SAGE: LLM-Based Evaluation Through Selective Aggregation for Free-Form Question-Answering

Anonymous ACL submission

Abstract

Evaluating Large Language Models (LLMs) free-form generated responses remains a challenge due to their diverse and open-ended nature. Traditional supervised signal-based automatic metrics fail to capture semantic equivalence or accommodate the variability of open-ended responses, while human evaluation, though reliable, is resource-intensive at scale. Leveraging LLMs as evaluators offers a promising alternative due to their strong language understanding and instruction-following capabilities. To harness these strengths efficiently, we propose the Selective Aggregation for Generative Evaluation (SAGE), which employs two primary LLMs as judges and engages a third judge only in cases of disagreement. This selective aggregation prioritizes evaluation reliability while reducing unnecessary computational demands compared to conventional majority voting. SAGE incorporates task-specific reference answers to improve judgment accuracy, leading to substantial gains in evaluation metrics such as Macro F1 and Cohen's Kappa. Through experiments, including human evaluation, we demonstrate SAGE's ability to provide consistent, scalable, and resource-efficient assessments, establishing it as a robust framework for evaluating free-form model outputs.

1 Introduction

The rapid advancements in Large Language Models (LLMs) have propelled the field of natural language processing forward, yet their evaluation remains a challenge (Laskar et al., 2024). In particular, free-form model responses are difficult to evaluate because their correctness depends on understanding the broader context and underlying meaning (Si et al., 2021). Many benchmarks, such as MMLU (Hendrycks et al., 2021), often simplify evaluation by focusing on structured outputs (e.g., multiple-choice questions) (Chen et al., 2024). Although effective for certain tasks, such methods rely on the model's probability distribution over predefined options (Thakur et al., 2024). By selecting the highest-probability response, they constrain the evaluation to closed-ended outputs and narrow the scope for assessing broader model capabilities (Chang et al., 2024). The rigid, predefined options in such evaluations not only limit the scope of assessment but also overlook the diversity of potential correct responses in free-form tasks (Li et al., 2023; Zhang et al., 2024).

While automatic metrics offer scalability for evaluating free-form model outputs, they face notable limitations. For instance, Exact Match (EM) requires strict lexical alignment (e.g., failing to equate "nuclear weapon" and "atomic bomb") and ignores semantic equivalence. N-gram metrics, such as BLEU, ROUGE, prioritize surface-level similarity, struggling with structural or lexical diversity that preserves meaning (Papineni et al., 2002; Lin, 2004; Zhu et al., 2023). Neural metrics like BERTScore (Zhang et al., 2020) address this via contextual embeddings but remain brittle: overly dependent on reference quality (Liu et al., 2024), sensitive to domain shifts and text length (Zhu et al., 2023). Additionally, BERTScore produces continuous scores (i.e, from 0 to 1), which are not well-suited for binary evaluations where a clear true or false decision is required (Xu et al., 2023). These shortcomings are exacerbated with instruction-tuned models (Doostmohammadi et al., 2024), which generate verbose, diverse outputs (Saito et al., 2023; Wang et al., 2024b).

Contrary to automatic metrics, human evaluation provides a more reliable assessment (Chiang and Lee, 2023). However, despite being the "gold standard", it has limitations. LLMs' growing complexity and scale have made recruiting and coordinating multiple human raters increasingly resource-intensive and time-consuming (Mañas et al., 2024). Furthermore, the reliability of human evaluation is additionally challenged by variations in rater expertise and inherent subjectivity that affect re-

producibility (Clark et al., 2021; Chiang and Lee, 2023).

Recently, a paradigm shift has emerged where LLMs are utilized to judge the candidate model generations for given tasks (Zheng et al., 2024). This model-based method leverages the instruction-following capabilities of LLMs through evaluation prompts or, in some cases, fine-tuned versions of LLMs that are specifically optimized for evaluation. Existing studies using LLM as judges primarily focus on subjective pairwise comparison (Zheng et al., 2024; Wang et al., 2023a; Vu et al., 2024) and single-answer scoring (Verga et al., 2024) (Chiang and Lee, 2023; Hu et al., 2024; Liu et al., 2023; Chan et al., 2024; Chu et al., 2024). However, to the best of our knowledge, objective evaluation using LLM judges, particularly for free-form question-answering, remains largely unexplored.

