RESEARCH ARTICLE

MAD-Fact: A Multi-Agent Debate Framework for Long-Form Factuality Evaluation in LLMs

Yucheng NING[1,2], Xixun LIN[1]✉, Fang FANG[1,2], Yanan CAO[1,2]

1. Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100085, China
2. School of Cyber Security, University of Chinese Academy of Sciences, Beijing 100049, China

Abstract

The widespread adoption of Large Language Models (LLMs) raises critical concerns about the factual accuracy of their outputs, especially in high-risk domains such as biomedicine, law, and education. Existing evaluation methods for short texts often fail on long-form content due to complex reasoning chains, intertwined perspectives, and cumulative information. To address this, we propose a systematic approach integrating large-scale long-form datasets, multi-agent verification mechanisms, and weighted evaluation metrics. We construct LongHalluQA, a Chinese long-form factuality dataset; and develop MAD-Fact, a debate-based multi-agent verification system. We introduce a fact importance hierarchy to capture the varying significance of claims in long-form texts. Experiments on two benchmarks show that larger LLMs generally maintain higher factual consistency, while domestic models excel on Chinese content. Our work provides a structured framework for evaluating and enhancing factual reliability in long-form LLM outputs, guiding their safe deployment in sensitive domains.

■ 1 Introduction

Large Language Models (LLMs) have demonstrated remarkable capabilities in text understanding and generation [1–6], leading to their widespread adoption across a variety of domains [7–14]. However, LLMs remain prone to generating inaccurate, incorrect, or biased content that deviates from factual reality, a phenomenon commonly referred to as hallucination [15–20]. This issue is particularly critical in domains where a high level of factual accuracy is required, such as biomedicine, law, and finance [21–23], as hallucinated content in these contexts can lead to the spread of misinformation and potentially serious consequences [24]. Therefore, evaluating the factuality of LLMs has become a core research priority.
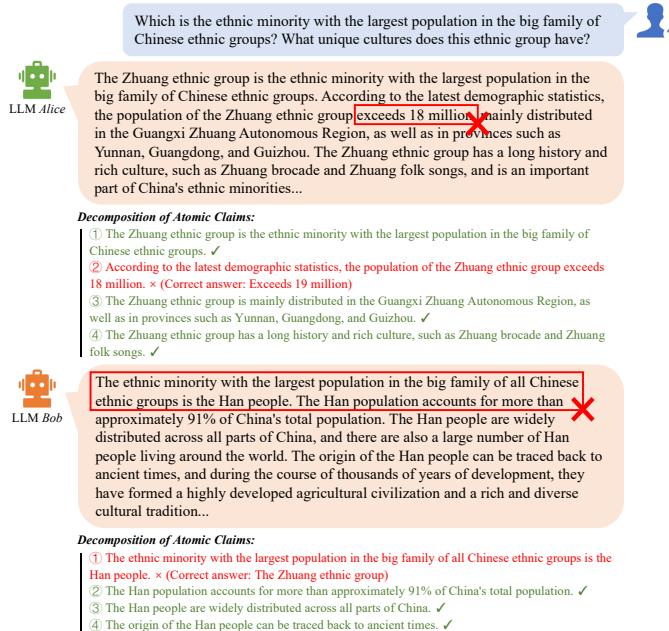
Most existing research on factuality evaluation has focused on short-form text, especially in the context of question-answering tasks (QA tasks) [25–28]. However, many real-world applications require models to generate long-form content [29], sometimes extending to several hundred or even thousands of words. Unlike short-form texts (typically single-viewpoint), long-form generation often incorporates multiple perspectives and complex logical structures [30], greatly increasing the challenge of factuality evaluation. As a result, traditional short-text evaluation methods are often ill-suited to long-form scenarios, highlighting the need for a more fine-grained evaluation mechanism that exceeds binary judgments.

Research on long-form factuality evaluation seeks to develop tech-

niques that can accurately evaluate the factual reliability of extended textual content, thereby enhancing the factual accuracy of outputs generated by LLMs and mitigating hallucinations. Current approaches generally adopt atomic claim decomposition with knowledge-base verification. While methods such as FActScore [31] and SAFE [32] have pushed this direction forward, several critical challenges remain:

- **Scarcity of Chinese Long-Form Benchmarks.** Most current long-form factuality benchmarks are designed for English [31–33], while comprehensive resources for Chinese long-form evaluation are severely lacking. Mainstream English datasets often overlook culturally specific entities, historical events, and linguistic nuances unique to Chinese, thereby hindering the objective evaluation of LLM performance in Chinese generation tasks. Building a multi-topic Chinese long-form factuality benchmark has thus become a foundational and urgent task.

- **Biases in Single-Model Evaluation Frameworks.** Existing evaluation frameworks typically rely on a single model for factual verification, assuming it can consistently identify inaccuracies and weigh evidence [31, 32, 34, 35]. However, even leading LLMs exhibit substantial hallucination problems [?], which can produce incorrect or inconsistent judgments. As a result, single-model evaluation architectures are prone to systematic biases, potentially misrepresenting the factuality of generated content.

- **Neglect of Fact Importance in Metrics.** Existing metrics typ-

ically treat all claims equally [31–34], without considering differences in their relative importance. As illustrated in Figure 1, central claims (e.g., *"The Zhuang are the largest ethnic minority in China"*) and auxiliary claims (e.g., *"The Zhuang are known for brocade"*) receive the same weight, despite substantial disparities in user relevance and task impact. Such uniform treatment undermines the ability of current evaluation methods to capture the true factual quality of long-form outputs, particularly in high-stakes domains such as medicine, law, or education.

Fig. 1   Examples of core factual claims and auxiliary claims. Both responses from LLM *Alice* and LLM *Bob* contain a single incorrect atomic claim. However, from the user's perspective, the answer provided by LLM *Bob* is clearly inferior, as its error occurs within a core factual claim, whereas LLM *Alice*'s mistake affects only an auxiliary claim.

Our solution addresses the limitations of existing resources and evaluation methods through three complementary components. First, to overcome the scarcity of Chinese long-form benchmarks, we extend existing short-text datasets (HalluQA [36] and ChineseSimpleQA [37]) into a large-scale, multi-topic dataset suitable for long-form factuality evaluation. Second, to mitigate systematic biases inherent in single-model verification, we employ a multi-agent debate framework. By leveraging the complementary capabilities of multiple models, this approach enables structured cross-validation among diverse evaluators and enhances factual reasoning [38]. Third, to account for the varying importance of facts, we introduce a hierarchical strategy inspired by the Pyramid Method [39, 40], which captures the relative significance of claims and enables weighted evaluation. Together, these components provide a more reliable framework for evaluating long-form factuality.

We summarize the main contributions of this work as follows:

- We introduce **LongHalluQA**, a Chinese long-form factuality dataset. It comprises 2,746 high-quality samples spanning 7 topics, such as Chinese culture, natural sciences, social sciences, among others, providing a foundational resource for evaluating Chinese long-form generation.

- We develop **MAD-Fact**, a multi-agent debate system for factual verification, which mitigates single-model biases and improves reasoning reliability through structured interactions among the Clerk, Jury, and Judge modules. Experiments show that MAD-Fact consistently outperforms strong baselines such as SAFE [32] and FIRE [35] on multiple long-form factuality benchmarks.

- We propose a **fact importance hierarchy model** to capture the varying significance of claims in long-form texts. Using this model, we design weighted evaluation metrics that correlate strongly with human judgments ($r = 0.701$, $p = 0.036$), effectively reflecting the true factual quality of generated content.

- We benchmark 9 mainstream LLMs from 7 model families on LongFact [32] and LongHalluQA, showing that larger models generally perform better, while Chinese-specific models excel on Chinese tasks. These results provide practical guidance for model selection and optimization in long-form factuality evaluation.

■ 2 Related Work

2.1 Factuality Evaluation of LLMs

Evaluating the factuality of LLMs has been explored through diverse benchmarks, methods, and metrics. This section reviews recent progress along these three dimensions, highlighting current limitations in long-form generation scenarios.

Factuality Evaluation Benchmark. In practice, LLMs are primarily applied to long-form generation tasks, whereas existing factuality benchmarks predominantly focus on short-form, manually constructed QA tasks. Benchmarks such as TruthfulQA [41], HaluEval [42], PopQA [43], FreshQA [44], and LLM-Oasis [45] assess factual consistency in English short-form QA settings. FactScore [31] was an early attempt to evaluate factuality in long-form generation but focused only on simple biographical QA. LongFact [32] expanded coverage to 38 human-curated topics across diverse domains, while Fact-Bench [33] extracted hallucination-inducing prompts from real conversations. Nevertheless, these benchmarks remain limited to English. In the Chinese context, HalluQA [36] targets hallucination evaluation with 450 short prompts across misleading and knowledge-based categories. ChineseSimpleQA [37] systematically evaluates models on short factual questions, including culturally specific content. However, both datasets are restricted to short-form QA, underscoring the urgent need for high-quality Chinese benchmarks for evaluating factuality in long-form generation tasks.

Factuality Evaluation Method. Existing factuality evaluation methods for LLMs primarily follow two technical directions. One approach builds automated verifiers using advanced models such as GPT-4o [46], leveraging their strong reasoning and anomaly detection capabilities. The other approach employs retrieval-augmented generation (RAG) [47], which retrieves up-to-date knowledge to compensate for the static nature of model parameters. Recently, long-form factuality evaluation has gained increasing attention due to the structural complexity and partial correctness of extended outputs. FactScore [31] and

SAFE [32] decompose text into atomic claims for fine-grained RAG-based verification. VeriScore [34] filters unverifiable claims according to predefined criteria, and FIRE [35] introduces confidence-based retrieval to reduce computational costs. However, most existing methods depend on a single verifier, which makes them vulnerable to systematic errors when the model hallucinates or encounters knowledge gaps, posing an open challenge that remains unresolved.

Factuality Evaluation Metrics. Traditional evaluation paradigms rely on binary judgments at the atomic claim level and aggregate metrics such as accuracy and F_1 at the text level. FactScore [31] introduces a long-form evaluation framework by computing the proportion of factually correct atomic claims, but it does not incorporate a recall component. SAFE [32] addresses this limitation by introducing a hyperparameter K to define recall; however, its fixed value restricts adaptability across different domains and content types. FactBench [33] further accounts for claim verifiability and penalizes irrelevant content, offering a more nuanced evaluation of long-form outputs. However, all of these metrics treat facts equally, failing to account for their varying importance and relevance, which reduces sensitivity to critical errors in high-stakes applications.

2.2 Factual Verification

Factual verification has evolved from traditional automated pipelines to LLM-powered systems. Classic systems follow a three-step process: extracting verifiable claims, retrieving evidence from sources (like Wikipedia), and verifying claims using binary or multi-class classification models [31, 34, 48]. With the rise of LLMs and retrieval-augmented generation (RAG), recent systems offer stronger end-to-end performance [49]. Models like RARR [50] prompt LLMs to generate queries, retrieve evidence, and make factual judgments. Other works integrate tool use [32, 51], enabling access to external databases for dynamic knowledge updates. Multimodal extensions such as RA-GAR [52] incorporate both textual and visual evidence through structured retrieval strategies like CoRAG and ToRAG. More recently, agent-based approaches have emerged. ReAct-based agents [53] dynamically coordinate retrieval and reasoning, while multi-agent debate frameworks [54] use iterative evidence exchange to improve decision quality. These methods signal a shift from static rule-based pipelines to more adaptive, collaborative verification systems.

