}(f_i - o_i)^2$. For reference, an unskilled forecaster with a constant value of 0.5 yields a Brier score of 0.25. These metrics are averaged across all retrieval dates.

4.2 Experimental Results

Main Results. Table 1 presents the detailed comparisons between MCA and other methods. Figure 4 further shows a comparison across more LLMs. The experiments demonstrate that MCA consistently outperforms other methods by a significant margin across four LLMs, with an average accuracy improvement of 4.7% and a decrease of 0.008 in Brier score compared to the second-best results. Notably, when using Llama-3.1-70B, MCA surpasses the challenging *human crowd* (82.3% vs 79.5% in accuracy). Additionally, we observe that: (1) Compared to single-agent baselines (CoT and ExpertPrompting), MCA achieves substantial performance gains, outperforming CoT by 11.4% and ExpertPrompting by 9.7% across four LLMs, highlighting the necessity of MCA. (2) MCA excels in predicting events with complex cognition, achieving the highest accuracy gains in the "Security" category, which involves diverse countries and organizations with varying cognition. (3) Ensemble methods (self-consistency, GPT4+3CoT) consistently outperform vanilla CoT. (4) Debating method, MAD, surprisingly yields the poorest performance, as also demonstrated in Smit et al. (2024). We check the debating process and find a decline in accuracy as debate rounds increase, particularly in the first round, when the opposing side rebuts the affirmative side. (5) ExpertPrompting and SPP exhibit a performance decline over CoT. Additionally, we observe negligible variations in accuracy for SPP across conversation rounds, probably due to shared cognitive biases among domain experts.

Ablation Results. In the upper section of Table 2, replacing multi-cognition agents with domain experts, replacing multi-cognition retrieval with English news retrieval, removing experience memory, and replacing group-level aggregation with vanilla averaging all lead to a decline in performance, demonstrating their effectiveness. Additionally, in the lower section of the table, there is a consistent performance improvement after incrementally incorporating four modules.

4.3 Discussion

RQ1: Can MCA alleviate the cognitive biases? Using CogForecast, we employ e_i^1 and e_i^2 as event participants (agents) to perform perspective-taking and treat them as two groups for group-level aggregating. As depicted in the first (e_i^1) and the second rows (e_i^2) of Figure 5, LLMs exhibit significant cognitive preferences to given identities, demonstrating the perspective-taking capabilities

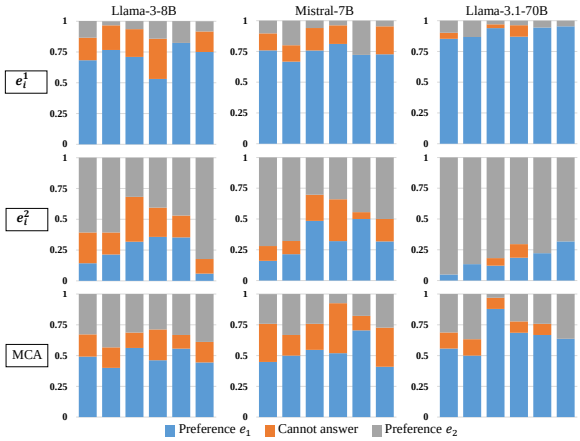

Figure 5: The cognitive preference analysis of MCA.

Var	Setting	Llama-3-8B Kappa	Llama-3-8B Acc	Mistral-7B Kappa	Mistral-7B Acc
Profiles	ABC	0.479	71.2	0.624	71.0
	Domain Experts	0.443	69.5	0.485	72.4
	Debater	0.168	68.1	0.298	73.1
	MCA	0.401	**73.4**	0.412	**80.3**
Info	No RAG	0.255	60.9	0.264	70.2
	YouTube	0.383	65.4	0.331	71.2
	News	0.384	69.6	0.357	78.9
	News+YouTube	0.402	70.9	0.420	79.0
	Multilingual	0.401	**73.4**	0.412	**80.3**
Certainty	Absolute	0.372	70.7	0.408	78.2
	Strong	0.363	70.0	0.373	79.3
	Balanced	0.401	**73.4**	0.412	**80.3**
	Low	0.391	72.4	0.364	77.2

Table 3: Analysis of three influencing factors.

of LLMs. After aggregation (third row), LLMs are prompted to ignore inherent cognition and answer objectively according to the rationality of e_i^1 and e_i^2, thereby alleviating the cognitive biases of LLMs compared to other methods (Figure 2).

RQ2: The influencing factors of cognitive biases and forecasting performance. We investigate three factors influencing cognitive biases in LLMs and multi-agent forecasting systems as follows. Except for prediction accuracy, we incorporate Fleiss' kappa to assess the degree of agreement among agents and conduct experiments across four challenging event categories: security, politics, economics, and technology.