As discussed earlier, one practical limitation to objective evaluation is the lexical–semantic mismatch between instruction-tuned LLM outputs and the terse "gold" strings in many free-form QA benchmarks. For the query *"Who wrote 1984?"*, the dataset may list simply *"George Orwell"*, while a helpful model replies: *"It was penned by the British author Eric Arthur Blair."* Although the sentence refers to the same person, it shares no surface tokens with the reference, so EM, n-gram, and even embedding-based metrics assign it an unduly low score. A reference-aware LLM-as-a-judge can instead reason over meaning, recognize that the candidate entails the gold fact, and deliver a reliable verdict—thereby overcoming this lexical–semantic gap. However, LLM-based judging itself lies on a cost–quality spectrum. Querying a single judge is efficient but is less reliable due to the known limitations such as prompt sensitivity, inconsistency, and bias (Ye et al., 2024), as no individual model captures the full diversity of reasoning styles, long-tail knowledge, and user values (Feng et al., 2025). Multi-judge ensembles improve robustness through diversity and majority voting, yet invoking several large models per instance increases cost and latency, limiting practicality for large-scale or continuous evaluation (Jung et al., 2024; Adlakha et al., 2024).

To address these trade-offs, we propose the Selective Aggregation for Generative Evaluation (SAGE)—a scalable framework that balances the reliability and efficiency of using LLMs as judges. SAGE employs two primary judges for initial assessments and invokes a third judge only when disagreements occur. By minimizing redundant calls in the fixed majority-based voting, SAGE reduces computational overhead by 80–95% while achieving near-human alignment (Macro-F1: 0.95–0.98). Our key contributions include: 1) introducing LLM-as-a-judge for objective evaluation with reference-guided assessment, 2) proposing selective aggregation that maintains evaluation quality while achieving substantial efficiency, 3) comprehensive empirical validation across five free-form QA datasets and multiple state-of-the-art models, and 5) systematic analysis of LLM-as-judge failure cases.

2 Methodology

This section briefly describes the key components of our proposed framework.

2.1 Candidate LLMs

A candidate LLM $\mathcal{C}_{\text{llm}}$ generates output $\bar{y}$ for the given input x.

2.2 LLMs-as-a-Judge

A judge $\mathcal{J}_{\text{llm}}$ LLM delivers evaluation or verdict V on candidate LLMs $\mathcal{C}_{\text{llm}}$ outputs $\bar{y}$. The $\mathcal{J}_{\text{llm}}$ evaluates output when prompted with x (i.e., $x \rightarrow \mathcal{C}_{\text{llm}}$) and $\bar{y}$. We utilize the reference answer r and prompt P the $\mathcal{J}_{\text{llm}}$ as:

$$P = \{x, \bar{y}, r\}$$

Utilizing P, $\mathcal{J}_{\text{llm}}$ performs the evaluation and delivers a decision as $V = J(P)$. The structure of V depends on the instructions provided in P. For instance, if a binary V is required, J assesses whether $\bar{y}$ is aligned with r given the context x and returns `True` if $\bar{y}$ is deemed correct, or `False` if it is not. The evaluation P may vary from zero-shot, where $\mathcal{J}_{\text{llm}}$ receives no prior examples, to few-shot, which includes several related examples, or a chain of thought, encouraging $\mathcal{J}_{\text{llm}}$ to reason stepwise through the problem.

2.3 Selective Aggregation for Generative Evaluation (SAGE)

In traditional human evaluation settings, when two annotators disagree on a judgment, a third expert is often called upon to resolve the dispute. Drawing inspiration from this efficient practice, we propose SAGE. Rather than immediately employing three LLMs (Badshah and Sajjad, 2024; Verga et al.,

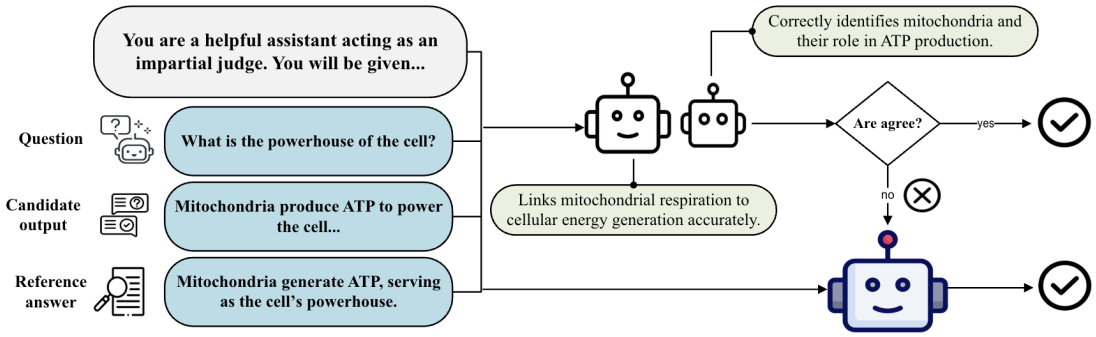

Figure 1: **Our proposed Selective Aggregation for Generative Evaluation (SAGE).** Two primary judges, J_1 and J_2, first provide verdicts V_{i_1} and V_{i_2} for an instance i. If agree, that consensus V_i is the final decision D_i. If disagree, a third model J_t independently produces a verdict V_t. The final decision D_i is then determined via majority voting among $\{V_{i_1}, V_{i_2}, V_t\}$.