2.3 Multi-agent Systems for Evaluation

With the rapid maturation of single-agent technologies, research on LLM–driven multi-agent systems (MAS) has grown exponentially [55–57]. Multi-agent systems have recently emerged as promising tools for evaluating generative language models. Compared to static evaluation methods, multi-agent systems offer more adaptive and dynamic evaluation mechanisms, while mitigating risks such as data leakage [57]. For example, ChatEval [58] leverages multi-agent debate to assess text generation quality, significantly outperforming single-agent baselines. JudgeBlender [59] introduces an agent-based voting strategy for evaluating information retrieval systems. M-MAD [60] enhances the robustness of machine translation evaluation through multidimensional debates. Building on this line of work, our proposed

system adopts a multi-agent debate framework for factuality verification, achieving notable improvements in both precision and recall over existing approaches.

■ 3 LongHalluQA

Building upon HalluQA [36] and ChineseSimpleQA [37], we systematically develop LongHalluQA, a comprehensive benchmark for Chinese long-form factuality evaluation, by employing a structured three-stage construction pipeline (Figure 2). The pipeline involves three key steps: (1) constructing a structured factual knowledge base to provide reliable information for question expansion, (2) generating and combining factuality evaluation questions to form semantically rich long-form queries, and (3) screening and editing the resulting samples to ensure high quality and factual accuracy.

3.1 Construction of Factual Knowledge Base

In this stage, we begin with the original questions from the HalluQA [36] and ChineseSimpleQA [37] datasets and prompt a large language model to acquire relevant background knowledge through multi-round web searches.[1] The retrieval process combines exact matching with semantic expansion. For example, for a question like "*What are the representative works of the Tang dynasty poet Li Bai?*", the system retrieves not only a list of Li Bai's works but also contextual information about relevant literary schools. The retrieved results are cleaned and deduplicated, then stored in a structured format as a triple-based knowledge base (original question – related knowledge – category, with categories such as "literature", "geography", etc.), ultimately forming a reference knowledge system that comprehensively covers the target question.

This knowledge base restricts the expansion scope to trusted knowledge directly relevant to the original question. Compared with prompting large language models to directly expand questions, this approach effectively mitigates the generation of factually incorrect, unanswerable, or irrelevant questions due to hallucinations, providing a solid factual foundation for subsequent text expansion.

3.2 Generation and Combination of Factuality Questions

Based on the structured triple entries in the knowledge base, we first perform credibility and controversy verification using a verification model[2] driven by a retrieval-augmented generation (RAG) framework. The verification process follows a confidence-threshold mechanism. For instance, for the factual entry "*The Zhuang ethnic group are indigenous people and descendants of the Baiyue*", if the model identifies alternative views such as "*The Zhuang formed through migration from other provinces into Guangxi*", it assigns a low confidence score and filters out the entry, thereby ensuring factual verifiability.

Verified entries then proceed to the question generation stage. For each individual fact, the model generates a corresponding sub-question. For example, given the knowledge "*The Zhuang population is distributed across all 31 provinces, autonomous regions, and municipalities in China, with Guangxi being the primary region*", the model

---

[1] We implement the retrieval functionality using Serper, a Google Search API.

[2] We adopt DeepSeek-V3 for this task due to its strong performance in Chinese semantic understanding and generation.
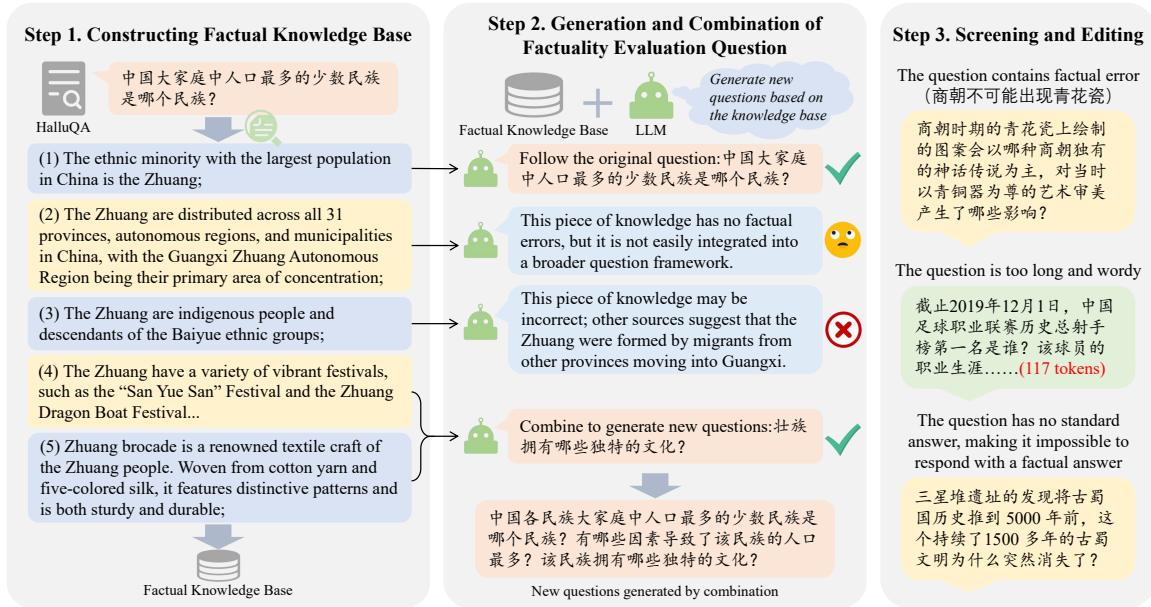
Fig. 2   The construction process of the LongHalluQA dataset.

generates the sub-question "*In which regions of China is the Zhuang ethnic group distributed?*".

When multiple knowledge entries share a common theme, the model performs topic clustering and synthesizes composite questions. For example, given the facts "*The Zhuang celebrate a variety of festivals...*" and "*Zhuang brocade is a renowned traditional textile*", both related to Zhuang culture, the model generates the higher-level sub-question "*What unique cultural features are associated with the Zhuang ethnic group?*" Each original question is thus expanded into a set of semantically relevant sub-questions.

In the sub-question fusion stage, the model selects 1–3 candidate questions from the generated sub-question set using multidimensional evaluation criteria, including factual accuracy, semantic coherence, and logical consistency. Since the chosen verification model is also employed in our subsequent factuality assessment experiments, we explicitly add an exploratory selection instruction in the prompt to mitigate potential selection bias stemming from its training data preferences. This encourages the model to generate sub-questions beyond its knowledge comfort zone, rather than defaulting to familiar or easily answerable content. These selected sub-questions are then dynamically integrated and restructured with the original question.

3.3 Screening and Editing the Resulting Samples

To ensure data quality, we established a review committee consisting of two trained professionals holding a bachelor's degree. A two-stage quality control process was implemented. The first stage involves factual consistency checks, where samples containing factual contradictions or erroneous statements are removed. The second stage focuses on answerability assessment, filtering out samples with semantic ambiguity, redundancy, or unclear questions. The final dataset, LongHalluQA, is composed of samples that pass both stages of manual review, ensuring a high level of accuracy and reliability.

3.4 Dataset Statistics and Comparison

The resulting LongHalluQA benchmark contains 2,746 high-quality Chinese long-form samples, with a retention rate of approximately 86%, demonstrating the robustness and completeness of the proposed construction method. Figure 3 illustrates the distribution of topics covered in the LongHalluQA dataset.
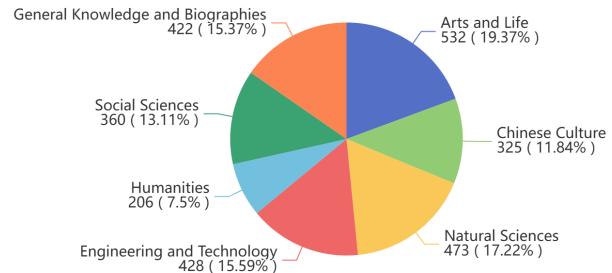

Fig. 3   The topic distribution of the LongHalluQA dataset.

A comparison between sample entries from LongHalluQA and the original datasets is shown in Figure 4. In an experiment with 100 randomly sampled questions, the average response length increased by 9.4 times compared to the original datasets, effectively alleviating the scarcity of Chinese long-form factuality evaluation resources.

■ 4 MAD-Fact

This section introduces MAD-Fact, a **M**ulti-**A**gent **D**ebate system for **Fact**ual verification, as illustrated in Figure 5. We first provide a formal definition of the task that MAD-Fact aims to address, followed by a detailed introduction of the architecture of each module within the system.
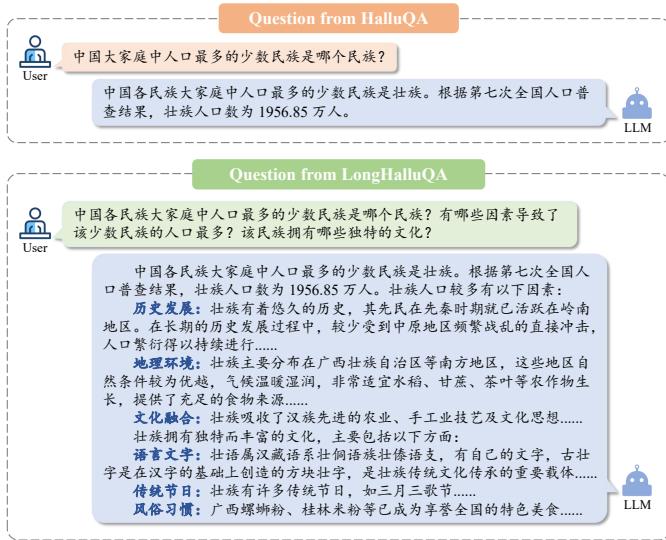
4.1 Task Definition

Given a question q_i from a factuality evaluation dataset, the target LLM generates a long-form response a_i. The goal of MAD-Fact is to output a factuality evaluation score s_i for this response. The overall process can be represented as:

$$s_i = \text{MAD-Fact}(conf, q_i, a_i), \tag{1}$$

where $conf$ denotes the system configuration, including role definitions, system prompts, and other operational settings.

4.2 System Architecture

The MAD-Fact system consists of three types of agents:

- **The Clerk Agent**, responsible for decomposing the long-form response a_i generated by the evaluated model into multiple atomic claims $\{c_{i,j}\}_{j=1}^{T}$;
- **The Jury**, composed of **the Evaluator Agents** assuming various professional roles, who assess the factuality of each atomic claim $c_{i,j}$ (assigning TRUE or FALSE) through external retrieval and multi-agent debate;
- **The Judge Agent**, which aggregates the individual evaluations from the jury for each atomic claim and produces a final factuality predict $p_{i,j}$ (either TRUE or FALSE). Based on the set of judgments $\{p_{i,j}\}_{j=1}^{T}$, the system calculates the overall factuality score s_i for the response a_i.