(1) Agent Profiling. To make a comprehensive comparison, we employ three additional agent profiling methods: (1) vanilla expert ABC, including four agents with the name "1-4"; (2) domain experts, where four human-crafted experts are assigned to each category, such as "Security & Defense Scholars" and "Politics & Governance Analysts". (3) debater, including three agents representing the affirmative, negative, and neutral sides. For a fair comparison, the prompt template (except for profile prompt for agent) and information source (multilingual) remain consistent. As shown in Table 3, MCA achieves the highest accuracy, whereas domain experts yield moderate performance. Furthermore, in the Fleiss' Kappa columns, debater agents exhibit the lowest inter-agent agreement, as they are deliberately assigned opposing positions. In contrast, domain experts and vanilla ABC agents inherit the cognitive biases of LLMs, thus demonstrating higher agreement levels. For MCA, agents are profiled as multi-cognition participants, such as the US government and Russian troops, and these agents change the inherent cognition of LLMs, thus offering diverse perspectives (with low agreement). **(2) Information Source.** In the experiments, there is a continuous improvement in performance after progressively adding more information sources, underscoring the necessity of background information. Additionally, in social cognition theory, increased exposure to information with certain cognition will result in an enhanced cognitive identity. Such phenomenon is reflected in the increase of inter-agent agreement between multi-cognition agents from "No RAG" to monolingual "News+YouTube". Notably, after incorporating multilingual information, continuous improvements in accuracy and reduced agreements are observed. The multilingual information exhibits various cognition, thus facilitating diverse thinking and further alleviating cognitive bias.

(3) Cognitive Certainty refers to the degree of confidence a person has in their cognition. To investigate its impact, we examine four certainty degrees using prompts: (1) absolute certainty, fully aligned with the given identity; (2) strong certainty, permitting the incorporation of some objective perspectives; (3) balanced certainty, analyzing from the given perspective and objectively ; (4) low certainty, adopting a completely objective viewpoint. As depicted in Table 3, in certainty-enhanced settings, absolute and strong certainty levels yield lower accuracies, as agents overestimate their judgments and ignore conflicting evidences. Despite increased objectivity, low certainty setting leads to performance degradation. Therefore, a balanced cognitive certainty is recommended, as it offers optimal performance by combining perspectives

Methods	Llama-3-8B		Mistral-7B	
	Brier	Acc	Brier	Acc
Average	0.200	71.9	**0.180**	73.7
Plurality	0.279	72.0	0.258	74.2
Plurality_score	0.195	72.0	0.186	74.2
Dictatorial	0.202	67.0	0.212	50.2
Dictatorial_group	0.198	68.7	0.187	63.7
Debate	0.210	70.1	0.213	59.1
Ours	**0.194**	**74.3**	0.181	**76.0**
Δ	-0.001	+2.3	+0.001	+1.8

Table 4: Comparison of various CDM methods.

beyond the inherent cognition and objectivity.

RQ3: The impact of CDM mechanisms in prediction performance? Except for averaging method and our group-level aggregating, we examine three CDM mechanisms in MCA: (1) plurality voting, which selects the option (True/False) of the first-preference votes, and its variant, which adopts the averaged score from the selected group; (2) dictatorial, where a judge agent determines the final prediction based on all agents or aggregated groups; (3) debate, which involves two-round debates between aggregated groups before the final judge. Results in Table 4 show that our method outperforms other methods. Additionally, both dictatorial and debate methods rely on a judge and thus obtain accuracies close to CoT.

5 Related Works

Cognitive biases in LLMs. Studies have extensively explored social biases towards protected groups in LLMs, such as gender and religious bias. Differently, cognitive biases focus on decision-making. Talboy and Fuller (2023) demonstrate the presence of various cognitive biases in LLMs. Echterhoff et al. (2024) develop a dataset to evaluate three categories of cognitive biases in campus enrollment task, such as sequential bias. Bang et al. (2024) investigate the biases of LLMs regarding political issues. Xie et al. (2024) construct MindScope, a cognitive bias evaluation dataset that incorporates multi-turn dialogue scenarios. Mina et al. (2025) demonstrate that cognitive biases in LLMs tend to be more pronounced as task complexity increases. Beyond prompt or option sequence, cognitive biases in event forecasting are influenced by intricate and underexplored factors, necessitating investigation and effective mitigation strategies.

Event Forecasting. Early studies address event forecasting as a text classification task, modeling event chains (Wang et al., 2021), event graphs (Du et al., 2022), and unstructured text (Jin et al., 2021) through small language models or graph neural networks (Zhang et al., 2023). Recently, LLM-based forecasting methods have arisen. Lee et al. (2023); Shi et al. (2023) introduce various prompting methods to leverage the reasoning ability of LLMs. To augment LLMs with current information, researchers retrieve structured events (Liao et al., 2024) or news (Guan et al., 2024; Halawi et al., 2024). Instruction tuning methods are also employed to enhance the reasoning ability (Tao et al., 2024a,b) and interpretability of LLMs (Yuan et al., 2024). Additionally, LLM-based agent frameworks (Ye et al., 2024; Cheng and Chin, 2024) profile LLMs as agents with various capabilities. Despite their significant contributions, these studies treat LLMs as objective analysts, a premise that is proven invalid in our work.

