The Power of Simplicity in LLM-Based Event Forecasting

Meiru Zhang[♠] **Auss Abbood**[♠] **Zaiqiao Meng**[♠♢] **Nigel Collier**[♠]
[♠]Language Technology Lab, University of Cambridge
[♢]School of Computing Science, University of Glasgow
[♠]{mz468, aa2613, nhc30}@cam.ac.uk
[♢]zaiqiao.meng@glasgow.ac.uk

Abstract

Event forecasting is a challenging task that requires temporal reasoning over historical data. Although iterative reasoning agents following the ReAct paradigm bring improvements to event forecasting tasks, they also increase the cost of each prediction and bring challenges in tracing the information that contributes to the prediction. In this study, we simplify the ReAct framework into a retrieval-augmented generation (RAG) pipeline. Surprisingly, the RAG outperforms ReAct with only 10% of the token costs. Furthermore, our experiments reveal that structured statistical contexts significantly enhance forecasting accuracy, whereas introducing unstructured semantic information (e.g., news article titles) negatively impacts performance. In-depth analyses further highlight that the iterative reasoning traces impair forecasting accuracy in smaller-scale models but benefit larger models (e.g., 70B) in the event forecasting task. These insights underscore existing limitations in large language models' temporal and semantic reasoning abilities, providing critical guidance for developing more cost-effective and reliable forecasting systems.

1 Introduction

Temporal event forecasting, the capability to anticipate future events based on historical and current data, is crucial across domains such as climate change (Gillingham et al., 2018), finance (Christensen et al., 2018), and policy-making (Savio and Nikolopoulos, 2013), where timely and accurate predictions directly influence decision-making and strategic planning (Anastassopoulou et al., 2020).

Traditional forecasting approaches predominantly employ statistical techniques such as autoregression (Makridakis et al., 2008) or machine learning-based time-series models (Triebe et al., 2021). Recent advancements in Large Language Models (LLMs) have enabled novel approaches in this task, leveraging extensive textual resources

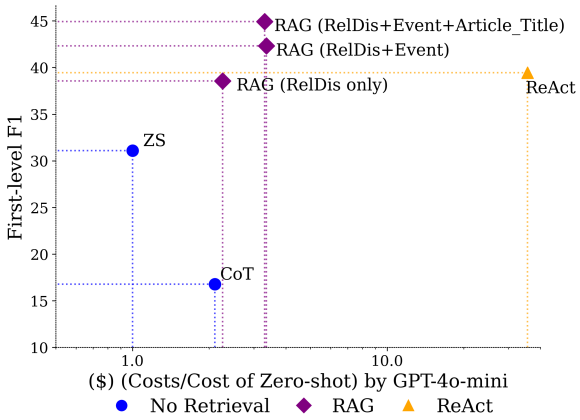

Figure 1: F1 score versus token cost per query (log scale) for different approaches using `GPT-4o-mini` as the backbone model. Costs are expressed as a ratio to the zero-shot (ZS) cost for comparability. RelDis, Event, and Article_Title represent information types provided to the RAG system (details in Section 2.2).

to predict international events and their potential outcomes (Wang et al., 2024; Chang et al., 2024; Wang et al., 2025). Lee et al. (2023) explored in-context learning for event forecasting, and retrieval-augmented generation (RAG) has also been applied and demonstrated the effectiveness of LLMs in solving this task (Sun et al., 2023; Liao et al., 2023). Agentic methods, which enable LLMs to autonomously interact with external knowledge sources and dynamically retrieve information, have become increasingly popular for event forecasting tasks. Wang et al. (2025) proposed a pipeline that incorporates both a reasoning agent and an evaluation agent for time series forecasting. Ye et al. (2024) introduced the MIRAI benchmark, which includes a contextual information database and a pre-defined API that allows LLM agents to interact with and retrieve data from the database.

The iterative API-based interaction between the LLM agent and the database enables step-by-step planning, allowing the model to rethink its reasoning based on additional information retrieved at

Proceedings of the 1st Workshop for Research on Agent Language Models (REALM 2025), pages 454–470

each step. However, despite this flexibility, our preliminary experiments show that ReAct incurs substantial token inference costs without yielding more accurate event forecasting predictions compared to RAG. As shown in Figure 1, `GPT-4o-mini`, when paired with RAG using the same retrieved data, achieves comparable or superior performance at only about 10% of the inference cost relative to ReAct on the MIRAI benchmark.[1] For visualization, the zero-shot cost is used as the baseline to scale the token costs of RAG and ReAct.

Building on the observation that RAG can outperform ReAct with significantly lower token costs, we systematically evaluate its generalizability across a range of LLM architectures. Additionally, we investigate the factors contributing to the performance differences between RAG and ReAct on the MIRAI benchmark.

The contribution of this work is as follows:

- We demonstrate that RAG, when combined with different types of contextual information and LLM backbones, consistently achieves comparable or superior forecasting accuracy at reduced inference costs.
- Our result illustrates that structured event data significantly enhances predictive performance, while semantic information holds a less important role in making accurate predictions.
- Our experimental results indicate that larger models (e.g., 70B parameters) effectively leverage enriched semantic contexts such as artitle titles and reasoning cues generated by themselves, while smaller models struggle with excessive contextual information.

2 Experiment Setup

2.1 Preliminary

The MIRAI benchmark. This paper focuses on the task of temporal event forecasting, with our experiments conducted using the MIRAI benchmark (Ye et al., 2024). Figure 2 (a) provides an overview of the task, the interaction between the API and the database, and the expected output. In particular, it visually summarizes how country-pair queries, historical event data, and the CAMEO ontology (Boschee et al., 2015)[2] converge to form the event forecasting pipeline. The international event is represented as $e_t = (t, s, r, o)$, where t denotes the event's timestamp, s and o are the subject and object countries, respectively, and $r \in \mathcal{R}$ denotes the relation type defined by the CAMEO ontology.

The forecasting task query is formalized as $(t + l, s, r?, o)$, where $r?$ represents the unknown relation to be predicted, aiming to predict the international relational events between a pair of countries occurring l days after the current time t. The current time, referred to as the *Cutoff Date*, up to which historical data are available. The interval $\texttt{Timediff} = l$ encompasses different forecasting challenges; for instance, predicting events 90 days ahead is naturally more difficult than forecasting those for the next day. Longer horizons require models to integrate and interpret information across broader temporal windows.

The expected prediction output includes the CAMEO codes of all anticipated events, presented in JSON format (e.g., `'01': ['011', '012'], '02': ['021']`). These codes span both first-level and second-level CAMEO classifications, allowing for coarse- and fine-grained accuracy assessment.[3] Predictions are evaluated using F1 scores where positive predictions are forecasted CAMEO codes that match ground-truth event relations.

Query and Database. The benchmark's data comprises **country-pair queries** (e.g., forecasting relation CAMEO codes between Australia and China on November 3, 2023: (2023-11-03, AUS, ?, CHN)) and a **database** containing both structured information (e.g., event relation distributions, event counts between country pairs) and unstructured information (e.g., news articles). Structured historical events are drawn from the Global Database of Events, Language, and Tone (GDELT) (Leetaru and Schrodt, 2013), while unstructured news articles are sourced and filtered from OBELICS (Laurençon et al., 2023). The CAMEO ontology is also included in the database, allowing models to access the parent-child hierarchy connections of the relation types. Overall, the dataset contains 59,161 unique (t, s, r, o) events with timestamps between January 1, 2023 and November 30, 2023. The test query set, based on November 2023 events, contains 705 queries with corresponding answers. Additionally, a balanced subset of 100 queries with

[1] We describe the details of the benchmark in Section 2.1.

[2] Conflict and Mediation Event Observations (CAMEO) is a well-established ontology for categorizing international political events.