2024), SAGE adopts an efficient approach by beginning with two open-source models as primary judges. When these judges reach a consensus, no further evaluation is needed. Only in cases of disagreement, the third LLM is engaged, whose decision then creates a majority verdict. This selective approach maintains evaluation quality while minimizing reliance on expensive models (e.g., GPT-4).

Formally, let V_{i_1} and V_{i_2} denote the verdicts from the two primary judges for the i-th evaluation instance. We define the agreement status A_i as:

$$A_i = \begin{cases} 1 & \text{if } V_{i_1} = V_{i_2}, \\ 0 & \text{otherwise.} \end{cases}$$

If $A_i = 1$, the final decision D_i is simply V_i, the agreed-upon verdict of the primary judges. If $A_i = 0$, a third model provides an additional verdict V_t. The final decision D_i is then obtained via majority voting among $\{V_{i_1}, V_{i_2}, V_t\}$. Formally:

$$D_i = \begin{cases} V_i & \text{if } A_i = 1, \\ \text{majority}(\{V_{i_1}, V_{i_2}, V_t\}) & \text{if } A_i = 0. \end{cases}$$

The majority operation selects the verdict that appears at least twice among $\{V_{i_1}, V_{i_2}, V_t\}$. Since there are three votes, at least two must coincide for a majority.

2.4 Judges Inclusion and Exclusion Criteria

To systematically select suitable judges for SAGE, we evaluate various LLMs (see Figure 2) using 100 random instances from HotpotQA. For each model, we compare binary verdicts against human annotations and compute Cohen's Kappa (κ) and Macro F1. We interpret κ following the commonly used guideline where values between 0.61–0.80 indicate *substantial agreement*, and values above 0.80 indicate *near-perfect agreement* (McHugh, 2012). However, since κ is known to be sensitive to class imbalance (Cicchetti and Feinstein, 1990), we jointly consider Macro F1 to ensure balanced evaluation across both classes. Formally:

$$\text{status}(\mathcal{J}) = \begin{cases} (V_{i_1} \;\&\; V_{i_2}) & \text{if } \kappa \geq 0.6 \\ & \wedge\, \text{F1} \geq 0.85, \\ V_t & \text{if } \kappa \geq 0.8 \\ & \wedge\, \text{F1} \geq 0.9, \\ \text{excluded} & \text{otherwise.} \end{cases}$$

Here, κ and F1 represent agreement metrics between judge $\mathcal{J}$ and the human majority.

3 Experiments

We utilize the following settings to examine the performance and reliability of individual LLM judges and SAGE.

3.1 Models

We select open and closed-source instruct models to serve as candidates in our experiment. These include Llama-3.1 70B[1] (Meta AI, 2024), GPT-3.5-turbo (Brown et al., 2020), Mistral 7B[2] (Jiang et al., 2023), Mixtral 8x7B[3] (Jiang et al., 2024) and DeepSeek-R1 (Team, 2025).

Based on our criteria for judges selection in Section 2.4, we found that Mistral 7B consistently met the required agreement thresholds ($\kappa \geq 0.6$,

[1] https://huggingface.co/meta-llama/Meta-Llama-3.1-70B-Instruct

[2] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3

[3] https://huggingface.co/mistralai/Mixtral-8x7B-Instructv0.1

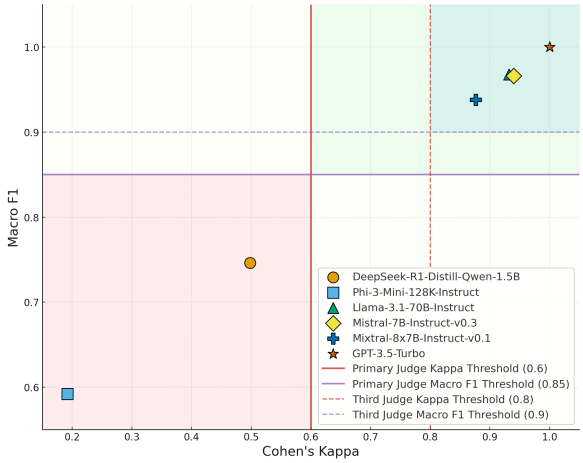

Figure 2: Judges selection based on the defined criteria in Section 2.4.