LLM-based Multi-Agent Systems. Compared to single-agent systems, multi-agent systems leverage the collective intelligence of multiple agents, yielding superior performance on complex tasks such as software development (Qian et al., 2024; Hong et al., 2024), society simulation (Kaiya et al., 2023; Jin et al., 2024), and gaming (Wang et al., 2023). In agent profiling, agents are defined as roles tailored to specific tasks (Qian et al., 2024; Cheng and Chin, 2024), domain experts (Xu et al., 2023; Wang et al., 2024b), simulated personas (Kaiya et al., 2023), etc. In agent communication, Hong et al. (2024) simulate the software development workflow, Wang et al. (2024b,a) facilitate the cooperation of agents for a shared goal, and Park et al. (2024); Liang et al. (2024) introduce multi-agent debate systems to enhance reasoning capabilities.

6 Conclusion

In this work, we propose a dataset, CogForecast, and reveal the cognitive biases in LLM-based forecasting methods. To alleviate this issue, we propose a multi-cognition agentic framework, characterized by facilitating LLMs in perspective-taking as event participants and comprehensive perspectives. Extensive experiments demonstrate the superior performance of MCA and the effectiveness in mitigating cognitive biases. Additionally, we investigate three influencing factors in cognitive biases, shedding light on future research. Future work will focus on eliminating the inherent cognitive biases in LLMs and improving perspective-taking ability.

Limitations

In this section, we discuss several limitations in our works. First, to alleviate the cognitive biases in LLMs, MCA profiles agents as multi-cognition event participants, which perform perspective-taking to provide perspective beyond inherent cognitive patterns. As demonstrated in Figure 5, the perspective-taking ability is proved effective across various LLMs. However, weaker LLMs, such as Mistral-7B, might struggle to simulate roles with seriously opposing cognition, such as simulating "Russia" in "Russia-Ukraine" topic. Therefore, future work will focus on enhancing role-playing capabilities and further reducing the inherent cognitive biases in LLMs. Second, MCA introduces additional computational overhead compared to single-agent approaches. While it achieves significant performance improvements and effectively mitigates cognitive biases, the increased cost remains a concern. To address this, future work will explore strategies to reduce computational burden, such as leveraging lightweight LLMs for specific sub-tasks like multilingual information retrieval and multi-cognition reasoning.

Ethics Statement

In our study, we investigate the cognitive biases in LLM-based forecasting methods and introduce a multi-cognition agentic framework to alleviate these biases. Cognitive biases are systematic deviations from normative or rational decision-making processes. Through our framework, LLMs can offer a more comprehensive and objective perspective on event forecasting, thereby mitigating the risk of cognitive biases regarding various topics, such as politics, economics, and international relations. We emphasize the importance of maintaining objectivity throughout our research, adhering to the ethical principle of impartiality in scientific inquiry. Our goal is to contribute responsibly and constructively to the advancement of AI technologies.

References

Yejin Bang, Delong Chen, Nayeon Lee, and Pascale Fung. 2024. Measuring political bias in large language models: What is said and how it is said. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 11142–11159. Association for Computational Linguistics.

Junyan Cheng and Peter Chin. 2024. Sociodojo: Building lifelong analytical agents with real-world text and time series. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Li Du, Xiao Ding, Yue Zhang, Ting Liu, and Bing Qin. 2022. A graph enhanced BERT model for event prediction. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 2628–2638. Association for Computational Linguistics.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2024. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Jessica Maria Echterhoff, Yao Liu, Abeer Alessa, Julian J. McAuley, and Zexue He. 2024. Cognitive bias in decision-making with llms. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 12640–12653. Association for Computational Linguistics.

Alberto García-Durán, Sebastijan Dumancic, and Mathias Niepert. 2018. Learning sequence encoders for temporal knowledge graph completion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 4816–4821. Association for Computational Linguistics.

Mark Granroth-Wilding and Stephen Clark. 2016. What happens next? event prediction using a compositional neural network model. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, pages 2727–2733. AAAI Press.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based TF-IDF procedure. *CoRR*, abs/2203.05794.

Yong Guan, Hao Peng, Xiaozhi Wang, Lei Hou, and Juanzi Li. 2024. Openep: Open-ended future event prediction. *CoRR*, abs/2408.06578.

Danny Halawi, Fred Zhang, Chen Yueh-Han, and Jacob Steinhardt. 2024. Approaching human-level forecasting with language models. *CoRR*, abs/2402.18563.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. 2024. Metagpt: Meta programming for A multi-agent collaborative framework. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Woojeong Jin, Rahul Khanna, Suji Kim, Dong-Ho Lee, Fred Morstatter, Aram Galstyan, and Xiang Ren.

2021. Forecastqa: A question answering challenge for event forecasting with temporal text data. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 4636–4650. Association for Computational Linguistics.