4.2.1 Clerk: Atomic Claim Decomposition

Given a question q_i and its corresponding long-form response a_i generated by the evaluated model, the Clerk Agent decomposes a_i into T atomic claims $\{c_{i,j}\}_{j=1}^{T}$. This process is formalized as:

$$\{c_{i,j}\}_{j=1}^{T} = \text{Clerk}(conf_{Clerk}, q_i, a_i), \tag{2}$$

where $conf_{Clerk}$ specifies the role description, prompt settings, and other configuration details for the Clerk agent.

Importantly, the Clerk agent is designed to extract only fact-checkable atomic claims, systematically filtering out unverifiable content (e.g., instructions, suggestions, or subjective statements).

4.2.2 Jury: Fact Verification via Retrieval and Multi-Agent Debate

For each atomic claim $c_{i,j}$, the jury conducts a multi-round debate involving multiple Evaluator Agents playing different roles. This process generates a set of factuality judgments $\{p_{i,j}^{n}\}_{n=1}^{N}$ and corresponding explanations $\{e_{i,j}^{n}\}_{n=1}^{N}$ from the N Evaluators. The procedure can be formalized as:

$$[\{p_{i,j}^{n}\}_{n=1}^{N}, \{e_{i,j}^{n}\}_{n=1}^{N}] = \text{Debate}(\{\text{Evaluator}^{n}\}_{n=1}^{N}), \tag{3}$$

where N denotes the number of Evaluator Agents in the jury.

In the following sections, we describe the key components of the debate process that govern how agents interact and produce judgments. Specifically, we detail the **Response Strategies** employed by the agents, the **Debate Rules** that structure their interactions, and the **Agent Role-Playing** mechanism that ensures diverse perspectives in the evaluation.

- Response Strategies

The jury organizes the debate in a sequential manner, where Evaluator Agents speak in a predefined order. Before responding, each agent can review the shared knowledge base $K_{i,j}$, examine previous statements in the message pool $M_{i,j}$, and decide whether to invoke external retrieval tools to supplement its knowledge. During each debate round t, the selected agent determines its response strategy from three options: a direct response, a retrieval-based response, or a conditional retrieval response, as illustrated in Figure 5:

(1) Direct Response. Based on the current message history $M_{i,j,t-1}$ and reference knowledge $K_{i,j,t-1}$, the Evaluator Agent provides its judgment and explanation for the atomic claim $c_{i,j}$:

$$[p_{i,j,t}, e_{i,j,t}] = \text{Evaluator}^{m}(conf^{m}, c_{i,j}, M_{i,j,t-1}, K_{i,j,t-1}), \tag{4}$$

where $m = t \bmod N$ denotes the agent's index, and $conf^{m}$ specifies its configuration including role definition, prompts, and operational parameters. After the response, the shared message pool is updated as:

$$M_{i,j,t} = \text{Context}(M_{i,j,t-1}, [p_{i,j,t}, e_{i,j,t}]), \tag{5}$$

while the shared knowledge base remains unchanged:

$$K_{i,j,t} = K_{i,j,t-1}. \tag{6}$$

(2) Retrieval-Based Response. If additional evidence is needed, the Evaluator Agent first formulates a new search query $query_{i,j,t}$ based on the current message and knowledge context, ensuring it does not duplicate any existing query $\{query_{i,j,k}\}_{k=1}^{t-1}$:

$$query_{i,j,t} = \text{Evaluator}^{m}(conf^{m}, c_{i,j}, M_{i,j,t-1}, K_{i,j,t-1}), \tag{7}$$

the agent then invokes an external search tool to retrieve knowledge:

$$k_{i,j,t} = \text{Search}(query_{i,j,t}), \tag{8}$$

and the retrieved information is appended to the shared knowledge base:

$$K_{i,j,t} = \text{Context}(K_{i,j,t-1}, query_{i,j,t}, k_{i,j,t}). \tag{9}$$
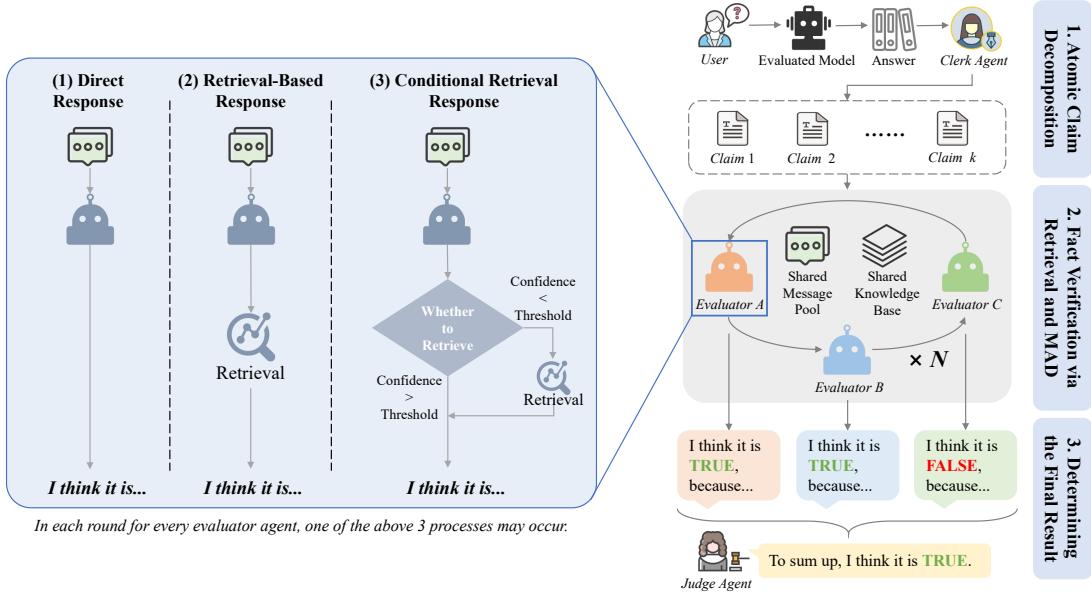
Fig. 5   The overall framework of the MAD-Fact system.

Then, the agent responds based on the updated knowledge:

$$[p_{i,j,t}, e_{i,j,t}] = \text{Evaluator}^m(conf^m, c_{i,j}, M_{i,j,t-1}, K_{i,j,t}). \quad (10)$$

Finally, the message pool is updated:

$$M_{i,j,t} = \text{Context}(M_{i,j,t-1}, [p_{i,j,t}, e_{i,j,t}]). \quad (11)$$

(3) Conditional Retrieval Response. The Evaluator Agent first estimates its confidence $c_{i,j,t}$. If $c_{i,j,t} \geq \theta$, it proceeds with a **direct response** as above. Otherwise, if $c_{i,j,t} < \theta$, it switches to the **retrieval-based response** workflow, invoking external search to enhance factual grounding before issuing a judgment.
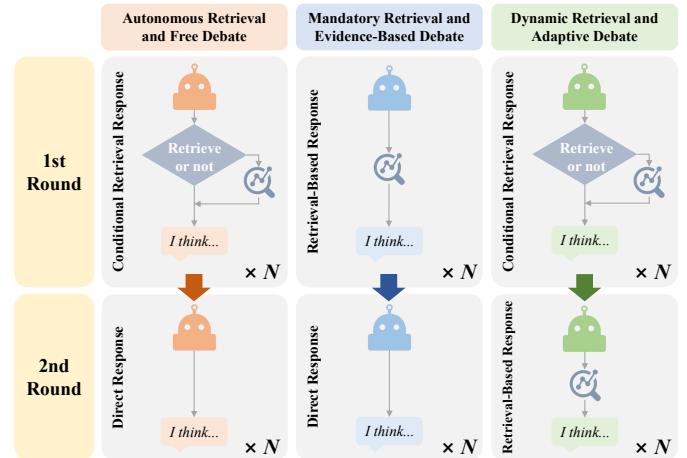
• Debate Rules

Based on the three response strategies discussed earlier and inspired by existing works [58, 61], we further design three types of debate rules to regulate how Evaluator agents interact during the verification process, as illustrated in Figure 6. These rules enable a systematic exploration of the trade-offs between retrieval cost, prior knowledge utilization, and factual reliability, thus facilitating a comparison of debate strategies within MAD-Fact.

Rule 1: Autonomous Retrieval and Free Debate. In the first round, each agent may autonomously decide, based on its confidence, whether to invoke external retrieval before speaking, so as to balance knowledge utilization and retrieval cost. In the second round, agents directly deliver their statements while also engaging in discussion with others by drawing on peers' suggestions or correcting mistakes.

Rule 2: Mandatory Retrieval and Evidence-Based Debate. In the first round, retrieval is mandatory before speaking, ensuring that agents compensate for possible knowledge gaps and mitigate overconfidence, a phenomenon where they assign overly high confidence to themselves. In the second round, agents debate with others based on multiple retrieved references, making the discussion more evidence-grounded and reliable.

Rule 3: Dynamic Retrieval and Adaptive Debate. In the first round, each agent may autonomously retrieve information before speaking. The debate process in later rounds is then adjusted according to the jury's consensus: if consensus is reached in the first round, the results are directly output to avoid redundant costs; if not, retrieval becomes mandatory before the second round, so that authoritative references can help resolve conflicts and guide the jury toward agreement.

Fig. 6   Three types of debate rules of the MAD-Fact system. For the convenience of systematic analysis, we set the number of debate rounds among multi-agent to 2, and the number of agents N to 3.

• Agent Role-Playing

In addition, to improve the diversity of the evaluation system, the jury adopts a heterogeneous role assignment mechanism, where each agent is assigned a distinct professional role. This differentiated setup encourages specialized focuses and enables cross-checking from multiple perspectives, thereby reducing factual evaluation biases that may arise

Table 1   Different role descriptions adapted to MAD-Fact.

Name	Description
General Public	*You are now General Public*, one of the referees in this task. As a member of the general public, you are interested in the claim and eager to get updates on the investigation. You can precisely capture the main meaning of the text rather than fixating on every single word.
Critic	*You are Critic*, one of the referees in this task. You are adept at questioning the judgment of others by searching through chains of evidence. In addition, you pay attention to the rigor of the data and are keen to pick up small differences in claims.
News Author	*You are News Author*, one of the referees in this task. You focus on the factual basis of the claim and the latest developments, verifying the accuracy of the claim through extensive access to information. When information is insufficient, you tend to search rather than jump to conclusions.
Scientist	*You are Scientist*, one of the referees in this task. As a data science research professional, you have a deep background in critical thinking and problem solving skills and are sensitive to data. You are adept at verifying the accuracy of claims by looking at references.
Psychologist	*You are Psychologist*, one of the referees in this task. Your job is to study human behavior and mental processes to understand and explain human behavior. Assist others in determining which response is the better one among the available options.
Data Analyst	*You are now Data Analyst*, one of the referees in this task. Specializing in dissecting complex datasets, you approach the claim with a quantitative lens. You have a knack for gathering relevant data from diverse sources, cleaning and organizing it to extract meaningful insights.

from single-perspective blind spots. Following the setup in ChatEval [58], we define six roles for the factuality evaluation task: *Public*, *Critic*, *News Author*, *Scientist*, *Psychologist*, and *Data Analyst*, as shown in Table 1.