F1 ≥ 0.85) while offering the advantages of lower computational cost and faster inference. Llama 3.1 70B, as a much larger model, demonstrated higher overall agreement but at greater computational expense. By choosing both a strong lightweight model (Mistral 7B) and a state-of-the-art large model (Llama 3.1 70B) as primary judges, we cover a meaningful spectrum of capabilities rather than arbitrarily selecting models of different sizes (Feng et al., 2025; Liang et al., 2024; Sun et al., 2024). For the third judge role, we chose GPT-3.5-turbo because it exceeded the stricter agreement criteria ($\kappa \geq 0.8$, F1 ≥ 0.9), is widely accessible via API, and offers favorable pricing compared to larger closed models such as GPT-4o (see Figure 2). In addition, we utilize GPT-4o (Team, 2023) and DeepSeek-R1 (Team, 2025) in our ablation experiments.

To ensure the reproducibility of our experiments, we set the temperature to 0 for all models under study, as the performance of LLM-based evaluators has been shown to drop when temperature increases (Hada et al., 2024).

3.2 Datasets

We focus on free-form question-answering (QA) since it has widespread practical applications and the critical importance of truthfulness in this domain (Gou et al., 2024a; Evans et al., 2021). We utilize five free-form QA datasets: AmbigQA (Min et al., 2020), FreshQA (Vu et al., 2023), HotpotQA (Yang et al., 2018), Natural Questions (Kwiatkowski et al., 2019), and TriviaQA (Joshi et al., 2017). See Appendix A for details.

3.3 Prompts

We design generalized (i.e., with minimum instructions) zero-shot prompts with role-playing (Kong et al., 2024) for both candidates and judges. Initially, we prompt candidate LLMs to elicit outputs for the given random samples associated with each dataset.

To evaluate the outputs of candidate LLMs, we prompt judge LLMs for binary verdicts (i.e., True or False) using $P = \{x, \bar{y}, r\}$ and instructed to provide a brief explanation for their verdicts (see Appendix E for examples). Binary verdicts explicitly differentiate between correct and incorrect answers, minimize subjective interpretations, and simplify the evaluation process, thus facilitating automatic evaluation. We chose not to use few-shot or chain-of-thought prompting strategies to keep the solution robust to a variety of tasks. Previous studies have also shown that in-context examples do not significantly improve the performance of model-based evaluators (Hada et al., 2024; Min et al., 2022).

3.4 Baselines

We compare individual LLM judges, and specifically, SAGE, against the following baseline (details in Appendix A): (1) Exact Match (EM), adapted for verbose LLM outputs by marking an answer correct if any gold span appears anywhere in the response; (2) BERTScore (Zhang et al., 2020), computed with microsoft/deberta-xlarge-mnli and thresholded at $\tau = 0.5$ to yield binary decisions; (3) Majority Voting, which always invokes a fixed trio of LLM judges and returns the model verdict; (4) Self-Consistency (Wang et al., 2023b), which samples five judgments from the same model (T = 0.7) and takes their majority; and (5) Prometheus 2 (Kim et al., 2024), we run in its reference-based "direct assessment" using GPT-4 Turbo.

Human Evaluation: We recruit three graduate students from our academic network, all specialized in natural language processing, to serve as annotators. We provide the input given to the candidate LLMs, reference answers, and candidate LLMs responses. This format, while similar, is distinct from the judge LLMs prompts which additionally require formatted decisions. We anonymize the origin of model responses to reduce potential bias linked to model familiarity or reputation. The annotators are asked to score the candidate LLMs outputs on a binary scale: '1' for 'True' and '0'

for 'False' based on alignment with the reference answer and relevance (see Appendix C for details).

4 Evaluation Metrics

We compute **Fleiss' Kappa** (κ) (Fleiss and Cohen, 1973) and **percent agreement** to assess inter-rater reliability among human annotators. Similarly, we use **Cohen's kappa** (McHugh, 2012) to find the agreement between each evaluator and the human majority to obtain instance-level comparison. Due to the high-class imbalance in TriviaQA, kappa scores can be misleadingly low despite high raw agreement - a known limitation called the *"kappa paradox"* (Cicchetti and Feinstein, 1990). Therefore, we treat the evaluation as a binary classification task where we consider each evaluator's predictions against the human majority and report **Macro-F1** scores which give equal weight to both classes regardless of their frequency in the selected random samples.

To quantify the efficiency of our selective aggregation, we report the **disagreement rate** between the two primary judges that indicates how often the third model is required, thereby revealing the reduction in third-model usage compared to always employing three judges. Formally,

$$\text{Disagreement rate (\%)} = \left(\frac{1}{N} \sum_{i=1}^{N} \mathbb{I}[V_{i_1} \neq V_{i_2}] \right) \times 100$$

where N is the total number of evaluation instances and $\mathbb{I}[\cdot]$ is the indicator function that equals 1 when the condition is satisfied and 0 otherwise.

