Tsinghua Science and Technology
Doi: 10.26599/TST.2025.9010160

Research Article

MAEPS: Multi-Agent Event Prediction System Based on Human Expert Team Collaboration Simulation

Chengyuan Jin, Tong Zhou, Yubo Chen (✉), Kang Liu, and Jun Zhao

Abstract Event prediction (EP), the accurate forecasting of future events, is vital for strategic planning and risk management in both governmental and business contexts. The rapid advancement of large language models (LLMs) has positioned AI-based automated prediction methods at the forefront of academic and industrial research. However, current LLM prediction systems exhibit several shortcomings. Firstly, their information retrieval mostly searches based on the question itself, failing to gather relevant data from multiple perspectives as human expert teams do. Secondly, their temporal analysis is inadequate, as the collected information often includes subjective opinions or speculations and lacks the ability to reconcile contradictory information across different time points during real-time prediction. To address these issues this paper introduces MAEPS (Multi-Agent Event Prediction System), which emulates the collaborative efforts of human expert teams through 12 specialized agents. Each agent collects data from a specific professional dimension. The system automatically identifies and resolves conflicting information, ensuring that predictions prioritize recent and consistent facts. Experiments on EP datasets from real prediction platforms demonstrate that MAEPS significantly outperforms existing LLM prediction systems by 7% in accuracy, thereby validating the efficacy of simulating expert team collaboration for prediction purposes.

Keywords Event Prediction, Multi-Agent Systems, Information Retrieval, Large Language Models

1 Introduction

Event prediction (EP) refers to the systematic and methodical forecasting of future events based on comprehensive analysis of available information, historical patterns, and sophisticated analytical methods. EP plays a crucial role in numerous critical fields, including social policy formulation [1] and financial market analysis [2], where stakeholders require reliable insights to navigate complex decision-making processes. Accurate forecasts provide a robust scientific foundation for strategic decision-making, significantly enhancing both the rationality and effectiveness of policy implementations and business strategies. However, achieving consistently high accuracy in complex, dynamic, and rapidly evolving environments remains a formidable challenge due to the inherent complexity of information sources, intricate temporal dependencies between events, and the pervasive uncertainty that characterizes real-world scenarios. Consequently, improving prediction accuracy while maintaining practical applicability has emerged as a pivotal research focus that continues to attract substantial attention from both academic researchers and industry practitioners. Prediction methodologies are broadly categorized into two fundamental approaches: statistical methods and judgmental approaches, each with distinct strengths and limitations. While statistical methods, which are primarily based on sophisticated time series modeling techniques [3], tend to falter when confronted with sparse data availability or rapid environmental changes that disrupt established patterns, judgmental prediction—which represents our primary focus in this research—emphasizes the strategic integration of human expertise, accumulated historical experience, and intuitive reasoning capabilities. This judgmental approach demonstrates particular value in its ability to compensate for the inherent deficiencies of purely statistical methods, especially in challenging scenarios characterized by limited data availability or exceptionally high levels of uncertainty.

The remarkable rise of artificial intelligence technologies has transformed LLM-based automated prediction into a particularly prominent and rapidly evolving research area [4–8], attracting significant attention from researchers worldwide. These methods typically follow a two-stage ”retrieval-then-reason” paradigm, aligned with Large Knowledge Models [**?**]. In the first stage, an LLM issues targeted queries to retrieve

• Chengyuan Jin, Tong Zhou, Yubo Chen, Kang Liu, and Jun Zhao are with The Key Laboratory of Cognition and Decision Intelligence for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, and University of Chinese Academy of Sciences, School of Artificial Intelligence, Beijing 100049, China. E-mail: {jinchengyuan2024, tong.zhou}@ia.ac.cn, {yubo.chen, kliu, jzhao}@nlpr.ia.ac.cn
* Corresponding author: Yubo Chen. E-mail: yubo.chen@nlpr.ia.ac.cn
Manuscript received: 2025-09-27; accepted: 2025-10-18

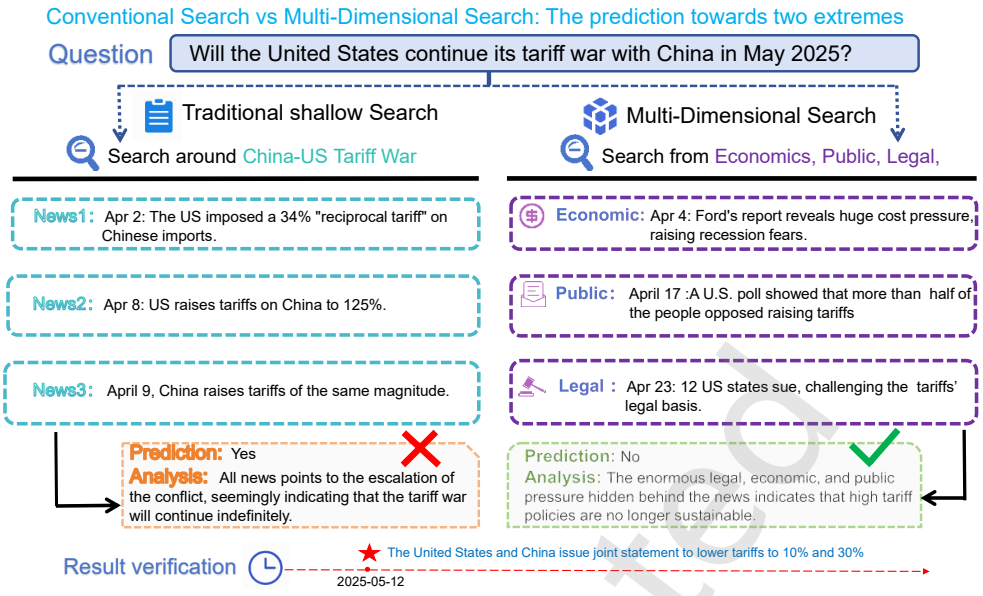


Fig. 1 Comparison of information retrieval approaches: traditional unidimensional vs. multi-agent multi-dimensional approach

external sources; in the second, it integrates the evidence for analysis and prediction.

However, despite their promising potential, current high-quality prediction methods continue to exhibit three fundamental and interconnected issues that significantly limit their effectiveness:

Unidimensional Information Retrieval: Existing methodologies predominantly focus on searching for information that is directly and explicitly related to the prediction question itself, systematically failing to gather comprehensive information from multiple professional perspectives and diverse analytical dimensions. As clearly illustrated in Figure 1, this narrow approach inevitably leads to incomplete information coverage that undermines prediction quality. For instance, accurately predicting the complex dynamics of the US-China tariff war requires sophisticated insights from multiple interconnected dimensions, including economic indicators, public sentiment analysis, and legal regulatory frameworks, rather than relying solely on direct tariff-related news coverage.

Insufficient Temporal Conflict Processing: Current prediction systems systematically neglect the critical temporal attributes of information sources, leading to problematic conflicts where newer reports directly contradict older ones regarding the same event or phenomenon. This oversight creates two significant problems: first, excessive reliance on outdated information substantially reduces prediction timeliness, as critical decisions become based on stale facts that no longer accurately reflect current reality; second, unresolved contradictions fundamentally undermine prediction accuracy because the underlying model cannot effectively determine which information sources should be trusted and prioritized in the decision-making process.