Yiqiao Jin, Qinlin Zhao, Yiyang Wang, Hao Chen, Kaijie Zhu, Yijia Xiao, and Jindong Wang. 2024. Agentreview: Exploring peer review dynamics with LLM agents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 1208–1226. Association for Computational Linguistics.

Zhao Kaiya, Michelangelo Naim, Jovana Kondic, Manuel Cortes, Jiaxin Ge, Shuying Luo, Guangyu Robert Yang, and Andrew Ahn. 2023. Lyfe agents: Generative agents for low-cost real-time social interactions. *CoRR*, abs/2310.02172.

Dong-Ho Lee, Kian Ahrabian, Woojeong Jin, Fred Morstatter, and Jay Pujara. 2023. Temporal knowledge graph forecasting without knowledge using in-context learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 544–557. Association for Computational Linguistics.

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, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 17889–17904. Association for Computational Linguistics.

Ruotong Liao, Xu Jia, Yangzhe Li, Yunpu Ma, and Volker Tresp. 2024. Gentkg: Generative forecasting on temporal knowledge graph with large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 4303–4317. Association for Computational Linguistics.

Ruilin Luo, Tianle Gu, Haoling Li, Junzhe Li, Zicheng Lin, Jiayi Li, and Yujiu Yang. 2024. Chain of history: Learning and forecasting with llms for temporal knowledge graph completion. *CoRR*, abs/2401.06072.

Yunshan Ma, Chenchen Ye, Zijian Wu, Xiang Wang, Yixin Cao, Liang Pang, and Tat-Seng Chua. 2023. Structured, complex and time-complete temporal event forecasting. *CoRR*, abs/2312.01052.

Mario Mina, Valle Ruíz-Fernández, Júlia Falcão, Luis Vasquez-Reina, and Aitor Gonzalez-Agirre. 2025. Cognitive biases, task complexity, and result intepretability in large language models. In *Proceedings of the 31st International Conference on Computational Linguistics, COLING 2025, Abu Dhabi, UAE, January 19-24, 2025*, pages 1767–1784. Association for Computational Linguistics.

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, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 20963–20987. Association for Computational Linguistics.

Chen Qian, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Jiahao Li, Cheng Yang, Weize Chen, Yusheng Su, Xin Cong, Juyuan Xu, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2024. Chatdev: Communicative agents for software development. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 15174–15186. Association for Computational Linguistics.

Philipp Schoenegger, Indre Tuminauskaite, Peter S. Park, and Philip E. Tetlock. 2024. Wisdom of the silicon crowd: LLM ensemble prediction capabilities rival human crowd accuracy. *CoRR*, abs/2402.19379.

Xiaoming Shi, Siqiao Xue, Kangrui Wang, Fan Zhou, James Y. Zhang, Jun Zhou, Chenhao Tan, and Hongyuan Mei. 2023. Language models can improve event prediction by few-shot abductive reasoning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Andries P. Smit, Nathan Grinsztajn, Paul Duckworth, Thomas D. Barrett, and Arnu Pretorius. 2024. Should we be going mad? A look at multi-agent debate strategies for llms. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Alaina N. Talboy and Elizabeth Fuller. 2023. Challenging the appearance of machine intelligence: Cognitive bias in llms. *CoRR*, abs/2304.01358.

Zhengwei Tao, Xiancai Chen, Zhi Jin, Xiaoying Bai, Haiyan Zhao, and Yiwei Lou. 2024a. EVIT: event-oriented instruction tuning for event reasoning. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 8966–8979. Association for Computational Linguistics.

Zhengwei Tao, Zhi Jin, Junqiang Huang, Xiancai Chen, Xiaoying Bai, Yifan Zhang, and Chongyang Tao. 2024b. MEEL: multi-modal event evolution learning. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 8912–8925. Association for Computational Linguistics.

Lihong Wang, Juwei Yue, Shu Guo, Jiawei Sheng, Qianren Mao, Zhenyu Chen, Shenghai Zhong, and Chen Li. 2021. Multi-level connection enhanced representation learning for script event prediction. In *WWW '21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021*, pages 3524–3533. ACM / IW3C2.

Shenzhi Wang, Chang Liu, Zilong Zheng, Siyuan Qi, Shuo Chen, Qisen Yang, Andrew Zhao, Chaofei Wang, Shiji Song, and Gao Huang. 2023. Avalon's game of thoughts: Battle against deception through recursive contemplation. *CoRR*, abs/2310.01320.

Xiaolong Wang, Yile Wang, Sijie Cheng, Peng Li, and Yang Liu. 2024a. DEEM: dynamic experienced expert modeling for stance detection. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 4530–4541. ELRA and ICCL.

Zhenhailong Wang, Shaoguang Mao, Wenshan Wu, Tao Ge, Furu Wei, and Heng Ji. 2024b. Unleashing the emergent cognitive synergy in large language models: A task-solving agent through multi-persona self-collaboration. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 257–279. Association for Computational Linguistics.