5 Results

In this section, we briefly report the experimental results and refer the reader to Appendix D for detailed results.

5.1 Alignment with Human Evaluation

As evidenced by consistently high Cohen's kappa and Macro F1 scores in Table 1 and 2, SAGE maintains a strong alignment with human evaluation. This represents a substantial improvement over individual model performance, where individual judges generally showed varying levels of agreement with human evaluation. Overall, LLM-as-a-judge works better with larger models. This is particularly noticeable in Llama and GPT, which achieve greater performance across AmbigQA, HotpotQA, and NQ-Open compared to smaller models. This indicates an important scaling law in evaluation capability (Kaplan et al., 2020; Zheng et al., 2024; Team, 2024). However, we also found that the most advanced models are not always guaranteed to be the best evaluators. We observed slightly comparable performance through the small open-source Mistral7B. For instance, when evaluating candidate Mixtral 8x7B on AmbigQA (see Table 2), Mistral 7B as-a-judge outperformed (0.944) judge GPT-3.5-turbo (0.891). Regardlesss, we observe relatively lower Macro-F1 scores for all LLM judges in TriviaQA.

Interestingly, lexical matching EM typically accomplishes better alignment with human evaluation on the instance-level in Table 2 than neural-based BERTScore. EM's strict and conservative nature leads to lower overall performance, but its high-precision characteristics ensure that when it identifies a match, it strongly aligns with human annotations. In contrast, BERTScore takes a more lenient approach to semantic matching. Although

		Evaluators						
LLMs	**Tasks**	**EM**	**BS**	**Llama**	**GPT**	**Mixtral**	**Mistral**	**SAGE**
Llama	AmbigQA	0.518	0.283	0.888	0.844	0.824	0.858	0.911
	HotpotQA	0.577	0.498	0.877	0.899	0.820	0.832	0.953
	NQ-Open	0.381	0.437	0.833	0.793	0.816	0.738	0.927
	TriviaQA	0.281	0.564	0.547	0.439	0.396	0.299	0.684
GPT	AmbigQA	0.561	0.252	0.944	0.897	0.861	0.853	0.967
	HotpotQA	0.604	0.300	0.953	0.973	0.873	0.933	0.987
	NQ-Open	0.453	0.218	0.884	0.824	0.824	0.829	0.956
	TriviaQA	0.335	0.364	0.650	0.401	0.580	0.467	0.775
Mixtral	AmbigQA	0.546	0.337	0.896	0.781	0.909	0.887	0.951
	HotpotQA	0.546	0.349	0.940	0.933	0.859	0.940	0.973
	NQ-Open	0.371	0.301	0.879	0.728	0.899	0.815	0.913
	TriviaQA	0.317	0.390	0.625	0.605	0.678	0.436	0.764
Mistral	AmbigQA	0.599	0.254	0.893	0.893	0.893	0.860	0.953
	HotpotQA	0.605	0.383	0.937	0.902	0.895	0.937	0.958
	NQ-Open	0.484	0.291	0.851	0.838	0.878	0.840	0.953
	TriviaQA	0.467	0.239	0.758	0.725	0.645	0.470	0.854

Table 1: Cohen's Kappa scores displaying the agreement levels of individual and multiple (SAGE) evaluators with human judgments across candidate models and tasks.

		Evaluators						
LLMs	**Tasks**	**EM**	**BS**	**Llama**	**GPT**	**Mixtral**	**Mistral**	**SAGE**
Llama	AmbigQA	0.744	0.641	0.944	0.922	0.912	0.929	0.955
	HotpotQA	0.778	0.745	0.939	0.949	0.910	0.916	0.976
	NQ-Open	0.653	0.718	0.916	0.896	0.907	0.869	0.964
	TriviaQA	0.612	0.782	0.772	0.717	0.695	0.640	0.842
GPT	AmbigQA	0.792	0.622	0.972	0.949	0.930	0.927	0.984
	HotpotQA	0.794	0.623	0.977	0.987	0.936	0.966	0.993
	NQ-Open	0.703	0.606	0.942	0.911	0.911	0.914	0.978
	TriviaQA	0.646	0.681	0.824	0.700	0.789	0.730	0.887
Mixtral	AmbigQA	0.760	0.666	0.948	0.891	0.955	0.944	0.975
	HotpotQA	0.761	0.657	0.970	0.966	0.930	0.970	0.987
	NQ-Open	0.650	0.649	0.939	0.863	0.950	0.908	0.956
	TriviaQA	0.625	0.695	0.812	0.803	0.838	0.716	0.882
Mistral	AmbigQA	0.792	0.622	0.947	0.947	0.947	0.930	0.977
	HotpotQA	0.796	0.673	0.969	0.951	0.947	0.969	0.979
	NQ-Open	0.726	0.639	0.925	0.919	0.939	0.920	0.976
	TriviaQA	0.718	0.608	0.879	0.863	0.822	0.735	0.927