[3] For example, '01' indicates 'Make public statement', while '012', as a finer category of '01', which refers to 'Make pessimistic statement'.

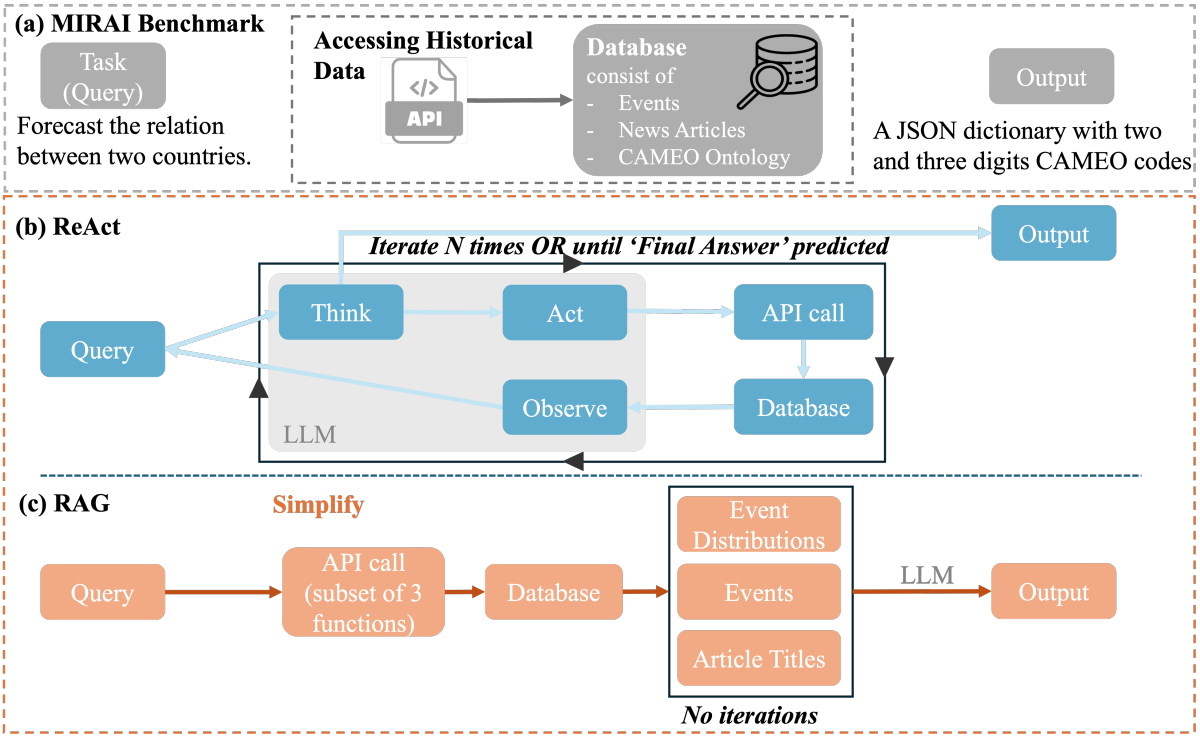

Figure 2: Comparison of RAG and ReAct in the event forecasting scenario. (a) Overview of the event forecasting task and dataset in MIRAI. (b) ReAct Forecasting Framework. (c) Our simplified framework.

uniform date coverage throughout November is provided to enable comprehensive ablation studies while maintaining temporal representativeness and computational tractability.

API for Data Retrieval. The benchmark also provides an API that contains the essential data classes and a suite of functions designed to interact with various types of information within the database. These functions cover various types of information, including country/relation code mappings and hierarchies, event statistics (counts, listings, distributions), event and news article retrieval. Unlike traditional RAG, which retrieves data into context based on the similarity between query and data in the database, MIRAI's API allows LLM agents to retrieve data flexibly by passing various parameters. For example, the model could ask for events beyond the given pair of countries for more advanced geopolitical considerations. The agent could also access the CAMEO ontology to retrieve the 3-digit second-level codes given the 2-digit first-level code and vice versa if deemed helpful.

2.2 RAG-based Simplification of the ReAct Framework

As illustrated in Figure 2, the ReAct paradigm relies on iterative inferences by an LLM to generate API calls and continuously integrate new information retrieved from the API and database until reaching a specified step limit or final answer. Although ReAct can theoretically access comprehensive data, each additional interaction increases the cost and complexity of forecasting. To address these limitations, we propose a simplified RAG approach that constrains ReAct along two dimensions: the interaction pipeline and the scope of retrieved data. We re-propose MIRAI's API to perform a single-turn retrieval operation using the *Cutoff Date*, *subject country*, and *object country*, as specified by the forecasting task. Unlike ReAct, which can leverage the full API and retrieve any information in the database, our approach focuses on three data types deemed most beneficial for event forecasting:

- **Relation Distribution (RelDis)**: Statistical frequencies of CAMEO relation codes summarizing historical interaction patterns between country pairs (retrieved by *get_relation_distribution* function).
- **Event Data (Uni-directional/Bi-directional)**: Structured representations of historical events either uni-directionally (from one country to another) or bidirectionally. We fix the number of retrieved events at 30 (retrieved by

get_events function), maintained the default setting of the benchmark.

- **News Article Titles**: Titles from recent news articles related to the specified country pairs. We fix the number of titles retrieved at 15 (retrieved by *get_news_articles* function).

2.3 Task Settings

To fully investigate the necessity of iterative reasoning and retrieval, we compared the effectiveness of the simplified RAG with iterative ReAct by evaluating on both the full-and sub-set test splits of the benchmark. Due to limited resources, we show the full analysis on the sub-set only and the full-set results with *Timediff=7* (details in Appendix A.4).

We adopt the same system prompts, query prompts, and extractor prompts defined in the original MIRAI paper for the ReAct Strategy. For RAG, we minimally modify these prompts only to remove the iterative thought-action loops to immediately retrieve information described in Section 2.2. All experiments utilize the same computational resources, evaluation metrics, and temporal horizons (*Timediff*=1, 7, 30, 90 days).

2.4 Models

We conduct our experiment on three open-source LLMs, i.e. `Llama-3.1-8B` (Meta AI Research, 2023), `Llama-3-70B`[4] (Meta AI Research, 2023) and `Mistral-7B-v0.2`, and `GPT-4o-mini`, a reference closed-source model[5]. The hardware and inference setup details are in Appendix A.1.

3 Results and Analysis

3.1 Retrieval Strategy Comparison and Information Type Analysis

RAG comparison with ReAct. Before comparing retrieval strategies, we evaluate the necessity to access historical information. Baseline experiments without retrieval show substantial performance degradation across all models (detailed analysis in Appendix A.2), confirming that models require explicit access to historical patterns rather than relying on memorized training data. Building on this methodological foundation, we systematically compare our simplified RAG approach against iterative ReAct across multiple model scales and analyze the effectiveness of different information types in forecasting performance.

Our findings in Table 1 confirm the earlier observation – illustrated in Figure 1 – that RAG performs on par with or better than ReAct and that this generalizes effectively across open-source LLMs at a `Timediff` of 1. Additionally, structured data formats, such as relation distributions (RelDis) and historical information as lists of event code as context, consistently outperform richer yet less structured semantic contexts like raw news article titles.

Specifically, providing structured relation distributions and a list of past event codes consistently improves performance across all models. For instance, `GPT-4o-mini` and `Llama-3-70B` achieved significant F1 gains when supplied with structured event graphs and relational distributions, highlighting the importance of structured information in facilitating effective tremporal event forecasting. Conversely, incorporating raw article titles typically decreased accuracy, indicating that unstructured semantic content may introduce more noise than being beneficial. Notably, with only the article titles, all the models fail to predict future events correctly, suggesting a challenge in mapping the conceptual international relations to their CAMEO code labels. We discuss the effect of the in-context label bias in Section 3.3.

