SAGE: A Search-AuGmented Evaluation of Large Language Models on Free-Form QA

Anonymous ACL submission

Abstract

As Large Language Models (LLMs) become increasingly used for question-answering (QA), relying on static, pre-annotated references for evaluation poses significant challenges in cost, scalability, and completeness. We propose Search-AuGmented Evaluation (SAGE), a framework to assess LLM outputs without predetermined ground-truth answers. Unlike conventional metrics that compare to static references or depend solely on LLM-as-a-judge knowledge, SAGE acts as an agent that actively retrieves and synthesizes external evidence. It iteratively generates web queries, collects information, summarizes findings, and refines subsequent searches through reflection. By reducing dependence on static reference-driven evaluation protocols, SAGE offers a scalable and adaptive alternative for evaluating the factuality of LLMs. Experimental results on multiple free-form QA benchmarks show that SAGE achieves substantial to perfect agreement with human evaluations.

1 Introduction

Free-form Question Answering (QA) requires models to generate precise natural language responses to broad, open-ended queries (Wang et al., 2023a). As such, it serves as a key benchmark for evaluating the factuality of Large Language Models (LLMs), which are increasingly integrated into real-world applications such as online search engines and virtual assistants. However, LLMs are prone to hallucination (Gou et al., 2024), and evaluating their factuality with standard protocols remains difficult.

Traditional evaluation methods, including lexical matching metrics such as Exact Match (EM) and F1, rely on comparisons to static ground-truth references. While convenient and efficient, these methods fall short of capturing the diversity of free-form QA outputs and are often infeasible to scale due to the high cost of human annotations (Chiang and Lee, 2023; Mañas et al., 2024; Zhu et al., 2023). More critically, instruction-tuned LLMs produce outputs that are often unpredictable, context-dependent, and non-deterministic, making it impractical to pre-annotate reference answers for every possible response (Yehudai et al., 2025; Li et al., 2024). As a result, static, reference-driven evaluation protocols are fundamentally misaligned with the nature of free-form QA, where answers are open-ended and often lack a single definitive ground truth.

An emerging alternative reference-based evaluation is the LLM-as-a-judge approach (Zheng et al., 2024; Chen et al., 2024), where one model, for instance, is prompted to assess the output of another based on task-specific criteria such as relevance, depth, or creativity (Verga et al., 2024). This method has shown promise in subjective tasks such as summarization, dialogue, and instruction following, where quality is shaped by style or user preference and multiple interpretations are often equally valid (Gu et al., 2025; Son et al., 2024). However, its reliability deteriorates when the goal shifts to objective correctness (Krumdick et al., 2025; Badshah and Sajjad, 2025; Gu et al., 2025).

For objective, fact-centric tasks such as free-form QA, an unguided (i.e., reference-free) judge is forced to lean solely on its frozen pre-trained knowledge. This constraint exposes several recurring failure modes: i) knowledge staleness where the judges confidently endorse answers that became outdated after their training cut off (Vu et al., 2024; Cheng et al., 2024; Badshah and Sajjad, 2025); ii) length or verbosity bias, where longer or more detailed answers are overrated even when they contain errors (Li et al., 2025b; Ye et al., 2024); iii) prompt sensitivity, in which small variations to the prompt or the order of candidate answers flip a correct/incorrect verdict (Ye et al., 2024; Thakur et al., 2025); and iv) hallucinated rationales, where the judge develops supporting evidence to justify its decision (Kamalloo et al., 2023).

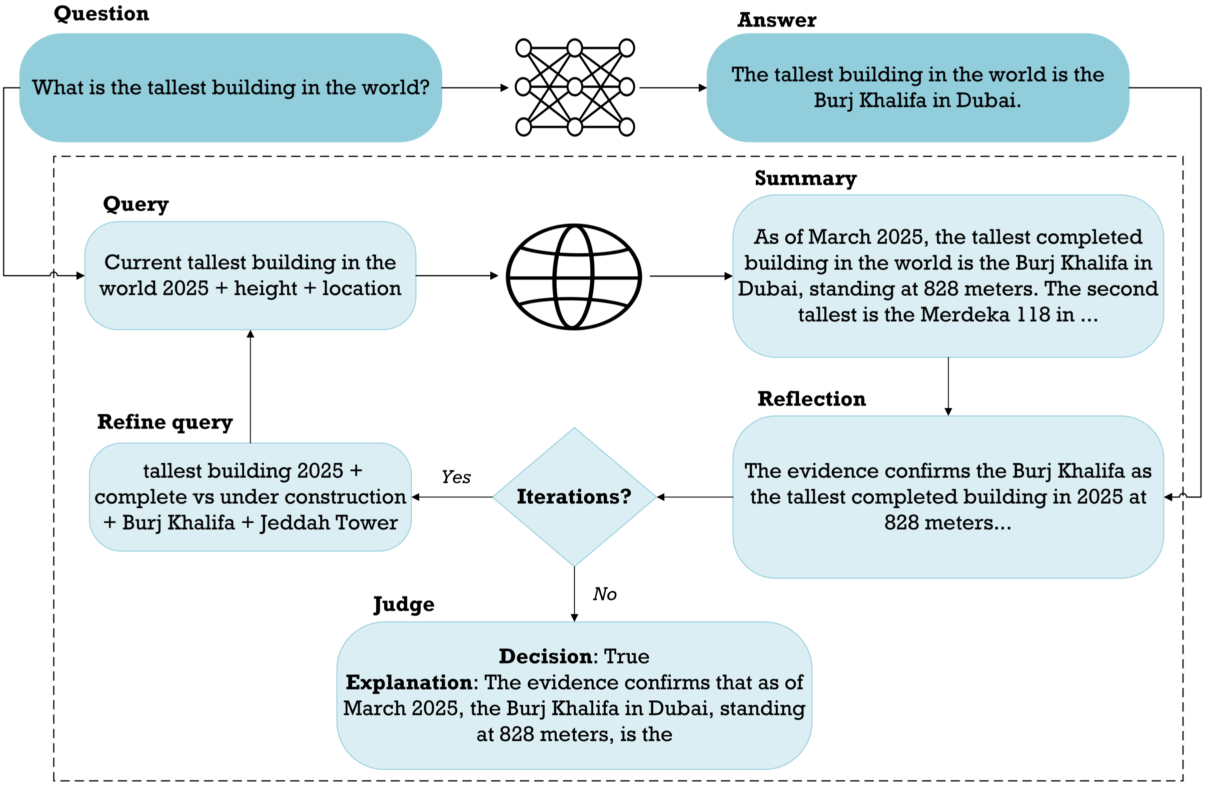
Question
What is the tallest building in the world?
Answer
The tallest building in the world is the Burj Khalifa in Dubai.
Query
Current tallest building in the world 2025 + height + location
Summary
As of March 2025, the tallest completed building in the world is the Burj Khalifa in Dubai, standing at 828 meters. The second tallest is the Merdeka 118 in ...
Reflection
The evidence confirms the Burj Khalifa as the tallest completed building in 2025 at 828 meters...
Refine query
tallest building 2025 + complete vs under construction + Burj Khalifa + Jeddah Tower
Yes
Iterations?
No
Judge
Decision: True
Explanation: The evidence confirms that as of March 2025, the Burj Khalifa in Dubai, standing at 828 meters, is the

Figure 1: Given the question and candidate answer, SAGE begins with ***initial query*** from the question. The query triggers ***web searches*** across multiple sources, followed by ***evidence summarization*** to extract key insights. The ***reflection module*** assesses evidence sufficiency and relevance, triggering ***query refinement*** if needed. After N iterations, the ***judge*** synthesizes the evidence to provide a final decision with rationale.

Given that pre-annotated reference answers at scale are impractical and reference-free LLM-as-a-judge setups remain unreliable for objective tasks, we argue that, unlike subjective evaluation, objective correctness cannot be assessed solely through an LLM's parametric knowledge or its preferences. Therefore, we propose Search-AuGmented Evaluation (SAGE), a novel framework that bridges the stated gap by equipping LLM judges with the ability to actively collect and synthesize external evidence. Instead of costly human-annotated reference answers, SAGE dynamically and iteratively generates output-specific web queries, retrieves information, reflects on findings, and refines its search strategy to verify the correctness of model outputs.

SAGE offers four key advantages: (1) it substantially reduces dependency on the judge's parameter knowledge, (2) avoids the need for human-annotated reference answers, making evaluation more scalable, (3) grounds evaluations in up to date, verifiable information, and (4) enables assessment of novel or rapidly evolving topics where parameter knowledge may be outdated or incomplete. Through experiments on free-form QA, we find that SAGE is aligned with reference-based evaluation and achieves substantial to perfect agreement with human evaluators.

2 Methodology

We introduce Search-AuGmented Evaluation (SAGE), a reference-free framework for evaluating LLM responses. Unlike conventional approaches that rely on fixed reference answers or human-annotated ground truths, SAGE autonomously gathers and integrates external evidence to assess the correctness of free-form responses. Figure 1 illustrates the process of SAGE.

Let $x \in \mathcal{X}$ denote an input question of which a candidate LLM C, such as GPT-3.5-turbo, produces a response $\hat{y} \in \mathcal{Y}$ according to $\hat{y} = C(x)$. In our framework, an LLM is employed as a judge, denoted by J, to assess the correctness of $\hat{y}$ based on external evidence, i.e., augmented search. Algorithm 1 presents the procedure of **SAGE**. It comprises of the following modules, where the LLM agent acts in different roles, except for memory.

Query Generation (Q): SAGE first generates a search query with a goal to retrieve relevant information from the web, which is then augmented with a judge to evaluate the correctness of the answer $\hat{y}$. It achieves this in an iterative process. In the first iteration ($i = 1$), the query q_1 is constructed solely from the input question x: $q_1 = Q(x)$. The goal is to retrieve background information relevant to the question domain, without assuming the correctness of any particular answer. In subsequent iterations ($i > 1$), queries are refined using accumulated evidence and reflections stored in the short-term memory: $q_i = Q'(\mathcal{M})$. The query refinement resolves remaining uncertainty by retrieving more targeted information needed to verify the candidate answer $\hat{y}$.

Web Search (S): The query q_i is submitted to a web search via the Serper API,[1] to return real-time results. For each query, we retrieve the top $k = 3$ search snippets or URLs, which serve as raw external evidence for the subsequent steps. The search results typically include titles, snippets, and source URLs, providing diverse perspectives from multiple unique sources. We refer to the web search results as S.

Summarization (Σ): This condenses the retrieved web search results S into a focused evidence segment: $E_i = \Sigma(S(q_i))$. Σ extracts salient factual content while filtering out redundant or irrelevant information. The resulting evidence E_i serves as a concise, interpretable knowledge unit that informs reflection.

Reflection (R): The summarized evidence E_i is evaluated in relation to the input question x and the candidate answer $\hat{y}$ to assess its relevance, sufficiency, and factual alignment. This step yields a reflection: $R_i = R(x, \hat{y}, E_i)$, which captures whether the current evidence supports, contradicts, or is inconclusive with respect to $\hat{y}$. The reflection module also identifies missing information or ambiguities that can guide subsequent query refinement. By explicitly reasoning over the current evidence, reflection enables SAGE to improve its evaluation behavior over multiple steps.