Table 2: Macro-F1 scores of individual and multiple (SAGE) evaluators applied to different candidate LLMs and associated tasks.

this leniency produces higher raw scores, it introduces more false positives, consequently reducing instance-level agreement with human judgments.

5.2 Selective Aggregation vs. Majority Voting

Our selective aggregation approach provides evaluation quality comparable to full majority voting, while substantially reducing computational cost. As presented in Table 3, SAGE matches or closely approaches the Macro F1 and Cohen's Kappa scores of the three-judge majority across almost all tasks and candidate LLMs. For example, on HotpotQA, evaluating candidate Llama with SAGE achieves a Macro F1 of 97.6% (compared to 97.6% for majority voting) and a Cohen's Kappa of 0.95, while for GPT-3.5 on AmbigQA, SAGE reaches a Macro F1 of 98.4% (versus 98.3% for majority voting). By invoking the third judge only when disagreements occur, SAGE reduces usage by roughly 80–95% (averaging about 88%) compared to always-on majority voting across tasks, making it practical for large-scale deployments (see Table 11).

5.3 SAGE vs. Prometheus 2

As given in Table 4, SAGE outperformed the fine-tuned Prometheus 2 on every LLM–task pair. Crucially, SAGE achieves these gains while calling a third judge selectively, whereas Prometheus 2 runs a GPT-4 Turbo, after fine-tuning with scalar ratings.

5.4 SAGE vs. Self-consistency

As shown in Table 5, SAGE consistently outperforms self-consistency (Wang et al., 2023b) across all tasks and LLMs. While self-consistency relies on five model calls per instance to stabilize decisions, SAGE achieves higher accuracy with fewer calls, invoking a third judge only when needed. This demonstrates that leveraging diverse models is not only more effective but also substantially more efficient than repeated sampling from a single model.

5.5 Evaluation with One Strong LLM-as-a-judge

While a single state-of-the-art (i.e., based on it's leaderboard performance) evaluator can achieve strong performance in many cases, the dual-LLM framework remains critical for ensuring robustness, particularly in high-stakes or ambiguous scenarios.

LLMs	Tasks	Majority Voting		Disagr. (%)	SAGE	
		Macro F1	Kappa		Macro F1	Kappa
Llama	AmbigQA	95.5	0.91	10.0	95.5	0.91
	HotpotQA	97.6	0.95	13.0	97.6	0.95
	NQ-Open	96.3	0.93	18.0	96.4	0.92
	TriviaQA	84.1	0.68	17.0	84.2	0.68
GPT	AmbigQA	98.3	0.97	7.0	98.4	0.96
	HotpotQA	99.3	0.99	5.7	99.3	0.98
	NQ-Open	97.8	0.96	13.0	97.8	0.95
	TriviaQA	90.5	0.81	15.7	88.7	0.77
Mixtral	AmbigQA	98.9	0.98	9.0	97.5	0.95
	HotpotQA	98.6	0.97	4.7	98.7	0.97
	NQ-Open	98.3	0.97	13.0	95.6	0.91
	TriviaQA	95.0	0.90	17.0	88.2	0.76
Mistral	AmbigQA	97.6	0.95	11.7	97.7	0.95
	HotpotQA	97.9	0.96	6.0	97.9	0.95
	NQ-Open	97.6	0.95	14.7	97.6	0.95
	TriviaQA	93.5	0.87	20.3	92.7	0.85

Table 3: Comparison between Majority Voting (Llama+GPT-3.5+Mistral) and SAGE. Disagr. refers to disagreement.

LLMs	Tasks	Prometheus	SAGE
Llama	AmbigQA	0.894	0.955
	HotpotQA	0.891	0.976
	NQ-Open	0.855	0.964
	TriviaQA	0.804	0.842
GPT	AmbigQA	0.937	0.984
	HotpotQA	0.942	0.993
	NQ-Open	0.843	0.978
	TriviaQA	0.796	0.887

Table 4: Macro-F1 comparison between Prometheus 2 and SAGE in the reference-based setting. See Table 10 for complete results.