Smaller models, such as `Llama-3.1-8B` and `Mistral-7B-v0.2`, demonstrated notably variable sensitivity to information types. `Llama-3.1-8B` showed strong improvements when provided with only the distribution of past events, whereas `Mistral-7B-v0.2` performed best when given full event lists and experienced performance drops with only relation distributions for context. This reveals that even within structured data, the optimal information granularity and type must be tailored to individual model capabilities.

The influence of different types of context information. Having demonstrated RAG's superior efficiency, we analyze which information types drive this performance. Providing models solely with relation distributions of previous events resulted in a surprisingly strong performance in all `Timediff` settings on the test subset (details in Appendix A.3.1), demonstrating the significance of statistical reasoning in event forecasting. However, despite the effectiveness of statistical signals, purely statistical information alone was insufficient for optimal performance. Particularly evident in

[4]In GPTQ format due to resource constraints: https://huggingface.co/TechxGenus/Meta-Llama-3-70B-Instruct-GPTQ.

[5]https://platform.openai.com/docs/models/gpt-4o-mini, point to gpt-4o-mini-2024-07-18.

Model	RelDis	Event (Uni_dir)	Event (Bi_dir)	Article (Title)	First-level F1	Second-level F1
					± w.r.t ReAct	
GPT-4o-mini	✓				$39.95_{+1.43}$	$28.41_{-1.11}$
		✓			$41.20_{+2.68}$	$27.44_{-2.08}$
				✓	$14.11_{-24.41}$	$7.22_{-22.30}$
	✓	✓			$40.67_{+2.15}$	$31.71_{+2.18}$
	✓		✓		$42.74_{+4.22}$	$\mathbf{32.93}_{+3.41}$
	✓	✓		✓	$43.63_{+5.11}$	$32.27_{+2.75}$
	✓		✓	✓	$\mathbf{43.79}_{+5.28}$	$32.44_{+2.92}$
Llama-3.1-8B	✓				$\mathbf{41.12}_{+13.05}$	$\mathbf{26.63}_{+10.08}$
		✓			$38.66_{+10.60}$	$25.31_{+8.75}$
				✓	$13.21_{-14.85}$	$6.63_{-9.92}$
	✓	✓			$34.85_{+6.78}$	$22.31_{+5.76}$
	✓		✓		$36.01_{+7.94}$	$21.64_{+5.09}$
	✓	✓		✓	$28.92_{+0.85}$	$16.62_{+0.07}$
	✓		✓	✓	$28.06_{-0.00}$	$16.32_{-0.23}$
Mistral-7B-v0.2	✓				$24.86_{+2.30}$	$16.51_{+3.87}$
		✓			$\mathbf{31.54}_{+8.97}$	$\mathbf{18.16}_{+5.52}$
				✓	$11.07_{-11.50}$	$4.03_{-8.60}$
	✓	✓			$26.07_{+3.50}$	$14.13_{+1.49}$
	✓		✓		$25.55_{+2.98}$	$17.27_{+4.64}$
	✓	✓		✓	$24.45_{+1.88}$	$15.84_{+3.20}$
	✓		✓	✓	$25.79_{+3.22}$	$15.26_{+2.63}$
Llama-3-70B	✓				$44.10_{+2.31}$	$31.26_{+1.73}$
		✓			$46.71_{+4.91}$	$\mathbf{33.05}_{+3.52}$
				✓	$14.47_{-27.33}$	$4.60_{-24.93}$
	✓	✓			$\mathbf{47.12}_{+5.32}$	$32.03_{+2.50}$
	✓		✓		$45.99_{+4.19}$	$32.12_{+2.59}$
	✓	✓		✓	$46.31_{+4.51}$	$32.49_{+2.96}$
	✓		✓	✓	$46.58_{+4.78}$	$32.68_{+3.15}$

Table 1: Performance (First-level and Second-level F1s) comparison of different LLMs on the test subset between RAG with different data retrieval and ReAct at Timediff of 1. The ✓represents that this information is provided to the LLM as retrieved content. '± w.r.t ReAct' represents the difference in F1 score w.r.t. ReAct.

GPT-4o-mini and Llama-3-70B, performance was highest when structured statistical information was complemented by event semantics, suggesting that larger models possess the capacity to extract useful signals from semantic content that smaller models cannot utilize.

Thus, effective event forecasting requires structured information such as past events combined with robust statistical signals, reinforcing the need for precise information structuring rather than merely increasing the contextual verbosity.

3.2 Analyzing Limitations of the ReAct Framework in Event Forecasting

We first assess the action execution success rate of API calls generated by the ReAct agent, observing an execution success rate consistently exceeding 95%, with an average of three to four functions executed per query across all four Timediffs. The function distribution indicates that the *get_relation_distribution* function is invoked at least once per query on average. These results suggest that the observed lower performance is not due to functional limitations of using the API. Having eliminated implementation artifacts, we investigate whether reasoning traces themselves impair performance.

Impact of iterative thoughts on model performance. We furthermore examined whether iterative thoughts generated within the ReAct paradigm enhance or impede model performance. As shown

in Table 2, we compared two one-step generation scenarios: (1) "observation-only" and (2) "observation-with-thoughts." For the observation-only scenario, observations collected during ReAct iterations were concatenated and appended to the query input for the LLMs to perform a one-step generation. In the observation-with-thoughts scenario, we preserved the thoughts preceding each action and observation and conducted a one-step generation, allowing for a direct comparison with the observation-only scenario.

The results vary depending on model capacity and task complexity. Notably, structured observations alone consistently matched or outperformed scenarios where thoughts were included. Smaller models, such as `GPT-4o-mini` and `Llama-3.1-8B`, exhibited a lower performance when thoughts were incorporated. Specifically, `GPT-4o-mini` showed a significant performance decline for second-level predictions with thoughts included (e.g., from 30.79% down to 27.33% at `Timediff=1`). The inconsistent and generally low performance of `Mistral-7B-v0.2` points to possible limitations arising from reduced model capacity or restricted context windows.

Figure 4 clearly visualizes these performance trends across models. Smaller models demonstrate a substantial divergence in performance between observation-only and thought-enhanced contexts, highlighting their limited capacity for utilizing reasoning traces. Conversely, `Llama-3-70B` displayed consistent robustness and minimal performance fluctuations across all experimental conditions, maintaining high F1 scores even with integrated thoughts. This indicates that smaller models struggle to benefit from reasoning traces, likely due to their limited capacity to filter useful signals from noise.

Temporal sensitivity and challenges in long-term forecasting. Analyzing temporal cutoffs in Table 2 revealed notable patterns. Models occasionally achieved higher performance at intermediate cutoffs (e.g., `Timediff=7` days) compared to the shortest interval (`Timediff=1` day). For example, `GPT-4o-mini` had a slightly better second-level prediction performance at `Timediff=7` days than at `Timediff=1` day (30.79% vs. 29.64% in observation-only scenarios). The performance difference may be an artifact that short-term predictions might suffer from data sparsity compared to predictions over longer intervals. However, performance was consistently worse at longer intervals (`Timediff=30` and 90 days), emphasizing the intrinsic challenges associated with long-term forecasting.

Cross-context comparison and quality of generated thoughts. We conducted cross-model experiments using the thoughts and observations generated by `Llama-3-70B` (Table 3) to clarify if the observed performance degradation was due to the reasoning quality of small models or their inherent limitations to utilize reasoning traces. Smaller models, which received reasoning and observations generated by the larger `Llama-3-70B` model, demonstrated consistent performance improvements as compared to their self-generated reasoning traces. Specifically, models such as `GPT-4o-mini` and `Mistral-7B-v0.2` showed significant performance gains, particularly at the first-level predictions, indicating that enhanced reasoning quality alleviates the necessity for smaller models to generate high-quality reasoning themselves.