Zhentao Xie, Jiabao Zhao, Yilei Wang, Jinxin Shi, Yanhong Bai, Xingjiao Wu, and Liang He. 2024. Mindscope: Exploring cognitive biases in large language models through multi-agent systems. In *ECAI 2024 - 27th European Conference on Artificial Intelligence, 19-24 October 2024, Santiago de Compostela, Spain - Including 13th Conference on Prestigious Applications of Intelligent Systems (PAIS 2024)*, volume 392 of *Frontiers in Artificial Intelligence and Applications*, pages 3308–3315. IOS Press.

Benfeng Xu, An Yang, Junyang Lin, Quan Wang, Chang Zhou, Yongdong Zhang, and Zhendong Mao. 2023. Expertprompting: Instructing large language models to be distinguished experts. *CoRR*, abs/2305.14688.

Wen Yang. 2024. Information cocoons on social media: Why and how should the government regulate algorithms. *CoRR*, abs/2404.15630.

Chenchen Ye, Ziniu Hu, Yihe Deng, Zijie Huang, Mingyu Derek Ma, Yanqiao Zhu, and Wei Wang. 2024. MIRAI: evaluating LLM agents for event forecasting. *CoRR*, abs/2407.01231.

Chenhan Yuan, Qianqian Xie, Jimin Huang, and Sophia Ananiadou. 2024. Back to the future: Towards explainable temporal reasoning with large language models. In *Proceedings of the ACM on Web Conference 2024, WWW 2024, Singapore, May 13-17, 2024*, pages 1963–1974. ACM.

Mengqi Zhang, Yuwei Xia, Qiang Liu, Shu Wu, and Liang Wang. 2023. Learning long- and short-term representations for temporal knowledge graph reasoning. In *Proceedings of the ACM Web Conference 2023, WWW 2023, Austin, TX, USA, 30 April 2023 - 4 May 2023*, pages 2412–2422. ACM.

Liang Zhao. 2022. Event prediction in the big data era: A systematic survey. *ACM Comput. Surv.*, 54(5):94:1–94:37.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, and 3 others. 2023. A survey of large language models. *CoRR*, abs/2303.18223.

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, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 2712–2727. Association for Computational Linguistics.

Pengpeng Zhou, Bin Wu, Caiyong Wang, Hao Peng, Juwei Yue, and Song Xiao. 2022. What happens next? combining enhanced multilevel script learning and dual fusion strategies for script event prediction. *Int. J. Intell. Syst.*, 37(11):10001–10040.

A Appendix

A.1 Implementation Details

Except for the self-consistency method (0.7), the decoding temperature is set to 0.0 to ensure reproducibility. Experiments are conducted on four NVIDIA Tesla A100 GPUs with 80GB of RAM each.

A.2 The construction of CogForecast

As illustrated in Figure 6, the construction of CogForecast includes two stages: question generation and cognitive preference annotation. In the question generation, for the topic selection in CogForecast, we employed an expert in international relation analysis to list entity pairs exhibiting significant cognitive discrepancies. From these, the expert selected those that had attracted substantial international attention and remained relatively recent, resulting in pairs including "US-China", "US-Iran", "Ukraine-Russia", "Palestine-Israel", "South Korea-North Korea", and "Syrian-HTS". For each selected entity pair, the expert collected controversial issues spanning political, economic, cultural, and military domains, , leveraging diverse sources such as news media and Wikipedia. Subsequently, for each issue, the expert designed multiple event forecasting questions, each offering three options—with option "A" representing a neutral stance. For example: *"Question: In 2024, the Syrian opposition HTS succeeded in overthrowing the Assad government. Will Syria gain more freedom and democracy? Options: (A) Cannot answer; (B) Yes; (C) No"*. To ensure the quality of the dataset, a second expert was engaged to review and filter the generated questions. The evaluation dimensions are outlined as follows:

- *Avoiding Knowledge Leakage*: The resolution date of a question must not precede the knowledge cutoff date of the evaluated LLMs.
- *Question Relevance*: Question should pertain directly to a significant, controversial issue associated with the specified entity pair.
- *Question Clarity*: This criterion assesses whether the question clearly contextualizes the background of the associated event.
- *Cognitive Diversity*: Options "B" and "C" should reflect divergent cognitive preferences, with one aligning with entity 1 and the other with entity 2.