Short-Term Memory: After each iteration, the tuple (q_i, E_i, R_i) consisting of the generated query, corresponding evidence, and reflection is appended to the short-term memory buffer: $\mathcal{M} \leftarrow \mathcal{M} \cup \{(q_i, E_i, R_i)\}$. This memory $\mathcal{M}$ serves as an evolving trace of the evaluation process, accumulating context that informs subsequent query refinement and reasoning. The memory is append-only and scoped to a single evaluation episode, ensuring that each step builds upon the full history of prior actions, observations, and reflections.

Algorithm 1 SAGE

```
Require: Question $x$, candidate answer $\hat{y}$, number of iterations $N$
Ensure: Verdict $v \in \{0, 1\}$ and rationale $r_J$
Initialize short-term memory $\mathcal{M} \leftarrow []$ {Memory buffer for query-evidence-reflection trace}
$q_1 \leftarrow Q(x)$ {Generate initial query from input question}
$E_1 \leftarrow \Sigma(S(q_1))$ {Retrieve and summarize external evidence}
$R_1 \leftarrow R(x, \hat{y}, E_1)$ {Reflect on evidence relevance and sufficiency}
Append $(q_1, E_1, R_1)$ to memory $\mathcal{M}$
for $i \leftarrow 2$ to $N$ do
    $q_i \leftarrow Q'(\mathcal{M})$ {Refine query using accumulated memory}
    $E_i \leftarrow \Sigma(S(q_i))$
    $R_i \leftarrow R(x, \hat{y}, E_i)$
    Append $(q_i, E_i, R_i)$ to memory $\mathcal{M}$
end for
$E_{\text{total}} \leftarrow \bigoplus_{i=1}^{N} E_i$ {Aggregate evidence summaries}
$R_{\text{total}} \leftarrow \bigoplus_{i=1}^{N} R_i$ {Aggregate reflections}
$(v, r_J) \leftarrow J(x, \hat{y}, E_{\text{total}}, R_{\text{total}})$ {Final verdict and rationale}
return $(v, r_J)$
```

Judge (J): After N iterations, the judge module synthesizes all collected evidence to assess the correctness of the candidate answer. Specifically, it includes the aggregated evidence summary and reflections. Based on such prior components, J produces a binary verdict $v \in \{0, 1\}$ indicating whether $\hat{y}$ is factually correct, along with a natural language rationale r_J: $(v, r_J) = J(x, \hat{y}, E_{\text{total}}, R_{\text{total}})$. The overall procedure is summarized in Algorithm 1.

Each of the above module uses a one-shot prompt template that demonstrates the expected input-output behavior.

3 Experimental Setup

3.1 Models

In our experiments, we utilize: Gemini-1.5-pro (Team, 2024), GPT-3.5-turbo (Brown et al., 2020), and GPT-4o-mini (Team, 2023) both as candidates (C) and as judges (J) within the SAGE framework, allowing us to assess both their ability to generate accurate responses and to evaluate re-

[1] https://serper.dev/

sponses from other models. We also explore the potential of a small LLM as a judge, including Mistral 7B (Jiang et al., 2023). All experiments are conducted with a temperature of 0 to maximize determinism and reliability, as increasing the temperature degrades the performance of LLM-based evaluators (Hada et al., 2024). For brevity, we refer to these models as Gemini, GPT-3.5, GPT-4o, and Mistral throughout our analysis.

3.2 Datasets

We evaluate SAGE on widely used free-form question-answering datasets that represent different question types, knowledge domains, and complexity levels. These includes AmbigQA (Min et al., 2020), HotpotQA (Yang et al., 2018), TriviaQA (Joshi et al., 2017), and Natural Questions (NQ-Open) (Kwiatkowski et al., 2019). Free-form question-answering underpins a broad range of practical applications, where maintaining accuracy and ensuring truthfulness are paramount (Gou et al., 2024). We also used FreshQA (Vu et al., 2023) to evaluate SAGE's ability in detecting outdated knowledge. Due to computational constraints, we randomly sample 300 instances from each dataset, ensuring balanced representation across question types and difficulty levels. Each dataset provides reference answers that serve as ground truth for evaluators (see Appendix 8.1).

3.3 Prompts

Our prompting strategy uses templates for both response generation and evaluation. For candidate models, we use few-shot Chain-of-Thought (CoT) prompts with 6 examples per instance to elicit detailed, reasoning-based responses. This strategy is well-suited for free-form QA as it encourages explicit reasoning toward a conclusion (Gou et al., 2024). In SAGE, we design module-specific prompts that combine role instructions with a step-by-step reasoning guide (see Appendix 8.2).

3.4 Baselines

We compare SAGE against several established evaluation approaches, including reference-based and reference-free methods. Moreover, we conduct a human evaluation using two QA datasets. In the following, we summarize each baseline method. Appendix 8.3 provides further details on them.

Reference-based evaluation. We consider ***Exact Match (EM) and F1*** for reference-based evaluation. Kamalloo et al. (2023) shows that standard automatic metrics can be misleading for free-form QA (Kamalloo et al., 2023). Therefore, following prior work (Wang et al., 2023a), we also employ GPT-4 as an evaluator that assesses candidate answers by comparing them to gold answers. To avoid confusion with other models and methods, we refer to this GPT-4 evaluator as **RefGPT**. As depicted in Table 1, RefGPT demonstrates consistently substantial agreement with human evaluations compared to EM and F1. In the paper, we mainly consider RefGPT as a reference-based evaluation baseline, unless specified. Moreover, due to its high agreement with human evaluation, we consider it as a silver human evaluation to calculate the agreement of SAGE over all QA datasets considered in the paper.

Candidate	**Task**	**Cohen's κ**			**Macro F1**		
		EM	**F1**	**RefGPT**	**EM**	**F1**	**RefGPT**
GPT-3.5	AmbigQA	0.54	0.66	**0.76**	0.76	0.83	**0.88**
	HotpotQA	0.60	0.76	**0.90**	0.79	0.88	**0.95**
GPT-4o	AmbigQA	0.48	0.55	**0.70**	0.73	0.77	**0.85**
	HotpotQA	0.54	0.66	**0.77**	0.76	0.83	**0.88**
Gemini	AmbigQA	0.56	0.57	**0.71**	0.77	0.78	**0.85**
	HotpotQA	0.49	0.66	**0.76**	0.73	0.83	**0.88**

Table 1: Agreement of reference-based metrics with human majority. F1 scores are converted to binary using a $\tau = 0.5$.

Reference-free evaluation. We adapt ***Judge without search*** as a baseline, following the approach from Liu et al. (2023). In this setting, the judge relies entirely on its pre-trained knowledge to determine factual correctness.

Human Evaluation. We invite three graduate researchers to evaluate model outputs on AmbigQA and HotpotQA. Due to budget constraints, we limited the human evaluation to two datasets. Annotators were presented with input questions, corresponding reference answers, and anonymized model responses in a randomized order to prevent position or model identity bias. Each response is evaluated using a binary scoring system: 1 ("True") for responses that accurately aligned with reference answers and demonstrated contextual relevance, and 0 ("False") for responses that deviated from these criteria. The majority vote determines the final judgment (see Appendix 8.3.3).

3.5 Evaluation Metrics

To assess SAGE's performance, we use ***Accuracy*** as the proportion of instances where the judge's binary verdict aligns with the ground truth derived from reference answers through automatic metrics and RefGPT. We also utilize ***Macro-F1*** to assess judges' agreement with reference-based metrics. For AmbigQA and HotpotQA that include human annotations, we calculate ***Cohen's Kappa*** ***(*** κ ***)*** and ***Macro-F1*** scores to evaluate SAGE's alignment with human majority votes. Additionally, we conduct ablation studies to quantify the ***impact of specific SAGE components***, using changes in agreement with human judgments.

4 Results

As mentioned earlier, given the high agreement of RefGPT to human evaluation (Table 1), we consider it as a silver human evaluation to compare the results of SAGE with other baselines. For additional results and analysis, we refer the readers to Appendices 9 and 10.

4.1 Main results

By integrating external evidence into pre-trained knowledge, SAGE outperforms judges without external evidence. In Table 2, SAGE shows consistently higher agreement with RefGPT compared to judges without access to references. For instance, GPT-3.5, as a judge within SAGE, shows an accuracy of 0.80 when evaluating itself on AmbigQA. In contrast, the same model without reference support achieves only 0.67, indicating lower agreement with GPT-4 reference-based judgments. Similar gains are observed for GPT-4o and Gemini, demonstrating that external evidence improves both precision and recall.

SAGE strongly agrees with human evaluations To evaluate correlation with human judgment, we compared SAGE and baseline methods using majority vote annotations from three expert annotators on AmbigQA and HotpotQA. As depicted in Table 3, judges without access to reference answers rely on their pre-trained knowledge, which often confirms the candidate's answer as correct. As a result, their agreement with human annotations is low, with Cohen's kappa scores often below 0.40. In contrast, SAGE achieves substantially higher agreement with human annotations. For instance, when GPT-4o evaluates itself without reference yields a κ of only 0.38 on HotpotQA, while the same judge model under SAGE reaches 0.70.

Cohen's κ measures agreement beyond chance but can mislead under class imbalance, known as the *kappa paradox* (Cicchetti and Feinstein, 1990). Therefore, we report Macro F1, which treats both classes equally and provides a balanced view of evaluation performance. In Table 3, LLM-as-a-judge without access to reference answers shows competitive macro F1 scores, but analysis reveals a tendency to over-estimate correctness, leading to inflated recall at the expense of precision (see Figure 2). In contrast, SAGE delivers the highest Macro F1 across models and tasks.

Cand.	Task	Judge without search			SAGE		
		GPT-3.5	GPT-4o	Gemini	GPT-3.5	GPT-4o	Gemini
GPT-3.5	AmbigQA	0.67	0.73	0.81	0.80	**0.83**	0.81
	FreshQA	0.51	0.70	0.83	0.79	**0.91**	0.89
	HotpotQA	0.58	0.64	0.66	0.70	**0.76**	0.73
	NQ-Open	0.61	0.70	0.70	0.70	0.72	**0.74**
	TriviaQA	0.80	0.85	0.84	0.84	**0.89**	0.82
GPT-4o	AmbigQA	0.70	0.70	0.79	0.80	**0.83**	**0.83**
	FreshQA	0.54	0.59	0.68	0.76	0.78	**0.81**
	HotpotQA	0.57	0.63	0.62	0.70	**0.77**	**0.77**
	NQ-Open	0.59	0.65	0.71	0.71	0.74	**0.75**
	TriviaQA	0.82	**0.86**	0.81	0.84	**0.86**	0.80
Gemini	AmbigQA	0.68	0.72	0.70	0.75	**0.83**	0.76
	FreshQA	0.64	0.64	0.65	0.73	0.75	**0.76**
	HotpotQA	0.61	0.63	0.61	0.75	**0.76**	0.75
	NQ-Open	0.61	0.62	0.61	0.64	**0.72**	0.67
	TriviaQA	0.81	0.82	0.79	0.83	**0.85**	0.80

Table 2: Judge without search and SAGE agreement with RefGPT across candidates (cand.) and tasks. The performance is measured as the accuracy of their agreement with reference-based RefGPT judgments. Higher scores indicate better agreement with RefGPT.