However, absolute scores still remained lower than those of `Llama-3-70B`, confirming intrinsic limitations in smaller models' temporal reasoning capabilities. `Llama-3.1-8B`, interestingly, exhibited a greater improvement in fine-grained predictions (e.g., second-level) than other models, suggesting differential sensitivities to the thought based on prediction granularity.

Overall, our analyses emphasize that the effectiveness of iterative reasoning depends on the quality of generated thoughts and the models' intrinsic cognitive capacities. Larger models, such as `Llama-3-70B`, benefit from iterative reasoning, whereas smaller models' performance suffers.

3.3 Reasonability of Event Forecasting Task

We measured the ratio of predicted codes being in the top-k most common event codes in the context to investigate the extent to which predictions are influenced by label occurrence frequency. As illustrated in Figure 3, models exhibit a high ratio (60%–80%) in predicting the single most frequent CAMEO code (top-1), consistent with the expectation that frequent events are likely to recur. However, the ratio sharply declines when considering broader sets of labels (top-5), typically dropping below 50%. This trend indicates that although event frequency heavily influences predictions, models do not rely solely on frequency-based information.

Model	Training Data Cutoff Date	F1 (%) First-level ‖ Second-level			
		Timediff=1	Timediff=7	Timediff=30	Timediff=90
ReAct					
GPT-4o-mini	2023-10	38.52 ‖ 29.52	38.18 ‖ **29.83**	37.61 ‖ **28.20**	**38.94** ‖ **27.64**
Llama-3.1-8B	2023-12	28.07 ‖ 16.55	32.77 ‖ 18.03	30.76 ‖ 17.88	25.67 ‖ 15.45
Mistral-7B-v0.2	—	22.57 ‖ 12.63	24.00 ‖ 13.93	20.82 ‖ 14.69	20.66 ‖ 11.60
Llama-3-70B	2023-12	**41.79** ‖ **29.53**	**41.54** ‖ 26.57	**38.56** ‖ 26.84	38.92 ‖ 26.52
Observation-only					
GPT-4o-mini	2023-10	40.63 ‖ 30.79	41.41 ‖ 29.64	37.23 ‖ **28.71**	39.72 ‖ 26.29
Llama-3.1-8B	2023-12	33.60 ‖ 20.86	34.01 ‖ 20.57	27.73 ‖ 17.48	29.83 ‖ 19.06
Mistral-7B-v0.2	—	19.17 ‖ 6.92	22.70 ‖ 10.19	18.62 ‖ 7.17	20.10 ‖ 11.48
Llama-3-70B	2023-12	**45.81** ‖ **32.73**	**44.10** ‖ **30.49**	**40.26** ‖ 28.59	**40.19** ‖ **28.97**
Observation with Thought					
GPT-4o-mini	2023-10	39.24 ‖ 27.33	38.44 ‖ 28.65	35.86 ‖ 26.53	36.45 ‖ 24.99
Llama-3.1-8B	2023-12	28.00 ‖ 14.24	29.79 ‖ 13.40	27.32 ‖ 15.51	27.01 ‖ 15.37
Mistral-7B-v0.2	—	21.50 ‖ 10.34	17.47 ‖ 9.38	17.29 ‖ 9.90	24.08 ‖ 13.31
Llama-3-70B	2023-12	**44.68** ‖ **32.51**	**46.40** ‖ **31.62**	**40.13** ‖ **28.03**	**39.88** ‖ **29.73**

Table 2: Event forecasting performance on **test-subset** under four generation strategies: (i) *ReAct* allows the ReAct agent to access all functions in the API; (ii) *Observation-only* employs one-step generation with retrieved observations during ReAct process; and (iii) *Observation with Thought* augments one-step generation with a thought component and the corresponding observations. Reported are first-level and second-level F1 scores (%) across temporal cutoffs (Timediff=1, 7, 30, and 90 days). Bold and underlined values indicate the best and second-best performances within each setting, respectively.

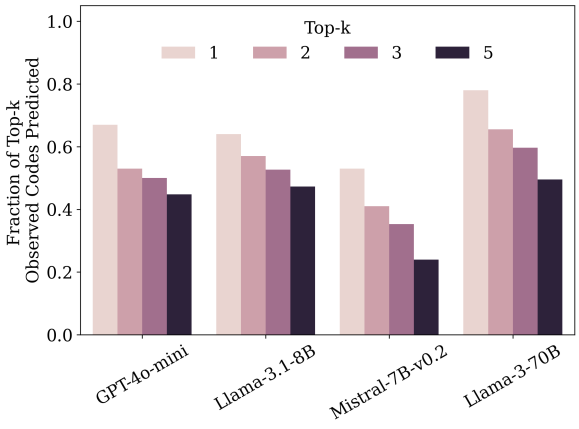

Figure 3: Top-k prediction ratios across models. Each bar represents the proportion of top-k most frequent observed codes predicted by each model, for $k \in 1, 2, 3, 5$.

4 Related Work

4.1 LLMs for Temporal Reasoning and Event Forecasting

Temporal reasoning involves processing and interpreting time-dependent information, which is crucial for accurate forecasting and decision-making in a dynamic environment (Xiong et al., 2024; Ge et al., 2025; Yuan et al., 2024). Several studies have explored the use of LLMs for event forecasting. While zero-shot or few-shot prompting can elicit some temporal reasoning (Yu et al., 2023; Lee et al., 2023), studies suggest that fine-tuning is particularly beneficial when incorporating raw text for complex events (Chang et al., 2024). Wang et al. (2025) proposed a framework that integrates news events into time series forecasting by fine-tuning an LLM. Chang et al. (2024) conducted a comprehensive evaluation of LLMs on temporal event forecasting, highlighting the importance of incorporating raw texts in specific complex events and fine-tuning LLMs. RAG provides an alternative direction to leverage large historical datasets or knowledge bases (Zhang et al., 2024b). Ye et al. (2024) introduced MIRAI as a benchmark for evaluating LLM agents in event forecasting, emphasizing the use of API to automate the data retrieval using LLMs' agentic ability. Although Chang et al. (2024) compared the effectiveness of in-context learning, finetuning, and RAG approaches in temporal forecasting, they did not investigate agentic frameworks and their potential to build a more cost-effective forecasting system.

Model	First-Level (Coarse)				Second-Level (Fine)			
	With 70B Context	With Self-produced Context	$\bar{\Delta}$ F1	Max Δ (Timediff)	With 70B Context	With Self-produced Context	$\bar{\Delta}$ F1	Max Δ (Timediff)
Observation with Thought								
GPT-4o-mini	40.19	37.50	2.69	5.92 (1d)	27.91	26.88	1.03	3.76 (1d)
Mistral-7B-v0.2	24.16	20.09	4.07	7.72 (7d)	13.39	10.73	2.66	5.70 (30d)
Llama-3.1-8B	31.00	28.03	2.97	5.67 (1d)	17.74	14.63	3.11	6.12 (7d)
Observation Only								
GPT-4o-mini	38.92	39.75	-0.83	0.94 (1d)	27.38	28.86	-1.48	0.34 (1d)
Mistral-7B-v0.2	25.03	20.15	4.89	9.55 (30d)	14.17	8.94	5.23	9.33 (30d)
Llama-3.1-8B	31.90	31.29	0.61	3.47 (30d)	18.75	19.49	-0.75	0.41 (30d)

Table 3: Cross-model reasoning trace transfer analysis (**test-subset**). Performance comparison when smaller models use reasoning traces generated by Llama-3-70B versus self-produced traces. $\bar{\Delta}$ F1 is the average of four difference values (one per `Timediff` setting: 1, 7, 30, 90 days). Max Δ indicates the maximum difference and corresponding `Timediff` setting.