In addition, at the role description level, we introduce a retrieval incentive mechanism to encourage agents to use external tools, thereby reducing overconfidence and enhancing the reliability of factual verification.

4.2.3   Judge: Determining the Final Result

The output opinions and statements from the last round of the jury debate, $\{p_{i,j}^n, e_{i,j}^n\}_{n=1}^N$, are submitted to the Judge agent. The Judge agent aggregates these outputs and determines the final result based on the majority voting principle. This process can be represented as:

$$p_{i,j} = \text{Judge}(\{p_{i,j}^n, e_{i,j}^n\}_{n=1}^N). \tag{12}$$

If the number of agents N is even and results in a tie, the output opinion $p_{i,j}^N$ of the last-speaking Evaluator agent is chosen as the final result, since this agent has access to the complete debate process and its output best represents the overall deliberation.

For a long-form response a_i generated by the evaluated model, along with its corresponding T atomic claims $\{c_{i,j}\}_{j=1}^T$, the Judge agent integrates the final evaluation results of each atomic claim $\{p_{i,j}\}_{j=1}^T$, to calculate the score s_i of the response. This process can be expressed as:

$$s_i = \text{Calculate}(\{p_{i,j}\}_{j=1}^T). \tag{13}$$

The overall score s of the evaluated model on the entire dataset is defined as the arithmetic mean of the scores s_i of all long-form responses, which can be formulated as $s = \frac{1}{|D|} \sum_{i=1}^{|D|} s_i$, where $|D|$ denotes the total number of samples in the dataset. The specific calculation method will be elaborated in the next chapter.

◼ 5 Evaluation Metrics Based on Factual Importance

In this section, we introduce evaluation metrics that account for the varying importance of facts in long-form text. We first describe the construction of a fact importance hierarchy model, which quantifies the relative significance of individual claims, and then present weighted evaluation metrics built upon this model to provide a more nuanced factuality assessment.

5.1 Fact Importance Hierarchy Model

As illustrated in Figure 7, for a given question q_i, we introduce G powerful closed-source models $\{Model_j\}_{j=1}^G$ as expert models, each generating a reference answer $\{r_{i,j}\}_{j=1}^G$. This process can be expressed as:

$$r_{i,j} = \text{Model}_j(q_i). \tag{14}$$

Each reference answer is then decomposed into a set of atomic claims $\{[c_{i,j,k}]_{k=1}^{K_j}\}_{j=1}^G$, where the decomposition process can be expressed as:

$$[c_{i,j,k}]_{k=1}^{K_j} = \text{Clerk}(r_{i,j}), \tag{15}$$

with K_j denoting the number of atomic claims in the j-th set.

Next, semantically equivalent atomic claims across different sets are merged to form a single golden set of atomic claims $\{g_{i,k}\}_{k=1}^{K_g}$, which is regarded as exhaustive. This merging process can be expressed as:

$$\{g_{i,k}\}_{k=1}^{K_{gold}} = \text{set}\left(\{[c_{i,j,k}]_{k=1}^{K_j}\}_{j=1}^G\right), \tag{16}$$

where K_{gold} denotes the number of atomic claims in the golden set.

A G-level pyramid model $P_{i,G}$ is then constructed, where the levels are ordered from the first layer at the top to the G-th layer at the bottom. For each atomic claim in the golden set, we count its frequency f of occurrence across the G reference answers, and place it into the $(G - f + 1)$-th layer of the pyramid model. The higher the layer an atomic claim belongs to, the more frequently it is mentioned by expert models, and the greater its assigned factual importance weight. The weight of the m-th layer in the pyramid model is defined as ω_m, satisfying $\omega_i > \omega_j$ (if $1 \le i < j \le G$).

5.2 Weighted Evaluation Metrics

Given a question q_i, we construct the pyramid model $P_{i,G}$ following the aforementioned procedure and obtain the golden set of atomic claims $\{g_{i,k}\}_{k=1}^{K_{gold}}$. For a long-form response a_i to be evaluated, we calculate the weight $\omega_{i,j}$ of each decomposed atomic claim $c_{i,j}$ according to the pyramid model and define the following weighted evaluation metrics.

5.2.1   Weighted Precision

Let $c_{i,j}^T$ denote factual atomic claims, with a total count of $|S|$, and $c_{i,j}^F$ denote non-factual atomic claims, with a total count of $|N|$. The
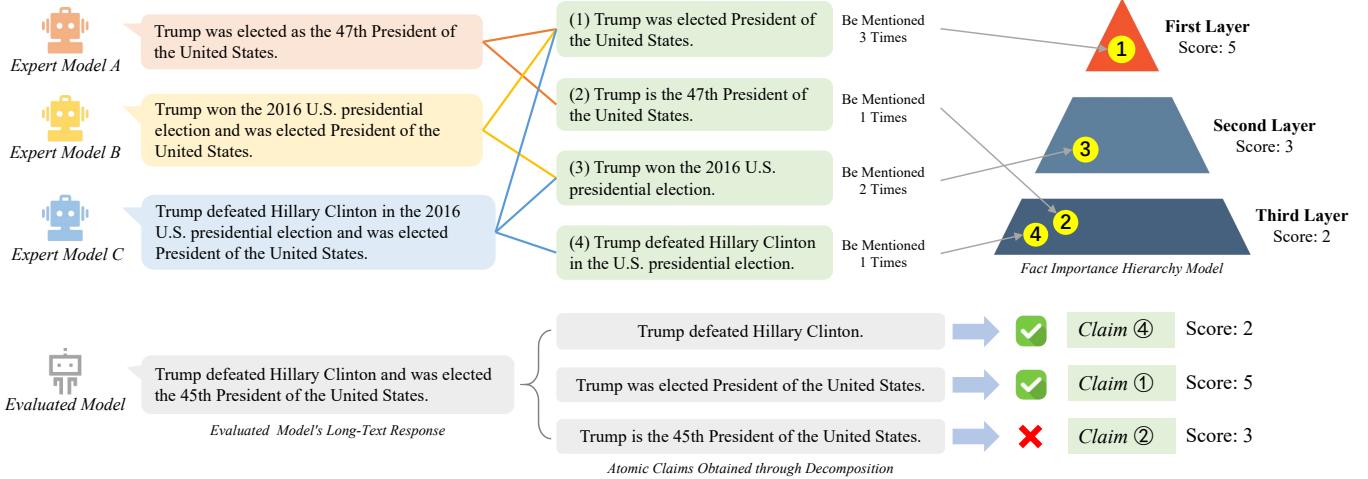
Fig. 7   Overview of the fact importance hierarchy model.

weighted factual precision score is defined as:

$$
\begin{aligned}
Prec_w(a_i) &= \frac{\sum_{j=1}^{|S|} \omega_{i,j}\, c_{i,j}^T}{\sum_{j=1}^{|S|} \omega_{i,j}\, c_{i,j}^T + \sum_{j=1}^{|N|} \omega_{i,j}\, c_{i,j}^F} \\
&= \frac{\sum_{j=1}^{|S|} \omega_{i,j}\, c_{i,j}^T}{\sum_{j=1}^{|S|+|N|} \omega_{i,j}\, c_{i,j}}.
\end{aligned}
\tag{17}
$$

5.2.2 Weighted Recall

For a given question q_i, the golden set $\{g_{i,k}\}_{k=1}^{K_{gold}}$ is regarded as an exhaustive reference answer. However, responses generated by LLMs often contain redundant information, and the golden set may amplify the error caused by such redundancy. To mitigate this issue, we introduce a hyperparameter γ ($\gamma \leq 1$) to account for the gap between the golden set and a perfect answer, thereby alleviating evaluation bias. The weighted factual recall score is defined as:

$$
R_w@\gamma(a_i) = \min\left(\frac{1}{\gamma} \cdot \frac{\sum_{j=1}^{|S|} \omega_{i,j}\, c_{i,j}^T}{\sum_{k=1}^{K_{gold}} \omega_{i,k}\, g_{i,k}},\ 1 \right).
\tag{18}
$$

5.2.3 Weighted F1-Score

Based on the weighted precision and weighted recall, the weighted F1-score is defined as:

$$
F_1@\gamma(a_i) = \begin{cases} \frac{2 \cdot Prec_w(a_i) \cdot R_w@\gamma(a_i)}{Prec_w(a_i) + R_w@\gamma(a_i)}, & \text{if } |S| > 0, \\ 0, & \text{if } |S| = 0. \end{cases}
\tag{19}
$$

■ 6 Experiments

Our experiments comprise two parts. In Section 6.1, we evaluate the MAD-Fact system on five fact-checking datasets and assess the effects of different debate rules, initialization strategies, and individual system components on its performance. In Section 6.2, we evaluate nine mainstream LLMs using MAD-Fact (Section 4) on LongFact [32] and our newly constructed LongHalluQA (Section 3), employing weighted evaluation metrics (Section 5).

6.1 Evaluation of MAD-Fact

6.1.1   Datasets

We evaluate the performance of the MAD-Fact system on four fact-checking datasets: FacTool [62], FELM [63], Factcheck-Bench [51], and BingCheck [64]. To ensure fair comparison with prior studies, we follow the settings of FIRE [35] and select from the multi-domain data of FacTool and FELM the subsets that require factual knowledge for verification, which we denote as FacToolQA and FELM-WK. To address the class imbalance issue in BingCheck (3,581 atomic claims labeled as TRUE versus 42 labeled as FALSE), we adopt a stratified sampling strategy for balancing. Specifically, we construct the test set by sampling 100 representative examples from the TRUE class while retaining all 42 FALSE examples. To further validate the fact-checking capability of MAD-Fact in the Chinese context, we construct a dataset called FactEval-CN by extracting one representative model statement from each sample in the ChineseFactEval [65], resulting in a total of 125 Chinese factual claims.

All five datasets contain multiple long-form factual QA pairs, along with the corresponding atomic claims decomposed from the long responses and their binary factuality labels. Detailed statistics are shown in Table 2.

Table 2   Statistics of the Fact-Checking Dataset.

Dataset	Language	#TRUE claims	#FALSE claims	All claims
FacToolQA	English	177	56	233
FELM-WK	English	99	85	184
Factcheck-Bench	English	472	159	631
BingCheck	English	100	42	142
FactEval-CN	Chinese	70	55	125

6.1.2   Baselines

To ensure a rigorous performance comparison, we select two representative methods from the fact-checking domain as baselines:

(1) SAFE [32]: Utilizes LLM-based agents for long-form factuality evaluation, outperforming crowdsourced human annotators by reach-

ing 72% agreement and prevailing in 76% of 100 randomly sampled disagreement cases.

(2) FIRE [35]: Iteratively integrates retrieval and verification, triggering external retrieval only when model confidence falls below a threshold. This design exploits the verifier's internal knowledge while substantially reducing computational costs without sacrificing accuracy.