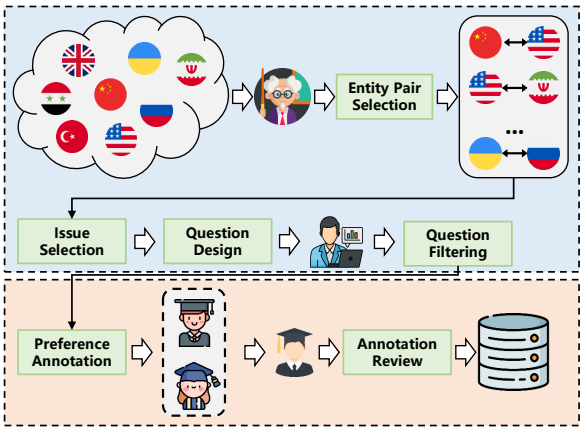

Figure 6: Illustration of the construction pipeline of CogForecast

In the cognitive preference annotation stage, excluding the neutral option, we engage two independent annotators[1] to determine the cognitive preference labels p_j^b for option "B" and p_j^c for option "C" from $\{e_i^1, e_i^2\}$. For each instance, annotators are required to investigate the topic background through Wikipedia and web searching. The annotation process adheres to the following criteria:

- *Background Familiarization*: Annotators must thoroughly investigate the background of given question and understand cognitive divergences between the entities.
- *Perspective-Taking Analysis*: For each entity, annotators perform perspective-taking to determine the option most aligned with the entity's stance. Justifications must reflect the official or mainstream position of the entity, rather than non-mainstream views.
- *Minimization of Personal Bias*: Annotators must ensure that the assigned labels represent the cognitive preferences of the entities themselves, independent of the annotators' personal beliefs or biases.

After annotation, we calculate the Fleiss' Kappa score to assess inter-group agreement, obtaining a score of 96.7%, which indicates substantial consistency. To resolve discrepancies between two annotators, a third annotator is employed to review and eliminate their discrepancies. The distribution of six topics in CogForecast is depicted in Table 5. See examples of CogForecast in Table 8.

[1]Graduate students specializing in event forecasting.

Index	Entity 1	Entity 2	Number
1	the US	China	61
2	the US	Iran	30
3	Ukraine	Russia	33
4	Palestine	Israel	54
5	South Korea	North Korea	18
6	Syrian	HTS	22

Table 5: The distribution of six topics in CogForecast.

Methods		Llama3-8b	Mistral-7b
Self-Consistency		69.9	69.3
MCA	100%	74.3	76.0
	75%	74.1	76.1
	50%	74.0	75.9
	25%	73.9	75.6

Table 6: The robustness of MCA on the size of agent collection.

	AR	MIR	SR	CDM	Total	Acc
CoT	/	11	1	/	12	66.6
ExpertPrompting	1	11	1	/	13	68.3
Self-Consistency	/	11	10	/	21	73.2
SPP	/	11	9	1	21	65.2
MAD	/	11	8	1	20	59.2
MCA	1	11	9.5	3	24.5	78.0

Table 7: The comparison of computational cost across various methods. The averaged accuracy across four LLMs is reported in the last column.

A.3 Details of Multilingual Information Retrieving

To retrieves multilingual, multi-cognition information from news websites and YouTube, the retrieval assistant employs the following steps:

(1) Search Query Generation. To provide comprehensive information coverage, following Halawi et al. (2024), the assistant leverages LLM to generate three English search queries based on the given question q_i and its background.

(2) Information Retrieval. To obtain multilingual search queries, MCA collects all official languages of agents A_i and translates English queries with Google Translation API. Subsequently, using these queries, the assistant retrieves articles from news APIs (NewsCatcher and Google News) and metadata of videos from YouTube Data API. All APIs are set with a cutoff date of $date_{retrieval}$ to avoid knowledge leakage.

(3) Information Processing. Given the limitations in multimodal and multilingual capabilities of LLMs, the assistant downloads YouTube audio and performs speech transcription using Whisper-large-v3-turbo. Subsequently, non-English articles from YouTube and news websites are identified and translated into English through Google Translation.

(4) Information Filtration. To eliminate articles of low relevance, MCA employs text embedding model *bge-large-en-v1.5* to generate embeddings for each question and retrieved articles. Subsequently, assistant computes the cosine similarities between question embedding and article embeddings and discards those articles with similarities below 0.65.

(5) Information Summarizing. Assistant retains the top-10 articles based on their similarity scores and prompts LLM to summarize related information to reduce context length.

The prompt templates of these steps are provided in Table 10.

A.4 Robustness of MCA on New Domain

In Step 1 (Multi-Cognition Agent Retrieving), MCA incorporates three agent types—affirmative, negative, and neutral to promote diversity. Those unretrieved agents will be created and added to the agent collection. Therefore, this strategy will ensure adaption for unseen forecasting scenarios and has negligible computing cost. Furthermore, to evaluate the robustness of MCA on new domain, we conducted experiments using a subset of high-frequency agents from the original set. As depicted in Table 6, the accuracy of MCA using different sizes of agent collection achieves similar accuracy, even with only 25% agents. Therefore, MCA exhibit good robustness on agent collection size, ensuring quick adaption to new domain.