Lack of Factual Screening: Existing prediction systems demonstrate a concerning absence of robust mechanisms for systematic factual screening and information quality assessment. While advanced LLMs possess demonstrated capabilities to extract factual information [9], which has proven particularly valuable in specialized domains such as financial analysis [10], these systems are frequently fed an indiscriminate mixture of factual content and non-factual material, including subjective opinions, unverified speculation, and potentially misleading commentary. Effective and reliable prediction must be fundamentally grounded in objective, independently verifiable data [11]. The existing method described in [12], which attempts to filter articles by content type (distinguishing events from commentary), relies heavily on manually crafted rules and consequently lacks the scalability necessary for practical implementation in diverse real-world scenarios.

Human experts consistently demonstrate their ability to improve complex task processing capabilities through systematic and strategic collaboration [13], leveraging the complementary strengths of diverse professional backgrounds and specialized knowledge domains. Different categories of experts, such as experienced policy analysts and seasoned economists, systematically gather and analyze information from their respective specialized domains, bringing unique perspectives and analytical frameworks to the prediction process. For example, when tasked with predicting significant trade policy changes, policy analysts typically focus their

attention on legislative processes, regulatory frameworks, and political dynamics, while economists simultaneously analyze market indicators, economic impacts, and broader macroeconomic trends. Their collaborative efforts effectively integrate regulatory feasibility assessments with comprehensive economic consequence analysis, ultimately producing more accurate, nuanced, and actionable predictions. This sophisticated multi-dimensional approach provides a comprehensive factual foundation for prediction activities and serves as a highly successful model that demonstrates the power of coordinated expertise.

Inspired by this proven collaborative model, we propose MAEPS (Multi-Agent Event Prediction System), an innovative framework where multiple specialized agents work in coordinated collaboration to effectively simulate the complex dynamics of human expert teams. By maintaining a strategic focus on high-quality information collection and sophisticated processing methodologies, MAEPS systematically addresses each of the three previously identified critical issues, thereby enhancing prediction comprehensiveness, improving temporal accuracy, and increasing overall prediction reliability.

This paper's main contributions are:

- We propose MAEPS, a novel and comprehensive multi-agent framework that effectively simulates human expert team collaboration for event prediction, directly addressing the significant limitations of existing single-dimensional information retrieval approaches.
- We design and implement 12 specialized agents that systematically collect information from different professional dimensions and analytical perspectives, ensuring comprehensive coverage of all relevant factors necessary for accurate prediction.
- We introduce sophisticated temporal conflict resolution and fact extraction mechanisms that automatically filter high-quality information and systematically resolve contradictory data across different time points.
- Extensive experiments conducted on real prediction datasets demonstrate that MAEPS significantly outperforms existing LLM prediction systems by 7% in accuracy, providing strong empirical validation of the effectiveness of our proposed approach.

2 Related Work

With the remarkable advances in large language models (LLMs) and their increasingly sophisticated reasoning capabilities, LLM-based forecasting systems have emerged as a particularly vibrant and rapidly evolving research hotspot, attracting significant attention from both academic researchers and industry practitioners. Halawi et al. [4] develop a comprehensive retrieval-augmented language model system that systematically and automatically searches for relevant information across multiple data sources, generates well-informed forecasts through sophisticated reasoning processes, and intelligently aggregates predictions from multiple perspectives, ultimately achieving performance levels remarkably close to those of experienced human forecasters on highly competitive prediction platforms. RTF [5] leverages a sophisticated architecture of hierarchical ReAct agents that work collaboratively to retrieve and process information from diverse sources, while simultaneously employing specialized computational tools for numerical simulation and quantitative analysis to realize comprehensive forecasting capability across various domains. OpenEP [6] proposes an innovative open-ended future event prediction task that addresses the limitations of closed-domain forecasting and constructs a corresponding comprehensive dataset; in its sophisticated retrieval pipeline, key stakeholders and relevant historical events are systematically identified and analyzed, while advanced clustering techniques applied to news text snippets are strategically used to clarify complex inter-event dependencies and significantly reduce information redundancy.

Time series analysis and temporal modeling play an increasingly crucial and foundational role in modern prediction systems, providing essential mathematical frameworks for understanding temporal patterns and dependencies. Chen and Cheng [14] propose CLSTM-BPR, an innovative fusion method that effectively combines the temporal modeling capabilities of Long Short-Term Memory networks with the ranking optimization strengths of Bayesian Personalized Ranking, demonstrating significant improvements in prediction accuracy. Early forecasting benchmarks include MCNC [15], SCT [16], and CoScript [17], which primarily focus on script learning and common-sense reasoning in carefully controlled synthetic scenarios but unfortunately lack comprehensive tracking and analysis of real-world events with their inherent complexity and unpredictability. ForecastBench [18] introduces a dynamic and continuously evolving benchmark that systematically evaluates machine learning systems on automatically generated and regularly updated forecasting questions, providing a more realistic assessment of prediction capabilities in rapidly changing environments. The PROPHET benchmark [7] further advances the field by categorizing forecasting problems according to the likelihood and potential impact of causal intervention, offering a novel and theoretically grounded evaluation perspective; its sophisticated temporal assumptions provide robust theoreti-

Table 1 Specialized Agent Configuration

Agent Name	Functional Introduction
Key People	Search dynamics, positions, and influence of key figures
Key Organizations	Search official positions and plans of key organizations
Concepts Entities	Explain professional concepts and entity definitions
Historical Precedents	Search similar historical events as references
Related News	Collect latest related news, construct factual background
Policy Regulation	Interpret policies, regulations, and industry standards
Economic Indicators	Collect quantified economic data and market indicators
Expert Opinions	Integrate expert, analyst, and institutional viewpoints
Social Sentiment	Analyze public emotions, attitudes, and public opinion
External Related Events	Search external shocks or black swan events
Technology Innovation	related technological progress and supply chain dynamics
Competitive Landscape	Analyze competitors' strategies and market reactions

cal grounding for handling complex causal relations among interconnected news events and their temporal dependencies. Document-level financial event datasets complement multi-agent retrieval [**?**]. Inspired by these theoretical foundations and practical insights, our MAEPS system strategically applies comprehensive temporal analysis to practical forecasting scenarios to systematically detect, analyze, and resolve information conflicts that arise from temporal inconsistencies.

Other lines of research enhance LLM forecasting. Human–AI teaming improves performance [19], and LLMs work well as expert tools [**?**]. Prompt robustness matters for domain NER [**?**]. DPO-based fine-tuning aligns outputs [8] but is resource-intensive. Consistency/arbitrage scoring enables instant evaluation [12]. Active semi-supervised debiasing supports low-budget labeling [**?**].

Knowledge graph embedding techniques and their sophisticated mathematical foundations have demonstrated significant and growing potential in complex prediction tasks, offering powerful approaches for representing and reasoning about structured knowledge. Han et al. [20] developed an innovative temporal knowledge graph embedding model based on variable translation mechanisms for link prediction and triplet classification tasks, which offers valuable insights and methodological foundations for multi-agent information integration approaches that require sophisticated handling of temporal relationships and entity interactions. Similarly, Duan et al. [21] proposed a novel approach for inductive relation prediction using carefully designed disentangled subgraph structures, providing robust methodological foundations and theoretical frameworks for handling complex entity relationships and their intricate dependencies within the comprehensive MAEPS framework.