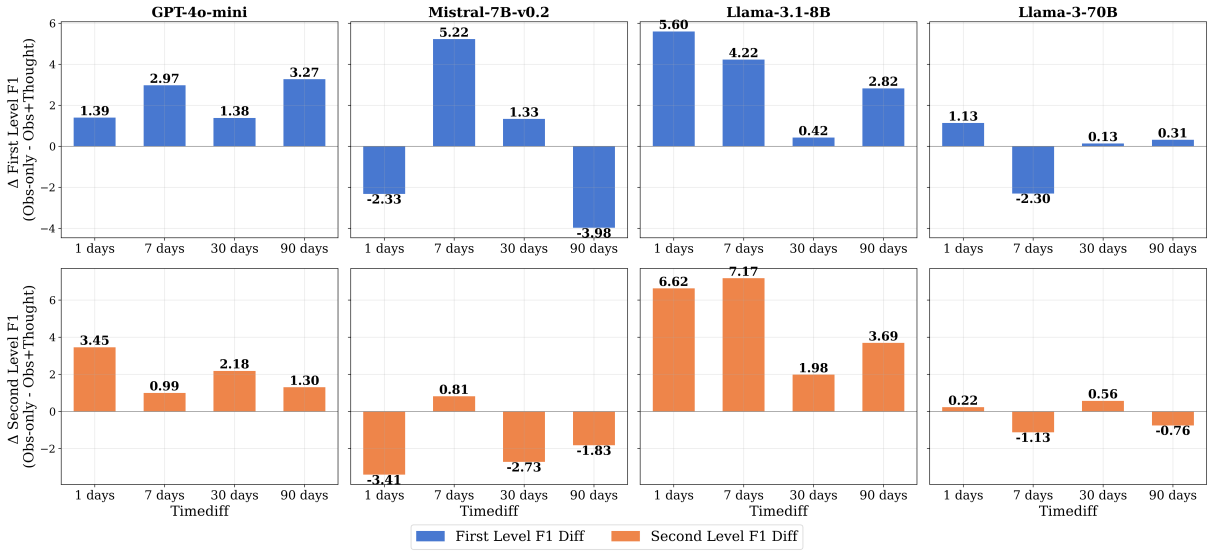

Figure 4: Performance (F1 score) difference of LLMs when thoughts generated are augmented together with factual information collected during ReAct logs. Each panel corresponds to a model.

4.2 Effectiveness of Automate LLM Agents

Large Language Models (LLMs) as agents have received increasing attention and are applied in many domains and applications, including code generation (Islam et al., 2024; Zhang et al., 2024a), formal math reasoning (Wang et al., 2023; Song et al., 2023), and commonsense reasoning (Zhao et al., 2023). These advanced AI systems enable LLMs to perform complex reasoning and interact dynamically with external environments and tools (Inaba et al., 2023), and have demonstrated the potential to achieve human-like decision making capabilities by collecting and processing various types of information. Yao et al. (2023) introduced ReAct as a framework for combining reasoning and acting in LLMs, enabling them to interact with external sources to generate more reliable responses. HuggingGPT (Shen et al., 2023) operates as an LLM-based agent controller that interfaces with the Hugging Face Hub to address complex user requests. Reflexion (Shinn et al., 2023) introduces a framework for agents to learn from past failures through verbal self-reflection. Despite promising performance gains observed in reasoning-intensive tasks, the efficacy and role of reasoning traces within agentic frameworks remain under-explored (Wu et al., 2025). Prior studies demonstrated that explicit reasoning processes did not yield significant benefits in Audio QA tasks (Li et al., 2025), while Verma et al. (2024) suggested that performance improvements may be due to exemplar-query similarity rather than enhanced reasoning abilities. In our study, we isolate the observations and thoughts

generated during the iterations and provide insights into the effectiveness of traces on the performance of different LLMs.

5 Conclusion

Our investigation revealed that RAG can outperform the agentic framework ReAct in the MIRAI temporal event forecasting task. Besides, models achieve higher robust performances when using structured relation distributions or graphs as context, whereas raw news articles, despite their semantic richness, negatively affect forecasting accuracy. In addition, our findings highlight the current limitation of small-scale LLMs in aligning intricate, semantically rich contexts to specific event predictions, suggesting a need for further research into semantic grounding and structured contextual representation. This research reveals the practical advantages of simplifying event forecasting frameworks, suggesting that strategic, structured data retrieval within RAG methods can yield more efficient and accurate predictive systems than the more elaborate ReAct paradigm.

6 Limitations and Future Work

Our evaluation framework presents several inherent limitations. First, we employ fixed retrieval counts (30 events, 15 article titles) based on MIRAI benchmark defaults rather than systematic optimization, which ensures fair comparison with existing ReAct implementations but may not represent optimal configurations. Second, we do not compare against iterative RAG or GraphRAG methods due to fundamental incompatibilities: iterative RAG approaches rely on semantic similarity-based retrieval that conflicts with MIRAI's structured API design, while GraphRAG methods target document chunking scenarios rather than structured database interaction. Our preliminary experiments using similarity-based search yielded substantially degraded performance, confirming these incompatibilities.

Our analysis focuses exclusively on international event forecasting within the MIRAI benchmark, limiting the generalizability of our findings to other temporal reasoning tasks or forecasting domains. The structured nature of CAMEO event representations may not extend to more open-ended forecasting scenarios. Additionally, our ReAct analysis centers on standard reasoning-acting paradigms without exploring advanced agentic strategies incorporating reflection, self-refinement, or multi-agent coordination. We did not systematically assess hallucination patterns in generated reasoning traces due to their verbosity, potentially overlooking important failure modes. The observed performance degradation with semantic information (news article titles) suggests underlying noise introduction mechanisms that warrant deeper theoretical investigation.

Future work should pursue several promising directions. First, conducting fine-grained interpretability analysis—including attention studies and token-level contribution analysis—could elucidate the root causes of semantic noise in unstructured contexts. Second, employing LLM-as-a-Judge frameworks would enable systematic quantification of hallucination patterns and reasoning coherence in iterative agent traces. Third, investigating thought-action alignment through causal analysis could reveal specific mechanisms by which reasoning quality affects downstream performance. Finally, developing adaptive retrieval strategies that dynamically adjust information types and quantities based on query complexity represents a natural extension of our structured retrieval approach. Extending evaluation to additional temporal reasoning tasks would strengthen generalizability claims across diverse forecasting domains.

7 Ethics Statement

This research focuses on methodological improvements to event forecasting systems using established public datasets (GDELT, OBELICS) within the MIRAI benchmark framework. We acknowledge that event forecasting technologies carry potential dual-use risks, including possible applications in market manipulation or political interference. Our work advances scientific understanding of temporal reasoning rather than developing deployment-ready systems. All experiments utilize existing pre-trained models and established benchmarks to ensure reproducibility and minimize computational overhead.

8 Acknowledgments

The authors gratefully acknowledge the support of the funding from UKRI under project code ES/T012277/1.

References

Cleo Anastassopoulou, Lucia Russo, Athanasios Tsakris, and Constantinos Siettos. 2020. Data-based analysis, modelling and forecasting of the covid-19 outbreak. *PloS one*, 15(3):e0230405.

Elizabeth Boschee, Jennifer Lautenschlager, Sean O'Brien, Steve Shellman, James Starz, and Michael Ward. 2015. Cameo. cdb. 09b5. pdf. *ICEWS Coded Event Data. Harvard Dataverse*, 3.