A.5 Computational Cost

The primary computational overhead arises from LLM inference for LLM-based forecasting methods. As shown in Table 7, we categorize the inference cost into four stages: agent retrieving (AR), multilingual information retrieval (MIR), single-agent reasoning (SR), and collective decision-making (CDM), reporting the number of LLM inference calls required for each sample. The role of multi-cognition information is pivotal in mitigating

cognitive biases, as evidenced by results in Table 2 and Table 3. Compared to well-optimized single-agent baselines—-CoT (with the best-performing prompt from prior work) and ExpertPrompting (which simulates a domain expert)—-MCA introduce higher costs in SR and CDM stages. Nevertheless, it achieves substantial performance gains, outperforming CoT by 11.4% and ExpertPrompting by 9.7% across four LLMs, thereby highlighting the necessity and effectiveness of the multi-agent framework. Compared to other multi-agent approaches exhibiting similar reasoning costs, such as self-consistency, SPP, and MAD, MCA consistently yields superior performance. Additionally, as discussed in section 5.3 *Discussion——RQ1*, MCA employs e_i^1 and e_i^2 as event participants (agents) to perform perspective-taking. The first (e_i^1) and the second rows (e_i^2) of Figure 5 demonstrate that LLMs exhibit significant cognitive preferences to given identities, demonstrating the perspective-taking capabilities of LLMs. After aggregation, MCA reduces the cognitive biases compared to other baselines. These improvements indicate that the additional computational cost is justified by significant gains in both accuracy and bias mitigation. Notably, event forecasting is an important and challenging task, making it worthy of additional resources to replace costly human forecasters.

A.6 Prompt Template

The prompt templates for the multi-cognition agent construction stage are provided in Table 9. The prompt templates for the multi-cognition event forecasting stage are provided in Table 10 and Table 11.

Topic	Question	Options (Preference Labels)
Syrian government-HTS	In 2024, the Syrian anti-government armed group HTS won the war. Will HTS's coming to power improve the living standards of the Syrian people?	(A): Cannot answer (Neutral), (B): Yes (HTS), (C): No (Syrian government)
Russia-Ukraine	In 2024, Ukraine announced that it would strengthen its military cooperation with NATO, which Russia sees as a threat to its security. Which side is more likely to take military or diplomatic action to further aggravate the conflict between the two countries?	(A): Cannot answer (Neutral), (B): Russia (Ukraine), (C): Ukraine (Russia)
the US-China	In 2024, as China signs new military cooperation agreements with other countries, the United States portrays it as China's global military expansion and increases arms sales to the Philippines and Taiwan. Which country's actions will endanger world peace and stability?	(A): Cannot answer (Neutral), (B): the US (China), (C): China (the US)

Table 8: Examples of CogForecast.

Task	Prompt
Agent Generation	{question} \n Instructions: \n 1.Based on the above questions and backgrounds, please analyze which entities, including countries, organizations, political parties, companies, groups, and individual are related to it. Be careful not to overlook seemingly irrelevant but actually important entities, such as: the United States and China are important in international politics, powerful competitors in sports, competitors in business. \n 2.Output their entity types from country, organization, political party, company, group, and individual. \n 3.Briefly output their descriptions, each limited to a maximum of 50 words. For example, the description for "United states" is "a country primarily located in North America"; the description for "Elon Musk" is "a businessman and investor known for his key roles in the space company SpaceX and the automotive company Tesla, Inc. Other involvements include ownership of X Corp, the Boring Company, xAI, Neuralink, and OpenAI." \n The output format for each entity should be Name: xxx; Type: xxx; Description: xxx" such as "1.Name: Russia; Type: country; Description: a country spanning Eastern Europe and North Asia and is the largest country in the world by area; \n 2.Name: the Democratic Party of the United States; Type: political party; Description: one of the two major contemporary political parties in the United States \n".
Language Code Generation	{agent name} \n Instructions: \n Based on the above entity, please analyze the country to which the entity belongs and its 2-letter language code. If the entity is an international political organization and doesn't belong to any country, such as NATO, the country code should be "None". The language code should not be "None". The output format should be "Country:xxx; Language code:xxx" such as "Countries:Russia; Language code:RU".

Table 9: Prompt templates for the multi-cognition agent construction.