To explore the potential of a more powerful single LLM, we evaluated GPT-3.5-turbo on HotpotQA and TriviaQA using GPT-4o as a judge. With this configuration, GPT-4o as the evaluator achieved a Macro-F1 score of 0.946 on HotpotQA, demonstrating its exceptional capability. However, the same GPT-4o judge achieved only 0.784 on TriviaQA, which falls short of SAGE's performance of 0.887. This shows that even the most advanced models show inconsistencies when evaluating free-form QA. This is particularly critical in precision-sensitive domains where minor errors can have outsized consequences.

5.6 Analysis

In our main experiments, candidate LLMs generated 6000 outputs for the given tasks, with each evaluator producing corresponding evaluations. We randomly sampled 100 error cases (50 false positives and 50 false negatives) from each evaluator to understand their behavior. Given EM had only 11 false positives, we included all of them in our analysis. Due to length constraints, we moved the detailed analysis of EM and BERTScore to Ap-

LLMs	Tasks	SC (Llama)	SC (GPT)	SAGE
Llama	AmbigQA	0.933	0.914	0.955
	HotpotQA	0.925	0.896	0.976
Mistral	AmbigQA	0.957	0.943	0.977
	HotpotQA	0.965	0.922	0.979

Table 5: Comparison of SAGE and Self-Consistency (SC) Macro F1. SC results are based on 5 samples per instance.

pendix D and focused exclusively on the LLM-as-a-judge method here.

LLM-based evaluators demonstrate strong abilities in recognizing semantic variations while maintaining the core meaning, especially when assessing responses that use different terminology or structural approaches to convey the same information. For instance, in the evaluation examples, evaluators correctly identified that *"Salma Hayek"* and *"Salma Hayek Pinault"* refer to the same individual, acknowledging the semantic equivalence despite differences in phrasing. Similarly, when assessing responses that use different terms for the same entity, such as recognizing *"Nick Fury, Agent of S.H.I.E.L.D."* as part of the broader *"Marvel"* universe, the evaluators effectively maintain the core meaning and contextual relevance. Their explanations show systematic assessment patterns that combine multiple evaluation criteria including factual accuracy, logical coherence, and contextual relevance.

LLMs are prone to hallucination in justification (Zhang et al., 2023), where they fabricate reasoning to support their evaluations, produce detailed but incorrect explanations, or reference nonexistent criteria or standards. In LLM judges, false positives and negatives often result from overlooking critical distinctions between candidate LLM outputs and failing to account for the specificity required by the reference answer. This pattern is particularly noticeable in Mistral 7B, where the model disregards the ground truth and provides evaluations influenced by unknown factors. For example, when evaluating candidate GPT-3.5's response *"The foreign minister of Germany who signed the Treaty of Versailles was Hermann Müller."* which is correct according to the reference answer *"Hermann Müller"* and human evaluation, Mistral 7B as-a-judge incorrectly marked this response as false and fabricated reasoning *"Hermann Müller was the Chancellor of Germany, not the Foreign Minister. The Foreign Minister of Germany who signed the Treaty of Versailles was Gustav Stresemann."* in support of its decision. The same problem can also be attributed to inconsistent evaluations. Because when Mistral 7B acted as a candidate for the same question, its response to the question is completely different: *"The Treaty of Versailles was signed by Matthias Erzberger, a German politician who served as the President of the German National Assembly at the time"*. There are also alternative interpretations of this issue, such as ambiguity in the question, but we leave a deeper exploration of these aspects to future work.

We observe a different pattern in some judges, specifically, GPT-3.5 and Mixtral 8x7B which focuses more on specificity. This approach shifts the evaluation towards false negatives by missing semantically similar but structurally different answers. We found many cases when such evaluators failed to account for valid variations in phrasing or granularity, focusing instead on rigid adherence to the reference answer. Compounding these issues are reasoning errors within the evaluators' own explanations, which often contain fabrications, circular logic, or overconfident assertions. By insisting on correctness derived strictly from the reference, evaluators disregard valid alternative perspectives and can even mischaracterize or invert the facts in their attempts to justify their decisions. This dynamic leaves little room for nuance or ambiguity, and it pushes the evaluation process away from fair, context-sensitive assessment toward rigid, and sometimes inaccurate, verdicts.