He Chang, Chenchen Ye, Zhulin Tao, Jie Wu, Zhengmao Yang, Yunshan Ma, Xianglin Huang, and Tat-Seng Chua. 2024. A comprehensive evaluation of large language models on temporal event forecasting. *CoRR*.

Peter Christensen, Kenneth Gillingham, and William Nordhaus. 2018. Uncertainty in forecasts of long-run economic growth. *Proceedings of the National Academy of Sciences*, 115(21):5409–5414.

Yubin Ge, Salvatore Romeo, Jason Cai, Raphael Shu, Monica Sunkara, Yassine Benajiba, and Yi Zhang. 2025. Tremu: Towards neuro-symbolic temporal reasoning for llm-agents with memory in multi-session dialogues. *arXiv preprint arXiv:2502.01630*.

Kenneth Gillingham, William Nordhaus, David Anthoff, Geoffrey Blanford, Valentina Bosetti, Peter Christensen, Haewon McJeon, and John Reilly. 2018. Modeling uncertainty in integrated assessment of climate change: A multimodel comparison. *Journal of the Association of Environmental and Resource Economists*, 5(4):791–826.

Tatsuro Inaba, Hirokazu Kiyomaru, Fei Cheng, and Sadao Kurohashi. 2023. Multitool-cot: Gpt-3 can use multiple external tools with chain of thought prompting. *arXiv preprint arXiv:2305.16896*.

Md Ashraful Islam, Mohammed Eunus Ali, and Md Rizwan Parvez. 2024. Mapcoder: Multi-agent code generation for competitive problem solving. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4912–4944.

Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander Rush, Douwe Kiela, and 1 others. 2023. Obelics: An open web-scale filtered dataset of interleaved image-text documents. *Advances in Neural Information Processing Systems*, 36:71683–71702.

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*, pages 544–557.

Kalev Leetaru and Philip A Schrodt. 2013. Gdelt: Global data on events, location, and tone, 1979–2012. In *ISA annual convention*, volume 2, pages 1–49. Citeseer.

Gang Li, Jizhong Liu, Heinrich Dinkel, Yadong Niu, Junbo Zhang, and Jian Luan. 2025. Reinforcement learning outperforms supervised fine-tuning: A case study on audio question answering. *arXiv preprint arXiv:2503.11197*.

Ruotong Liao, Xu Jia, Yunpu Ma, and Volker Tresp. 2023. Gentkg: Generative forecasting on temporal knowledge graph. In *Temporal Graph Learning Workshop@ NeurIPS 2023*.

Spyros Makridakis, Steven C Wheelwright, and Rob J Hyndman. 2008. *Forecasting methods and applications*. John wiley & sons.

Meta AI Research. 2023. Meta LLaMA 3: Improving Instruction Following and Few-Shot Learning. Accessed: 30 April 2024.

Nicolas D Savio and Konstantinos Nikolopoulos. 2013. A strategic forecasting framework for governmental decision-making and planning. *International Journal of Forecasting*, 29(2):311–321.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2023. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. *Advances in Neural Information Processing Systems*, 36:38154–38180.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36:8634–8652.

Peiyang Song, Kaiyu Yang, and Anima Anandkumar. 2023. Towards large language models as copilots for theorem proving in lean. In *The 3rd Workshop on Mathematical Reasoning and AI at NeurIPS'23*.

Jiashuo Sun, Chengjin Xu, Lumingyuan Tang, Saizhuo Wang, Chen Lin, Yeyun Gong, Lionel Ni, Heung-Yeung Shum, and Jian Guo. 2023. Think-on-graph: Deep and responsible reasoning of large language model on knowledge graph. In *The Twelfth International Conference on Learning Representations*.

Oskar Triebe, Hansika Hewamalage, Polina Pilyugina, Nikolay Laptev, Christoph Bergmeir, and Ram Rajagopal. 2021. Neuralprophet: Explainable forecasting at scale. *arXiv preprint arXiv:2111.15397*.

Mudit Verma, Siddhant Bhambri, and Subbarao Kambhampati. 2024. On the brittle foundations of react prompting for agentic large language models. *arXiv preprint arXiv:2405.13966*.

Haiming Wang, Huajian Xin, Chuanyang Zheng, Zhengying Liu, Qingxing Cao, Yinya Huang, Jing Xiong, Han Shi, Enze Xie, Jian Yin, and 1 others. 2023. Lego-prover: Neural theorem proving with growing libraries. In *The Twelfth International Conference on Learning Representations*.

Ruijie Wang, Yutong Zhang, Jinyang Li, Shengzhong Liu, Dachun Sun, Tianchen Wang, Tianshi Wang, Yizhuo Chen, Denizhan Kara, and Tarek Abdelzaher. 2024. Metahkg: Meta hyperbolic learning for few-shot temporal reasoning. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 59–69.

Xinlei Wang, Maike Feng, Jing Qiu, Jinjin Gu, and Junhua Zhao. 2025. From news to forecast: Integrating event analysis in llm-based time series forecasting with reflection. *Advances in Neural Information Processing Systems*, 37:58118–58153.

Junde Wu, Jiayuan Zhu, and Yuyuan Liu. 2025. Agentic reasoning: Reasoning llms with tools for the deep research. *arXiv preprint arXiv:2502.04644*.

Siheng Xiong, Ali Payani, Ramana Kompella, and Faramarz Fekri. 2024. Large language models can learn temporal reasoning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10452–10470.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*.

Chenchen Ye, Ziniu Hu, Yihe Deng, Zijie Huang, Mingyu Derek Ma, Yanqiao Zhu, and Wei Wang. 2024. Mirai: Evaluating llm agents for event forecasting. *arXiv preprint arXiv:2407.01231*.

Xinli Yu, Zheng Chen, and Yanbin Lu. 2023. Harnessing llms for temporal data-a study on explainable financial time series forecasting. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 739–753.

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 Web Conference 2024*, pages 1963–1974.

Kechi Zhang, Jia Li, Ge Li, Xianjie Shi, and Zhi Jin. 2024a. Codeagent: Enhancing code generation with tool-integrated agent systems for real-world repo-level coding challenges. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13643–13658.

Zhihan Zhang, Yixin Cao, Chenchen Ye, Yunshan Ma, Lizi Liao, and Tat-Seng Chua. 2024b. Analyzing temporal complex events with large language models? a benchmark towards temporal, long context understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1588–1606.

Zirui Zhao, Wee Sun Lee, and David Hsu. 2023. Large language models as commonsense knowledge for large-scale task planning. *Advances in Neural Information Processing Systems*, 36:31967–31987.

Appendix

A Appendix

A.1 Hardware and inference setup

We run the open-source models using the **vLLM** library for efficient inference with multi-GPU support. All inference is performed in **FP16** precision. The Llama-3.1-8B and Mistral-7B models are served on a machine with **2× NVIDIA RTX 3090** GPUs (24GB VRAM each), which is sufficient for these smaller models. The larger Llama-3-70B (GPTQ 4-bit) model is hosted on a single **NVIDIA A100 80GB** GPU to accommodate its higher memory requirements. For **GPT-4o-mini**, we leverage the model's **remote API** endpoint. To ensure **reproducibility**, we fix the random seed to 0 for all runs and set the generation **temperature to 0**, yielding deterministic outputs. All other decoding hyperparameters follow the MIRAI benchmark defaults. Under these settings, our experiments are fully deterministic and can be replicated exactly. All models are evaluated on the same set of queries (the 100-query subset described above) to enable direct, apples-to-apples comparison of their forecasting performance.