Task	Prompt
Agent Generation	Question: {question} \n Background: {background} \n Instructions: \n 1.Based on the above question and background, please identify which entities are relevant to the answer of given question, including countries, organizations, political parties, companies, groups, and individual. \n 2.Please identify the relevant entities from three stance, including (1) Positive stance (argue that the given event is more likely to occur, those who may benefit from the event), (2) neutral positions (no obvious interests or stance), and (3) Negative stance (argue that the given event is less likely to occur, those who may be harmed by the event, competitors). Be careful not to overlook seemingly irrelevant but actually important entities, such as: the United States and China are important in international politics, powerful competitors in sports, competitors in business. \n 3.Entities such as places, buildings, objects, concepts, etc. cannot answer the given question and should not be output. \n 4.Output their entity types from country, organization, political party, company, group, and individual. \n 5.Briefly output their descriptions, each limited to a maximum of 50 words. For example, the description for "United states" is "a country primarily located in North America"; the description for "Elon Musk" is "a businessman and investor known for his key roles in the space company SpaceX and the automotive company Tesla, Inc. Other involvements include ownership of X Corp, the Boring Company, xAI, Neuralink, and OpenAI." \n The output format for each entity should be Name: xxx; Type: xxx; Description: xxx" such as "1.Name: Russia; Type: country; Description: a country spanning Eastern Europe and North Asia and is the largest country in the world by area \n 2.Name: the Democratic Party of the United States; Type: political party; Description: one of the two major contemporary political parties in the United States".
Search query Generation	I will provide you with a forecasting question and the background information for the question. I will then ask you to generate short search queries (no more than 3 words each) that I'll use to find articles (using exact matching) on Google News to help answer the question. \n Question: \n {question}\n Question Background: \n {background} \n You must generate this exact amount of queries: 3 \n Start off by writing down sub-questions. Then use your sub-questions to help steer the search queries you produce. \n Your response should take the following structure: \n Thoughts: \n {{ Insert your thinking here. }} \n Search Queries:\n {{ Insert the queries here. Use semicolons to separate the queries. }}
Information Summarizing	I want to make the following article shorter (condense it to no more than 100 words). \n Article: \n —\n {article} \n — \n When doing this task for me, please do not remove any details that would be helpful for making considerations about the following forecasting question. \n Forecasting Question: {question} \n Question Background: {background}

Table 10: Prompt templates for the agent construction (step 1) and multilingual information retrieving (step 2) of multi-cognition event forecasting.

Task	Prompt
Single-Agent Prediction	You are an AI agent who specializes in event forecasting, and here's your profile. \n Name: {name} \n Type: {type} \n Description: {description} \n Professional field: {domain} \n Please answer the following question from your perspective and objectively. \n Question: \n {question} \n Question Background: {background} \n Resolution Criteria: \n {resolution_criteria} \n Today's date: {date_begin} \n Question close date: {date_end} \n We have retrieved the following information for this question: \n {retrieved_info} \n Instructions: \n 1. Provide reasons why the answer might be no. \n Insert your thoughts \n 2. Provide reasons why the answer might be yes. \n Insert your thoughts \n 3. Aggregate your considerations. \n {{ Insert your aggregated considerations }} \n 4. Output your answer (a number between 0 and 1) with an asterisk at the beginning and end of the decimal. \n {{ Insert your answer }}
Opinion Aggregation	I need your assistance with aggregating the reasoning from multiple AI agent forecasters. Here is the question and its metadata. \n Question: {question} \n Background: {background} \n Resolution criteria: {resolution_criteria} \n Today's date: {date_begin} \n Question close date: {date_end} \n The reasoning from AI agent forecasters: \n {reasoning} \n Instructions: \n Your goal is to aggregate the above reasonings, ensuring to merge similar analyses into one. \n The aggregated reasoning should be concise, capturing the essential elements. \n Be careful to output only the aggregated reasoning and not the answer. \n The output format should be like "Here is the aggregated reasoning: 1.The available information suggests that the cause of the plane crash that killed Yevgeny V. Prigozhin is still unknown, and the Russian authorities have not released any official findings on the matter. 2.While hand grenade fragments were found in the bodies of the victims, which suggests that the crash may have been intentional, the Kremlin has rejected US allegations that the crash was an assassination. 3.The Russian authorities have confirmed Prigozhin's death through genetic tests, but the cause of the crash remains unclear. 4.The Kremlin's statements have not provided any clear indication of Prigozhin's death, and the investigation is ongoing. 5.Considering the lack of conclusive evidence and the ongoing investigation, it is unlikely that Prigozhin's death will be confirmed as due to any cause before November 2023."
Reliability Scoring	I need your assistance with making a reliability analysis. Here is the question and its metadata. \n Question: \n {question} \n Question Background: {background} \n Resolution Criteria: \n {resolution_criteria} \n Today's date: {date_begin} \n Question close date: {date_end} \n In addition, I have generated a collection of predictions from two forecasters groups: \n Group 1 (likely to occur, prediction probability higher than 0.5), {group1_info} \n Group 2 (unlikely to occur, prediction probability lower than 0.5), {group2_info} \n Your goal is to score the reliability of two agent groups. \n Note: Reliability scores should follow the following definitions: \n 0.0 0.25: Extremely low reliability \n 0.25 0.5: Low reliability \n 0.5 0.75: Moderate reliability \n 0.75 0.9: High reliability \n 0.9 1.0: Very high reliability \n 1.If the reliability score is equal to 0.7 then the weight of the prediction will not be changed, if the reliability score is greater than 0.7 then the weight will be increased, if the reliability score is less than 0.7 then the weight will be decreased. \n 2.The sum of the reliability scores of the two groups need not equal 0. \n Rules: rules \n \n The output format should follow "Group 1: {{ insert the reliability score of group 1}}; Group 2: {{ insert the reliability score of group 2}}".

Table 11: Prompt templates for step 3 and 4 in multi-cognition event forecasting.