6.1.3 Evaluation Metrics

We evaluate the system in terms of precision (Prec), recall, and F1 score, separately for the positive class (atomic claims labeled as TRUE) and the negative class (atomic claims labeled as FALSE). Precision refers to the proportion of atomic claims predicted as factual that are indeed labeled TRUE, while recall measures the proportion of TRUE-labeled claims that are correctly identified as factual. Let TP, FP, FN, and TN denote true positives, false positives, false negatives, and true negatives, respectively. The evaluation metrics for the positive class are defined as:

$$\text{Prec} = \frac{TP}{TP+FP}, \text{ Recall} = \frac{TP}{TP+FN}, \ F_1 = \frac{2 \cdot \text{Prec} \cdot \text{Recall}}{\text{Prec} + \text{Recall}}. \quad (20)$$

The evaluation metrics for the negative class are defined analogously.

6.1.4 Comparative Experiments

● *Q1. How Does MAD-Fact Perform Compared to Baselines?*

Table 3 presents the fact-checking evaluation results of the MAD-Fact system initialized with `GPT-4o-mini` compared with baseline methods.

Table 3 Results of the MAD-Fact system based on `GPT-4o-mini` and baselines. We set the number of Evaluator agents in the MAD-Fact system to three.

Dataset	Method	LABEL=True			LABEL=False		
		Prec	Recall	F1	Prec	Recall	F1
FactcheckBench	SAFE	0.89	0.83	0.86	**0.70**	0.79	**0.74**
	Fire	0.91	**0.84**	**0.87**	0.61	0.74	0.67
	MAD-Fact	**0.94**	0.77	0.84	0.55	**0.84**	0.67
FacToolQA	SAFE	**0.92**	0.82	0.87	0.58	**0.79**	**0.67**
	Fire	0.87	**0.88**	0.87	**0.60**	0.59	0.59
	MAD-Fact	0.91	0.85	**0.88**	**0.60**	0.75	**0.67**
BingCheck	SAFE	0.86	0.81	0.84	0.60	0.69	0.64
	Fire	0.87	**0.91**	**0.88**	**0.74**	0.67	0.70
	MAD-Fact	**0.90**	0.86	**0.88**	0.70	0.76	**0.73**
FELM-WK	SAFE	0.61	0.76	0.68	0.61	0.44	0.51
	Fire	0.63	**0.82**	0.71	0.67	0.44	0.53
	MAD-Fact	**0.69**	0.79	**0.74**	**0.70**	**0.59**	**0.64**
FactEval-CN	SAFE	0.85	**0.83**	0.84	0.60	0.72	0.65
	Fire	0.86	0.82	0.74	0.63	0.73	0.68
	MAD-Fact	**0.89**	0.82	**0.85**	0.64	**0.76**	**0.70**

In Table 3, the MAD-Fact system adopts the simplest Rule 1 (Autonomous Retrieval and Free Debate). The bold numbers denote the best results for each dataset and label category. The F1-score, which balances both precision and recall, serves as a comprehensive metric for model performance. As shown in Table 3, the MAD-Fact system

achieves the best F1-scores in 8 out of 10 comparisons across datasets and label categories, yielding a win rate of 80%. This demonstrates the superior fact-checking capability of the MAD-Fact system.

● *Q2. How Do Debate Rules Influence MAD-Fact's Performance?*

Further investigation is conducted to explore the impact of different multi-agent debate rules on the performance of the MAD-Fact system. Table 4 presents the fact-checking evaluation results of the MAD-Fact system initialized with `GPT-4o`, compared against the baseline methods.

Table 4 Results of the MAD-Fact system based on `GPT-4o` and baselines. The three debate rules introduced in Section 4.2.2.2 are denoted as follows: Rule 1 (Autonomous Retrieval and Free Debate) is referred to as MAD-Fact-*free*; Rule 2 (Mandatory Retrieval and Evidence-Based Debate) is referred to as MAD-Fact-*search*; and Rule 3 (Dynamic Retrieval and Adaptive Debate) is referred to as MAD-Fact-*adapt*.

Dataset	Method	LABEL=True			LABEL=False		
		Prec	Recall	F1	Prec	Recall	F1
FactcheckBench	SAFE	**0.94**	0.74	0.83	0.62	**0.90**	0.74
	Fire	0.92	**0.79**	**0.85**	0.56	0.79	0.66
	MAD-Fact-*free*	0.92	0.76	0.83	0.53	0.80	0.64
	MAD-Fact-*search*	0.92	**0.79**	**0.85**	**0.65**	0.87	**0.74**
	MAD-Fact-*adapt*	0.91	0.72	0.81	0.60	0.85	0.70
FacToolQA	SAFE	**0.92**	0.88	0.90	0.66	**0.77**	**0.71**
	Fire	**0.92**	0.88	0.90	0.65	0.71	0.68
	MAD-Fact-*free*	0.90	0.88	0.89	0.64	0.68	0.66
	MAD-Fact-*search*	0.91	0.88	0.90	**0.66**	0.73	0.69
	MAD-Fact-*adapt*	0.91	0.85	0.88	0.61	0.73	0.67
BingCheck	SAFE	0.86	0.81	0.84	0.71	0.60	0.65
	Fire	0.86	0.88	0.87	0.70	0.67	0.68
	MAD-Fact-*free*	**0.87**	0.90	**0.89**	0.74	**0.69**	**0.72**
	MAD-Fact-*search*	0.85	0.88	0.87	0.69	0.64	0.67
	MAD-Fact-*adapt*	**0.87**	0.92	**0.89**	0.78	0.67	**0.72**
FELM-WK	SAFE	0.70	0.80	0.75	0.72	**0.60**	0.65
	Fire	0.70	0.86	0.77	0.77	0.54	0.63
	MAD-Fact-*free*	**0.72**	0.89	0.79	0.82	0.59	**0.68**
	MAD-Fact-*search*	0.70	0.86	0.77	0.77	0.56	0.65
	MAD-Fact-*adapt*	0.70	0.84	0.76	0.76	0.59	0.66
FactEval-CN	SAFE	0.85	0.85	0.85	0.64	0.72	0.68
	Fire	0.86	0.83	0.84	0.62	0.70	0.66
	MAD-Fact-*free*	0.90	0.83	**0.86**	0.66	0.76	0.71
	MAD-Fact-*search*	0.87	0.83	0.85	0.64	0.71	0.67
	MAD-Fact-*adapt*	0.81	**0.87**	0.84	0.64	0.64	0.64

In Table 4, the bold numbers indicate the best results for each dataset and label category. The experimental results show that the MAD-Fact system employing Rule 1 and Rule 2 demonstrates superior fact-checking performance, validating the importance of synergizing external retrieval tools with the model's internal knowledge base. In contrast, the system based on Rule 3 performs less effectively. Error analysis reveals that its weakness mainly stems from erroneous consensus reached in the first round of debate, which, due to the lack of subsequent correction opportunities, becomes entrenched. Specifically, when the system arrives at an incorrect conclusion in the early stage, the dynamic termination mechanism prematurely halts the de-

bate process, depriving the multi-agent system of the opportunity for deeper reasoning through multi-round interactions.

• *Q3. Does Initializing with Different Model Families Affect MAD-Fact?*

Further exploration was conducted to investigate the impact of initializing the MAD-Fact system with models from different families. Specifically, we initialized the three Evaluator agents in the MAD-Fact system with `GPT-4o-mini`, `DeepSeek-V3`, and `Claude-3-Haiku`, denoted as MAD-Fact-*various*, and compared it with the MAD-Fact system initialized solely with `GPT-4o-mini`. Using the simplest Rule 1 (Autonomous Retrieval and Free Debate), Table 5 presents the fact-checking evaluation results of the MAD-Fact system based on multi-model initialization.

Table 5   Results of the MAD-Fact system based on multi-model initialization.

Dataset	Method	LABEL=True			LABEL=False		
		Prec	Recall	F1	Prec	Recall	F1
FactcheckBench	MAD-Fact	0.94	0.77	0.84	0.55	0.84	0.67
	MAD-Fact-*various*	0.92	0.78	0.84	0.55	0.79	0.64
FacToolQA	MAD-Fact	0.91	0.85	0.88	0.60	0.73	0.66
	MAD-Fact-*various*	0.89	0.87	0.88	0.62	0.66	0.64
BingCheck	MAD-Fact	0.90	0.86	0.88	0.70	0.76	0.73
	MAD-Fact-*various*	0.89	0.87	0.88	0.70	0.74	0.72
FELM-WK	MAD-Fact	0.69	0.79	0.74	0.70	0.59	0.64
	MAD-Fact-*various*	0.67	0.83	0.74	0.73	0.53	0.61

The experimental results indicate that the MAD-Fact system initialized with multiple models did not achieve significant performance improvement; instead, it showed a slight decline compared to the system initialized with a single model. A deeper analysis of the misclassified cases reveals that, relative to the single-model system, the agents in the multi-model initialization setting exhibited a markedly higher frequency of misleading each other, and reaching consensus within the predefined number of debate rounds became more difficult. Based on these findings, subsequent evaluations adopt the MAD-Fact system initialized with a single model.

6.1.5   Ablation Study

To validate the effectiveness of each module in the multi-agent debate system, we designed a series of ablation experiments. The multi-agent debate system was examined in the following three variants:

(1)w/o Role-Playing: Removing the diversified role configurations (i.e., modifying the agent profile $conf^m$ in Equation (7)), such that all Evaluator agents have exactly the same role settings;

(2)w/o Debate: Removing the multi-round debate process, where Evaluator agents deliver only one round of statements before the Evaluator agent aggregates the results;

(3)w/o Search: Removing the retrieval module (i.e., Equation (8)), prohibiting Evaluator agents from invoking external retrieval tools to access reference materials before their statements.

Table 6 presents the results of the ablation study based on the MAD-Fact system initialized with the gpt-4o-mini model. The experimental

results demonstrate that each module in the multi-agent debate system contributes to the overall performance.

Table 6   Results of the ablation experiment of the MAD-Fact system. We set the number of Evaluator agents to 3 and applied the simplest Rule 1.

Dataset	Method	LABEL=True			LABEL=False		
		Prec	Recall	F1	Prec	Recall	F1
BingCheck	MAD-Fact	0.90	0.86	0.88	0.70	0.76	0.73
	w/o Role-Playing	0.89	0.87	0.88	0.70	0.74	0.72
	w/o Debate	0.87	0.78	0.82	0.58	0.71	0.64
	w/o Search	0.85	0.83	0.84	0.63	0.52	0.57
FacToolQA	MAD-Fact	0.91	0.85	0.88	0.60	0.75	0.67
	w/o Role-Playing	0.91	0.83	0.87	0.58	0.75	0.66
	w/o Debate	0.89	0.85	0.87	0.57	0.66	0.61
	w/o Search	0.84	0.84	0.84	0.48	0.48	0.48

6.1.6   Case Study

During the experiments, we observed several exciting cases. In debate scenarios, multi-agent systems exhibited many human-like traits: they were able to adhere to their own viewpoints, actively correct their peers' mistakes, and continuously engage in self-reflection. Such human-like behaviors may be one of the key factors contributing to the superior performance of multi-agent systems compared to single-agent systems. Below, we present several interesting results we observed:

(1) Agents do not easily change their stance when facing differing opinions, but instead adhere to their own viewpoints. As shown in Figure 8: in the first round of debate, *Public* and *News Author* provided the correct answers, while *Critic* gave an incorrect one. In the second round, *Public* and *News Author* did not readily accept the view of the *Critic* but instead held to their own judgments. Eventually, during the voting process, the multi-agent system reached the correct conclusion.