A.2 Baseline Performance Without Retrieval

Model	Zero-Shot (ZS)		Chain-of-Thought (CoT)	
	First-Level F1	Second-Level F1	First-Level F1	Second-Level F1
`GPT-4o-mini`	31.05	8.84	9.89	4.38
`Llama-3.1-8B`	10.80	4.35	12.75	5.11
`Mistral-7B-v0.2`	11.15	3.49	10.10	3.38
`Llama-3-70B`	9.96	3.14	13.11	4.17

Table 4: Baseline experimental results of different LLMs on the test full set under zero-shot and chain-of-thought prompting without historical information retrieval (`Timediff` is set to 7). These results demonstrate the necessity of structured information retrieval by evaluating prediction performance without access to retrieved historical events.

To address potential concerns regarding model memorization due to training cutoff proximity to test events, we evaluate baseline performance without historical information retrieval. Table 4 presents zero-shot (ZS) and chain-of-thought (CoT) results across all models under Timediff=7.

The results reveal substantial performance degradation without retrieval-augmented context. Even GPT-4o-mini, which achieves 31.05% first-level F1 in zero-shot mode, falls significantly short of RAG performance (43.90% with relation distributions, Table 8). Notably, CoT prompting further degrades performance across most models, indicating that reasoning chains without factual grounding are counterproductive for temporal forecasting.

These findings directly address memorization concerns by demonstrating that models require explicit access to historical patterns and relation distributions to generate reliable predictions. The performance gap between baseline and retrieval-augmented approaches validates our experimental framework's integrity and the necessity of information retrieval components.

A.3 Additional Results on Test Subset

A.3.1 RAG compared to ReAct on different Timediff

Model	RelDis	Event (Uni_dir)	Event (Bi_dir)	Article (Title)	First-level F1	Second-level F1
					± w.r.t ReAct	
GPT-4o-mini	✓				$39.38_{+1.20}$	$28.39_{-1.44}$
		✓			$39.56_{+1.38}$	$25.95_{-3.88}$
				✓	$13.46_{-24.72}$	$7.05_{-22.78}$
	✓	✓			$40.61_{+2.43}$	$30.44_{+0.60}$
	✓		✓		$39.25_{+1.07}$	$29.81_{-0.03}$
	✓	✓		✓	$41.99_{+3.81}$	$\mathbf{32.08}_{+2.25}$
	✓		✓	✓	$\mathbf{42.34}_{+4.16}$	$31.78_{+1.94}$
Llama-3.1-8B	✓				$\mathbf{42.14}_{+9.37}$	$\mathbf{27.48}_{+9.45}$
		✓			$38.56_{+5.79}$	$26.71_{+8.68}$
				✓	$14.19_{-18.58}$	$6.55_{-11.48}$
	✓	✓			$36.61_{+3.84}$	$23.64_{+5.61}$
	✓		✓		$36.47_{+3.70}$	$22.66_{+4.62}$
	✓	✓		✓	$27.95_{-4.82}$	$15.88_{-2.15}$
	✓		✓	✓	$28.91_{-3.86}$	$16.46_{-1.57}$
Mistral-7B-v0.2	✓				$23.53_{-0.48}$	$13.54_{-0.39}$
		✓			$\mathbf{30.45}_{+6.45}$	$\mathbf{19.91}_{+5.98}$
				✓	$10.43_{-13.57}$	$4.20_{-9.73}$
	✓	✓			$21.79_{-2.22}$	$12.30_{-1.63}$
	✓		✓		$15.35_{-8.65}$	$9.78_{-4.15}$
	✓	✓		✓	$29.39_{+5.39}$	$17.60_{+3.67}$
	✓		✓	✓	$25.97_{+1.97}$	$15.36_{+1.43}$
Llama-3-70B	✓				$44.31_{+2.77}$	$32.43_{+5.86}$
		✓			$45.38_{+3.84}$	$32.20_{+5.64}$
				✓	$16.15_{-25.39}$	$4.90_{-21.66}$
	✓	✓			$45.18_{+3.64}$	$31.67_{+5.10}$
	✓		✓		$44.89_{+3.35}$	$31.36_{+4.80}$
	✓	✓		✓	$\mathbf{46.10}_{+4.56}$	$31.79_{+5.23}$
	✓		✓	✓	$45.37_{+3.83}$	$\mathbf{32.76}_{+6.20}$

Table 5: Experimental results of different LLMs on the test subset under the setting of simple one-step generation by providing different types of information (Timediff is set to 7). The ✓represents that this information is provided to the LLM using the same retrieval function defined by MIRAI Agentic Framework.

Model	RelDis	Event (Uni_dir)	Event (Bi_dir)	Article (Title)	First-level F1	Second-level F1
					± w.r.t ReAct	
GPT-4o-mini	✓				$37.42_{-0.19}$	$27.83_{-0.37}$
		✓			$35.97_{-1.64}$	$23.96_{-4.25}$
				✓	$13.03_{-24.58}$	$6.89_{-21.31}$
	✓	✓			$37.41_{-0.20}$	$27.65_{-0.56}$
	✓		✓		$37.98_{+0.37}$	$28.44_{+0.24}$
	✓	✓		✓	$\mathbf{39.06}_{+1.45}$	$\mathbf{29.07}_{+0.86}$
	✓		✓	✓	$38.85_{+1.24}$	$28.61_{+0.41}$
Llama-3.1-8B	✓				$\mathbf{36.84}_{+6.08}$	$21.47_{+3.59}$
		✓			$34.65_{+3.89}$	$\mathbf{21.56}_{+3.68}$
				✓	$13.50_{-17.26}$	$6.05_{-11.84}$
	✓	✓			$33.12_{+2.36}$	$19.85_{+1.97}$
	✓		✓		$32.71_{+1.95}$	$19.59_{+1.71}$
	✓	✓		✓	$28.09_{-2.66}$	$15.46_{-2.42}$
	✓		✓	✓	$28.37_{-2.39}$	$16.50_{-1.38}$
Mistral-7B-v0.2	✓				$22.18_{+1.36}$	$9.82_{-4.87}$
		✓			$26.16_{+5.34}$	$13.68_{-1.02}$
				✓	$12.17_{-8.65}$	$4.06_{-10.64}$
	✓	✓			$27.51_{+6.70}$	$11.90_{-2.79}$
	✓		✓		$15.56_{-5.25}$	$7.61_{-7.09}$
	✓	✓		✓	$\mathbf{28.17}_{+7.35}$	$16.17_{+1.47}$
	✓		✓	✓	$27.80_{+6.98}$	$\mathbf{17.96}_{+3.27}$
Llama-3-70B	✓				$40.84_{+2.28}$	$\mathbf{30.30}_{+3.46}$
		✓			$37.26_{-1.30}$	$24.78_{-2.06}$
				✓	$14.02_{-24.54}$	$5.36_{-21.48}$
	✓	✓			$42.18_{+3.62}$	$28.16_{+1.33}$
	✓		✓		$\mathbf{44.13}_{+5.57}$	$28.64_{+1.80}$
	✓	✓		✓	$41.89_{+3.33}$	$28.58_{+1.75}$
	✓		✓	✓	$42.21_{+3.66}$	$28.00_{+1.17}$

Table 6: Experimental results of different LLMs on the test subset under the setting of simple one-step generation by providing different types of information (`Timediff` is set to 30). The ✓represents that this information is provided to the LLM using the same retrieval function defined by MIRAI Agentic Framework.