SAGE works better with larger models. SAGE's performance improves significantly when using more capable judge models (i.e., based on their leaderboard performance). In Table 3, GPT-4o outperforms GPT-3.5 and Gemini across both AmbigQA and HotpotQA. For instance, GPT-4o reaches a macro F1 of 0.90 on AmbigQA when judging GPT-3.5, higher than GPT-3.5's 0.78 and Gemini's 0.83.

SAGE can detect untruthful facts and outdated knowledge. SAGE excels at identifying false facts by cross-referencing claims with external evidence. For example, when evaluating the question *"Who sings the theme song for the show Half & Half?"*, a candidate model incorrectly answered *"Erica Campbell."* LLM judges without tools falsely evaluated the answer as correct. However, SAGE retrieved accurate evidence confirming that *"Melonie Daniels"* performed the theme song. Using this information, SAGE correctly rejected the candidate's response, providing a clear rationale grounded in the evidence.

Candid.	Task	Judge without search						SAGE					
		GPT-3.5		GPT-4o		Gemini		GPT-3.5		GPT-4o		Gemini	
		κ	Mac-F1	κ	Mac-F1	κ	Mac-F1	κ	Mac-F1	κ	Mac-F1	κ	Mac-F1
GPT-3.5	AmbigQA	0.23	0.61	0.39	0.70	0.58	0.79	0.57	0.78	**0.80**	**0.90**	0.66	0.83
	HotpotQA	0.16	0.53	0.26	0.61	0.35	0.67	0.41	0.71	0.56	**0.78**	0.53	0.76
GPT-4o	AmbigQA	0.24	0.59	0.38	0.68	0.57	0.78	0.60	0.80	**0.91**	**0.96**	**0.91**	**0.96**
	HotpotQA	0.17	0.53	0.38	0.67	0.36	0.68	0.54	0.77	0.70	**0.85**	0.68	0.84
Gemini	AmbigQA	0.20	0.58	0.35	0.67	0.26	0.61	0.64	0.82	0.75	**0.87**	0.67	0.84
	HotpotQA	0.17	0.55	0.27	0.62	0.27	0.62	0.63	0.82	0.69	**0.85**	0.59	0.80

Table 3: Cohen's κ and Macro F1 between human majority and reference-free methods.

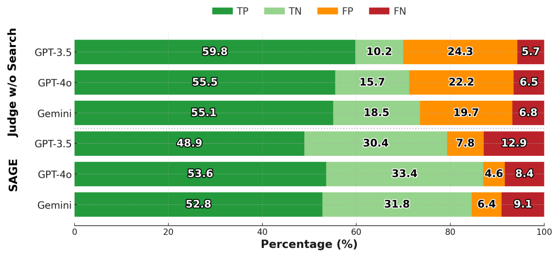

Figure 2: Distribution of true positive (TP), true negative (TN), false positive (FP), and false negative (FN) rates for each LLM judge in both Judge without search and SAGE, averaged across AmbigQA and HotpotQA and candidate models.

On outdated knowledge, the FreshQA (Vu et al., 2023) results show that SAGE consistently identified outdated information in candidate outputs. For instance, when asked *"Where is EMNLP this year?"*, candidate models often provided outdated responses based on their training data. SAGE retrieved current information indicating the correct location, *"Suzhou, China."* (see Appendix 9.4).

SAGE fixes incorrect reasoning traces. We analyzed cases where candidate models produced logically inconsistent or unsupported reasoning. SAGE's reflection enables it to detect these inconsistencies (see Table 10 in the Appendix).

4.2 Error analysis

To better understand the limitations of SAGE in evaluating candidate responses, we conducted a manual error analysis. We randomly sampled 100 evaluation cases from the AmbigQA and HotpotQA datasets, focusing on instances where SAGE disagreed with human annotators. We categorized the errors into the following categories: 1) **Contextual misunderstanding (23%):** SAGE generates inaccurate or incomplete queries when it misinterprets the intent of the candidate's question. This is particularly evident in AmbigQA, where questions are often intentionally ambiguous or lack sufficient context, leading to the retrieval of irrelevant or contradictory evidence. Although SAGE's components help clarify intent by reformulating the context, challenges remain when the question itself is open to multiple interpretations. 2) **Incomplete evidence (15%):** SAGE fails when the retrieved evidence is insufficient or lacks relevant information, specifically for recent events with limited online coverage. 3) **Reasoning error (32%):** Despite accurate evidence, the judge model misinterprets the information or applies flawed reasoning. 4) **Hallucination (13%):** In cases where evidence is ambiguous or inconclusive, SAGE relies on its pre-trained knowledge, resulting in hallucinated rationales. 5) **Conflicting evidence (7%):** In some cases, SAGE encounters conflicting evidence across multiple search iterations. The framework is designed to iteratively refine its understanding; however, judges sometimes over-rely on earlier sources or fail to appropriately weigh the of conflicting information (details in Appendix 10.3).

4.3 Ablation study

Effect of iterations Figure 3 shows that effect of number of iterations on SAGE's performance. We observe a consistent increase in performance over zero iterations. Iteration 3, which serves as our default configuration, consistently provides the best trade-off between performance and efficiency. We observed a slight drop in performance of some models in the case of 2 and 4 iterations. In the case of 2 iterations, the drop was due to occasional off-topic queries during the refinement stage which is later fixed in the subsequent iteration. The decline in performance at iteration 4 across all models was due to the overabundance of sources, resulting in the accumulation of redundant or irrelevant information, increased context length, and potential

Candidate	Task	Judge w/o Query	SAGE
GPT-3.5	AmbigQA	0.84	0.90
	HotpotQA	0.76	0.78
GPT-4o	AmbigQA	0.84	0.96
	HotpotQA	0.82	0.85
Gemini	AmbigQA	0.85	0.87
	HotpotQA	0.80	0.85

Table 4: Macro F1 scores between Judge w/o Query and SAGE using GPT-4o as the judge. Note that directly using the input question without generating refined queries is still considered a form of querying, but for clarity, we refer to this setting as w/o Query.

Category	Init. Query	2. Query	3. Query
Sufficient (On-topic)	77	69	70
Partial sufficient	15	25	22
Insufficient (Off-topic)	8	6	7

Table 5: Topic drift across 300 queries (100 inst × 3).

model confusion. Nevertheless, results using any reasonable number of iterations show a consistent improvement over not using any iterations.

Effect of query generation To evaluate the importance of the query generation module in SAGE, we remove that step and directly use the input question for evidence retrieval. Table 4 shows that SAGE consistently outperforms the query-free baseline, demonstrates the importance of the query generation step. The most notable improvements are observed on AmbigQA, where SAGE achieves a Macro F1 of 0.96 compared to 0.84 without query making. In HotpotQA, while the performance gain is smaller, SAGE still demonstrates clear advantages, particularly due to its ability to adaptively generate focused queries that facilitate multi-hop reasoning.

Robustness of query refinement We examined whether SAGE's query refinement introduces topic drift or irrelevant queries by manually analyzing 100 instances from our error analysis. Since our SAGE configuration involves up to three iterations, this requires analyzing the queries at each stage. Table 5 shows that initial queries generated directly from the input question are generally on-topic and relevant to the core question being asked. However, due to the ambiguity of some questions (e.g., AmbigQA), we observe some off-topic queries in this stage. Topic drift, though not the most dominant error, did occur during the refinement stages (iterations 2 and 3). Out of the 200 refined queries analyzed in the refinement stages, only 6.5% of queries were judged "insufficient."

Impact of iterative evidence gathering To isolate the value of iterative retrieval in SAGE, we introduce a minimal search-augmented baseline that performs only a single-pass retrieval. This baseline issues a web search using the input question and collects the top-3 snippets without further refinement. The judge model then evaluates the candidate's answer based on this context.

On AmbigQA, this naive baseline raises judge-without-search κ in Table 3 from 0.38 to 0.65, confirming that one round of external evidence already yields a substantial improvement. However, SAGE's pushes κ further to 0.91. A similar pattern appears on HotpotQA, where κ climbs from 0.38 (without reference) to 0.58 (single-pass) and then to 0.70 with SAGE. All configurations use GPT-4o as the candidate and judge.

4.4 Cost analysis

With three iterations, SAGE issues 3 Serper queries and processes 4,410 input and 989 output tokens per judgement. At GPT-4o-mini[2] and Serper rates[3], the per-instance cost is

$$\underbrace{4{,}410 \times \frac{\$0.15}{10^6} + 989 \times \frac{\$0.60}{10^6}}_{\text{LLM}=\$0.00126} + \underbrace{3 \times \frac{\$0.30}{10^3}}_{\text{Search}=\$0.00090} = \$\mathbf{0.00216}.$$

Evaluating 300 AmbigQA instances, therefore, costs \$0.65 and finishes in 2.3 h (27s/instance). A single annotator, given gold answers and averaging 1 min/instance, needs 5 h and \$75 (\$15h^{-1}). Thus SAGE is $\approx 2.2\times$ faster and $\approx 115\times$ cheaper, while matching three humans-level agreement ($\kappa \approx 0.91$). Note that throughout this cost analysis, GPT-4o-mini serves as both the candidate and the judge on AmbigQA.

5 Related Work

Evaluating LLMs is a critical yet challenging aspect of modern NLP research. We review existing approaches across the following categories.

Free-form QA. It is a valuable benchmark for ensuring the factuality of LLMs (Wang et al., 2023a). This type of task is traditionally evaluated through automatic metrics that rely on comparing model outputs against expert-annotated reference answers using metrics such as EM and F1 (Gou et al., 2024). While efficient, such methods cannot capture the

[2] \$0.15 per 10^6 input tokens; \$0.60 per 10^6 output tokens.

[3] \$0.30 per 10^3 queries.

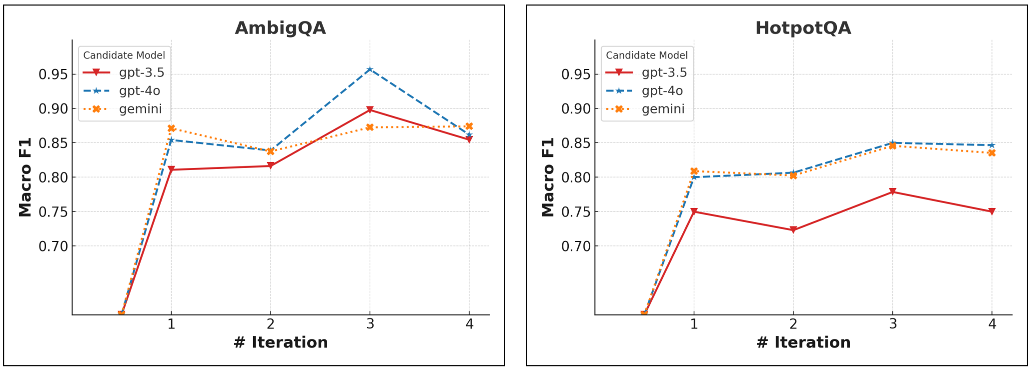

Figure 3: The effect of iterations. GPT-4o is used as a judge here.

diversity of responses, require costly reference annotations, and fail to adapt to evolving factual information (Kamalloo et al., 2023).

Reference-based LLM judge. Recent work has attempted to address such limitations by utilizing LLMs for evaluation. Specifically, given the model answer for a question, an LLM is prompted with the original question, the candidate answer, and the dataset reference answer to evaluate the correctness of the model response (Wang et al., 2023a; Kamalloo et al., 2023). This approach often returns a verdict in the form of a categorical label or a scalar score. Recent methods, for instance, PoLL (Verga et al., 2024) follow a similar template but utilize multiple LLMs for more reliable evaluations.

Reference-free LLM judge. To avoid the need for reference answers to improve scalability, subsequent work explored reference-free LLM judges (Zheng et al., 2023). G-Eval (Liu et al., 2023) implements direct evaluation by prompting models to assess outputs based on predefined criteria. Other methods include pairwise comparisons (Zheng et al., 2023), debate-style frameworks (Khan et al., 2024), and ensemble approaches (Zhang et al., 2024). These methods have demonstrated success in subjective evaluation tasks such as summarization or dialogue generation, where human preferences rather than factual correctness are the primary concern. However, for objective correctness, reference-free LLM judges often struggle with reliability (Badshah and Sajjad, 2025; Kim et al., 2024), because although we can provide them with detailed instructions at inference time, their factual grounding still depends entirely on the parametric knowledge encoded in their pretrained weights and thus inherits its limitations.

LLMs evaluation with search-augmentation. An emerging category attempts to overcome the limitations of LLM evaluators by incorporating external tools. More closely related to our work, FActScore (Min et al., 2023) decomposes long-form generated text into atomic facts and verifies each against Wikipedia pages. Similarly, SAFE (Wei et al., 2024) uses an LLM to split long-form responses into individual facts, issue a Google Search query for each, and reason about relevance. In contrast to these methods, which target long-form outputs and rely on splitting text into atomic facts, our focus is short-form responses, where the challenge is less about exhaustive claim coverage and more about precise, reference-free evaluation under uncertainty. Unlike prior search-augmented evaluators that typically issue a single-pass query, our method iteratively conducts output-specific searching, summarization, reflection, and refinement. This added iteration increases cost relative to a single pass but provides greater reliability, particularly in ambiguous cases.

6 Conclusion

We presented SAGE, a framework for evaluating LLMs that integrates external evidence through an iterative process. Our experiments demonstrate that SAGE achieves substantial to perfect agreement with human evaluations. SAGE offers interpretable, evidence-aware evaluations, making it a reliable alternative to evaluate free-form QA. In the future, we aim to reduce computational cost by using conformal prediction to invoke external tool calls only when the model is uncertain in its judgment. We also plan to extend SAGE to visual QA and code evaluation by incorporating additional tools. We believe SAGE can serve as a practical foundation for LLM evaluation in real-world applications.

7 Limitations

Although SAGE significantly improves the reliability and accuracy of LLM evaluation in reference-free settings, it has several limitations. Addressing these limitations will enhance SAGE's robustness, scalability, and effectiveness, making it more valuable for evaluating LLM outputs in diverse, real-world scenarios.

Context window constraints. While our short-term memory strategy reduces token input, limitations inherent in current LLM context windows still pose challenges, especially for longer evaluation sequences. SAGE short-term memory provides a persistent record of the evaluation process. While this enhances interpretability and traceability, it remains constrained by the model's ability to process information within its context window during subsequent reasoning. Future work could explore integrating a long-term memory with a recall-based mechanism that selectively retrieves relevant traces from past evaluations, such as episodic and semantic memory (Park et al., 2023). This selective calling of traces will also reduce computational cost.

Source bias and quality control. Although SAGE's reflection module actively identifies inconsistencies across multiple evidence sources, our current implementation does not explicitly quantify source credibility. SAGE may inherit biases from external data sources (Zhan et al., 2024; Li et al., 2025a), which can cause further biases, such as sycophancy bias (Sharma et al., 2023), where the model aligns with false claims instead of critically assessing them. To mitigate the impact of source biases and misinformation, future work should prioritize the integration of credibility scoring or weighting mechanisms that dynamically assess evidence reliability.

Dependency on Judge LLM capabilities. SAGE's performance depends on the capabilities of the underlying judge model. Our experiments show a noticeable drop in performance when using smaller models such as Mistral 7B (see Appendix 9.3). Achieving near-perfect agreement requires stronger, more capable LLMs.

Diminishing returns with iterative refinement. Our empirical analysis shows that performance gains plateau and sometimes degrade beyond a certain number of iterative evidence-gathering steps due to redundant information and increased strain on the LLM context window. Although we currently select an optimal iteration count empirically, future research can develop adaptive stopping criteria, leveraging real-time evidence sufficiency metrics to halt iterations dynamically and efficiently.

References

Sher Badshah and Hassan Sajjad. 2024a. Quantifying the capabilities of llms across scale and precision. *Preprint*, arXiv:2405.03146.

Sher Badshah and Hassan Sajjad. 2024b. Reference-guided verdict: Llms-as-judges in automatic evaluation of free-form text. *arXiv preprint arXiv:2408.09235*.

Sher Badshah and Hassan Sajjad. 2025. DAFE: LLM-Based Evaluation Through Dynamic Arbitration for Free-Form Question-Answering. *Preprint*, arXiv:2503.08542.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yinuo Liu, Yaochen Wang, Huichi Zhou, Qihui Zhang, Pan Zhou, Yao Wan, and Lichao Sun. 2024. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. *arXiv preprint arXiv:2402.04788*.

Jeffrey Cheng, Marc Marone, Orion Weller, Dawn Lawrie, Daniel Khashabi, and Benjamin Van Durme. 2024. Dated data: Tracing knowledge cutoffs in large language models. *Preprint*, arXiv:2403.12958.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada. Association for Computational Linguistics.

Domenic V. Cicchetti and Alvan R. Feinstein. 1990. High agreement but low kappa: Ii. resolving the paradoxes. *Journal of Clinical Epidemiology*, 43(6):551–558.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. 2024. Critic: Large language models can self-correct with tool-interactive critiquing. *Preprint*, arXiv:2305.11738.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. A survey on llm-as-a-judge. *Preprint*, arXiv:2411.15594.

Rishav Hada, Varun Gumma, Adrian de Wynter, Harshita Diddee, Mohamed Ahmed, Monojit Choudhury, Kalika Bali, and Sunayana Sitaram. 2024. Are large language model-based evaluators the solution to scaling up multilingual evaluation? *Preprint*, arXiv:2309.07462.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *Preprint*, arXiv:1705.03551.

Ehsan Kamalloo, Nouha Dziri, Charles Clarke, and Davood Rafiei. 2023. Evaluating open-domain question answering in the era of large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5591–5606, Toronto, Canada. Association for Computational Linguistics.

Akbir Khan, John Hughes, Dan Valentine, Laura Ruis, Kshitij Sachan, Ansh Radhakrishnan, Edward Grefenstette, Samuel R Bowman, Tim Rocktäschel, and Ethan Perez. 2024. Debating with more persuasive llms leads to more truthful answers. In *Proceedings of the 41st International Conference on Machine Learning*, pages 23662–23733.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024. Prometheus 2: An open source language model specialized in evaluating other language models. *Preprint*, arXiv:2405.01535.

Michael Krumdick, Charles Lovering, Varshini Reddy, Seth Ebner, and Chris Tanner. 2025. No free labels: Limitations of llm-as-a-judge without human grounding. *Preprint*, arXiv:2503.05061.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Ang Li, Yin Zhou, Vethavikashini Chithrra Raghuram, Tom Goldstein, and Micah Goldblum. 2025a. Commercial llm agents are already vulnerable to simple yet dangerous attacks. *Preprint*, arXiv:2502.08586.

Jiatong Li, Rui Li, Yan Zhuang, Kai Zhang, Linan Yue, Qingchuan Li, Junzhe Jiang, Qi Liu, and Enhong Chen. 2024. Dynaeval: A dynamic interaction-based evaluation framework for assessing LLMs in real-world scenarios.

Qingquan Li, Shaoyu Dou, Kailai Shao, Chao Chen, and Haixiang Hu. 2025b. Evaluating scoring bias in llm-as-a-judge. *Preprint*, arXiv:2506.22316.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522.

Oscar Mañas, Benno Krojer, and Aishwarya Agrawal. 2024. Improving automatic vqa evaluation using large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 4171–4179.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. 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*, pages 12076–12100.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. AmbigQA: Answering ambiguous open-domain questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5783–5797, Online. Association for Computational Linguistics.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. *Preprint*, arXiv:2304.03442.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. 2023. Towards understanding sycophancy in language models. *Preprint*, arXiv:2310.13548.

Guijin Son, Hyunwoo Ko, Hoyoung Lee, Yewon Kim, and Seunghyeok Hong. 2024. Llm-as-a-judge & reward model: What they can and cannot do. *Preprint*, arXiv:2409.11239.

Gemini Team. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

OpenAI Team. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Aman Singh Thakur, Kartik Choudhary, Venkat Srinik Ramayapally, Sankaran Vaidyanathan, and Dieuwke Hupkes. 2025. Judging the judges: Evaluating alignment and vulnerabilities in llms-as-judges. *Preprint*, arXiv:2406.12624.

Pat Verga, Sebastian Hofstatter, Sophia Althammer, Yixuan Su, Aleksandra Piktus, Arkady Arkhangorodsky, Minjie Xu, Naomi White, and Patrick Lewis. 2024. Replacing judges with juries: Evaluating llm generations with a panel of diverse models. *Preprint*, arXiv:2404.18796.

Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, and Thang Luong. 2023. Freshllms: Refreshing large language models with search engine augmentation. *Preprint*, arXiv:2310.03214.

Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, and Thang Luong. 2024. FreshLLMs: Refreshing large language models with search engine augmentation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13697–13720, Bangkok, Thailand. Association for Computational Linguistics.

Cunxiang Wang, Sirui Cheng, Qipeng Guo, Yuanhao Yue, Bowen Ding, Zhikun Xu, Yidong Wang, Xiangkun Hu, Zheng Zhang, and Yue Zhang. 2023a. Evaluating open-QA evaluation. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023b. Self-consistency improves chain of thought reasoning in language models. *Preprint*, arXiv:2203.11171.

Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Jie Huang, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, and 1 others. 2024. Long-form factuality in large language models. *Advances in Neural Information Processing Systems*, 37:80756–80827.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *Preprint*, arXiv:1809.09600.

Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, Nitesh V Chawla, and Xiangliang Zhang. 2024. Justice or prejudice? quantifying biases in llm-as-a-judge. *Preprint*, arXiv:2410.02736.

Asaf Yehudai, Lilach Eden, Alan Li, Guy Uziel, Yilun Zhao, Roy Bar-Haim, Arman Cohan, and Michal Shmueli-Scheuer. 2025. Survey on evaluation of llm-based agents. *Preprint*, arXiv:2503.16416.

Qiusi Zhan, Zhixiang Liang, Zifan Ying, and Daniel Kang. 2024. Injecagent: Benchmarking indirect prompt injections in tool-integrated large language model agents. *Preprint*, arXiv:2403.02691.

Xiaoyu Zhang, Yishan Li, Jiayin Wang, Bowen Sun, Weizhi Ma, Peijie Sun, and Min Zhang. 2024. Large language models as evaluators for recommendation explanations. In *Proceedings of the 18th ACM Conference on Recommender Systems*, pages 33–42.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2024. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Lianghui Zhu, Xinggang Wang, and Xinlong Wang. 2023. Judgelm: Fine-tuned large language models are scalable judges. *arXiv preprint arXiv:2310.17631*.

8 Experimental detail

8.1 Datasets

We evaluate SAGE on widely used free-form question-answering datasets that represent different question types, knowledge domains, and complexity levels. Free-form question-answering underpins a broad range of practical applications, where maintaining accuracy and ensuring truthfulness are paramount (Gou et al., 2024). Evaluating large-scale datasets can be costly; therefore, we randomly sample 300 instances from each dataset, ensuring a balanced representation across question types and difficulty levels. This sampling strategy provides a fair evaluation while maintaining computational feasibility. Each dataset reference answers serve as ground truth for our reference-based baseline metrics, allowing us to compare the performance of SAGE against established evaluation approaches. Our selected datasets are:

AmbigQA (Min et al., 2020) Contains questions with multiple valid answers due to inherent ambiguities, challenging evaluators to consider multiple interpretations.

HotpotQA (Yang et al., 2018) Features multi-hop reasoning questions that require synthesizing information from multiple sources.

Natural Questions (NQ-Open) (Kwiatkowski et al., 2019) Consists of real user queries from Google Search, representing naturally occurring information needs.

TriviaQA (Joshi et al., 2017) Includes trivia questions from various domains, testing breadth of knowledge and factual recall.

FreshQA (Vu et al., 2023) Contains questions about recent events occurring after most LLMs' training cutoff, specifically designed to test knowledge updating capabilities.
In the above datasets, we leverage their respective validation splits for evaluation: the standard validation sets for AmbigQA and Natural Questions, the validation split of the distractor subset for HotpotQA, and the unfiltered.nocontext validation subset for TriviaQA. For FreshQA, we adopt the version released on December 18, 2024.

8.2 Prompting

We employ a template-based prompting strategy for both response generation and evaluation. For candidate models, we utilize few-shot Chain-of-Thought (CoT) prompts (Gou et al., 2024), incorporating 6 examples per dataset to encourage detailed, reasoning-driven, and structured responses (see Figure 4).

SAGE prompts candidate LLMs using few-shot CoT reasoning to elicit faithful and interpretable outputs across each module. Each module in the evaluation pipeline is guided by a carefully constructed prompt, often in a one-shot format, that combines role-playing instructions with explicit reasoning goals. Below, we describe the prompting strategy for each component.

Query Generation. The query generation module converts an input question $x \in \mathcal{X}$ into an initial search query without referencing the candidate answer. The prompt instructs the model to reflect step-by-step on the most relevant aspects and keywords before proposing a final query.

Evidence Summarization. To reduce raw search results $S(q_i)$, the summarization module uses a Chain-of-Thought (CoT) prompt that walks the model through evaluating and synthesizing relevant content. The prompt emphasizes factual grounding and asks the model to avoid repetition and speculation.

Iterative Reflection. The reflection module analyzes the current evidence summary E_i in relation to the input question x and candidate answer $\hat{y}$. The prompt guides the model to assess whether the evidence supports, contradicts, or is inconclusive with respect to the answer, and highlights missing information.

Query Refinement. To improve the evidence retrieved in future steps, the query refinement module generates a new query by analyzing the short-term memory contents—specifically, the previous query q_i, evidence E_i, and reflection R_i. The Chain-of-Thought (CoT) prompt instructs the model to identify remaining uncertainties or gaps and generate a refined, more targeted query.

Judgment. Finally, the judgment module evaluates whether the candidate's answer $\hat{y}$ is factually correct based on the accumulated external evidence.

The Chain-of-Thought (CoT) prompt instructs the model to reason step-by-step using the evidence and produce a binary decision, True/False decision, along with rationale.

In cases where the evidence is insufficient or contradictory, the prompt explicitly instructs the model to either defer to its prior knowledge or explain uncertainty. The output is formatted as a JSON object with keys `"decision"` and `"explanation."` All modules are executed within a single LLM agent under a unified prompting interface.

8.3 Baselines

We compare SAGE against several established evaluation approaches:

8.3.1 Reference-Based Metrics.

We implement two widely-used automatic metrics that rely on comparison with dataset-specific reference answers:

- **Exact Match (EM)** measures whether the model's answer exactly matches any of the reference answers after normalization.

Candidate Answer Generation (Few-shot CoT)

Instructions: Answer each question step by step and conclude with: “So the answer is: [Answer].”

Q: What is the elevation range for the area that the eastern sector of the Colorado orogeny extends into? **A:** The eastern sector of the Colorado orogeny extends into the High Plains, which rise from around 1,800 to 7,000 ft. **So the answer is:** 1,800–7,000 ft.

Q: Musician and satirist Allie Goertz wrote a song about the "The Simpsons" character Milhouse, who Matt Groening named after who? **A:** Milhouse was named after U.S. president Richard Nixon. **So the answer is:** Richard Nixon.

Q: Which documentary is about Finnish rock groups, *Adam Clayton Powell* or *The Saimaa Gesture*? **A:** *Adam Clayton Powell* is about an African-American politician, so the documentary about Finnish rock groups must be *The Saimaa Gesture*. **So the answer is:** The Saimaa Gesture.

Q: What profession do Nicholas Ray and Elia Kazan have in common? **A:** Both were directors, screenwriters, and actors. **So the answer is:** director, screenwriter, actor.

Q: Which magazine was started first, *Arthur’s Magazine* or *First for Women*? **A:** *Arthur’s Magazine* (1844) was started before *First for Women* (1989). **So the answer is:** Arthur’s Magazine.

Q: Were Pavel Urysohn and Leonid Levin known for the same type of work? **A:** Both were mathematicians (Levin is also a computer scientist). **So the answer is:** Yes.

Figure 4: Examples of few-shot CoT (Gou et al., 2024) prompts for candidate answer generation.

- **F1 Score** computes the harmonic mean of precision and recall between the token sets of the model’s answer and the references, providing a softer measure of overlap.

- **RefGPT** compares the model output with the dataset reference answers. This method provides a context-aware evaluation beyond strict token-level matching (Badshah and Sajjad, 2024b).

8.3.2 Judge without Tool-Augmentation

Following the approach from Liu et al. (2023), we implement a reference-free baseline where the judge LLM evaluates candidate answers based solely on the question-answer pair, without access to external tools or reference answers. The judge relies entirely on its pre-trained knowledge to determine factual correctness. This baseline isolates the impact of tool augmentation in SAGE by maintaining the same judge model while removing the evidence retrieval mechanism.

8.3.3 Human Evaluation

We recruited three volunteer graduate researchers from our lab with expertise in natural language processing to evaluate model outputs on AmbigQA and HotpotQA. Annotators were presented with input questions, corresponding reference answers, and anonymized model responses in a randomized order to prevent position or model identity bias. Each response was evaluated using a binary scoring system: **1 (“True”)** for responses that accurately aligned with the reference answers and demonstrated contextual relevance, and **0 (“False”)** for responses that deviated from these criteria.

Evaluation Rationale Due to budget and resource constraints, we focused our human evaluation on AmbigQA and HotpotQA. These datasets were chosen because they represent challenging real-world scenarios involving multi-hop reasoning and ambiguous question-answering, making them ideal for assessing SAGE’s effectiveness. Evaluating additional datasets would have significantly increased the time and cost of human annotations.

Furthermore, we evaluated 300 randomly sampled instances from each dataset, resulting in 600 samples per model across the two tasks. With three candidate models, the total number of samples evaluated was resulted in 1,800. Conducting large-scale human evaluation beyond this would incur substantial annotation costs and additional cognitive load on annotators. By limiting the sample size, we maintained a balance between evaluation comprehensiveness and resource efficiency.

Evaluation Guidelines To ensure consistent assessments, annotators followed the guidelines inspired by established evaluation protocols. Annotators were instructed to evaluate responses based on the following principles:

- **Semantic equivalence:** A response is marked **True** if it conveys the same core information as the reference answer, even if phrased differently using synonyms, paraphrasing, or structural variations. Additional contextual information is acceptable as long as it is factually

Query generation

Your goal is to generate a targeted web search query.
Before producing the final query, carefully consider:

1. The question's key concepts or keywords (e.g., important names, dates).
2. Whether the question might be ambiguous or reference multiple possible answers (e.g., a book with the same title by different authors, or a modern text about a historical figure).

Question: {question}
Return your response as a JSON object with ALL three exact keys:

- `"query"`: The search query string.
- `"aspect"`: The specific aspect of the question to focus on.
- `"rationale"`: A brief explanation of why this query is relevant, including your chain-of-thought reasoning.

Example Output:

```
{
  "query": "Apollo 11 moon landing year + NASA + 1969",
  "aspect": "historical event",
  "rationale": "The question asks about Apollo 11's landing year,
                so I'm including NASA, year, and 1969 to get relevant info."
}
```

Figure 5: Prompt used for initial query generation, guiding the model to produce focused and relevant search queries.

correct and does not alter the original meaning.

- **Factual Accuracy:** Responses that contain factual errors, omit essential information, or introduce misleading content are marked **False**. If a response partially answers the question but excludes critical elements, it is considered incorrect.
- **Multiple Reference Answers:** In cases with multiple reference answers, a response is deemed correct if it is fully aligned with at least one reference.
- **Fact-Checking:** Annotators are allowed to consult external resources, such as search engines or online encyclopedias, to verify specific facts when uncertain. However, the reference answers served as the primary benchmark for correctness.
- **Documenting Ambiguity:** Annotators are encouraged to document cases where the evaluation is uncertain or requires further clarification. These cases were discussed collaboratively to ensure consensus.

By adhering to these guidelines, we ensured reliable and consistent human evaluations.

Inter Human Annotator Agreement We calculated **Fleiss' Kappa (**κ**)** and percent agreement to measure inter-annotator agreement. Fleiss' Kappa is defined as:

$$\kappa = \frac{\bar{P} - P_e}{1 - P_e},$$

where $\bar{P}$ is the average observed agreement among annotators, and P_e is the expected agreement by chance. Percent agreement (PA) is calculated as:

$$\text{PA} = \frac{\text{N.Agreements}}{\text{Total N.Annotations}} \times 100.$$

8.4 Evaluation Metrics

To assess SAGE's performance, we use multiple evaluation metrics:

Accuracy: We measure the proportion of instances where the judge's binary verdict (correct/incorrect) aligns with the ground truth derived from reference answers.

Agreement with Human Judgment: For the AmbigQA and HotpotQA subsets with human annotations, we calculate Cohen's Kappa (κ), majority voting, and Macro-F1 scores to assess agreement between SAGE's verdicts and human majority votes. These metrics were chosen because they account for both agreement beyond chance

Evidence summarization

You are a summarization assistant. Carefully review the raw search results and provide a concise summary of the key information relevant to the question.

Raw Search Results: {raw_results}

Return your summary as plain text:

- Keep it neutral and focused on the question.
- If results conflict, mention that briefly.
- Do not add extra commentary.

Example Output (Plain Text):

```
"Result 1 says X about the event date,
 Result 2 says Y but doesn't mention the exact date.
 Overall, it references 1969."
```

Figure 6: Prompt for evidence summarization, guiding the model to generate a concise, unbiased summary from raw search results.

(κ) and class balance (Macro-F1).

Cohen's Kappa: Cohen's Kappa measures the agreement between two annotators while correcting for chance agreement. It is defined as:

$$\kappa = \frac{P_o - P_e}{1 - P_e},$$

where P_o is the observed agreement, and P_e is the expected agreement by chance.

Majority Voting: In majority voting, the final decision is determined based on the majority of annotators' labels. Given n annotators and a binary classification, the majority label is defined as:

$$y_{\text{majority}} = \begin{cases} 1 & \text{if } \sum_{i=1}^{n} y_i > \frac{n}{2}, \\ 0 & \text{otherwise,} \end{cases}$$

where y_i represents the label assigned by the ith annotator.

Macro F1 Score: Macro F1 evaluates the balance between precision and recall for each class and averages the results. It is calculated as:

$$\text{Macro-F1} = \frac{1}{C} \sum_{c=1}^{C} \frac{2 \cdot \text{Precision}_c \cdot \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c},$$

where C is the number of classes, and Precision_c and Recall_c are the precision and recall for class c.

9 Additional results

In this section, we included additional results obtained through our experiments.

Task	Model	Percent Agreement (%)	Fleiss' Kappa	Samples
AmbigQA	GPT-3.5	98.3	0.972	300
	GPT-4	98.3	0.976	300
	Gemini	97.0	0.953	300
HotpotQA	GPT-3.5	98.3	0.978	300
	GPT-4	98.3	0.978	300
	Gemini	98.3	0.977	300

Table 6: Human annotator agreement results on AmbigQA and HotpotQA tasks.

9.1 Inter-human annotator agreement

Table 6 presents the human annotator agreement results for the AmbigQA and HotpotQA across three candidate models. The results indicate consistently high agreement among annotators.

9.2 SAGE agreement with reference-based metrics

Our proposed SAGE framework produces raw accuracy scores that closely align with reference-based metrics such as F1 and RefGPT. Unlike instance-level evaluator comparison, the raw accuracies indicate how candidate models are scored by different evaluators. For instance, as given in Table 7, GPT-3.5 acting as a judge within SAGE achieves an accuracy of 0.64 when evaluating its own answers on AmbigQA, which is close to its reference-based F1 of 0.63. In contrast, the same model operating without external evidence (Judge without search) reports a drastically inflated raw accuracy of 0.81, overestimating its own performance. This inflation is consistent across tasks and models. On the other hand, EM often underestimates model quality, as it fails to recognize valid paraphrases and alternative formulations. Overall, SAGE offers a balanced

Example prompt for iterative reflection

You are a research assistant tasked with analyzing the gathered evidence in relation to the question and candidate answer. Think step by step—explain your reasoning and note any gaps or additional details that might be needed. Do not provide a final decision; simply offer your chain-of-thought reflection.

Question: {question}
Candidate Answer: {candidate_answer}
Evidence Summary: {evidence_summary}

Return your response as a JSON object with a single key:

- `"reflection"`: Your chain-of-thought reflection summarizing your analysis.

Example Output:

```
{
  "reflection": "I observed that the evidence overwhelmingly confirms
                 that Apollo 11 landed on the moon in 1969, though there
                 is slight variation in the reported landing times across
                 sources. Additional authoritative sources might help
                 resolve these minor discrepancies."
}
```

Figure 7: Prompt for iterative reflection, instructing the model to analyze the relationship between the question, candidate answer, and evidence summary.

and evidence-aware evaluation, reducing overconfidence while maintaining better alignment with reference metrics.

9.3 SAGE with a small open-source model

We used Mistral 7B to investigate how smaller open-source models perform within our SAGE framework. Specifically, we employed Mistral 7B as a judge in SAGE to evaluate candidate GPT-3.5 on the AmbigQA and HotpotQA datasets. This evaluation provides insights into the effectiveness of smaller models in assessing complex reasoning and factual correctness.

Table 8 illustrates that Mistral 7B, despite its smaller size, demonstrates a reasonable capability in evaluating complex reasoning and factual correctness. On AmbigQA, Mistral 7B achieves a Cohen's Kappa of 0.60 and a Macro F1 of 0.80. On HotpotQA, which involves multi-hop reasoning, its Cohen's Kappa of 0.33 and Macro F1 of 0.67 point challenges in assessing factual accuracy and reasoning depth.

However, a notable limitation of Mistral 7B is its frequent difficulty in following instructions precisely, particularly when handling complex queries that require deep understanding. Additionally, its smaller context window limits its ability to maintain coherence across long reasoning chains. These issues are evident in scenarios where the model fails to properly parse iterative reflection responses or refines search queries incorrectly. Furthermore, Mistral 7B sometimes generates irrelevant reflections or fails to recognize when no supporting evidence is available, leading to errors in judgment. Despite these challenges, Mistral 7B remains a valuable option for resource-constrained environments where efficient evaluation is prioritized over peak accuracy (Badshah and Sajjad, 2024a).

Although Mistral 7B reveals limitations in complex reasoning and instruction following, it still achieves meaningful alignment with human judgments and surpasses Judge without search baselines. Since evaluation typically runs offline, users can select a judge model that balances performance and resource constraints. Furthermore, SAGE's model-agnostic design ensures that it remains compatible with ongoing improvements in smaller open-source LLMs, enhancing its accessibility and long-term utility.

9.4 SAGE can detect untruthful facts and outdated knowledge.

SAGE's iterative evidence-gathering and reflection process enables it to detect untruthful claims and identify outdated information. By continuously refining its search queries and critically evaluating retrieved evidence, SAGE can distinguish between correct and incorrect candidate answers, even when the misinformation is subtle. This capability is particularly valuable in dynamic domains where factual knowledge changes over time.

Table 9 presents an example where a candidate

Example prompt for query refinement

You are a research assistant. Before refining the search query, analyze the existing evidence and reflect on what keywords might be missing or need emphasis. Think step by step and then produce your final refined query.

Question: {question}
Current Search Query: {current_query}
Aggregated Evidence Summary: {evidence_summary}
Iterative Reflection: {iterative_reflection}

If the evidence still does not resolve the question or if there might be an alternative perspective, incorporate additional, more specific keywords to explore those possibilities. For instance:

- Add relevant dates or historical context.
- Use synonyms or alternate phrasings for ambiguous or repeated terms.
- Specify a domain or subject area (e.g., "film," "novel," "historical figure") if it reduces confusion.
- Highlight the location, time period, or any unique aspect not yet included in the current query.

Return your response as a JSON object with ALL three exact keys:

- `"query"`: The refined search query.
- `"aspect"`: The specific aspect being targeted with the refined query.
- `"rationale"`: A brief explanation of your reasoning (chain-of-thought) and why this refinement is needed.

Example Output:

```
{
  "query": "Apollo 11 detailed timeline moon landing 1969",
  "aspect": "chronological sequence",
  "rationale": "The initial query did not specify the temporal progression
                of events. I refined it to target a detailed timeline of
                the Apollo 11 mission in 1969 to capture the sequence of
                key events."
}
```

Figure 8: Prompt for query refinement, guiding the model to analyze evidence and generate more targeted queries.

answer incorrectly claims that the last perfect game in Major League Baseball was thrown by Félix Hernández in 2012. Through iterative search and reflection, SAGE discovers recent evidence confirming that Domingo Germán pitched a perfect game in 2023, successfully identifying the outdated information and concluding that the candidate's answer is incorrect.

We further evaluated SAGE's performance using FreshQA (Vu et al., 2023). In this evaluation, we used GPT-4o as the judge model within SAGE to assess the accuracy of candidate responses generated by GPT-3.5. The results demonstrate that SAGE performed notably well in these contexts, achieving an agent-based raw accuracy of 38.33%, which is significantly closer to reference-based metrics compared to its EM score of 25.00% and F1 score of 35.40%. This highlights SAGE's strength in adapting to evolving information and accurately identifying untruthful or outdated claims. From these raw scores, it is also evident that a pre-trained model like GPT-3.5 often struggles to accurately respond to factual questions in rapidly evolving domains.

9.5 SAGE fixes incorrect reasoning traces.

SAGE's iterative search and reflection process enables it to identify and correct flawed reasoning in candidate answers. Even when a final answer is correct, the candidate's reasoning may contain factual errors. By refining its search queries and critically analyzing the evidence, SAGE can highlight such errors and provide a more accurate rationale.

Table 10 presents an example where the candidate's answer correctly concludes that Sherwood Stewart was born before Javier Frana. However, the reasoning contains a factual inaccuracy, falsely stating Stewart's birth year as 1957 instead of the correct 1946. Through iterations, SAGE gathers

Prompt to the judge model

You are a critical evaluator. You have:
1. The question and the candidate answer,
2. The evidence summary from multiple iterative searches (which may contain overlapping or conflicting information),
3. The chain-of-thought reflection from prior steps,
4. Your own broad knowledge (only if the above are inconclusive).

Follow these guidelines:
- If the summarized evidence and reflections strongly conflict with the candidate answer, conclude `"False"`.
- If the evidence strongly confirms the candidate answer, conclude `"True"`.
- If the evidence is inconclusive or incomplete, but your own knowledge supports the answer, you may conclude `"True"` if confident. Otherwise, conclude `"False"` or state insufficient information.
- When the retrieved evidence is irrelevant, prioritize the chain-of-thought reflections and your own knowledge.

Produce your conclusion in JSON with:
- `"decision"`: `"True"` or `"False"`
- `"explanation"`: A concise reason (including your step-by-step reasoning) describing how you arrived at the verdict.

Input:
Question: {question}
Candidate Answer: {candidate_answer}
Evidence Summary: {evidence_summary}
Reflection: {reflection}

Example Output:

```
{
  "decision": "True",
  "explanation": "The evidence overwhelmingly confirms that Apollo 11
                  landed on the moon in 1969. While minor discrepancies
                  exist in the reported times, they do not undermine the
                  main conclusion. Additional verification is unnecessary."
}
```

Figure 9: Prompt for the judgment step, instructing the model to analyze evidence and reflections to generate a final verdict with justification.

evidence to correct this mistake while maintaining the correct conclusion.

10 Additional ablations and analysis

This section presents three extensions to our study. First, we compare SAGE with three stronger *non-iterative* baselines to evaluate its relative performance. Second, we assess the framework's robustness to prompt variations. Third, we analyze failure cases to better understand the limitations and potential areas for improvement in SAGE.

10.1 Comparison with stronger non-iterative baselines

We evaluate three non-iterative baselines that exclude SAGE's iterative summarise–reflect–refine process but retain the judgment prompt. These methods help disentangle the contributions of search augmentation, sampling-based self-consistency, and model diversity.

Across all baselines, we use GPT-4o to generate the candidate answer. For the *Single-Pass Tool-Augmented Judge* and *Self-Consistency* settings, GPT-4o also serves as the judge model. In the *Multi-LLM Majority Voting* setup, GPT-4o is the candidate model, while three different judge models, including GPT-4o, GPT-3.5, and Mistral 7B, independently evaluate the same input to support diverse model reasoning.

10.1.1 Single-Pass search-augmented judge

To isolate the effect of *web access* alone, we introduce a baseline that issues exactly one web query and performs no further reasoning or iteration. Given a question q and candidate answer $\hat{y}$, the judge LLM formulates a single Serper search, retrieves the top-3 snippets, and immediately produces a TRUE/FALSE verdict with rationale. All prompt instructions are identical to those in the full SAGE pipeline; the *only* change is the removal

Candidate	Task	Reference-based			Judge w/o Search (Acc.)			SAGE (Acc.)		
		EM	F1	RefGPT (Acc.)	GPT-3.5	GPT-4o	Gemini	GPT-3.5	GPT-4o	Gemini
GPT-3.5	AmbigQA	0.50	0.63	0.67	0.81	0.75	0.74	0.64	0.70	0.65
	FreshQA	0.25	0.35	0.35	0.74	0.53	0.39	0.43	0.37	0.34
	HotpotQA	0.34	0.47	0.50	0.86	0.76	0.70	0.50	0.54	0.53
	NQ-Open	0.36	0.53	0.56	0.91	0.83	0.78	0.69	0.70	0.62
	TriviaQA	0.74	0.81	0.81	0.89	0.86	0.82	0.81	0.85	0.78
GPT-4o	AmbigQA	0.47	0.61	0.63	0.88	0.79	0.76	0.63	0.63	0.63
	FreshQA	0.29	0.39	0.45	0.73	0.81	0.65	0.47	0.57	0.53
	HotpotQA	0.34	0.47	0.50	0.86	0.77	0.68	0.50	0.48	0.53
	NQ-Open	0.32	0.48	0.54	0.92	0.87	0.80	0.69	0.67	0.62
	TriviaQA	0.76	0.84	0.80	0.93	0.90	0.86	0.85	0.87	0.80
Gemini	AmbigQA	0.53	0.66	0.67	0.86	0.80	0.85	0.63	0.64	0.67
	FreshQA	0.33	0.44	0.54	0.65	0.82	0.81	0.49	0.54	0.61
	HotpotQA	0.35	0.50	0.53	0.83	0.79	0.75	0.50	0.51	0.55
	NQ-Open	0.36	0.53	0.56	0.91	0.86	0.91	0.73	0.71	0.72
	TriviaQA	0.79	0.86	0.82	0.91	0.92	0.89	0.87	0.88	0.82

Table 7: Raw performance of candidate LLMs obtained through different evaluators.

Candidate	Task	Accuracy	Cohen's Kappa	Macro F1
GPT-3.5	AmbigQA	0.74	0.5910	0.7955
GPT-3.5	HotpotQA	0.62	0.3282	0.6617

Table 8: Mistral 7B as a judge within SAGE evaluating GPT-3.5 answers on AmbigQA and HotpotQA.

of the query-summarize–reflect–refine loop. This baseline, therefore, quantifies the benefit of a one-shot evidence grab, sitting midway between the *Judge without Search-Augmentation* and the full iterative framework.

As shown in Table 11, on AmbigQA, this naive tool baseline raises Cohen's κ from 0.381 (judge without tools) to 0.658, confirming that one round of external evidence already yields a substantial improvement. However, SAGE's pushes κ further to 0.914. A similar pattern appears on HotpotQA, where κ climbs from 0.375 (no tool) to 0.583 (single-pass) and then to 0.701 with SAGE.

10.1.2 Self-consistency judge

For *Self-Consistency* (Wang et al., 2023b), we let the GPT-4o judge sample 10 independent verdicts at temperature 0.7 and take a simple majority. In self-consistency, we excluded access to the reference answer, so the model relies exclusively on its parametric knowledge.

Table 11 shows that self-consistency gives a modest boost over the vanilla judge (no search), but it still lags far behind the search-augmented baselines, confirming that external evidence, especially when gathered iteratively, is critical for reliable objective judgment.

10.1.3 Multi-LLM majority voting

Inspired by PoLL (Verga et al., 2024) and DAFE (Badshah and Sajjad, 2025), we instructed three different LLMs: GPT-4o, GPT-3.5, and Mistral 7B with the same question and candidate answer, asking each to provide a True/False verdict along with a brief explanation. By applying majority voting across the three model outputs, we determined the final answer. This ensemble approach leverages the diverse reasoning patterns of different models, increasing overall robustness and reducing individual model biases. This approach yields $\kappa = 0.609$ on AmbigQA and 0.527 on HotpotQA, outperforming self-consistency but falling short of the single-pass web call ($\kappa = 0.658/0.583$) and significantly below SAGE ($\kappa = 0.914/0.701$).

10.2 Prompt variations

To evaluate SAGE's robustness to prompt design, we re-ran the full framework (with GPT-4o as both candidate and judge) under three prompt variants:

- **Original (1-shot CoT):** Our default configuration, which includes detailed instructions and a single CoT example per module.

- **Few-shot CoT:** An extended version of the prompt containing three CoT examples instead of one.

Component	Content
Question	When was the last perfect game thrown in baseball?
Candidate answer	The last perfect game in Major League Baseball was thrown by Félix Hernández of the Seattle Mariners on August 15, 2012. So the answer is: August 15, 2012.
Initial query	last perfect game in MLB history
Evidence summary	The last perfect game in MLB history was thrown by Domingo Germán of the New York Yankees on June 28, 2023, against the Oakland Athletics.
Reflection	The candidate answer states that the last perfect game was thrown by Félix Hernández in 2012. However, the evidence confirms that the last perfect game was actually thrown by Domingo Germán in 2023. This indicates the candidate answer is outdated.
Refined query	perfect game baseball MLB after 2012
Final decision	**False** (True Negative)
Explanation	The candidate answer states that Félix Hernández threw the last perfect game on August 15, 2012. However, all evidence confirms that Domingo Germán pitched a perfect game on June 28, 2023. This makes the candidate answer outdated and incorrect.

Table 9: Example where SAGE detects an outdated fact in the candidate answer by identifying recent evidence through iterative search and reflection.

- **Simplified (0-shot w/o CoT):** Prompt that removes CoT examples for every module.

As shown in Table 12, the few-shot prompt yields a modest improvement over the original 1-shot setup. Removing CoT examples entirely results in a performance drop of at most 3 percentage points in both Cohen's κ and macro-F1. Interestingly, the default configuration remains highly stable: re-running SAGE with the original prompt reproduced the same κ and F1 scores as reported in prior experiments, underscoring its consistency.

10.3 SAGE failure cases

To better understand the limitations of SAGE in evaluating candidate responses, we conducted a manual error analysis. We randomly sampled 100 evaluation cases from the AmbigQA and HotpotQA datasets, focusing on instances where SAGE disagreed with human annotators. The errors were categorized into different categories, with representative examples presented in the Tables.

Contextual misunderstanding. SAGE generates inaccurate or incomplete queries when it misinterprets the intent of the candidate's question. This is particularly evident in AmbigQA, where questions are often intentionally ambiguous or lack sufficient context, leading to the retrieval of irrelevant or contradictory evidence.

Table 13 presents an example where SAGE faced a contextual misunderstanding. The question asked about the opening date of the Royal Liverpool Hospital before redevelopment. However, the candidate answer referred to the hospital's reopening in 2022, following redevelopment, rather than its original opening date in 1978. Despite iterative search and reflection, SAGE failed to align with the actual intent of the question, leading to a false final decision.

Incomplete evidence. SAGE fails when the retrieved evidence is insufficient or lacks relevant information, especially for verifying specific factual claims (see Table 14).

Reasoning error. Despite accurate evidence, the judge model misinterprets the information or applies flawed reasoning. Reasoning errors often occur when the model fails to apply appropriate logic to the available evidence or when it misrepresents the intent of the question. Table 15 illustrates a case where SAGE incorrectly concluded that Guglielmo Marconi was the first to achieve wireless telegraphy communication. The evidence suggested that while Marconi developed the first practical system, earlier demonstrations, including those by Jagadish Chandra Bose, may have preceded it.

Hallucination. In cases where evidence is unclear or inconclusive, SAGE relies on its pre-trained knowledge, resulting in hallucinated rationales. Hallucination errors often occur when the model confidently asserts false information without sufficient evidence. Table 16 illustrates a case where SAGE incorrectly concluded that Sarah Michelle Gellar, a cast member of Scream 2, was a star of the TV show Roseanne, despite the absence of supporting evidence.

At each round, SAGE's reflection module takes the aggregated evidence summary and generates a chain-of-thought reflection, pointing out which

Component	Content
Question	Who was born first, Javier Frana or Sherwood Stewart?
Answer	Let's think step by step. Javier Frana was born in 1966. Sherwood Stewart was born in 1957. 1957 (Sherwood Stewart) < 1966 (Javier Frana), so Sherwood Stewart was born first. So the answer is: Sherwood Stewart.
Query	Javier Frana birthday
Summary	Javier Frana was born on December 25, 1966. No evidence was found regarding Sherwood Stewart's birthdate.
Reflection	The candidate answer claims that Sherwood Stewart was born in **1957**, but the evidence only confirms Javier Frana's birth year. Further search is required to verify Stewart's actual birthdate.
Iteration	2
Query	Sherwood Stewart tennis birthdate
Summary	Sherwood Stewart was born on June 6, 1946, according to multiple sources. Javier Frana was born on December 25, 1966.
Reflection	While the candidate answer initially provides an **incorrect birth year for Stewart (1957), the evidence confirms that he was actually born in 1946.** The candidate's final conclusion that Stewart was born first remains correct despite the factual error.
Final decision	**True**(True Positive)
Explanation	The candidate answer correctly identifies Sherwood Stewart as being born first. Although it initially provided an incorrect birth year for Stewart, the evidence from iteration 2 confirms the correct birthdate as 1946. Therefore, the final answer is correct, but the reasoning was flawed.

Table 10: Example where SAGE detects and corrects an incorrect reasoning trace. While the candidate's final answer is correct, the system highlights the factual inaccuracy in the intermediate reasoning.

Task	Judge without search	Self-Consistency	Multi-LLMs	Single-Pass	SAGE
AmbigQA	0.381	0.515	0.609	0.658	**0.914**
HotpotQA	0.375	0.493	0.527	0.583	**0.701**

Table 11: Cohen's κ agreement between LLM-based evaluators and the human majority vote. SAGE achieves the highest alignment with human judgments across both AmbigQA and HotpotQA. *Single-Pass* denotes tool use without iteration or SAGE components (i.e., summarize, reflect, and refine).

attributes seem well-supported, which appear contradictory or missing, and where further detail is needed.

The next query is generated from that reflection, so contradictions are an explicit signal to search for clarifying evidence. We do not hard-code a credibility score; instead, SAGE relies on cross-source agreement and iterative follow-up. All retrieved snippets and their domains will be released so that future work can plug in credibility weighting without altering the loop.

Conflicting evidence. At each round, SAGE's reflection module takes the aggregated evidence summary and generates a CoT reflection, pointing out which attributes seem well-supported, which appear contradictory or missing, and where further detail is needed. The next query is generated from that reflection, so contradictions are an explicit signal to search for clarifying evidence. However, in some cases, SAGE encounters ***conflicting evidence*** across multiple search iterations. It sometimes struggles to appropriately resolve contradictory information. Table 17 presents a scenario where the model failed to handle conflicting evidence when assessing the number of band members in Lush and P.O.D.

Topic drift in query refinement. To analyze how often SAGE's query refinement step introduces topic drift or irrelevant queries, we examined this within the 100 instances from our error analysis. Since our SAGE configuration uses up to 3 iterations (1 initial query, up to 2 refined queries), this involved inspecting the query generation process across these stages.

11 Ethical considerations

SAGE aims to enhance evaluation transparency and factual accuracy by grounding judgments in externally retrieved, verifiable evidence. However, several ethical considerations require attention.

Reliance on external sources. The evidence quality in SAGE depends significantly on the retrieved external content, which may introduce

Task	Prompt Variant	Cohen's κ	Macro-F1
AmbigQA	Original (1-shot CoT)	0.914	0.957
	Few-shot CoT	0.920	0.960
	Simplified (0-shot)	0.882	0.932
HotpotQA	Original (1-shot CoT)	0.701	0.850
	Few-shot CoT	0.706	0.853
	Simplified (0-shot)	0.672	0.824

Table 12: SAGE performance under prompt variations. The default (1-shot CoT) remains robust, while removing CoT examples results in a small but consistent performance drop.

Component	Content
Question	When does the new Royal Liverpool Hospital open before redevelopment?
Answer	The new Royal Liverpool Hospital is scheduled to open in 2022, following delays in construction and redevelopment. So the answer is: 2022.
Search query	Royal Liverpool Hospital original opening date before redevelopment.
Summary	The new Royal Liverpool Hospital opened in 1978 after services were transferred from the old Liverpool Royal Infirmary.
Reflection	The candidate answer claims that the new hospital was scheduled to open in 2022. However, the evidence confirms its original opening was in 1978. The judge misinterpreted the question's intent, mistaking the redevelopment's reopening for the original opening date.
Iteration	2
Query	Royal Liverpool Hospital redevelopment opening date 2022.
Summary	Structural repairs were completed, allowing the hospital to reopen in 2022. The move to the new hospital began on September 28, 2022.
Reflection	The evidence confirms the hospital reopened in 2022, but the question specifically asks about the original opening before redevelopment. Despite the accurate evidence about the redevelopment timeline, the candidate answer misaligned with the question intent.
Final decision	**True** (False Positive)
Explanation	SAGE concluded that the candidate answer was correct, accepting the 2022 reopening as the answer. However, the correct answer was 1978, referring to the hospital's original opening. This was a case of contextual misunderstanding.

Table 13: Example of a contextual misunderstanding where SAGE misinterpreted the question's intent by focusing on the hospital's reopening in 2022 instead of its original opening in 1978.

unintended biases or inaccuracies. Although our iterative refinement explicitly mitigates biases by cross-checking multiple sources, SAGE does not currently include explicit credibility scoring. Researchers and practitioners should remain cautious about inherent biases in external sources and proactively manage credibility assessments.

Human oversight and accountability. SAGE substantially improves automated LLM evaluation accuracy, but human oversight remains critical, especially in high-stakes decision-making contexts. It is imperative to recognize the inherent limitations of automated evaluations, maintain human accountability and critical review, particularly when evaluating sensitive or consequential content.

Computational costs and accessibility. SAGE's iterative, tool-augmented process significantly increases computational costs and inference latency compared to simpler evaluation methods, as explicitly shown in our analysis. Although this computational overhead is justified by gains in accuracy for offline evaluation scenarios such as benchmarking and auditing, researchers must consider the environmental and resource accessibility implications. We encourage future research to optimize computational efficiency and ensure broad accessibility.

By continuously addressing these ethical dimensions, SAGE can support more reliable, transparent, and socially responsible AI evaluation practices.

Component	Content
Question	Wing Chun starred Michelle Yeoh and a Hong Kong actor, martial artist, and film director who won what tournament multiple times?
Candidate answer	The Hong Kong actor in Wing Chun is Donnie Yen. Donnie Yen won the multiple times wushu tournament champion. So the answer is: wushu tournament.
Evidence summary	The film "Wing Chun," released in 1994, stars Michelle Yeoh and Donnie Yen, directed by Yuen Woo-ping. The search results do not provide evidence that he won multiple wushu tournaments.
Reflection	The evidence summary lacks any mention of specific wushu tournament victories by Donnie Yen. The absence of evidence to validate the candidate's claim indicates incomplete information. Additional reliable sources would be required to confirm the accuracy of the answer.
Final decision	**False** (False Negative)

Table 14: Example of incomplete evidence where SAGE could not retrieve sufficient information to verify the candidate's claim.

Component	Content
Question	Who made communication by wireless telegraphy for the first time?
Answer	The first successful communication by wireless telegraphy was achieved by Guglielmo Marconi. He is credited with developing the first practical system for wireless telegraphy in the late 19th century. So the answer is: Guglielmo Marconi.
Query	first successful wireless telegraphy transmission + inventor + date
Summary	Guglielmo Marconi is known for creating a practical radio wave-based wireless telegraph system. He began working on wireless telegraphy in the 1890s.
Reflection	The candidate answer asserts that Guglielmo Marconi made the first successful wireless telegraphy communication. While the evidence supports Marconi's development of the first **practical** wireless telegraph system, the question explicitly asks who made the **first communication**, not the first practical one. Further evidence is required to investigate whether anyone transmitted information wirelessly before Marconi.
Iteration	2
Query	pre-Marconi wireless experiments OR early electromagnetic communication OR history of radio waves before 1900 OR wireless telegraphy before Marconi
Summary	While Marconi is credited with the first practical wireless telegraphy system, there were earlier demonstrations of electromagnetic wave transmission, including Jagadish Chandra Bose's 1895 experiment. Wireless telegraphy was also reported on ships in 1891.
Reflection	The evidence indicates the existence of earlier wireless communication attempts, particularly by Jagadish Chandra Bose. Despite this, the candidate answer **incorrectly concluded** that Marconi was the first without addressing these earlier cases. The system failed to recognize the distinction between "first communication" and "first practical system."
Final decision	**False** (False Negative)
Explanation	While the evidence confirms that Guglielmo Marconi developed the first practical wireless telegraphy system, the evidence also suggests prior communications using wireless methods. The distinction between a practical system and the first-ever communication was not correctly reasoned, leading to an incorrect final decision.

Table 15: Example of a reasoning error where SAGE misinterpreted the question's intent, failing to account for earlier instances of wireless telegraphy communication.

Component	Content
Question	Which "Roseanne" star is in Scream 2?
Answer	Let's think step by step. **"Roseanne" star in Scream 2 is Sarah Michelle Gellar.** So the answer is: Sarah Michelle Gellar.
Query	Roseanne Barr Scream 2 cast
Summary	Laurie Metcalf, who played Jackie Harris on the TV series Roseanne, also appeared in the film Scream 2 in 1997. **Sarah Michelle Gellar, while a star in Scream 2, is not a Roseanne star.**
Reflection	The candidate answer incorrectly claims that Sarah Michelle Gellar was a star on Roseanne. **This statement is a hallucination as no evidence supports this claim.** The evidence correctly identifies Laurie Metcalf as the only Roseanne star in Scream 2.
Iteration	2
Query	Roseanne cast members in Scream 2
Summary	Laurie Metcalf, who starred as Jackie Harris in Roseanne, is confirmed to be in Scream 2. Sarah Michelle Gellar also appeared in the film, but she has no connection to Roseanne.
Reflection	The evidence confirms Laurie Metcalf's participation in both Roseanne and Scream 2. **Despite further evidence confirming no connection between Gellar and Roseanne, the model continued to assert a false claim.**
Final decision	**False** (False Negative)
Explanation	The candidate answer confidently asserted that Sarah Michelle Gellar was a Roseanne star, despite all evidence indicating that Laurie Metcalf was the only Roseanne star in Scream 2. This case is a clear example of hallucination, where the model generated a false fact with misplaced confidence.

Table 16: Example of hallucination where SAGE falsely identified Sarah Michelle Gellar as a star of Roseanne.

Component	Content
Question	Do Lush and P.O.D. both consist of four band members?
Answer	Let's think step by step. Lush is a British rock band with four members. P.O.D. is an American rock band with four members. So both Lush and P.O.D. consist of four band members. So the answer is: Yes.
Query	number of members in Lush band + number of members in P.O.D. band
Summary	Lush originally consisted of four members: Miki Berenyi, Emma Anderson, Steve Rippon, and Chris Acland. P.O.D. currently consists of four members: Sonny Sandoval, Traa Daniels, Marcos Curiel, and Wuv Bernardo. However, another source mentions only two members of Lush, suggesting a potential inconsistency.
Reflection	The evidence confirms P.O.D. has four members, but Lush's member count is inconsistent across sources. **The model failed to resolve the contradiction, leading to a flawed decision.**
Iteration	2
Query	Lush band members timeline 1987-1998 + P.O.D. band members lineup history
Summary	One source claims Lush had five members initially, including Meriel Barham, but another states they had four, with Steve Rippon later replaced by Phil King. P.O.D. remains consistently reported as a four-member band.
Reflection	The conflicting evidence remains unresolved. **Instead of recognizing the uncertainty, the model inaccurately concluded that both bands had four members.**
Final decision	**False** (False Negative)

Table 17: Example of conflicting evidence where SAGE failed to resolve contradictions in band member counts. While P.O.D.'s four-member structure is consistent, the model ignored Lush's membership changes over time and incorrectly concluded both bands consist of four members.