Multi-agent systems and collaborative computational approaches are becoming increasingly important and strategically valuable in complex prediction scenarios that require coordination of diverse information sources and analytical perspectives. Du et al. [22] presented a sophisticated deep reinforcement learning approach for scheduling low-latency medical services in distributed healthcare cloud environments, demonstrating the remarkable effectiveness of intelligent resource allocation strategies and coordinated decision-making processes. This research aligns closely with the fundamental design principles of the MAEPS framework, where specialized agents collaborate systematically and coordinate their information-gathering efforts to achieve superior collective performance. Furthermore, Shang et al. [23] explored innovative ensemble knowledge distillation techniques in federated semi-supervised learning environments, providing valuable theoretical support and practical methodologies for multi-model aggregation strategies in prediction generation, which directly informs our approach to combining insights from multiple specialized agents.

Nevertheless, existing empirical findings and theoretical analyses suggest that when LLMs adopt traditional and conventional forecasting strategies—such as systematically decomposing complex events into simpler components or methodically analyzing pros and cons through structured reasoning—their overall performance may surprisingly underperform simple baseline prompting approaches that rely on more direct and intuitive reasoning processes [24]. Notably, that comprehensive study focuses primarily on closed datasets containing independent events with limited temporal dependencies and does not integrate real-time information aggregation capabilities or dynamic knowledge updating mechanisms. Thus, a central and persistent challenge remains: how to design and implement LLM-based forecasting systems that can effectively leverage real-time information streams while simultaneously maintaining high predictive accuracy and reliability across diverse domains and time horizons. Recent advances in future prediction benchmarks, such as the

comprehensive FutureX framework [25], provide systematic and rigorous evaluation frameworks for LLM agents in prediction tasks, highlighting the critical importance of systematic assessment and standardized evaluation methodologies in this rapidly evolving and increasingly complex field.

3 Preliminary

3.1 Core Notation and Definitions

For binary event prediction tasks where the target event occurrence is denoted as $Y \in \{0, 1\}$, we define the following core concepts:

Effective Information Set $\mathcal{I}$: The collection of information atoms v such that Y and v are not independent, i.e., $P(Y \mid v) \neq P(Y)$ (equivalently, the mutual information $I(Y; v) > 0$). This set consists of verifiable objective facts, excluding subjective opinions and speculative content.

Accessible Information Subset $\mathcal{V}$: We define a universe of accessible information $\mathcal{U}$, and let $\mathcal{V} \subseteq \mathcal{U}$ denote the actually obtainable set by the model. Write $\mathcal{V} = \mathcal{V}_{\text{rel}} \cup \mathcal{V}_{\text{irr}}$ with $\mathcal{V}_{\text{rel}} = \mathcal{V} \cap \mathcal{I}$ and $\mathcal{V}_{\text{irr}} = \mathcal{V} \setminus \mathcal{I}$. In MAEPS, $\mathcal{V}$ can be decomposed as $\mathcal{V} = \bigcup_{k=1}^{12} \mathcal{V}_k$, where $\mathcal{V}_k$ denotes the retrieval results of the k-th specialized agent.

Loss Function L: We adopt the Brier loss, defined as $L(f(\mathcal{V}), Y) = (f(\mathcal{V}) - Y)^2$, where $f(\mathcal{V}) \in [0, 1]$ is the model's predicted probability of $Y = 1$.

Theoretical Risk $R(\mathcal{V})$: The minimum expected loss of the optimal predictor given $\mathcal{V}$, defined as:

$$R(\mathcal{V}) = \inf_{f \in \mathcal{F}(\mathcal{V})} \mathbb{E}[(f(\mathcal{V}) - Y)^2]. \tag{1}$$

By the L^2 projection property, the Bayes optimal predictor is $f^*(\mathcal{V}) = \mathbb{E}[Y \mid \mathcal{V}]$, and the minimum Brier risk equals $R(\mathcal{V}) = \mathbb{E}[\mathrm{Var}(Y \mid \mathcal{V})]$. Here $\mathcal{F}(\mathcal{V})$ denotes the class of all square-integrable measurable functions with respect to $\sigma(\mathcal{V})$.

3.2 Theoretical Foundation of Multi-Agent Collaboration

Principle 1: Monotonic Information Sufficiency

Theorem: If $\mathcal{V}_a \subseteq \mathcal{V}_b$, then $R(\mathcal{V}_b) \leqslant R(\mathcal{V}_a)$.

Proof: Since $\mathcal{V}_a \subseteq \mathcal{V}_b$, we have $\mathcal{F}(\mathcal{V}_a) \subseteq \mathcal{F}(\mathcal{V}_b)$ (predictors based on $\mathcal{V}_a$ are contained in the predictor set for $\mathcal{V}_b$). Taking the infimum over expected loss:

$$R(\mathcal{V}_b) = \inf_{f \in \mathcal{F}(\mathcal{V}_b)} \mathbb{E}[L(f, Y)] \leqslant \inf_{f \in \mathcal{F}(\mathcal{V}_a)} \mathbb{E}[L(f, Y)] = R(\mathcal{V}_a). \tag{2}$$

Equality Condition: $R(\mathcal{V}_b) = R(\mathcal{V}_a)$ if and only if $\mathcal{V}_b \setminus \mathcal{V}_a$ is conditionally independent of Y given $\mathcal{V}_a$, i.e., $P(Y \mid \mathcal{V}_a, \mathcal{V}_b \setminus \mathcal{V}_a) = P(Y \mid \mathcal{V}_a)$. Equivalently, $\mathbb{E}[Y \mid \mathcal{V}_b] = \mathbb{E}[Y \mid \mathcal{V}_a]$ almost surely (a.s.).

Principle 2: Finite Sample Degradation from Irrelevant Information

Theorem: While theoretically introducing $v \notin \mathcal{I}$ (conditionally independent of Y) does not change $R(\mathcal{V})$, in finite samples, v increases parameter estimation variance, leading to degraded prediction performance.

Proof: Consider the true model $\mathrm{logit}(P(Y = 1 \mid X_{\text{rel}})) = X_{\text{rel}}^T \beta$, where X_{rel} corresponds to relevant features from $\mathcal{V}$. Introducing irrelevant features X_{irr} (corresponding to $v \notin \mathcal{I}$), the augmented feature matrix becomes $X = [X_{\text{rel}}, X_{\text{irr}}]$. The asymptotic covariance of the parameter estimate $\hat{\beta}$ is $\mathcal{J}(X)^{-1}$, where $\mathcal{J}(X)$ denotes the Fisher information matrix. We have, in the Loewner (positive semidefinite) order for the subvector corresponding to the relevant coefficients,

$$\mathrm{Var}(\hat{\beta}) \succeq \mathrm{Var}(\hat{\beta}_{\text{rel}}). \tag{3}$$

Equality holds only when $X_{\text{rel}}^T W X_{\text{irr}} = 0$ and X_{irr} carries no information about Y (where W is the GLM weight matrix, e.g., $\mathrm{diag}(p_i(1 - p_i))$ in logistic regression), which is practically unattainable. Despite the difference in loss functions, this conclusion is consistent with the general bias–variance tradeoff underlying the Brier risk analysis above. Note that this degradation concerns implementable finite-sample estimators rather than the Bayes risk limit.

This theoretical foundation motivates our multi-agent approach: each specialized agent focuses on collecting information from $\mathcal{I}$ within their domain expertise, maximizing information relevance while minimizing noise introduction.

Principle 3: Existence of a Global Optimum for Coordinated Retrieval

Theorem: Define the coordination objective over a finite ground set of candidate facts Ω as

$$\max_{S \subseteq \Omega} F(S) := -R(S) - \lambda C(S), \quad \lambda \geqslant 0, \tag{4}$$

where $R(S)$ is the minimum Brier risk given facts S and $C(S)$ is a nonnegative, monotone cost (e.g., total LLM token cost or number of API calls). Then a global maximizer $S^\star$ exists. Moreover, if Ω is finite (as in any practical run), existence is trivial; if Ω is infinite but compact under an appropriate topology and F is upper semicontinuous, a maximizer also exists by Weierstrass' theorem.

Proof (sketch): For finite Ω, F attains its maximum as it is evaluated over a finite set. In the infinite but compact case, F being upper semicontinuous ensures the existence of a maximizer. In practice, MAEPS works with a finite candidate pool per round, so an optimum exists at each iteration and for the union across iterations under a fixed budget.

Principle 4: Convergence and Approximation of Greedy Coordination

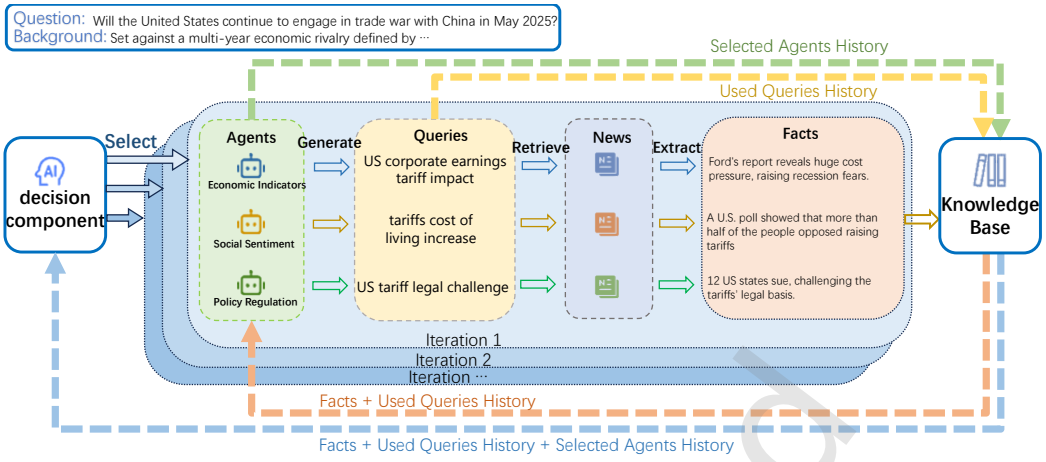


Fig. 2 The retrieval phase of a multi-agent event forecasting framework, where each round involves the decision component selecting dimensions for search, and the extracted facts and intermediate processes are saved to a knowledge base to guide subsequent iterations.

Let $G(S) := R(\emptyset) - R(S)$ denote the risk reduction when conditioning on S. Assume G is monotone and submodular with respect to the set of unique facts extracted into the knowledge base and that C is modular. Consider a greedy policy that, in each iteration, selects the next agent/query bundle maximizing the marginal gain-per-cost ratio $\Delta F/\Delta C$ until budget exhaustion or no positive gain.

Theorem: The greedy policy terminates in finitely many steps and achieves a $(1 - 1/e)$-approximation to the optimal value under a cardinality or knapsack budget. With a diminishing improvement threshold (or damping), the iterative MAEPS loop converges to a fixed point where no further positive-gain additions exist.

Proof (sketch): Finite termination follows from a finite candidate pool per iteration and strictly nonnegative costs. The $(1 - 1/e)$ guarantee follows from classical results for maximizing monotone submodular functions under budget constraints. Convergence of the outer loop follows from the monotonicity of G and bounded budget—after finitely many insertions, no action yields positive marginal gain, and the process stops.

4 Method

MAEPS (Multi-Agent Event Prediction System) is an iterative framework that simulates expert team collaboration for complex prediction. It comprises three core components: **Coordination and Decision** for process control, **Information Collection** with 12 specialized agents, and **Knowledge Processing** for analysis and prediction generation. Through iterative cycles of information gathering, integration, and evaluation, MAEPS produces a comprehensive and reliable final prediction.

4.1 Multi-Agent Information Retrieval

Prediction research and practice show that effective prediction requires multi-angle analysis. For instance, the PESTEL model emphasizes examining the environment from six macro-dimensions. Similarly, "Superforecasting" [26] notes that top forecasters have a "dragonfly-eye" view, integrating information from many domains. Based on this, we design 12 specialized agents, each responsible for collecting information in a specific dimension. These agents can execute search tasks autonomously and collaborate through a shared knowledge base. The knowledge base is the system's core, storing and managing all information to support unified retrieval, updates, and conflict detection. Figure 2 illustrates the retrieval phase of this multi-agent framework, showing how each round involves the decision component selecting dimensions for search, with extracted facts and intermediate processes saved to guide subsequent iterations. Table 1 introduces each agent's function.

Each agent can perform comprehensive autonomous analysis and sophisticated reasoning processes, generating highly targeted and contextually relevant search queries based on the specific nuances and complexities of the problem's context. For example, regarding a complex US-China trade war question, the Key People Agent would strategically query "Donald Trump" to gather information about key decision-makers and their policy positions, while the Competitive Landscape Agent would systematically query "US tech company supply chain transfer progress" to understand the broader economic implications and strategic shifts. This intelligent and adaptive search strategy ensures both the depth and relevance of information collection, maximizing the quality and

comprehensiveness of the gathered data.

The system utilizes two primary and complementary information sources: News API for accessing real-time news updates and current events, and Wikipedia for retrieving structured encyclopedic knowledge and historical context. To ensure complete fairness and prevent any potential information leakage that could compromise the integrity of the prediction process, all search operations are strictly restricted to the time period before the prediction question's resolution date. News API returns only news articles published within the specified date range, and Wikipedia returns carefully curated historical versions from a specific time point, ensuring temporal consistency. This rigorous temporal constraint ensures the authenticity and reliability of the prediction process while maintaining the integrity of the experimental setup.

Fact Extraction: To systematically address the critical lack of factual screening and information quality control, the system integrates a sophisticated fact extraction feature that serves as a crucial filtering mechanism. This component is particularly important because non-factual opinions (e.g., author's subjective judgments, speculative comments, and unverified claims) often conflict with each other and can significantly impair model judgment and decision-making capabilities. After obtaining a document through the retrieval process, the system calls an LLM, following the optimal settings and configurations from [4], to provide the document title and the first 250 words for analysis. The model then systematically extracts key facts while carefully filtering out subjective opinions and speculative content, with citation-aware generation for traceability [**?**]. This comprehensive filtering process ensures that the information stored in the knowledge base is objective, verifiable, and factually accurate, providing a reliable and trustworthy foundation for prediction generation.

Fact Extraction Prompt: The system employs the following prompt for extracting verifiable facts from documents:

> *"Given the document title and first 250 words below, extract 3-5 key facts that are most relevant to the prediction task. Focus on verifiable, objective information and clearly distinguish facts from opinions or predictions. For each fact, provide: (1) the factual statement, (2) source attribution, and (3) timestamp if available. Exclude subjective judgments, speculative comments, and unverified claims. Format: [Fact 1: statement — Source: attribution — Time: timestamp]."*

4.2 Temporal Conflict Processing

Real-world events are inherently dynamic and constantly evolving, and information about these events continuously changes and develops over time. Old reports and initial assessments can be systematically superseded by new developments and updated information, leading to potentially outdated or contradictory information in the collected dataset. To comprehensively address this critical challenge, the framework includes a sophisticated **temporal consistency guarantee** module with two essential and interconnected functions: systematically assessing information timeliness and quality, and effectively resolving factual conflicts by constructing detailed and chronologically accurate event timelines.

Information Quality Assessment and Timeliness: The system first conducts a comprehensive and multi-dimensional assessment of the content quality of new information based on rigorous criteria including accuracy, credibility, objectivity, and source reliability [27, 28]. More importantly and critically, it systematically evaluates the information's **timeliness** and temporal relevance. In prediction tasks, information value naturally decays over time due to changing circumstances and evolving contexts; newer and more recent information is inherently more valuable and relevant for accurate predictions. We quantify this temporal decay phenomenon with a sophisticated "timeliness score" calculation:

$$\text{Timeliness Score} = 5.0 \times e^{-\lambda \times \Delta t} \tag{5}$$

where Δt represents the number of days between the information's publication date and the prediction deadline, and λ is the carefully calibrated decay constant (0.01). This timeliness score is then systematically combined with the comprehensive content quality score to calculate a final weighted score, effectively prioritizing newer facts while maintaining consideration for information quality and reliability.

Temporal Conflict Resolution and Timeline Generation: The key and most critical aspect of handling information from different time points is systematically resolving factual conflicts and inconsistencies. For example, an early report might state that a policy is "under review," while a later and more recent report confirms that it is "rejected." Without carefully considering temporal relationships and chronological context, these facts appear contradictory and conflicting. The system resolves this complex challenge through the following systematic and comprehensive approach:

(1) **Conflict Identification**: The system systematically identifies and analyzes facts describing the same event or entity with conflicting or contradictory content, using sophisticated natural language processing techniques to detect semantic inconsistencies and temporal discrepancies.

(2) **Conflict Marking**: The system effectively resolves conflicts based on comprehensive analysis of publication

timestamps and source credibility. The latest and most recent fact is marked as the **current valid state**, while older and potentially outdated information is systematically marked as **"outdated,"** forming a comprehensive event evolution record that preserves historical context.

(3) **Dynamic Timeline Generation**: The system organizes all verified and marked facts chronologically to construct a complete and detailed **event development timeline**. This timeline serves as a dynamic narrative showing how events evolve and develop over time. Presenting this **complete story with evolution and contradictions** to the LLM helps it make more reliable and well-informed predictions by understanding the full context and temporal progression of events.

Conflict Detection Prompt: We also employ an LLM to explicitly detect contradictions and recency overrides:

> *"Given a list of timestamped facts about the same entity or event, identify any contradictions or inconsistencies. Apply the following resolution rules: (1) Prefer newer facts over older ones when timestamps differ, (2) If timestamps are identical, prioritize higher-credibility sources (official statements > major news outlets > blogs > social media), (3) Mark contradicted older facts as 'outdated' while preserving them for timeline context, (4) Merge consistent facts that complement each other. Return your analysis in the format: (kept_facts, outdated_facts, resolution_rationale)."*

4.3 Prediction Generation

The system systematically terminates and generates a comprehensive final prediction when it reaches the predetermined maximum number of iterations or when it determines that no new information gaps or knowledge deficiencies can be identified and addressed. The prediction generation process follows the established methodology of [4] but incorporates significantly different and enhanced input content. It utilizes the quality-assessed and temporally processed knowledge base content, combined with the detailed event development timeline and comprehensive temporal context, along with a sophisticated aggregation model for generating accurate and reliable predictions.

To ensure maximum robustness and reliability while maintaining computational efficiency, we employ a sophisticated multi-model aggregation strategy that avoids the complexity and high computational cost of fine-tuning specialized models. Referencing and building upon the proven approach in [4], the system strategically uses five identical base LLMs to generate diverse candidate predictions, each leveraging the same high-quality knowledge base but with slightly different reasoning approaches. Each model uses the best-verified and optimized reasoning prompt, with slight optimizations and variations for different instances, which introduces beneficial diversity in the reasoning processes and prompt formulations. Finally, the system integrates and synthesizes the five individual predictions using a robust trimmed mean aggregation algorithm to output the final probability estimate and confidence interval, ensuring that the final prediction is both accurate and well-calibrated.

5 Experiment

We conduct comprehensive experimental evaluation of MAEPS by systematically comparing it against strong baseline methods and conducting detailed analysis of its core components and individual contributions.

5.1 Experimental Setup

We conduct thorough evaluation on two distinct and complementary datasets: (1) our carefully curated dataset of 306 binary prediction questions, constructed by rigorously following the established protocol in [4]. We systematically collect candidate questions from the well-established forecasting platforms Polymarket and Metaculus, and then carefully screen out ambiguous or niche topics that might introduce bias; all questions have resolution deadlines between January and May 2025, and were proposed after the models' knowledge cutoff to prevent any potential data leakage. The final comprehensive set spans diverse domains including politics, economics, technology, and sports. (2) the publicly available dataset from Halawi et al. [4], which we use as an external benchmark to rigorously test generalization capabilities and cross-dataset performance. Unless otherwise explicitly stated, the experimental settings and evaluation procedure are identical and consistent across the two datasets.

We systematically evaluate model performance using two core and complementary metrics: **Accuracy**, which measures the proportion of correct predictions and reflects the model's ability to make accurate binary decisions, and **Brier Score** [29], which comprehensively assesses the accuracy of probabilistic forecasts and measures both calibration and resolution. Lower Brier scores indicate better accuracy and calibration performance. A random guess baseline yields a 0.25 Brier score.

5.1.1 Data Leakage Prevention

To rigorously prevent ex-post leakage of outcomes into the inputs and ensure experimental integrity, we enforce strict and comprehensive constraints: (1) all retrieved news articles and information sources must have publication dates *strictly before* each question's resolution date; (2) for encyclopedic

Table 2 Main Performance Comparison

Base Model	Method	Accuracy		Brier Score (↓)	
		Ours	Halawi[4]	Ours	Halawi[4]
-	**Human Crowd**	**0.790**	**0.780**	**0.141**	**0.152**
GPT-4o	**MAEPS (ours)**	**0.740**	0.733	**0.165**	**0.169**
	StkFEP-Re	0.720	**0.736**	0.176	0.178
	Halawi et al.	0.672	0.670	0.182	0.186
GPT-4	**MAEPS (ours)**	**0.710**	**0.705**	**0.171**	**0.174**
	StkFEP-Re	0.680	0.676	0.174	0.178
	Halawi et al.	0.703	0.700	0.191	0.195
GPT-3.5-Turbo	**MAEPS (ours)**	**0.670**	0.665	**0.192**	**0.196**
	StkFEP-Re	0.660	**0.679**	0.202	0.205
	Halawi et al.	0.610	0.606	0.219	0.222

sources (e.g., Wikipedia), we use carefully selected historical snapshots prior to the resolution date; (3) every question is chosen such that its resolution date is *after* the base models' knowledge cutoff, ensuring the models had no prior knowledge of outcomes; (4) we simulate realistic prediction days and only allow information available up to that specific day. These rigorous constraints guarantee that the model never sees the true outcome during the forecasting process.

5.1.2 Reproducibility and External API Dependence

Since the original prediction datasets do not contain associated news articles, external APIs (e.g., news search services) are necessary for information retrieval. To significantly enhance reproducibility despite this API dependence, we implement comprehensive measures: logging and releasing exact queries, time windows, and sources used; caching all retrieved URLs and storing complete article snapshots with timestamps and checksums for verification; fixing API parameters and rate limits to ensure consistency. This comprehensive protocol allows independent re-execution and verification even if API indices evolve over time. We systematically compare MAEPS against four carefully selected and representative baselines to comprehensively evaluate its performance:

- **Human Crowd**: Aggregated predictions from experienced human forecasters on established real-world forecasting platforms, serving as our gold standard benchmark.
- **Halawi et al.** [4]: A sophisticated retrieval-augmented LLM system that generates a predetermined fixed number of search queries for information gathering.
- **StkFEP-Re**: A comprehensive hybrid retrieval method that systematically follows the OpenEP framework [6] by collecting detailed information on **stakeholders**, **relevant events**, and **similar events**. For ensuring fair and unbiased comparison, its subsequent processing pipeline is carefully matched to MAEPS.

Table 3 MAEPS Ablation Experiment Results

Method	Brier Score (↓)
MAEPS (Full)	**0.165**
w/o Multi-Agent	0.190
w/o Timeline	0.173
w/o Fact Extract	0.187
w/o Search	0.224

5.1.3 Models

To rigorously test for general applicability and robustness across different model architectures, we comprehensively evaluate all methods on three distinct base models with varying capabilities: GPT-4o, GPT-4, and GPT-3.5-Turbo.

5.1.4 Evaluation Process and Parameter Settings

Our comprehensive experimental process is carefully designed and unified across all evaluations to ensure methodological rigor and complete reproducibility. We systematically compare model performance against the Human Crowd, which serves as our authoritative gold standard benchmark for validation. We meticulously simulate real-world prediction scenarios by strictly following [4]'s well-established methodology of using "simulated prediction days" to rigorously limit information retrieval and comprehensively prevent any potential data leakage. The final performance score is calculated through a carefully structured two-step averaging process: first averaging across all simulated days for each individual question, then averaging across all questions in the dataset. MAEPS is configured with 3 iterations for optimal performance, with 3-4 agents strategically selected per round, each retrieving exactly 10 articles to maintain consistency. For ensuring fairness and eliminating bias, all retrieved articles undergo comprehensive quality-filtering by an LLM before being used for prediction.

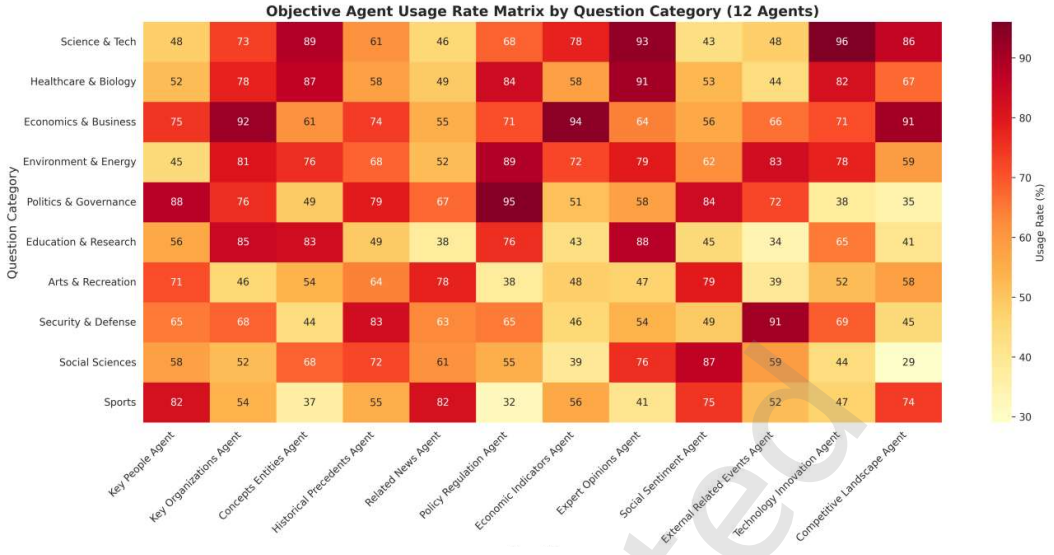


Fig. 3 This heatmap illustrates the usage rates of 12 objective agents across 10 different question categories, with color intensity representing the percentage of usage for each agent - category combination.

5.2 Experimental Results and Analysis

5.2.1 Main Performance Comparison

To comprehensively verify generalizability and robustness across different model architectures and datasets, we systematically evaluate all methods on three distinct base models (GPT-4o, GPT-4, GPT-3.5-Turbo) across two carefully curated datasets (our comprehensive dataset and the established benchmark from [4]). Table 2 (two-column format) provides a detailed summary of the results with strategically split columns under Accuracy and Brier for the two datasets, enabling clear cross-dataset comparison. Overall performance trends demonstrate remarkable consistency across datasets; MAEPS exhibits highly competitive performance—achieving results remarkably close to Human Crowd on our datasets—while showing dataset-specific variance on [4]'s benchmark that reflects inherent dataset characteristics rather than systematic limitations, thereby maintaining and validating the framework's fundamental advantages.

In terms of retrieval strategy effectiveness, StkFEP-Re consistently and significantly outperforms [4]'s baseline method, providing compelling evidence that a structured, multi-dimensional approach systematically and reliably improves prediction performance. MAEPS, leveraging its more sophisticated, flexible, and adaptive multi-agent search architecture, achieves even more impressive results, thereby providing strong empirical validation of our system's core philosophical approach and technical design principles. More powerful and advanced LLMs demonstrate superior capability in unlocking the framework's full potential; progressing from GPT-3.5-Turbo to GPT-4o, MAEPS's Brier score shows substantial improvement of 0.027, indicating significant performance gains. Notably, GPT-4 occasionally underperforms relative to GPT-4o due to its earlier knowledge cutoff date, a phenomenon that has been consistently observed and documented in [30].

5.2.2 Ablation Studies

Table 3 comprehensively shows the detailed ablation results, which systematically reveal each component's individual contribution to overall system performance. We use GPT-4o as the base model for consistency and optimal performance evaluation.

- **Search**: Removing external search capabilities causes the most substantial and significant performance drop (from 0.165 to 0.224), primarily because the model's internal knowledge becomes outdated and fundamentally lacks access to real-time, current information that is crucial for accurate predictions.
- **Multi-Agent**: Replacing our sophisticated multi-agent coordination mechanism with a single, broad search strategy following [4] significantly and substantially degrades performance (from 0.165 to 0.190), even when maintaining the same total number of retrieved articles, demonstrating the critical importance of specialized agent coordination.
- **Fact Extraction**: Removing the comprehensive fact extraction step and directly using the raw text extracted by the method from [4] for analysis causes a notable and concerning performance decline. The system completely

Table 4 Agent Comprehensive Assessment Ranking

Agent	Comprehensive Score	Unique Contribution	Novelty Contribution	Ablation Criticality
Related News	6.42	0.925	1.85%	8.29
External Related Events	5.91	0.820	1.45%	6.33
Key People	5.25	0.720	0.82%	7.95
Economic Indicators	5.08	0.695	0.55%	7.21
Social Sentiment	5.15	0.715	0.85%	6.79
Competitive Landscape	5.02	0.665	0.62%	6.29
Historical Precedents	4.83	0.655	0.22%	5.17
Policy Regulation	4.89	0.645	0.58%	7.08
Technology Innovation	4.85	0.635	0.72%	7.04
Key Organizations	4.72	0.610	0.28%	6.29
Expert Opinions	4.78	0.605	0.52%	5.33
Concepts Entities	4.18	0.495	0.25%	5.04

loses its crucial ability to filter for high-quality, relevant information and systematically remove irrelevant opinions and subjective content. Large amounts of subjective views and informational noise directly interfere with the model's judgment capabilities, which proves to be highly detrimental for a sophisticated multi-agent system.

- **Timeline**: Disabling the sophisticated temporal consistency guarantee module, where the system does not process temporal conflicts and the model is not provided with the crucial publication times of facts, also leads to a measurable performance decline (from 0.165 to 0.173) as predictions cannot effectively prioritize new, objective information over outdated or potentially contradictory data.

5.2.3 Agents and System Parameter Analysis

To comprehensively validate the theoretical rationale and practical complementarity of the 12 specialized agents in their sophisticated division of labor and to systematically find the optimal system configuration parameters, we conducted a series of detailed and rigorous analyses.

Agent Specialization Analysis To thoroughly assess each agent's individual value and contribution to the overall system performance, we carefully define four key metrics that capture different aspects of agent effectiveness:

- **Comprehensive Score**: Calculated via a sophisticated weighted model based on multi-criteria decision analysis methodology, incorporating various performance dimensions.
- **Unique Contribution Rate**: Precisely measures the proportion of information an agent retrieved that no other agent successfully discovered, quantified by advanced Jaccard similarity coefficients or semantic vector overlap analysis.
- **Novelty Contribution Rate**: Systematically measures an agent's specialized ability to find entirely new, previously unseen information that was not discovered in previous search rounds.
- **Ablation Criticality Score**: Quantitatively measures the substantial negative impact on overall system performance when a single agent is completely removed from the multi-agent ensemble.

The comprehensive analysis results (see Figure 3 and Table 4) clearly demonstrate that the agents have successfully formed a clear, well-defined specialized division of labor, and each individual agent possesses indispensable unique value, consistently providing novel, high-quality information to systematically enhance overall system performance.

System Parameter Sensitivity Analysis To systematically find the optimal configuration parameters that maximize performance while maintaining computational efficiency, we conduct a comprehensive sensitivity analysis on the number of agents selected per round and the total number of iterations. The detailed results (see Figure 4) clearly show that selecting 3-4 agents and iterating 3 times achieve the best balance between prediction performance and computational cost. Too few agents leads to insufficient information dimensions and coverage; too many agents introduces redundancy and noise that actually degrades performance. Therefore, we strategically chose this configuration as the system's default operational parameters.

6 Conclusion

This paper comprehensively introduces the innovative MAEPS framework to systematically address the critical issues of unidimensional information retrieval and inadequate temporal processing that significantly limit existing LLM-based prediction systems. By intelligently simulating

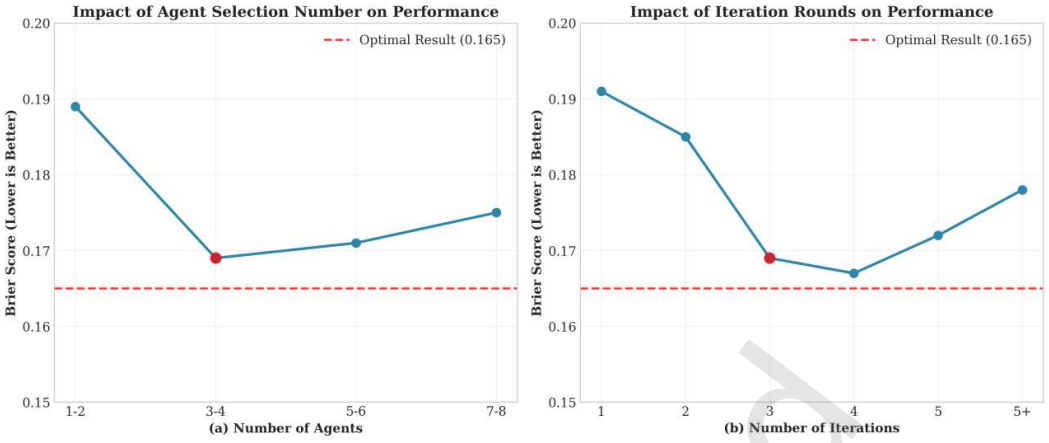


Fig. 4 These two graphs depict the impact of the number of agents per round (a) and iteration rounds (b) on predictive performance, with the optimal experimental outcome (0.165) indicated by a red dashed line.

human expert teams and their collaborative decision-making processes, the framework strategically employs 12 specialized agents for multi-dimensional, adaptive information gathering that covers diverse aspects of prediction scenarios. The sophisticated system seamlessly integrates advanced fact extraction and temporal consistency mechanisms to automatically filter high-quality facts and systematically resolve temporal conflicts, thereby ensuring that predictions are consistently based on reliable, up-to-date, and comprehensive information. Extensive experiments conducted across multiple datasets clearly demonstrate that the framework effectively and substantially improves prediction accuracy and reliability compared to existing baseline methods.

Future work will focus on exploring enhanced search strategies, integrating more diverse data sources, and extending the framework to additional prediction domains.

7 Acknowledgments

This work is supported by the National Natural Science Foundation of China (No. U24A20335, No. 62176257, No. 62576340). This work is sponsored by Beijing Nova Program (No. 20250484750) and supported by Beijing Natural Science Foundation (L243006). This work is also supported by the Youth Innovation Promotion Association CAS.

Reference

[1] V. Rotaru, Y. Huang, T. Li, J. Evans, and I. Chattopadhyay, Event-level prediction of urban crime reveals a signature of enforcement bias in us cities, *Nature human behaviour*, vol. 6, no. 8, pp. 1056–1068, 2022.

[2] X. Li, Z. Li, C. Shi, Y. Xu, Q. Du, M. Tan, J. Huang, and W. Lin, Alphafin: Benchmarking financial analysis with retrieval-augmented stock-chain framework, *arXiv preprint arXiv:2403.12582*, 2024.

[3] A. Shabani, A. Abdi, L. Meng, and T. Sylvain, Scaleformer: Iterative multi-scale refining transformers for time series forecasting, *arXiv preprint arXiv:2206.04038*, 2022.

[4] D. Halawi, F. Zhang, C. Yueh-Han, and J. Steinhardt, Approaching human-level forecasting with language models, *arXiv preprint arXiv:2402.18563*, 2024.

[5] E. Hsieh, P. Fu, and J. Chen, Reasoning and tools for human-level forecasting, *arXiv preprint arXiv:2408.12036*, 2024.

[6] Y. Guan, H. Peng, X. Wang, L. Hou, and J. Li, Openep: Open-ended future event prediction, *arXiv preprint arXiv:2408.06578*, 2024.

[7] Z. Tao, Z. Jin, B. Li, X. Bai, H. Zhao, C. Dou, X. Chen, J. Li, L. Li, and C. Tao, Prophet: An inferable future forecasting benchmark with causal intervened likelihood estimation, *arXiv preprint arXiv:2504.01509*, 2025.

[8] B. Turtel, D. Franklin, and P. Schoenegger, Llms can teach themselves to better predict the future, *arXiv preprint arXiv:2502.05253*, 2025.

[9] L. Golob and A. Sittar, Fact manipulation in news: Llm-driven synthesis and evaluation of fake news annotation, 2024.

[10] Q. Wang, Y. Gao, Z. Tang, B. Luo, and B. He, Enhancing llm trading performance with fact-subjectivity aware reasoning, *arXiv preprint arXiv:2410.12464*, 2024.

[11] L. Zhao, Event prediction in big data era: A systematic survey. arxiv preprint, *ArXivorg*, 2020.

[12] D. Paleka, A. P. Sudhir, A. Alvarez, V. Bhat, A. Shen, E. Wang, and F. Tramèr, Consistency checks for language model forecasters, *arXiv preprint arXiv:2412.18544*, 2024.

[13] D. Chen, H. J. Yoon, Z. Wan, N. Alluru, S. W. Lee, R. He, T. J. Moore, F. F. Nelson, S. Yoon, H. Lim *et al.*, Advancing human-machine teaming: Concepts, challenges, and applications, *arXiv preprint arXiv:2503.16518*, 2025.

[14] X. Chen and Q. Cheng, Acute complication prediction and diagnosis model clstm-bpr: A fusion method of time series deep learning and bayesian personalized ranking, *Tsinghua Science and Technology*, vol. 29,

no. 5, pp. 1509–1523, 2024. [Online]. Available: https://www.sciopen.com/article/10.26599/TST.2023.9010103

[15] M. Granroth-Wilding and S. Clark, What happens next? event prediction using a compositional neural network model, in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 30, no. 1, 2016.

[16] N. Mostafazadeh, M. Roth, A. Louis, N. Chambers, and J. F. Allen, Lsdsem 2017 shared task: The story cloze test, in *2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics*. Association for Computational Linguistics, 2017, pp. 46–51.

[17] S. Yuan, J. Chen, Z. Fu, X. Ge, S. Shah, C. R. Jankowski, Y. Xiao, and D. Yang, Distilling script knowledge from large language models for constrained language planning, *arXiv preprint arXiv:2305.05252*, 2023.

[18] E. Karger, H. Bastani, C. Yueh-Han, Z. Jacobs, D. Halawi, F. Zhang, and P. E. Tetlock, Forecastbench: A dynamic benchmark of ai forecasting capabilities, *arXiv preprint arXiv:2409.19839*, 2024.

[19] S. Lippert, A. Dreber, M. Johannesson, W. Tierney, W. Cyrus-Lai, E. L. Uhlmann, E. E. Collaboration, and T. Pfeiffer, Can large language models help predict results from a complex behavioural science study? *Royal Society Open Science*, vol. 11, no. 9, p. 240682, 2024.

[20] Y. Han, G. Lu, S. Zhang, L. Zhang, C. Zou, and G. Wen, A temporal knowledge graph embedding model based on variable translation, *Tsinghua Science and Technology*, vol. 29, no. 5, pp. 1554–1565, 2024. [Online]. Available: https://www.sciopen.com/article/10.26599/TST.2023.9010142

[21] G. Duan, R. Guo, W. Luo, G. Luo, and T. Huang, Inductive relation prediction by disentangled subgraph structure, *Tsinghua Science and Technology*, vol. 29, no. 5, pp. 1566–1579, 2024. [Online]. Available: https://www.sciopen.com/article/10.26599/TST.2023.9010154

[22] H. Du, M. Liu, N. Liu, D. Li, W. Li, and L. Xu, Scheduling of low-latency medical services in healthcare cloud with deep reinforcement learning, *Tsinghua Science and Technology*, vol. 30, no. 1, pp. 100–111, 2025. [Online]. Available: https://www.sciopen.com/article/10.26599/TST.2024.9010033

[23] E. Shang, H. Liu, J. Zhang, R. Zhao, and J. Du, Ensemble knowledge distillation for federated semi-supervised image classification, *Tsinghua Science and Technology*, vol. 30, no. 1, pp. 112–123, 2025. [Online]. Available: https://www.sciopen.com/article/10.26599/TST.2023.9010156

[24] S. Pratt, S. Blumberg, P. K. Carolino, and M. R. Morris, Can language models use forecasting strategies? *arXiv preprint arXiv:2406.04446*, 2024.

[25] Z. Zeng, J. Liu, S. Chen, T. He, Y. Liao, Y. Tian, J. Wang, Z. Wang, Y. Yang, L. Yin, M. Yin, Z. Zhu, T. Cai, Z. Chen, J. Chen, Y. Du, X. Gao, J. Guo, L. Hu, J. Jiao, X. Li, J. Liu, S. Ni, Z. Wen, G. Zhang, K. Zhang, X. Zhou, J. Blanchet, X. Qiu, M. Wang, and W. Huang, Futurex: An advanced live benchmark for llm agents in future prediction, 2025. [Online]. Available: https://arxiv.org/abs/2508.11987

[26] P. E. Tetlock and D. Gardner, *Superforecasting: The Art and Science of Prediction*. New York, NY: Crown Publishing Group, 2015.

[27] B. Stvilia, L. Gasser, M. B. Twidale, and L. C. Smith, A framework for information quality assessment, *Journal of the American society for information science and technology*, vol. 58, no. 12, pp. 1720–1733, 2007.

[28] R. Y. Wang and D. M. Strong, Beyond accuracy: What data quality means to data consumers, *Journal of management information systems*, vol. 12, no. 4, pp. 5–33, 1996.

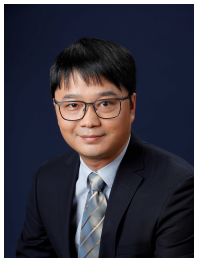
[29] G. W. Brier, Verification of forecasts expressed in terms of probability, *Monthly weather review*, vol. 78, no. 1, pp. 1–3, 1950.

[30] H. Dai, R. Teehan, and M. Ren, Are llms prescient? a continuous evaluation using daily news as the oracle, *arXiv preprint arXiv:2411.08324*, 2024.

Author biography

Chengyuan Jin received the BE degree from China University of Petroleum in 2024. He is currently a master student at School of Artificial Intelligence, University of Chinese Academy of Sciences, China. His research interests focus on natural language processing, retrieval augmentation, and event reasoning.

Yubo Chen is an associate professor at the Key Laboratory of Cognition and Decision Intelligence for Complex Systems,Institute of Automation, Chinese Academy of Sciences. He received his Ph.D. degree from Institute of Automation, Chinese Academy of Sciences, Beijing, in 2017. His research interests focus on natural language processing, knowledge graph construction and event extraction. He has published more than 40 papers in international conferences including ACL, EMNLP, COLING, IJCAI and AAAI.



Kang Liu is a professor at at the Key Laboratory of Cognition and Decision Intelligence for Complex Systems,Institute of Automation, Chinese Academy of Sciences. He received his Ph.D. degree from Institute of Automation, Chinese Academy of Sciences, Beijing, in 2010. His research interests focus on natural language processing, relation extraction, event extraction and dialogue system. He has published more than 80 papers in international journals and conferences including TKDE, ACL, EMNLP, IJCAI and AAAI.

Jun Zhao is a professor at at the Key Laboratory of Cognition and Decision Intelligence for Complex Systems,Institute of Automation, Chinese Academy of Sciences. He received his Ph.D. degree from Tsinghua University in 1998. His research interests focus on natural language processing, relation extraction, event extraction, question answering and dialogue system. He has published more than 100 papers in international journals and conferences including TKDE, ACL, EMNLP, IJCAI, AAAI, WWW and CIKM.