(2) Agents actively correct their peers' mistakes when they detect problems in their statements. As shown in Figure 9: in the first round of debate, *Public* gave an incorrect answer, while *News Author* and *Critic* provided the correct answers and attempted to correct *Public*, At the beginning of the second round, with the help of its peers, *Public* realized its mistake, revised its position, and gave the correct answer, thereby reaching consensus with *News Author* and *Critic*.

(3) Even when consensus is achieved among peers, agents continue to engage in self-reflection. As illustrated in Figure 10: in the first round of debate, *Public*, *Scientist*, and *News Author* all gave incorrect answers, reaching a wrong consensus. However, *Public* continuously reflected on its viewpoint and, in the second round, changed its stance, providing the correct answer. This also introduced a new explanation to *Scientist* and *News Author*, prompting them to engage in reflection as well. Eventually, the group reached the correct consensus.

6.2 Factuality Evaluation Results of LLMs
6.2.1   Datasets

This study employs two long-form factuality evaluation datasets: Long-Fact [32] and LongHalluQA. LongFact is the first systematically constructed cross-domain benchmark for long-form factuality evaluation,

Atomic claim: The Arctic Ocean covers an area of about 14.05 million square kilometers.
Label: TRUE

General Public — **I think it's True.** The statement claims the Arctic Ocean covers an area of about 14.05 million square kilometers. This is consistent with the knowledge that states the Arctic Ocean covers approximately 14,060,000 km to 14,090,000 km. The use of 'about' allows for slight variations, making the statement factually accurate.

News Author — **I think it's True.** The statement claims the Arctic Ocean covers 'about 14.05 million square kilometers.' This approximation is consistent with the range provided in the knowledge points, which state the Arctic Ocean covers 14,060,000 km to 14,090,000 km. The use of 'about' allows for this slight variation, making the statement factually accurate.

Critic — **I think it's False.** The statement claims the Arctic Ocean covers about 14.05 million square kilometers, which is below the range of 14.06 million km to 14.09 million km provided in the knowledge. The term 'about' allows for approximation, but the number is not within the acceptable range of approximation.

General Public — **I think it's True.** The statement claims that the Arctic Ocean covers an area of about 14.05 million square kilometers. According to the knowledge, the Arctic Ocean spans an area between 14,060,000 km and 14,090,000 km. The use of 'about' in the statement suggests an approximation, which allows for slight variations. Therefore, the statement is consistent with the knowledge provided, as 14.05 million km is close to the lower bound of the range. This makes the statement factually accurate.

News Author — **I think it's True.** The statement claims that the Arctic Ocean covers an area of about 14.05 million square kilometers. According to the knowledge points, the Arctic Ocean spans an area between 14,060,000 km and 14,090,000 km. The use of 'about' in the statement allows for an approximation, which can accommodate slight variations. Therefore, 14.05 million km is reasonably close to the lower bound of the range provided, making the statement factually accurate.

Critic — **I think it's False.** The statement claims the Arctic Ocean covers about 14.05 million square kilometers, which is below the range of 14.06 million km to 14.09 million km provided in the knowledge. The term 'about' allows for approximation, but the number is not within the acceptable range of approximation.

Result: TRUE

Fig. 8 Case 1 of multi-agent debate: agents adhere to their own views.

Atomic claim: The Pilgrims first landed in Plymouth, Massachusetts in 1620.
Label: FALSE

General Public — **I think it's True.** The statement that the Pilgrims first landed in Plymouth, Massachusetts in 1620 is historically accurate. The Pilgrims, a group of English settlers seeking religious freedom, are well-documented to have arrived on the Mayflower and established the Plymouth Colony in what is now Plymouth, Massachusetts, in December 1620. This event is a significant part of American history and is commonly known as the landing of the Pilgrims.

News Author — **I think it's False.** General Public is wrong. The statement is factually inaccurate because the Pilgrims first landed at Provincetown Harbor, not Plymouth, Massachusetts. While they did establish Plymouth Colony in 1620, it was not their initial landing site.

Critic — **I think it's False.** There are errors in General Public's account. The statement is inaccurate because the Pilgrims first anchored in Provincetown Harbor, not Plymouth, Massachusetts. The establishment of Plymouth Colony occurred after their initial landing.

General Public — **I think it's False.** The statement claims that the Pilgrims first landed in Plymouth, Massachusetts, in 1620. However, based on the knowledge provided by others, this is not accurate. The Pilgrims initially anchored in Provincetown Harbor before they went on to establish Plymouth Colony. While they did found Plymouth Colony in 1620, it was not their first landing site. This distinction is crucial, and thus the statement is factually incorrect.

News Author — **I think it's False.** The statement is factually inaccurate because the Pilgrims first landed at Provincetown Harbor, not Plymouth, Massachusetts. While they did establish Plymouth Colony in 1620, it was not their initial landing site.

Critic — **I think it's False.** The statement is inaccurate because the Pilgrims first anchored in Provincetown Harbor, not Plymouth, Massachusetts. The establishment of Plymouth Colony occurred after their initial landing.

Result: FALSE

Fig. 9 Case 2 of multi-agent debate: agents correct their peers' mistakes.

Atomic claim: MMR and chickenpox are vaccines that live but weakened versions.
Label: TRUE

General Public — **I think it's False.** The statement is slightly inaccurate in its wording. It suggests that MMR and chickenpox vaccines are 'live but weakened versions,' which is misleading. They are live attenuated vaccines, meaning they contain weakened forms of the viruses, but they are not 'live' in the sense that they can cause disease in healthy individuals. The phrasing could lead to confusion about the nature of these vaccines.

Scientist — **I think it's False.** The statement is misleading because it suggests that MMR and chickenpox vaccines are 'live but weakened versions,' which could imply they can cause disease. While they are live attenuated vaccines, the phrasing lacks clarity and may confuse individuals regarding their safety.

News Author — **I think it's False.** The statement is misleading due to its phrasing 'live but weakened versions,' which may imply a risk of causing disease that is not present for healthy individuals. The correct term is 'live attenuated vaccines,' and the phrasing could confuse people regarding the vaccines' safety.

General Public — **I think it's True.** I would like to correct my earlier statement, which focused on too many unnecessary details. Both MMR and chickenpox vaccines are live attenuated vaccines, meaning they contain weakened forms of the viruses.

Scientist — **I think it's True.** General Public is right. Both MMR and chickenpox vaccines are indeed live attenuated vaccines, which means they contain weakened forms of the viruses that do not cause disease in healthy individuals while still providing immunity.

News Author — **I think it's True.** I agree with Scientist. The statement accurately describes MMR and chickenpox vaccines as live attenuated vaccines, which contain weakened forms of the viruses that provide immunity without causing disease.

Result: TRUE

Fig. 10 Case 3 of multi-agent debate: agents continuously conduct self-reflection.

encompassing 38 manually curated topics and 2,280 factual questions, where models are required to generate long-form textual answers. Following the setup of SAFE [32], we apply stratified sampling to select 250 instances from LongFact with a balanced topic distribution as our test set. LongHalluQA, newly introduced in this paper, is a Chinese benchmark for long-form factuality evaluation, with details provided in Section 3. From this dataset, we sampled 250 high-quality instances for evaluation.

6.2.2 Evaluated Models

In this study, we carefully select 9 representative models from 7 different families, both domestic and international, to conduct a comprehensive and in-depth evaluation:

- **Closed-source models:** Claude series: `Claude-3-Sonnet`; OpenAI series: `GPT-3.5-Turbo`, `GPT-4-Turbo`; Doubao series: `Doubao-1.5-Pro-32k`; Kimi series: `Moonshot-V1-8k`.
- **Open-source models:** DeepSeek series: `DeepSeek-V3`; Llama series: `Llama3.1-70B`; Qwen series: `Qwen2.5-72B-Instruct`, `QwQ-32B`.

6.2.3 Evaluation Setup

In the MAD-Fact system, the Clerk and Evaluator agents are initialized based on `GPT-4o-mini`, while the Judge agent is initialized based on `Llama3.3-70B-instruct`. The number of Evaluator agents is set to 3, and the debate protocol follows Rule 1: Autonomous Retrieval and Free Debate.

We adopt the weighted evaluation metrics introduced in section 5.2. Among them, `GPT-4o`, `DeepSeek-R1`, and `Claude-3-Opus` are selected as expert models due to their comprehensive and strong performance. The factual importance score at the k-th level of the pyramid model is defined as $\omega = 5 - k$. Specifically, key points mentioned by all three expert models receive a score of 4, those mentioned by two expert models receive a score of 3, and so forth.

6.2.4 Results Analysis

Figure 11 illustrates the evaluation performance of the selected models on the LongFact dataset, ranked in descending order of $F_1@\gamma$. $F_1@\gamma$ is calculated under $\gamma = 1.0$ and $\gamma = 0.8$, and the model rankings remain relatively stable across different γ values. Detailed results are shown in Table 7.
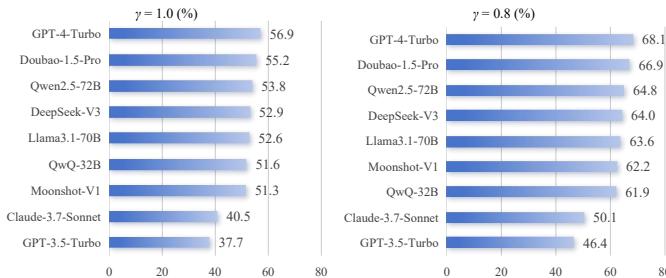

Fig. 11   The evaluation performance of the selected models on LongFact.

As shown in Figure 11 and Table 7, models with larger parameter sizes and more recent releases generally demonstrate better factual consistency in long-form scenarios. For instance, `GPT-4-Turbo` outperforms `GPT-3.5-Turbo`. Under the two selected γ values, the best-performing model in terms of factuality is `GPT-4-Turbo`, consistent with the findings reported by SAFE [32]. The performance gap between the domestic model `Doubao-1.5-Pro` and `GPT-4-Turbo` is minimal, only about 1–2 percentage points, highlighting the strong potential and advantages of domestic models. `GPT-3.5-Turbo`, as a relatively older model, is surpassed by most newer models. It is noteworthy that `Claude-3.7-Sonnet` and `QwQ-32B`, both widely regarded

Table 7   The detailed evaluation results of the selected models on LongFact. Bold numbers indicate the best results and underlined numbers denote the second-best results.