Model	RelDis	Event (Uni_dir)	Event (Bi_dir)	Article (Title)	First-level F1	Second-level F1
					± w.r.t ReAct	
GPT-4o-mini	✓				**38.40**$_{-0.53}$	**27.11**$_{-0.53}$
		✓			34.87$_{-4.07}$	20.97$_{-6.67}$
				✓	13.52$_{-25.42}$	6.78$_{-20.87}$
	✓	✓			35.38$_{-3.56}$	26.00$_{-1.65}$
	✓		✓		36.13$_{-2.80}$	26.58$_{-1.06}$
	✓	✓		✓	37.69$_{-1.25}$	26.59$_{-1.05}$
	✓		✓	✓	37.13$_{-1.81}$	26.27$_{-1.38}$
Llama-3.1-8B	✓				28.58$_{+2.90}$	18.15$_{+2.71}$
		✓			30.48$_{+4.81}$	**20.21**$_{+4.76}$
				✓	14.79$_{-10.88}$	7.19$_{-8.26}$
	✓	✓			**32.94**$_{+7.27}$	19.54$_{+4.09}$
	✓		✓		30.08$_{+4.41}$	18.40$_{+2.95}$
	✓	✓		✓	31.21$_{+5.54}$	17.46$_{+2.01}$
	✓		✓	✓	27.55$_{+1.88}$	14.45$_{-1.00}$
Mistral-7B-v0.2	✓				**28.95**$_{+8.29}$	14.05$_{+2.45}$
		✓			21.34$_{+0.68}$	12.16$_{+0.56}$
				✓	14.24$_{-6.42}$	5.47$_{-6.13}$
	✓	✓			20.83$_{+0.17}$	12.77$_{+1.17}$
	✓		✓		11.84$_{-8.82}$	6.99$_{-4.61}$
	✓	✓		✓	26.71$_{+6.05}$	**14.34**$_{+2.74}$
	✓		✓	✓	24.22$_{+3.56}$	12.20$_{+0.60}$
Llama-3-70B	✓				**40.43**$_{+1.51}$	**29.73**$_{+3.21}$
		✓			36.13$_{-2.80}$	23.46$_{-3.05}$
				✓	15.74$_{-23.18}$	4.78$_{-21.74}$
	✓	✓			39.45$_{+0.53}$	27.61$_{+1.10}$
	✓		✓		38.85$_{-0.08}$	27.20$_{+0.68}$
	✓	✓		✓	38.08$_{-0.85}$	27.29$_{+0.77}$
	✓		✓	✓	38.83$_{-0.09}$	27.95$_{+1.44}$

Table 7: Experimental results of different LLMs on the test subset under the setting of simple one-step generation by providing different types of information (`Timediff` is set to 90). The ✓represents that this information is provided to the LLM using the same retrieval function defined by MIRAI Agentic Framework.

A.4 Additional Results on Test Full Set

In this section, we present results for the full test set with a `Timediff` of 7 days, chosen to represent a moderate level of difficulty. Although `GPT-4o-mini` exhibits a slight decline in first-level F1 compared to ReAct, its second-level F1 performance continues to underscore the effectiveness of RAG. `Llama-3.1-8B` and `Mistral-7B-v0.2` show a similar pattern to that observed on the test subsets. Despite fluctuations, the findings of the full test set highlight the importance of structural information, particularly relation distributions, and emphasize the need for robust statistical signals to build an effective and efficient event forecasting system.

Model	RelDis	Event (Uni_dir)	Event (Bi_dir)	Article (Title)	First-level F1	Second-level F1
					± w.r.t ReAct	
GPT-4o-mini	✓				$43.80_{-4.69}$	$30.88_{-2.03}$
		✓			$40.14_{-8.35}$	$25.42_{-7.49}$
				✓	$9.17_{-39.32}$	$4.64_{-28.27}$
	✓	✓			$46.62_{-1.87}$	$\mathbf{33.68}_{+0.76}$
	✓		✓		$46.00_{-2.48}$	$33.62_{+0.70}$
	✓	✓		✓	$\mathbf{46.91}_{-1.57}$	$33.17_{+0.25}$
	✓		✓	✓	$46.39_{-2.09}$	$33.42_{+0.51}$
Llama-3.1-8B	✓				$38.85_{+6.71}$	$23.75_{+7.65}$
		✓			$39.80_{+7.65}$	$\mathbf{25.40}_{+9.29}$
				✓	$9.55_{-22.60}$	$4.90_{-11.21}$
	✓	✓			$\mathbf{40.77}_{+8.62}$	$24.43_{+8.33}$
	✓		✓		$40.26_{+8.11}$	$24.94_{+8.84}$
	✓	✓		✓	$30.89_{-1.26}$	$18.01_{+1.91}$
	✓		✓	✓	$30.93_{-1.22}$	$17.56_{+1.45}$
Mistral-7B-v0.2	✓				$27.69_{+0.15}$	$14.34_{-1.41}$
		✓			$31.83_{+4.29}$	$\mathbf{19.87}_{+4.12}$
				✓	$11.22_{-16.32}$	$3.40_{-12.35}$
	✓	✓			$29.15_{+1.61}$	$16.10_{+0.35}$
	✓		✓		$24.94_{-2.60}$	$13.25_{-2.50}$
	✓	✓		✓	$\mathbf{34.26}_{+6.72}$	$18.98_{+3.22}$
	✓		✓	✓	$33.10_{+5.56}$	$17.39_{+1.64}$
Llama-3-70B	✓				$44.58_{-1.83}$	$28.80_{+1.28}$
		✓			$43.80_{-2.61}$	$29.08_{+1.56}$
				✓	$12.76_{-33.64}$	$2.94_{-24.58}$
	✓	✓			$46.49_{+0.08}$	$29.91_{+2.39}$
	✓		✓		$\mathbf{47.41}_{+1.01}$	$\mathbf{31.15}_{+3.63}$
	✓	✓		✓	$47.33_{+0.92}$	$30.57_{+3.04}$
	✓		✓	✓	$47.11_{+0.71}$	$31.13_{+3.61}$

Table 8: Experimental results of different LLMs on the test full set under the setting of simple one-step generation by providing different types of information (`Timediff` is set to 7). The ✓represents that this information is provided to the LLM using the same retrieval function defined by MIRAI Agentic Framework.

A.5 Prompt for RAG

System prompt

You are an expert in forecasting future events based on historical data. The database contains news articles from January 1, 2023 to the current date {current_date_nlp} and the events extracted from these articles. The events are in the form of (date, subject country, relation, object country), where the countries are represented by ISO 3166-1 alpha-3 codes and the relations are represented by the CAMEO codes defined in the 'Conflict and Mediation Event Observations' ontology. The relations are hierarchical: first-level relations are general parent relations represented by two-digit CAMEO codes, while second-level relations are more specific child relations represented by three-digit CAMEO codes. Child relations have the same first two digits as their parent relations. For example, '01' is a first-level relation, and '010' and '011' are some of its second-level relations. The relations in the database are represented in the second-level form.

Your task is to forecast the future relations between two entities in a given query. You will be provided with the relevant events and news articles, as well as information about the ISO country codes, the CAMEO relation codes that allow you to analyze the historical events and statistics. The answer should be a JSON dictionary where the keys are the forecasted two-digit first-level CAMEO codes and the values are lists of forecasted three-digit second-level CAMEO codes that are child relations of the key. For example, 'Final Answer: {{"01": ["010", "011", "012"], "02": ["020", "023"]}}'.

The final answer will be evaluated based on the precision and recall of the forecasted first-level and second-level relations, so only include confident first-level and second-level CAMEO codes in your final forecast.

Query prompt

Query: Please forecast the relations that {actor1_name} will take towards {actor2_name} on {future_date_nlp} based on your knowledge up to {current_date_nlp}. I.e. forecast the relation CAMEO codes in query event Event(date={future_date}, head_entity=ISOCode({actor1_code}), relation=CAMEOCode(?), tail_entity=ISOCode({actor2_code})).

Here is the frequency of relation between {actor1_name} and {actor2_name} up to {current_date_nlp}: {relation_distribution}

Retrieved Events: {events}

Retrieved Articles: {article_titles}

Final Answer: