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

**Sher Badshah[1], Ali Emami[2], Hassan Sajjad[1]**
[1]Dalhousie University [2]Emory University
{sh545346, hsajjad}@dal.ca aemami@emory.edu

## 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. Meanwhile, using LLMs themselves as evaluators without external grounding remains unreliable for objective tasks, as they systematically over-accept incorrect answers, fabricate supporting rationales, and degrade sharply on questions that fall outside their training data. We propose Search-AuGmented Evaluation (SAGE), a framework to assess LLM outputs without fixed 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, 2025a; 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, 2025a); 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



*Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1466–1491
July 2-7, 2026 

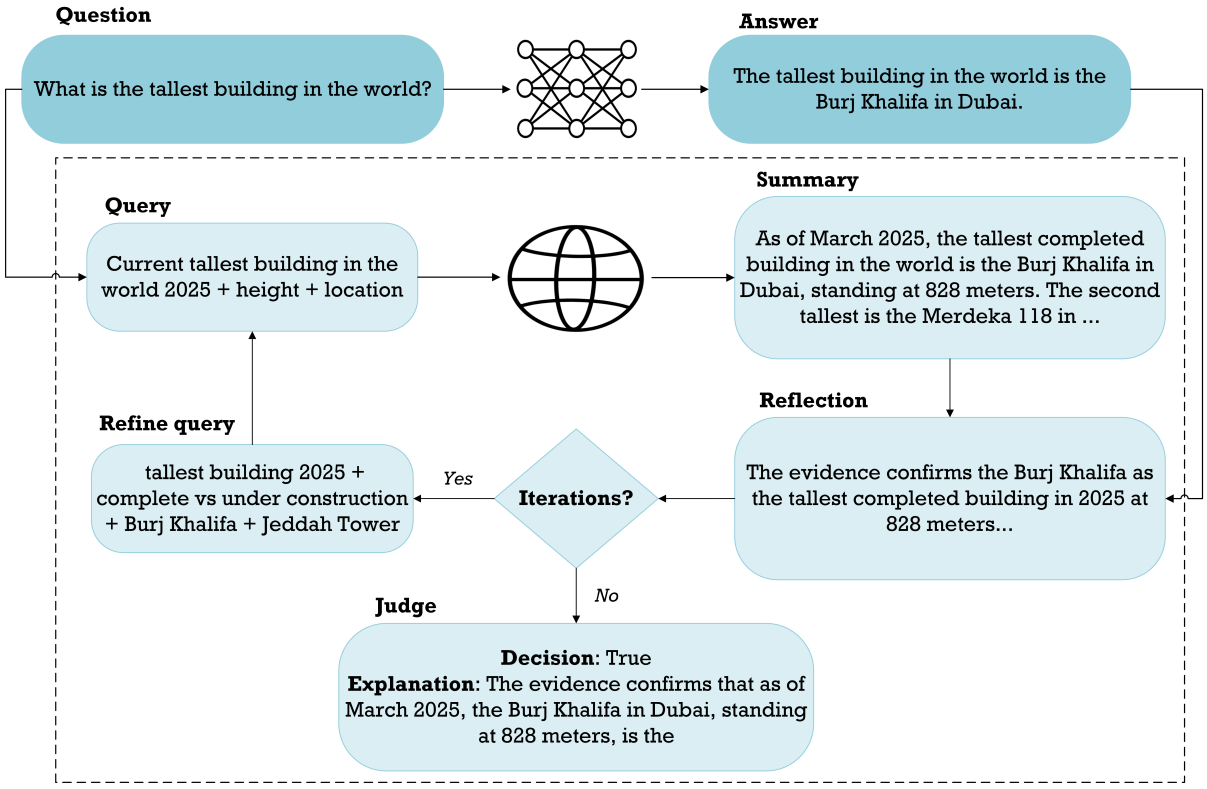


Figure 1: Given the question *"What is the tallest building in the world?"* and candidate answer *"The tallest building in the world is the Burj Khalifa in Dubai,"*, SAGE begins with ***initial query*** from the question. The query triggers ***web searches*** across multiple sources, followed by ***evidence 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.

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).

Given that pre-annotated reference answers at scale are impractical and reference-free LLM-as-a-judge setups remain largely 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 overall process.

Let $x$ denote an input question and let $C$ be a candidate LLM that produces a response $\hat{y} = C(x)$. Our goal is to decide, without access to a pre-

annotated reference answer, whether $\hat{y}$ is factually correct. SAGE addresses this by equipping an LLM judge $J$ to iteratively retrieve and reason over external evidence. Internally, $J$ plays four roles via distinct one-shot prompt templates: query generation ($Q$), summarization ($\Sigma$), reflection ($\Phi$), and final judgment. These roles are composed with an external web-search tool ($S$) and an append-only short-term memory buffer ($\mathcal{M}$) that records the trace of their interaction. We describe each component below.

**Query Generation ($Q$):** At iteration $i$, a query generator $Q$ conditions on the question $x$ and the short-term memory $\mathcal{M}_{i-1}$ accumulated in prior iterations:

$$q_i = Q(x, \mathcal{M}_{i-1}). \tag{1}$$

When $\mathcal{M}_{i-1} = \varnothing$ (i.e., $i = 1$), $Q$ produces an initial, topic-level query from $x$ alone, without assuming the correctness of $\hat{y}$. For $i > 1$, $Q$ refines the query using the accumulated evidence and reflections in $\mathcal{M}_{i-1}$, targeting information still needed to verify $\hat{y}$.[1] SAGE runs for a fixed budget of $N$ iterations; adaptive, reflection-driven stopping is discussed in Section 7.

**Web Search ($S$):** The query is submitted to a web search engine via the Serper API,[2] which returns real-time results. $S(q_i)$ denotes the set of up to $k = 3$ snippets (title, text, and source URL) returned for $q_i$; these raw results serve as the external evidence consumed by the next step.

**Summarization ($\Sigma$):** The retrieved results are condensed into a focused evidence segment,

$$E_i = \Sigma\big(S(q_i)\big), \tag{2}$$

where $\Sigma$ extracts salient factual content and filters redundant or irrelevant text. The resulting $E_i$ is a concise, interpretable evidence summary used by the reflection step.

**Reflection ($\Phi$):** The summarized evidence is assessed for relevance, sufficiency, and factual alignment with respect to $x$ and $\hat{y}$. This step yields a reflection

$$r_i = \Phi(x, \hat{y}, E_i), \tag{3}$$

which records whether the current evidence supports, contradicts, or is inconclusive with respect to $\hat{y}$, and flags missing information or ambiguities that should guide the next query. The reflection output is what drives query refinement in the following iteration.

**Short-Term Memory ($\mathcal{M}$):** We maintain an ordered buffer $\mathcal{M}$ whose state at the end of iteration $i$ is the sequence

$$\mathcal{M}_i = \big((q_1, E_1, r_1), \ldots, (q_i, E_i, r_i)\big), \tag{4}$$

with $\mathcal{M}_0 = \varnothing$. The update is append-only: $\mathcal{M}_i = \mathcal{M}_{i-1} \,\|\, (q_i, E_i, r_i)$, where $\|$ denotes sequence concatenation. $\mathcal{M}$ records the trace of a single $(x, \hat{y})$ episode: it conditions query refinement in $Q$ and is passed in full to the judge at the end.

**Judge ($J$):** After $N$ iterations, the judge is invoked on the full memory $\mathcal{M}_N$ and produces a binary verdict $v \in \{0, 1\}$ together with a natural-language rationale $\rho$:

$$(v, \rho) = J(x, \hat{y}, \mathcal{M}_N). \tag{5}$$

SAGE targets evidence-based verification of a single response: the judge asks whether the retrieved evidence supports or contradicts $\hat{y}$, not which of two responses is preferred nor how good $\hat{y}$ is on a scale. This is entailment-like by construction and distinct from pairwise preference judging or scalar quality rating; we therefore output a binary verdict and retain $\rho$ as the finer-grained reasoning trace for error analysis. The overall procedure is summarized in Algorithm 1.

## 3 Experimental Setup

### 3.1 Models

We use Gemini-1.5-pro (Team, 2024), GPT-3.5-turbo (Brown et al., 2020), and GPT-4o-mini (Team, 2023) as both candidates and judges within the SAGE framework, so that each model is assessed both on its ability to generate responses and on its ability to evaluate responses from others. All experiments use temperature 0 to maximize determinism, as higher temperatures degrade LLM-based evaluators (Hada et al., 2024). For brevity, we refer to these models as Gemini, GPT-3.5, and GPT-4o.

We further evaluate SAGE under smaller open-weight judges, including Mistral 7B (Jiang et al., 2023) and models from the Qwen family (Qwen Team, 2025a,b) such as Qwen2.5-7B, Qwen3-4B, Qwen3-8B (see Appendix 9.3).

[1] The initial and refinement prompts are templated variants of the same module; see Figures 5 and 8 in the Appendix.

[2] https://serper.dev/

**Algorithm 1** SAGE: Search-AuGmented Evaluation

**Require:** question $x$, candidate response $\hat{y}$, iteration budget $N$

**Ensure:** verdict $v \in \{0, 1\}$, rationale $\rho$

$\mathcal{M} \leftarrow \varnothing$ {short-term memory}
**for** $i = 1$ **to** $N$ **do**
  $q_i \leftarrow Q(x, \mathcal{M})$ {initial if $\mathcal{M} = \varnothing$, else refined}
  $E_i \leftarrow \Sigma(S(q_i))$ {retrieve & summarize}
  $r_i \leftarrow \Phi(x, \hat{y}, E_i)$ {reflect on sufficiency}
  $\mathcal{M} \leftarrow \mathcal{M} \,\|\, (q_i, E_i, r_i)$ {append to memory}
**end for**
$(v, \rho) \leftarrow J(x, \hat{y}, \mathcal{M})$ {final verdict}
**return** $(v, \rho)$

### 3.2 Datasets

We evaluate SAGE on widely used free-form question-answering datasets that span different question types, knowledge domains, and complexity levels. These include 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 in which accuracy and truthfulness are paramount (Gou et al., 2024). We also use FreshQA (Vu et al., 2023) to evaluate SAGE's ability to detect outdated knowledge. Due to computational constraints, we randomly sample 300 instances per dataset. Each dataset provides reference answers that serve as ground truth for reference-based 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 fixed exemplars per dataset to elicit detailed, reasoning-based responses, a strategy well-suited to free-form QA (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).

| Candidate | Task | Cohen's $\kappa$ | | | 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 $\tau = 0.5$.

### 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*** as lexical reference-based baselines. Because standard automatic metrics can be misleading for free-form QA (Badshah et al., 2025; Kamalloo et al., 2023), we also follow Badshah and Sajjad (2025b) and Wang et al. (2023a) and employ GPT-4 as a reference-based evaluator that compares candidate answers to gold answers; we refer to this evaluator as **RefGPT**. As shown in Table 1, RefGPT attains consistently substantial agreement with human annotators, well above EM and F1. We therefore adopt RefGPT as our primary reference-based baseline. On datasets without human annotations, we further use RefGPT as a *scalable proxy* ("silver human") when computing SAGE's agreement, and we explicitly treat it as a proxy—not a gold standard: its $\kappa \approx 0.71$ with human judges indicates substantial but imperfect alignment, and our primary claims are anchored to the human-labeled subsets of AmbigQA and HotpotQA.

**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, human evaluation is limited to these two datasets. Annotators are presented with input questions, reference

answers, and anonymized model responses in randomized order to prevent position or model-identity bias. Each response is rated on a binary scale: 1 ("True") for responses that align with the reference answer and demonstrate contextual relevance, and 0 ("False") otherwise. The majority vote determines the final judgment (see Appendix 8.3.3).

### 3.5 Evaluation Metrics

We report three metrics. ***Accuracy*** is the proportion of instances where the judge's binary verdict matches the reference label obtained from automatic metrics or RefGPT. ***Macro-F1*** measures a judge's agreement with reference-based metrics under class imbalance. For AmbigQA and HotpotQA, where human annotations are available, we additionally compute ***Cohen's*** $\kappa$ and ***Macro-F1*** against the human majority vote. We further run ablations to quantify the ***impact of specific SAGE components***, measured as changes in agreement with human judgments.

## 4 Results

Our primary comparisons use RefGPT as a scalable proxy for human judgments on datasets without human annotations, and human majority votes on the AmbigQA and HotpotQA subsets where annotations are available. Additional results and ablations are reported in Appendices 9 and 10.

### 4.1 Main results

**External evidence improves agreement with reference-based evaluation.** Table 2 shows that SAGE agrees with RefGPT substantially more often than a parametric-only judge. For instance, GPT-3.5 as a SAGE judge reaches 0.80 accuracy when evaluating itself on AmbigQA, compared to only 0.67 without external evidence. The same pattern holds for GPT-4o and Gemini across all five datasets, indicating that grounding verdicts in retrieved evidence improves both precision and recall of the judge.

**SAGE strongly agrees with human evaluations.** To evaluate alignment with human judgment, we compare SAGE and baseline evaluators against majority votes from three expert annotators on AmbigQA and HotpotQA. As shown in Table 3, judges without search 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 $\kappa$ often below 0.40. SAGE substantially closes this gap: for example, GPT-4o as a reference-free judge obtains $\kappa = 0.38$ on HotpotQA, whereas the same judge under SAGE reaches $\kappa = 0.70$. This pattern is notable because the judge model "knows" the correct answer as a candidate yet accepts contradictory claims at face value when acting as a reference-free judge—a failure mode that external evidence reliably corrects.

| 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: Agreement of Judge-without-search and SAGE with RefGPT across candidate models and tasks, measured as verdict accuracy against RefGPT's reference-based labels. Higher is better.

Cohen's $\kappa$ 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.

**SAGE works better with more capable judges.** SAGE's performance improves significantly when the judge is a more capable model, where "more capable" is defined by public leaderboard performance (e.g., MMLU, GSM8K) rather than by parameter count, which is undisclosed for several of the models we evaluate. In Table 3, GPT-4o outperforms GPT-3.5 and Gemini on both AmbigQA and HotpotQA; for example, GPT-4o as judge reaches a Macro-F1 of 0.90 on AmbigQA when evaluating GPT-3.5, higher than GPT-3.5's 0.78 and Gemini's 0.83.

| Candid. | Task | Judge without search | | | | | | SAGE | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | GPT-3.5 | | GPT-4o | | Gemini | | GPT-3.5 | | GPT-4o | | Gemini | |
| | | $\kappa$ | Mac-F1 | $\kappa$ | Mac-F1 | $\kappa$ | Mac-F1 | $\kappa$ | Mac-F1 | $\kappa$ | Mac-F1 | $\kappa$ | 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 $\kappa$ and Macro-F1 between the human majority vote and reference-free judges (with and without SAGE) on AmbigQA and HotpotQA.

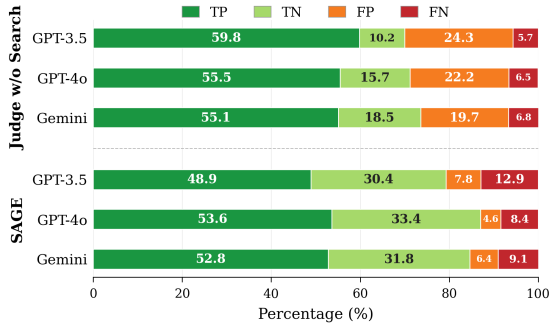


Figure 2: Distribution of true positive (TP), true negative (TN), false positive (FP), and false negative (FN) rates for each LLM judge in both Judge without search and SAGE, averaged across AmbigQA and HotpotQA and candidate models.

**SAGE detects untruthful facts and outdated knowledge.** SAGE flags false claims by cross-referencing them with retrieved evidence. For *"Who sings the theme song for the show Half & Half?"*, a candidate model answered *"Erica Campbell"* and tool-free judges accepted the claim; SAGE retrieved sources confirming that *"Melonie Daniels"* performed the theme song and correctly rejected the candidate's answer with a grounded rationale. On FreshQA (Vu et al., 2023), SAGE similarly catches stale knowledge: for *"Where is EMNLP this year?"*, candidates often returned outdated locations, while SAGE retrieved the correct current location *"Suzhou, China"*[3] (see Appendix 9.5).

**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 13 in the Appendix).

[3]These experiments were conducted in 2025, when EMNLP was held in Suzhou, China; EMNLP 2026 is hosted in Hungary.

## 4.2 Error analysis

To understand SAGE's limitations, we conducted a manual error analysis on 100 randomly sampled evaluation cases from AmbigQA and HotpotQA in which SAGE disagreed with the human majority. We group the errors into five categories: 1) **Contextual misunderstanding (23%):** SAGE generates inaccurate or incomplete queries when it misinterprets the candidate question's intent. This is particularly evident in AmbigQA, where questions are often intentionally ambiguous or underspecified, leading to irrelevant or contradictory evidence. 2) **Incomplete evidence (15%):** SAGE fails when the retrieved evidence is insufficient or lacks relevant information, especially 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%):** when evidence is ambiguous or inconclusive, the judge falls back on its pre-trained knowledge and produces 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 the effect of the number of SAGE iterations on judge performance. Any positive number of iterations improves over the zero-iteration (parametric-only) baseline, and three iterations—our default—consistently offers the best trade-off between performance and

| 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.

cost. A small dip at two iterations reflects occasional off-topic refinement queries that are typically corrected in the following round, while the decline at four iterations is driven by an overabundance of sources that inflates context length and introduces redundant or irrelevant evidence.

**Effect of query generation.** To evaluate the contribution of the query-generation module, we remove it and use the input question directly for evidence retrieval. Table 4 shows that SAGE consistently outperforms this query-free baseline, demonstrating the importance of the query-generation step. The most notable improvements are on AmbigQA, where SAGE reaches a Macro-F1 of 0.96 compared to 0.84 without query generation. On HotpotQA the gain is smaller but still positive, reflecting SAGE's 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. One 3rd-iteration query was excluded because annotators could not reach consensus on its category, leaving 299 queries in total. 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 199 refined queries analyzed in the refinement stages, only 6.5% of queries were judged "insufficient."

| 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 $\times$ 3).

**Impact of iterative evidence gathering.** To isolate the value of iterative retrieval, we compare against a minimal search-augmented baseline that performs a single-pass retrieval: it issues one web query from the input question, collects the top-3 snippets, and feeds them to the judge without summarization, reflection, or refinement. We write **SAGE-$k$** for SAGE restricted to $k$ iterations; our default configuration is SAGE-3. All configurations in this ablation use GPT-4o as both candidate and judge.

On AmbigQA, the single-pass baseline already raises the judge-without-search $\kappa$ from 0.38 to 0.65, confirming that one round of external evidence yields a substantial improvement over parametric-only judging. Restoring SAGE's summarize–reflect step under the same one-query budget (SAGE-1) lifts $\kappa$ further to 0.74 ($+0.09$), because the reflect step explicitly tests evidence sufficiency and rejects false-positive acceptances when the top-$k$ is off-topic. Running the full three-iteration loop (SAGE-3) pushes $\kappa$ to 0.91. HotpotQA shows the same trend: $\kappa$ climbs from 0.38 (no search) to 0.58 (single-pass), 0.64 (SAGE-1), and 0.70 (SAGE-3). These results isolate two distinct sources of gain: external evidence *per se*, and the reflect-and-refine loop that turns that evidence into a reliable verdict.

**Robustness to the search engine.** Because SAGE depends on an external retriever, we ask whether its gains persist when the search engine is replaced. We rerun the full pipeline with Brave and Tavily in place of Serper, keeping all other components fixed. Table 6 reports Cohen's $\kappa$ for the single-pass baseline, SAGE-1, and SAGE-3 on AmbigQA and HotpotQA with GPT-4o as candidate and judge. The iterative gain from SAGE is preserved across all three engines, and SAGE-3 remains the strongest configuration in every case.

### 4.4 Cost and latency

As given in Table 6, a full SAGE-3 evaluation makes 10 LLM calls and 3 web-search queries per instance. At April 2026 list prices for GPT-

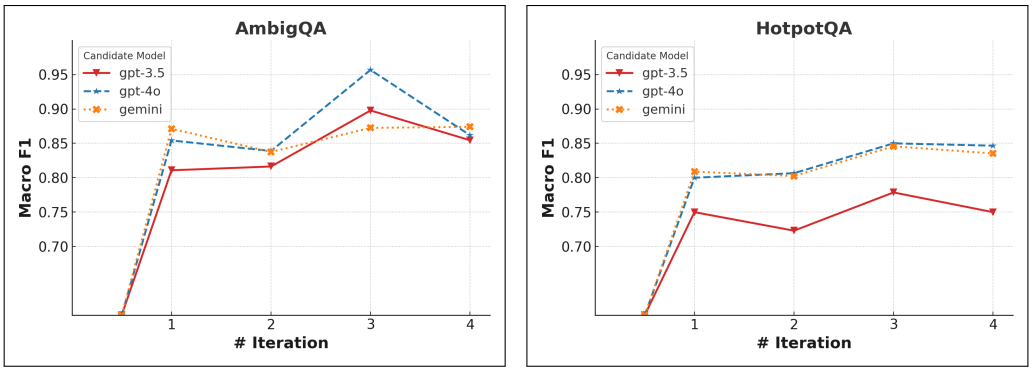


Figure 3: The effect of iterations. GPT-4o is used as a judge here.

| Task | Engine | 1-pass | SAGE-1 | SAGE-3 |
|---|---|---|---|---|
| **AmbigQA** | Serper | 0.65 | 0.74 | **0.91** |
| | Brave | 0.60 | 0.72 | 0.85 |
| | Tavily | 0.45 | 0.60 | 0.70 |
| **HotpotQA** | Serper | 0.58 | 0.64 | **0.70** |
| | Brave | 0.60 | 0.66 | **0.71** |
| | Tavily | 0.36 | 0.62 | 0.67 |

Table 6: Robustness of SAGE to the search engine. Cohen's $\kappa$ against the human majority on AmbigQA and HotpotQA, with GPT-4o as both candidate and judge.

| Setting | Total (s) | LLM (s) | Search (s) | #LLM |
|---|---|---|---|---|
| LLM-as-Judge | $1.65_{\pm 0.98}$ | — | — | 1 |
| 1-pass | $3.15_{\pm 1.24}$ | $2.18_{\pm 1.17}$ | $0.96_{\pm 0.40}$ | 1 |
| SAGE-3 | $26.32_{\pm 3.10}$ | $23.16_{\pm 3.06}$ | $3.13_{\pm 0.59}$ | 10 |

Table 7: End-to-end latency per instance on 50 AmbigQA items (GPT-4o-mini at $T$=0 via Serper). Values are mean $_{\pm \text{std}}$. LLM time dominates SAGE-3; the three search calls together contribute only $\sim$3 s.

4o-mini[4] and Serper's Starter plan,[5] this works out to about \$0.0043 per instance, so the full 300-instance AmbigQA subset is evaluated end-to-end for roughly \$1.28—orders of magnitude cheaper than manual annotation. Client-side wall-clock, measured on a 50-instance sample (Table 7), averages $\approx$ 27 s per instance, of which $\approx$ 23 s is LLM time spread across 10 sequential calls (network round-trip, queueing, and JSON handling) and only $\approx$ 3 s is search. Since instances are independent, the batch parallelizes trivially: a 300-item dataset can be evaluated in minutes with modest concurrency.

[4]\$0.15 / \$0.60 per $10^6$ input / output tokens; https://openai.com/api/pricing/.

[5]\$1.00 per $10^3$ queries; https://serper.dev/.

## 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 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 et al., 2025; Badshah and Sajjad, 2025a; 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 pre-trained 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 reference-free framework that equips an LLM judge with iterative web retrieval, evidence summarization, reflection, and query refinement to evaluate the factual correctness of free-form QA responses. Unlike reference-based metrics that require costly pre-annotated answers, and unlike parametric-only judges that are prone to hallucinated rationales, outdated knowledge, and uncritical acceptance of candidate claims, SAGE grounds every verdict in externally retrieved, verifiable evidence.

Experiments across five QA benchmarks and different model families demonstrate that SAGE achieves substantial to perfect agreement with human evaluators, consistently outperforming both lexical baselines and reference-free judges. Ablations show that the gains arise from two complementary sources: external evidence retrieval itself, and the iterative summarize–reflect–refine loop that converts raw snippets into reliable verdicts. These benefits hold across multiple search engines, prompt designs, and smaller open-weight judges, while remaining orders of magnitude cheaper than human annotation.

Looking ahead, we see two natural extensions. First, adaptive stopping based on reflection-derived sufficiency signals can reduce latency on straightforward instances without sacrificing reliability on harder ones. Second, SAGE's summarize–reflect–refine loop is domain-agnostic; by replacing the web-search tool with a code interpreter, a calculator, or a knowledge-base API, the same framework could support evaluation of code generation, mathematical reasoning, and structured-knowledge tasks.

## 7 Limitations

**Context window constraints.** SAGE's short-term memory grows with each iteration as (query, evidence, reflection) tuples accumulate, so the judge's context window bounds how many iterations can be chained before relevant earlier steps are truncated. Future work could replace the append-only buffer with a recall-based long-term memory that selectively retrieves past traces. For example, episodic and semantic memory (Park et al., 2023) to reduce both context pressure and computational cost.

**Source bias and quality control.** SAGE's reflection module detects inconsistencies across retrieved sources but does not explicitly model source credibility. When multiple sources agree on an incorrect or outdated claim, the judge has no mechanism to discount them, and the evaluation may inherit biases from the external data (Li et al., 2025a; Zhan et al., 2024; Yu et al., 2024). Integrating credibility scoring or provenance-aware weighting that dynamically assess evidence reliability is a natural extension.

**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 with smaller judges such as Mistral 7B (see Appendix 9.3). Deploying SAGE with very small judges therefore involves a reliability–cost

trade-off that users should weigh explicitly.

**Diminishing returns with iterative refinement.** Performance gains plateau and occasionally degrade beyond three iterations as redundant or off-topic evidence accumulates and the context window fills up; we therefore fix $N = 3$ empirically. In future work, we plan to explore adaptive stopping: a lightweight heuristic could use the existing reflection signal (evidence sufficiency combined with low evidence novelty across consecutive iterations) to terminate early, while more principled variants such as instruction-tuning the reflection module to output a calibrated sufficiency score, or conformal-prediction-based stopping when the estimated flip-risk falls below a target level could further improve the efficiency–reliability trade-off.

**Domain scope.** SAGE is validated on factual free-form QA, where claims can be verified against web-retrievable evidence. Extending SAGE to other evaluation domains such as code generation (via an interpreter or unit tests), mathematical reasoning (via a calculator or computer algebra system), or structured knowledge evaluation, would require swapping the web-search tool for domain-appropriate tools while retaining the summarize–reflect–refine loop. We leave this generalization to future work.

**Binary verdicts.** SAGE outputs a binary True/False verdict, which is well-suited to factual-correctness evaluation but does not support partial-correctness scoring, pairwise comparison, or rubric-based ratings. Tasks such as summarization quality or instruction-following preference, where judgments are inherently graded, would require extending SAGE's verdict schema beyond binary classification.

**Language coverage.** All experiments in this paper are conducted in English. SAGE's effectiveness on non-English QA where web-search quality, snippet relevance, and the judge LLM's multilingual reasoning capabilities all vary—remains untested.

## Acknowledgment

We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC), Canada Foundation for Innovation (CFI), and Research Nova Scotia. Advanced computing resources are provided by ACENET, the regional partner in Atlantic Canada, and the Digital Research Alliance of Canada.

## References

Sher Badshah, Moamen Moustafa, and Hassan Sajjad. 2025. CLEV: LLM-based evaluation through lightweight efficient voting for free-form question-answering. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 1513–1531, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

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. 2025a. DAFE: LLM-Based Evaluation Through Dynamic Arbitration for Free-Form Question-Answering. *Preprint*, arXiv:2503.08542.

Sher Badshah and Hassan Sajjad. 2025b. Reference-guided verdict: LLMs-as-judges in automatic evaluation of free-form QA. In *Proceedings of the 9th Widening NLP Workshop*, pages 251–267, Suzhou, China. Association for Computational Linguistics.

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.

Qwen Team. 2025a. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Qwen Team. 2025b. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

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.

Xuemin Yu, Fahim Dalvi, Nadir Durrani, Marzia Nouri, and Hassan Sajjad. 2024. Latent concept-based explanation of NLP models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12435–12459, Miami, Florida, USA. Association for Computational Linguistics.

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 span different question types, knowledge domains, and complexity levels. Evaluating large-scale datasets is costly, so we randomly sample 300 instances per dataset. Each dataset's reference answers serve as ground truth for our reference-based baselines. 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.

We use the respective validation splits of each dataset: the standard validation sets for AmbigQA and Natural Questions, the `distractor` subset validation split 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).

The *candidate* model generates responses under the few-shot CoT template above. The *SAGE* modules use a separate, one-shot role prompt per module that combines role-playing instructions with explicit reasoning goals; these prompts are shared across datasets and judges. Below, we describe the prompting strategy for each SAGE component.

**Query Generation.** The query generation module converts an input question $x$ 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 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 retrieval in subsequent iterations ($i > 1$), the query refinement prompt conditions on the accumulated short-term memory $\mathcal{M}_{i-1} = \big((q_1, E_1, r_1), \ldots, (q_{i-1}, E_{i-1}, r_{i-1})\big)$ and produces a new query $q_i$. Concretely, the prompt receives the question $x$, the most recent query, an aggregated evidence summary over all prior iterations, and the latest reflection, and is instructed via CoT to identify remaining uncertainties or gaps before generating a refined, more targeted query.

**Judgment.** Finally, the judgment module implements $J(x, \hat{y}, \mathcal{M}_N)$: given the question, the candidate answer, and the accumulated memory $\mathcal{M}_N$, it emits a binary verdict $v \in \{0, 1\}$ together with a natural-language rationale $\rho$. The CoT prompt conditions on the question, the candidate answer, the aggregated evidence summary, and the aggregated reflection string—all materialized from $\mathcal{M}_N$—and instructs the model to reason step by step before committing to its verdict. The output is formatted as a JSON object with keys "`decision`" and "`explanation`", corresponding to $v$ and $\rho$ respectively.

### 8.3 Baselines

We compare SAGE against several established evaluation approaches:

#### 8.3.1 Reference-based metrics

We implement three reference-based baselines 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** prompts GPT-4 with the question, candidate answer, and dataset reference answers, and asks it to issue a binary correctness verdict (Wang et al., 2023a). This provides a context-aware evaluation beyond strict token-level matching.

#### 8.3.2 Judge without search

Following Liu et al. (2023), we implement a reference-free baseline in which 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 search augmentation in SAGE by holding the judge model fixed and removing only the evidence-retrieval mechanism.

#### 8.3.3 Human evaluation

The main setup including annotators, randomization, and the binary 1/0 scoring scheme is described in §3. Here we add the budgetary rationale and the full annotator guidelines.

**Scale and rationale.** We limit human evaluation to AmbigQA and HotpotQA because these datasets best exercise SAGE's core use case (ambiguous or multi-hop questions) and because extending to all five datasets would have significantly increased annotation cost. For each of the two tasks we evaluate 300 instances per candidate model, and we evaluate three candidate models, for a total of $2 \times 300 \times 3 = 1{,}800$ annotated judgments.

**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 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 en-

**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.

gines 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** ($\kappa$) 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) matches the reference label obtained from automatic metrics or from RefGPT.

**Agreement with Human Judgment:** For the AmbigQA and HotpotQA subsets with human annotations, we calculate Cohen's Kappa ($\kappa$), 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 ($\kappa$) 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:

**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.

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

where $y_i$ represents the label assigned by the $i$th 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 $\text{Precision}_c$ and $\text{Recall}_c$ are the precision and recall for class $c$.

## 9 Additional results

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

### 9.1 Inter-human annotator agreement

Table 8 presents the human annotator agreement results for the AmbigQA and HotpotQA across three candidate models. The results indicate consistently high agreement among annotators.

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

Table 8: Human annotator agreement results on AmbigQA and HotpotQA tasks.

### 9.2 SAGE agreement with reference-based metrics

A reference-free judge is only useful if its verdicts track reference-based signals when the latter are reliable. Table 9 reports the per-evaluator accuracy for every (candidate, task) pair, alongside the lexical EM/F1 baselines and RefGPT. Two patterns stand out. First, SAGE's raw accuracy tracks the reference-based signals closely: for example, GPT-3.5 as a SAGE judge reaches 0.64 when evaluating its own answers on AmbigQA, essentially matching its reference-based F1 of 0.63. Second, judges without search exhibit a consistent self-enhancement bias—e.g., GPT-3.5 inflates its own AmbigQA accuracy to 0.81 when asked to judge without evidence, which is absent under SAGE. EM, by contrast, often under-estimates model quality because it misses valid paraphrases and alternative formulations. SAGE therefore sits between these extremes: it recovers the reliability of reference-based signals without requiring pre-annotated references.

### 9.3 SAGE with small open-source judges

To test whether SAGE's benefit persists outside frontier commercial judges, we evaluate four open-weight small LLMs as the judge model within SAGE: Mistral 7B (Jiang et al., 2023), Qwen2.5-7B (Qwen Team, 2025a), Qwen3-4B, and Qwen3-8B (Qwen Team, 2025b). The setup mirrors the main experiments: the judge's four roles (query generation, summarization, reflection, judgment)

**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.

are all played by the small open-weight model, while the candidate is held fixed.

**Mistral 7B.** We first use Mistral 7B as the judge to evaluate GPT-3.5 candidate answers on AmbigQA and HotpotQA. Table 10 shows that Mistral 7B reaches a Cohen's $\kappa$ of 0.59 / 0.33 and a Macro-F1 of 0.80 / 0.66 on AmbigQA / HotpotQA—substantially below frontier judges but still well above the corresponding Judge-without-search baselines. In practice we observe that Mistral 7B's main failure modes are (i) imprecise instruction following on the module-specific JSON prompts, (ii) a limited context window that truncates accumulated traces on HotpotQA's multi-hop questions, and (iii) occasional irrelevant reflections when no supporting evidence is available (Badshah and Sajjad, 2024a). HotpotQA's larger drop is consistent with these observations: the multi-hop traces stress both the context window and the judge's reasoning depth.

**Qwen family.** The Qwen models test whether a more recent instruction-tuned family closes this gap. We use GPT-4o as the candidate and evaluate Qwen2.5-7B, Qwen3-4B, and Qwen3-8B as judges on AmbigQA and HotpotQA. Three trends emerge from Table 10. First, Qwen3 variants clearly outperform Qwen2.5 of comparable size, consistent with Qwen3's reported improvements in instruction following and long-context handling. Second, scaling from Qwen3-4B to Qwen3-8B yields a modest improvement on AmbigQA but a small drop on HotpotQA, suggesting that 4B is already near the instruction-following threshold required for SAGE's templated roles on these tasks. Third, the 4B/8B Qwen3 judges match or exceed Mistral 7B's AmbigQA $\kappa$ while remaining below frontier commercial judges, confirming that SAGE is model-agnostic but that judge capability—specifically instruction following and context-window management—sets the achievable ceiling.

**Takeaway.** Small open-weight judges within SAGE consistently surpass Judge-without-search baselines, making SAGE usable in resource-constrained or fully-offline settings. Frontier judges remain preferable when near-perfect agreement with humans is needed, so users should choose a judge model that matches their reliability–cost trade-off.

### 9.4 Reproducibility across independent runs

SAGE's modules involve LLM sampling and real-time web retrieval, so a fair reproducibility check must verify that its headline numbers are stable across repeated runs rather than a one-shot artifact. We rerun the full SAGE-3 pipeline on AmbigQA and HotpotQA with GPT-4o as both candidate and judge, under identical settings ($T$=0, Serper, $N$=3), on two independent days. Table 11 reports Macro-F1 against the human majority vote for each run. Both runs reproduce the main-paper

**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.

numbers to within 0.00 Macro-F1 on both tasks, indicating that SAGE's behavior is stable at $T=0$ despite the underlying non-determinism of the web-search component (small day-to-day changes in retrieved snippets are absorbed by the summarize–reflect–refine loop).

### 9.5 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 12 presents an example where a candidate 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 evaluate SAGE on FreshQA (Vu et al., 2023) with GPT-4o as the SAGE judge and GPT-3.5 as the candidate. SAGE reaches an accuracy of 38.3%, closer to the reference-based evaluators (EM: 25.0%, F1: 35.4%) than a reference-free judge on the same candidate. The low absolute numbers across all evaluators reflect the well-known difficulty that pre-trained models like GPT-3.5 face on questions about rapidly evolving events; SAGE's retrieved evidence partially closes that gap at evaluation time.

**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.

### 9.6 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 13 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 evidence to correct this mistake while maintaining the correct conclusion.

## 10 Additional ablations and analysis

This appendix collects supplementary analyses referenced from the main paper: comparisons against stronger non-iterative baselines (§10.1), robustness to prompt design (§10.2), and qualitative failure cases (§10.3).

### 10.1 Comparison with stronger non-iterative baselines

We evaluate three non-iterative baselines that exclude SAGE's iterative summarize–reflect–refine loop but retain the judgment prompt. These baselines help disentangle the contributions of search augmentation, sampling-based self-consistency, and model diversity. Across all three, GPT-4o generates the candidate answer. For the *single-pass search-augmented judge* and *self-consistency* settings, GPT-4o also serves as the judge. In the *multi-LLM majority voting* setup, GPT-4o is the candidate and three judges—namely GPT-4o, GPT-3.5, and Mistral 7B—independently evaluate the same input to leverage diverse model reasoning.

| 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 9: Per-evaluator scores on each (candidate, task) pair. The **Reference-based** columns (EM, F1, RefGPT) compare candidate answers to gold references; the **Judge-without-search** and **SAGE** columns report each judge's verdict accuracy against the RefGPT reference label.

| Judge | Task | Cohen's $\kappa$ | Macro-F1 |
|---|---|---|---|
| Mistral 7B | AmbigQA | 0.59 | 0.80 |
| | HotpotQA | 0.33 | 0.66 |
| Qwen2.5-7B | AmbigQA | 0.51 | 0.74 |
| | HotpotQA | 0.30 | 0.63 |
| Qwen3-4B | AmbigQA | 0.60 | 0.80 |
| | HotpotQA | 0.53 | 0.76 |
| Qwen3-8B | AmbigQA | **0.67** | **0.83** |
| | HotpotQA | 0.50 | 0.74 |

Table 10: Small open-weight judges within SAGE, measured against the human majority vote on AmbigQA and HotpotQA. Mistral 7B evaluates GPT-3.5 candidate answers; the Qwen family evaluates GPT-4o candidate answers. Bold entries mark the strongest small-LLM judge on each task.

| Run | AmbigQA | HotpotQA |
|---|---|---|
| Run 1 | 0.96 | 0.85 |
| Run 2 | 0.96 | 0.85 |

Table 11: SAGE-3 Macro-F1 against the human majority vote on AmbigQA and HotpotQA across two independent runs with GPT-4o as both candidate and judge ($T$=0, $N$=3). Both runs reproduce the main-paper numbers exactly.

#### 10.1.1 Single-Pass search-augmented judge

To isolate the effect of *web access* alone, we use a baseline that issues exactly one web query and performs no further reasoning or iteration: given the question and candidate answer, the judge 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 the full SAGE pipeline; the only change is the removal of the summarize–reflect–refine loop. The main-paper comparison of this baseline against SAGE-1 and SAGE-3 is reported in Table 6; Table 14 below repeats the single-pass numbers alongside the additional non-iterative baselines for cross-comparison.

#### 10.1.2 Self-consistency judge

For *self-consistency* (Wang et al., 2023b), the GPT-4o judge samples 10 independent verdicts at temperature 0.7 and returns a simple majority. No external evidence is used; the judge relies entirely on its parametric knowledge, matching the original self-consistency setup. Table 14 shows that self-consistency gives a modest boost over the vanilla Judge-without-search, but 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 CLEV (Badshah et al., 2025; Badshah and Sajjad, 2024b), we instructed three different LLMs:

| 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 12: Example where SAGE detects an outdated fact in the candidate answer by identifying recent evidence through iterative search and reflection.

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

As shown in Table 15, 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 $\kappa$ and macro-F1. Interestingly, the default configuration remains highly stable: re-running SAGE with the original prompt reproduced the same $\kappa$ and F1 scores as reported in prior experiments, underscoring its consistency.

### 10.3 SAGE failure cases

The main-paper error analysis (§4.2) groups SAGE's disagreements with the human majority into five categories: contextual misunderstanding, incomplete evidence, reasoning error, hallucination, and conflicting evidence. This appendix reproduces one representative failure for each category, drawn from the same 100-instance error-analysis sample.

**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 16 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 17).

| 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 13: 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 w/o search | Self-consistency | Multi-LLM vote | Single-pass | SAGE-3 |
|---|---|---|---|---|---|
| AmbigQA | 0.38 | 0.52 | 0.61 | 0.66 | **0.91** |
| HotpotQA | 0.38 | 0.49 | 0.53 | 0.58 | **0.70** |

Table 14: Cohen's $\kappa$ against the human majority vote for five reference-free evaluators on AmbigQA and HotpotQA. *Single-pass* uses one web call with no summarize–reflect–refine loop; *SAGE-3* is our default three-iteration configuration. SAGE-3 achieves the highest alignment with humans across both tasks.

**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 18 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 pretrained knowledge, resulting in hallucinated rationales. Hallucination errors often occur when the model confidently asserts false information without sufficient evidence. Table 19 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 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. 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 iteration, SAGE's reflection module takes the aggregated evidence summary and produces a CoT reflection pointing out which attributes are well-supported, which appear contradictory or missing, and where further detail is needed. The next query is then generated from that reflection, so contradictions act as an explicit signal to search for clarifying evidence. In some cases, however, SAGE encounters *conflicting*

| Task | Prompt Variant | Cohen's $\kappa$ | Macro-F1 |
|---|---|---|---|
| AmbigQA | Original (1-shot CoT) | 0.91 | 0.96 |
| | Few-shot CoT | 0.92 | 0.96 |
| | Simplified (0-shot) | 0.88 | 0.93 |
| HotpotQA | Original (1-shot CoT) | 0.70 | 0.85 |
| | Few-shot CoT | 0.71 | 0.85 |
| | Simplified (0-shot) | 0.67 | 0.82 |

Table 15: SAGE-3 performance under prompt variations, using GPT-4o as both candidate and judge. The default 1-shot CoT configuration is robust: removing the CoT example entirely costs at most three points in Cohen's $\kappa$ on either task.

| 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 16: 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.

*evidence* across multiple iterations and struggles to resolve it. Table 20 presents an example where the model failed to handle conflicting reports on the number of band members in Lush and P.O.D.

## 11 Ethical considerations

SAGE grounds evaluation judgments in externally retrieved, verifiable evidence. This design raises three ethical considerations that require attention.

**Reliance on external sources.** SAGE's judgments depend on the retrieved external content, which may carry unintended biases or inaccuracies. Iterative refinement partially mitigates this by cross-checking multiple sources, but SAGE does not currently model source credibility explicitly. Researchers and practitioners deploying SAGE should remain cautious about biases inherited from search results and should consider adding credibility weighting where the evaluation target is high-stakes.

**Human oversight and accountability.** SAGE improves automated LLM evaluation accuracy but is not a substitute for human review in high-stakes decision-making contexts. Users should treat SAGE's verdicts as a scalable first pass and retain human accountability, particularly for sensitive or consequential content.

**Computational costs and accessibility.** SAGE's iterative, tool-augmented loop is inexpensive in dollar terms but does involve multiple LLM and web-search calls per instance. This is justified for offline evaluation and auditing use cases, where the cost is small relative to human annotation, but it does inherit the environmental footprint of the underlying LLM and search APIs. We encourage future work on adaptive stopping and parallel exe-

| 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 17: Example of incomplete evidence where SAGE could not retrieve sufficient information to verify the candidate's claim.

cution to further reduce this footprint (§7).

| 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 18: 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 19: 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 20: 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.