Dataset	Model	Precision	$\gamma = 1.0$		$\gamma = 0.8$	
			Recall	F1	Recall	F1
LongFact	GPT-4-Turbo	0.877	**0.445**	**0.569**	**0.557**	**0.681**
	Doubao-1.5-Pro	0.867	0.436	0.552	0.545	0.669
	Qwen2.5-72B	0.847	0.420	0.538	0.525	0.648
	DeepSeek-V3	0.838	0.414	0.529	0.518	0.640
	Llama3.1-70B	0.819	0.416	0.526	0.520	0.636
	QwQ-32B	0.763	0.416	0.516	0.520	0.619
	Moonshot-V1	0.837	0.396	0.513	0.495	0.622
	Claude-3.7-Sonnet	0.874	0.281	0.405	0.351	0.501
	GPT-3.5-Turbo	**0.908**	0.249	0.377	0.312	0.464

as strong and recently released models, did not achieve high rankings. This is primarily because `Claude-3.7-Sonnet` has a relatively low factual recall score (0.501 under $\gamma = 0.8$, ranking 8th among the selected models), while `QwQ-32B` has a relatively low factual precision score (0.763, ranking 9th). These results indicate that strong reasoning or coding capabilities do not necessarily guarantee superior factuality.
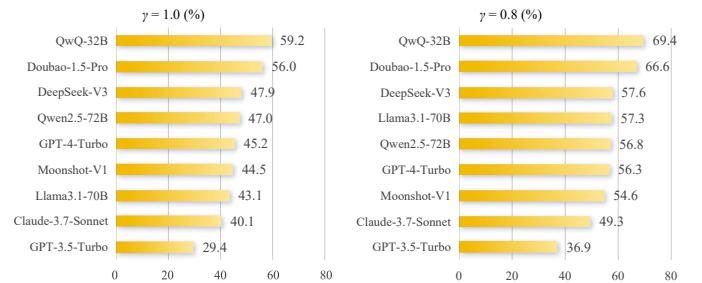

Fig. 12   The evaluation performance of the selected models on LongHalluQA.

Figure 12 presents the evaluation performance of the selected models on the LongHalluQA dataset, ranked in descending order of $F_1@\gamma$. $F_1@\gamma$ is calculated under $\gamma = 1.0$ and $\gamma = 0.8$, and the model rankings remain largely stable, with only slight variations in the middle range. Detailed results are provided in Table 8.
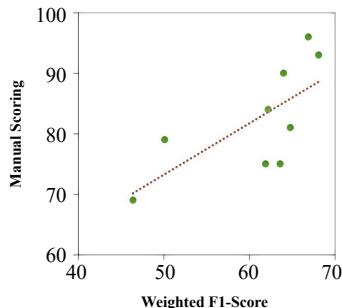
As shown in Figure 12 and Table 8, domestic models exhibit notable advantages in factuality evaluation for Chinese long-form responses. Under the two selected γ values, the top three models in terms of factuality are `QwQ-32B`, `Doubao-1.5-Pro`, and `DeepSeek-V3`, all domestic models. Among them, `QwQ-32B` and `Doubao-1.5-Pro` clearly outperform the others, exceeding the third-ranked `DeepSeek-V3` by nearly 10 percentage points. In contrast, `GPT-4-Turbo` performs worse on the Chinese dataset than on the English dataset, ranking 5th among the selected models when $\gamma = 1.0$ and 6th when $\gamma = 0.8$, indicating a notable issue of cultural bias [66] and an imbalance in multilingual capabilities.

To verify the effectiveness of the weighted evaluation metric based on the hierarchical importance of facts, we established a review panel consisting of two individuals with undergraduate degrees. From the LongFact dataset, 50 questions were sampled, and the outputs of differ-

Table 8 The detailed evaluation results of the selected models on LongHalluQA. Bold numbers indicate the best results and underlined numbers denote the second-best results.

Dataset	Model	Precision	$\gamma = 1.0$		$\gamma = 0.8$	
			Recall	F1	Recall	F1
	QwQ-32B	0.759	**0.511**	**0.592**	**0.639**	**0.694**
	Doubao-1.5-Pro	**0.823**	0.447	0.560	0.559	0.666
	DeepSeek-V3	0.773	0.367	0.479	0.459	0.576
	Qwen2.5-72B	0.807	0.351	0.470	0.439	0.568
LongHalluQA	GPT-4-Turbo	0.806	0.346	0.452	0.433	0.563
	Moonshot-V1	0.819	0.328	0.445	0.410	0.546
	Llama3.1-70B	0.787	0.360	0.431	0.450	0.573
	Claude-3.7-Sonnet	0.763	0.291	0.401	0.364	0.493
	GPT-3.5-Turbo	0.758	0.195	0.294	0.244	0.369

ent models were manually rated. The correlation between the human ratings and the weighted F_1 score ($\gamma = 0.8$) was then calculated, as shown in Figure 13. The results indicate that the Pearson correlation coefficient reached a statistically significant value of $r = 0.701$ ($p = 0.036$), demonstrating that the proposed weighted evaluation metric aligns well with human judgments in assessing the factuality of long-form responses.

Fig. 13 Correlation analysis between weighted evaluation metrics and human ratings.

■ 7 Discussion

This paper presents a systematic study on long-form factuality evaluation by LLMs. Nonetheless, limitations remain in dataset representativeness, multi-agent efficiency, and communication reliability. Future research can explore the following directions:

Expanding and grounding factuality evaluation datasets. While LongHalluQA extends existing short-text datasets to support long-form evaluation, the data are still derived from curated sources rather than real-world user-generated content. Future research should incorporate diverse, high-risk domains such as biomedical, finance, and law, and integrate practical real-world sources to improve coverage, generalization, and applicability of long-form factuality evaluation.

Optimizing the cost-effectiveness trade-off in multi-agent systems. The MAD-Fact system enhances fact-checking performance through multi-agent debate, but its reliance on multiple models increases token consumption and operational costs. Future work could explore lightweight collaboration strategies, such as dynamically adjusting the number of agents, optimizing debate rounds, or employing model distillation and parameter sharing to reduce computational

overhead. Integrating reinforcement learning to select optimal collaboration modes based on context could further improve efficiency without sacrificing evaluation accuracy, supporting deployment from experimental settings to real-world high-risk scenarios.

Mitigating communication hallucinations in multi-agent systems. During multi-agent debates, agents may reach misleading consensus due to confidently incorrect responses, causing erroneous conclusions. Yoffe et al. [67] attribute this issue to communication failures among agents, referred to as communication hallucinations. Future research can focus on: (i) confidence-based weighting of agent responses to downweight unreliable viewpoints; (ii) cross-agent consistency verification using logical reasoning to detect contradictions; (iii) adversarial training that injects incorrect viewpoints to test and enhance system robustness. Progress in these areas would strengthen the reliability of multi-agent systems in complex long-form evaluation tasks.

■ 8 Conclusion

We present a unified framework for long-form factuality evaluation of LLMs, combining large-scale benchmarks, multi-agent verification, and weighted evaluation metrics. We introduce LongHalluQA, a Chinese long-form factuality dataset, and MAD-Fact, a multi-agent debate system designed to mitigate single-model bias. Additionally, we propose hierarchical modeling of factual importance to guide weighted metrics that closely reflect human judgments. Extensive experiments demonstrate MAD-Fact's effectiveness, with evaluations on LongFact and LongHalluQA showing that larger models generally outperform smaller ones, while domestic models excel on Chinese-language tasks. We believe this paper can serve as a foundation for future research on long-form factuality evaluation and guide the development of more reliable LLMs.

■ Acknowledgement

■ References

[1] Brown T, Mann B, Ryder N, Subbiah M, Kaplan J D, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, others . Language models are few-shot learners. In: Advances in neural information processing systems. 2020, 1877–1901

[2] Wei J, Wang X, Schuurmans D, Bosma M, Xia F, Chi E, Le Q V, Zhou D, others . Chain-of-thought prompting elicits reasoning in large language models. In: Advances in neural information processing systems. 2022, 24824–24837

[3] Zhang S, Roller S, Goyal N, Artetxe M, Chen M, Chen S, Dewan C, Diab M, Li X, Lin X V, others . Opt: Open pre-trained transformer language models. arXiv preprint arXiv:2205.01068, 2022

[4] Touvron H, Lavril T, Izacard G, Martinet X, Lachaux M A, Lacroix T, Rozière B, Goyal N, Hambro E, Azhar F, others . Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971, 2023

[5] Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman F L, Almeida D, Altenschmidt J, Altman S, Anadkat S, others . Gpt-4 technical report. arXiv preprint arXiv:2303.08774, 2023

[6] Zhao W X, Zhou K, Li J, Tang T, Wang X, Hou Y, Min Y, Zhang B, Zhang J, Dong Z, others . A survey of large language models. arXiv preprint arXiv:2303.18223, 2023

[7] Yang R, Tan T F, Lu W, Thirunavukarasu A J, Ting D S W, Liu N. Large language models in health care: Development, applications, and challenges. Health Care Science, 2023, 2(4): 255–263

[8] Valeyre S, Aboura S. Llms for time series: an application for single stocks and statistical arbitrage. arXiv preprint arXiv:2412.09394, 2024

[9] Caballero W N, Jenkins P R. On large language models in national security applications. Stat, 2025, 14(2): e70057

[10] Johnson E, Wilson N. Enhancing agricultural machinery management through advanced llm integration. arXiv preprint arXiv:2407.20588, 2024

[11] Shu D, Zhao H, Liu X, Demeter D, Du M, Zhang Y. Lawllm: Law large language model for the us legal system. In: Proceedings of the 33rd ACM International Conference on Information and Knowledge Management. 2024, 4882–4889

[12] Wang L, Ma C, Feng X, Zhang Z, Yang H, Zhang J, Chen Z, Tang J, Chen X, Lin Y, Zhao W X, Wei Z, Wen J. A survey on large language model based autonomous agents. Frontiers of Computer Science, 2024, 18(6): 186345

[13] Qu C, Dai S, Wei X, Cai H, Wang S, Yin D, Xu J, Wen J r. Tool learning with large language models: a survey. Frontiers of Computer Science, 2025, 19(8): 198343

[14] Xi Z, Chen W, Guo X, He W, Ding Y, Hong B, Zhang M, Wang J, Jin S, Zhou E, others . The rise and potential of large language model based agents: A survey. Science China Information Sciences, 2025, 68(2): 121101

[15] Zhang Y, Li Y, Cui L, Cai D, Liu L, Fu T, Huang X, Zhao E, Zhang Y, Chen Y, Wang L, Luu A T, Bi W, Shi F, Shi S. Siren's song in the ai ocean: A survey on hallucination in large language models. Computational Linguistics, 2025, 1–46

[16] Rawte V, Sheth A, Das A. A survey of hallucination in large foundation models. arXiv preprint arXiv:2309.05922, 2023

[17] Huang L, Yu W, Ma W, Zhong W, Feng Z, Wang H, Chen Q, Peng W, Feng X, Qin B, others . A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. ACM Transactions on Information Systems, 2025, 43(2): 1–55

[18] Bai Z, Wang P, Xiao T, He T, Han Z, Zhang Z, Shou M Z. Hallucination of multimodal large language models: A survey. arXiv preprint arXiv:2404.18930, 2024

[19] Li H, Ye J, Wu J. Privacy dilemmas and opportunities in large language models: a brief review. Frontiers of Computer Science, 2025, 19(10): 1910356

[20] Lin X, Ning Y, Zhang J, Dong Y, Liu Y, Wu Y, Qi X, Sun N, Shang Y, Cao P, others . Llm-based agents suffer from hallucinations: A survey of taxonomy, methods, and directions. arXiv preprint arXiv:2509.18970, 2025

[21] Pal A, Umapathi L K, Sankarasubbu M. Med-HALT: Medical domain hallucination test for large language models. In: Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL). 2023, 314–334

[22] Cui J, Ning M, Li Z, Chen B, Yan Y, Li H, Ling B, Tian Y, Yuan L. Chatlaw: A multi-agent collaborative legal assistant with knowledge graph enhanced mixture-of-experts large language model. arXiv preprint arXiv:2306.16092, 2023

[23] Bhatia G, Nagoudi E M B, Cavusoglu H, Abdul-Mageed M. FinTral: A family of GPT-4 level multimodal financial large language models. In: Findings of the Association for Computational Linguistics: ACL 2024. 2024, 13064–13087

[24] Nan Q, Sheng Q, Cao J, Zhu Y, Wang D, Yang G, Li J. Exploiting user comments for early detection of fake news prior to users' commenting. Frontiers of Computer Science, 2025, 19(10): 1910354

[25] Lee N, Ping W, Xu P, Patwary M, Fung P N, Shoeybi M, Catanzaro B. Factuality enhanced language models for open-ended text generation. In: Advances in Neural Information Processing Systems. 2022, 34586–34599

[26] Gao L, Dai Z, Pasupat P, Chen A, Chaganty A T, Fan Y, Zhao V, Lao N, Lee H, Juan D C, Guu K. RARR: Researching and revising what language models say, using language models. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2023, 16477–16508

[27] Gao T, Yen H, Yu J, Chen D. Enabling large language models to generate text with citations. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. 2023, 6465–6488

[28] Manakul P, Liusie A, Gales M. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. 2023, 9004–9017

[29] Huang Y, Xu J, Lai J, Jiang Z, Chen T, Li Z, Yao Y, Ma X, Yang L, Chen H, others . Advancing transformer architecture in long-context large language models: A comprehensive survey. arXiv preprint arXiv:2311.12351, 2023

[30] Setty S, Thakkar H, Lee A, Chung E, Vidra N. Improving retrieval for rag based question answering models on financial documents. arXiv preprint arXiv:2404.07221, 2024

[31] Min S, Krishna K, Lyu X, Lewis M, Yih W t, Koh P, Iyyer M, Zettlemoyer L, Hajishirzi H. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. 2023, 12076–12100

[32] Wei J, Yang C, Song X, Lu Y, Hu N, Huang J, Tran D, Peng D, Liu R, Huang D, Du C, Le Q V. Long-form factuality in large language models. In: Advances in Neural Information Processing Systems. 2024, 80756–80827

[33] Fatahi Bayat F, Zhang L, Munir S, Wang L. FactBench: A dynamic benchmark for in-the-wild language model factuality evaluation. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2025, 33090–33110

[34] Song Y, Kim Y, Iyyer M. VeriScore: Evaluating the factuality of verifiable claims in long-form text generation. In: Findings of the Association for Computational Linguistics: EMNLP 2024. 2024, 9447–9474

[35] Xie Z, Xing R, Wang Y, Geng J, Iqbal H, Sahnan D, Gurevych I, Nakov P. FIRE: Fact-checking with iterative retrieval and verification. In: Findings of the Association for Computational Linguistics: NAACL 2025. 2025, 2901–2914

[36] Cheng Q, Sun T, Zhang W, Wang S, Liu X, Zhang M, He J, Huang M, Yin Z, Chen K, others . Evaluating hallucinations in chinese large language models. arXiv preprint arXiv:2310.03368, 2023

[37] He Y, Li S, Liu J, Tan Y, Wang W, Huang H, Bu X, Guo H, Hu C, Zheng B, Lin Z, Sun D, Zheng Z, Su W, Zheng B. Chinese SimpleQA: A Chinese factuality evaluation for large language models. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2025, 19182–19208

[38] Du Y, Li S, Torralba A, Tenenbaum J B, Mordatch I. Improving factuality and reasoning in language models through multiagent debate. In: Forty-first International Conference on Machine Learning. 2023

[39] Nenkova A, Passonneau R. Evaluating content selection in summarization: The pyramid method. In: Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004

[40] Chengqing Z, Rui X, Jiajun Z. Text Data Mining, 2nd edition. Beijing, China: Tsinghua University Press, 2022

[41] Lin S, Hilton J, Evans O. TruthfulQA: Measuring how models mimic human falsehoods. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2022, 3214–3252

[42] Li J, Cheng X, Zhao X, Nie J Y, Wen J R. HaluEval: A large-scale hallucination evaluation benchmark for large language models. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. 2023, 6449–6464

[43] Mallen A, Asai A, Zhong V, Das R, Khashabi D, Hajishirzi H. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2023, 9802–9822

[44] Vu T, Iyyer M, Wang X, Constant N, Wei J, Wei J, Tar C, Sung Y H, Zhou D, Le Q, Luong T. FreshLLMs: Refreshing large language models with search engine augmentation. In: Findings of the Association for Computational Linguistics: ACL 2024. 2024, 13697–13720

[45] Scirè A, Bejgu A S, Tedeschi S, Ghonim K, Martelli F, Navigli R. Truth or mirage? towards end-to-end factuality evaluation with llm-oasis. arXiv preprint arXiv:2411.19655, 2024

[46] Li D, Jiang B, Huang L, Beigi A, Zhao C, Tan Z, Bhattacharjee A, Jiang Y, Chen C, Wu T, others . From generation to judgment: Opportunities and challenges of llm-as-a-judge. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. 2025

[47] Zhang X, Gao W. Towards LLM-based fact verification on news

claims with a hierarchical step-by-step prompting method. In: Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers). 2023, 996–1011

[48] Guo Z, Schlichtkrull M, Vlachos A. A survey on automated fact-checking. Transactions of the Association for Computational Linguistics, 2022, 10: 178–206

[49] Xu D, Chen W, Peng W, Zhang C, Xu T, Zhao X, Wu X, Zheng Y, Wang Y, Chen E. Large language models for generative information extraction: a survey. Frontiers of Computer Science, 2024, 18(6): 186357

[50] Gao L, Dai Z, Pasupat P, Chen A, Chaganty A T, Fan Y, Zhao V, Lao N, Lee H, Juan D C, Guu K. Rarr: Researching and revising what language models say, using language models. In: Annual Meeting of the Association for Computational Linguistics. 2022

[51] Wang Y, Reddy R G, Mujahid Z M, Arora A, Rubashevskii A, Geng J, Afzal O M, Pan L, Borenstein N, Pillai A, Augenstein I, Gurevych I, Nakov P. Factcheck-bench: Fine-grained evaluation benchmark for automatic fact-checkers. In: Conference on Empirical Methods in Natural Language Processing. 2023

[52] Khaliq M A, Chang P Y C, Ma M, Pflugfelder B, Miletić F. RAGAR, your falsehood radar: RAG-augmented reasoning for political fact-checking using multimodal large language models. In: Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER). 2024, 280–296

[53] Quelle D, Bovet A. The perils and promises of fact-checking with large language models. Frontiers in Artificial Intelligence, 2023, 7

[54] Sun X, Li J, Zhong Y, Zhao D, Yan R. Towards detecting llms hallucination via markov chain-based multi-agent debate framework. In: ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2025, 1–5

[55] Han S, Zhang Q, Yao Y, Jin W, Xu Z, He C. Llm multi-agent systems: Challenges and open problems. arXiv preprint arXiv:2402.03578, 2024

[56] Guo T, Chen X, Wang Y, Chang R, Pei S, Chawla N V, Wiest O, Zhang X. Large language model based multi-agents: A survey of progress and challenges. In: Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24. 8 2024, 8048–8057

[57] Chen S, Liu Y, Han W, Zhang W, Liu T. A survey on multi-generative agent system: Recent advances and new frontiers. arXiv preprint arXiv:2412.17481, 2024

[58] Chan C M, Chen W, Su Y, Yu J, Xue W, Zhang S, Fu J, Liu Z. Chateval: Towards better llm-based evaluators through multi-agent debate. In: International Conference on Representation Learning. 2024, 9079–9093

[59] Rahmani H A, Yilmaz E, Craswell N, Mitra B. Judgeblender: Ensembling judgments for automatic relevance assessment. arXiv preprint arXiv:2412.13268, 2024

[60] Feng Z, Su J, Zheng J, Ren J, Zhang Y, Wu J, Wang H, Liu Z.

M-MAD: Multidimensional multi-agent debate for advanced machine translation evaluation. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2025, 7084–7107

[61] Bo X, Zhang Z, Dai Q, Feng X, Wang L, Li R, Chen X, Wen J R. Reflective multi-agent collaboration based on large language models. In: The Thirty-eighth Annual Conference on Neural Information Processing Systems. 2024

[62] Chern E, Chern S, Chen S, Yuan W, Feng K, Zhou C, He J, Neubig G, Liu P. Factool: Factuality detection in generative ai - a tool augmented framework for multi-task and multi-domain scenarios. arXiv preprint arXiv:2307.13528, 2023

[63] Zhao Y, Zhang J, Chern I, Gao S, Liu P, He J, others . Felm: Benchmarking factuality evaluation of large language models. In: Advances in Neural Information Processing Systems. 2023, 44502–44523

[64] Li M, Peng B, Galley M, Gao J, Zhang Z. Self-checker: Plug-and-play modules for fact-checking with large language models. In: Findings of the Association for Computational Linguistics: NAACL 2024. 2024, 163–181

[65] Wang B, Chern E, Liu P. Chinesefacteval: A factuality benchmark for chinese llms, 2023

[66] Xu S, Leng Y, Yu L, Xiong D. Self-pluralising culture alignment for large language models. In: Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). 2025, 6859–6877

[67] Yoffe L, Amayuelas A, Wang W Y. Debunc: mitigating hallucinations in large language model agent communication with uncertainty estimations. arXiv preprint arXiv:2407.06426, 2024

Yanan Cao obtained a Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences in 2012. She is a professor at the Institute of Information Engineering, Chinese Academy of Sciences. Her research interests include social network analysis and natural language processing. To date, she has published more than 70 papers, including AAAI, NeurIPS, IJCAI, ACL, WWW, and she won the best paper award at PAKDD-20.

Yucheng Ning received the B.S. degree in Artificial Intelligence from the University of Chinese Academy of Sciences (UCAS) in 2025, where he is currently pursuing the Ph.D. degree with the Institute of Information Engineering (IIE), Chinese Academy of Sciences. His research interests include natural language processing (NLP) and LLM safety.

Xixun Lin received his Ph.D. degree from the Institute of Information Engineering (IIE), Chinese Academy of Sciences in 2022. He is currently an Assistant Professor at IIE. His research focuses on trustworthy machine learning. He has published over 30 papers in top-tier academic journals and conferences, including TPAMI, TKDE, TNNLS, ICML, KDD, and WWW.

Fang Fang obtained her Ph.D. from the Institute of Computing Technology, CAS in 2019. Her research focuses on intelligent text content processing, including information extraction and knowledge graph in natural language processing. To date, she has published over 20 papers in CCF-A/B conferences like ACL and AAAI.