We found several temporal limitations in LLM-based evaluators. Although most of our datasets are older and the evaluator models are relatively up-to-date, we still observed instances where references to recent events, newly emerging terminology, or evolving contexts were misinterpreted. The FreshQA dataset (Vu et al., 2023), being recent, serves as a valuable testbed for assessing these temporal deficiencies. As shown in Table 6, LLM-based evaluators indicate deviation from human judgment on FreshQA compared to tasks that rely on older information, such as HotpotQA. Specifically, in dynamic or time-sensitive contexts, we found that LLM judges tend to hallucinate by consistently classifying candidate model responses as True, even when incorrect. For example, when presented with the question: *"On what date did the Patriots last play the Miami Dolphins?"* the LLM-generated response states: ‘*'The last time the*

	Evaluators				
LLMs	**Llama**	**GPT**	**Mixtral**	**Mistral**	**SAGE**
Llama	0.835	0.737	0.817	0.730	0.917
GPT	0.695	0.824	0.780	0.746	0.891
Mixtral	0.708	0.779	0.738	0.703	0.936
Mistral	0.665	0.802	0.818	0.723	0.880

Table 6: Performance (in Macro F1) of individual and multiple LLM judges on FreshQA.

New England Patriots played the Miami Dolphins was on January 1, 2023, during the NFL regular season." Despite the correct reference answer being *"November 24, 2024"* the LLM evaluator not only failed to recognize the inaccuracy but also hallucinated an erroneous justification, stating: *"The proposed answer correctly states the date the New England Patriots last played the Miami Dolphins as January 1, 2023, which matches the information provided."*

6 Related work

Free-form question-answering has traditionally relied on supervised signal-based metrics such as EM. Despite its simplicity and efficiency, EM overlooks semantically equivalent variations and often penalizes accurate responses that use different phrasing (Wang et al., 2024a; Kamalloo et al., 2023). Other commonly used metrics, including BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) primarily focus on n-gram overlap with reference texts. Although widely used, these metrics often fail to recognize correct answers that use different wording or sentence structure than the reference, limiting their ability to evaluate free-form responses accurately (Zhang et al., 2020).

Contextual metrics such as BERTScore (Zhang et al., 2020) and the learned regressor BLEURT (Sellam et al., 2020) increases robustness by comparing dense embeddings rather than n-grams. However, even BERTScore and similar embedding-based methods struggle to effectively evaluate open-ended generation (Zheng et al., 2024; Sun et al., 2022). More importantly, such methods return continuous similarity scores which are not well-suited for binary evaluations where a clear true or false decision is required.

Recent advances in LLMs have unlocked new opportunities for automatic and context-aware evaluation (Li et al., 2024b; Chiang and Lee, 2023; Zheng et al., 2024). A key strength of LLM-based evaluators lies in their ability to operate in reference-free settings, where evaluation does not rely on predefined answers but instead leverages subjective criteria such as helpfulness, relevance, and coherence. This capability makes LLM evaluators particularly well-suited for assessing tasks where multiple valid responses exist or where human-like judgment is required (Li et al., 2024a). For instance, LLMs are frequently used in subjective evaluations such as pairwise comparison or single-response scoring (Verga et al., 2024; Chan et al., 2024). LLM-based evaluators are specifically effective for tasks like summarization, where subjective criteria are central to evaluation (Liu et al., 2023). However, subjective evaluations are less useful for evaluating objective tasks such as free-form QA, where responses are either correct or incorrect and require explicit verification against reference answers.

LLM-based evaluators face several challenges, particularly in ensuring consistency and fairness (Ye et al., 2024; Khan et al., 2024). Recent studies recruit multiple LLMs and aggregate their votes to tackle such challenges. Self-consistency (Wang et al., 2023b) involves running the same model k times with diverse reasoning paths (e.g., $k = 40$) and then applying majority voting to the outputs, improving reliability but significantly increasing inference cost. Similarly, both the Reference-Guided Verdict (Badshah and Sajjad, 2024) and PoLL (Verga et al., 2024) methods employ multiple diverse LLM evaluators to mitigate intra-model bias and enhance alignment with human judgments. However, these studies rely on fixed majority voting across all instances, which increases unnecessary computational requirements.

7 Conclusion

We present SAGE, a framework designed to evaluate free-form question-answering by leveraging LLMs. Our findings demonstrate that individual LLM judges are reliable alternatives to traditional lexical and neural-based metrics, offering substantial alignment with human evaluations. However, relying solely on individual judges poses challenges, including inherent biases, inconsistencies, and prompt sensitivity, which can affect evaluation performance. SAGE addresses these limitations by combining the strengths of multiple LLMs through selective aggregation. It significantly improves evaluation accuracy over individual judges, while reducing redundancy and cost compared to full majority voting.

8 Limitations

We acknowledge certain limitations in our study. The accuracy of evaluations depends on the quality and clarity of reference answers, which serve as the basis for determining correctness. Incorrect or ambiguous references could affect evaluation outcomes. Similarly, this study primarily uses binary verdicts, which overlook detailed aspects of responses that could be captured through more comprehensive evaluation criteria. Furthermore, while we conducted an error analysis of LLM judges and automatic metrics, there may be error cases that were not identified during our manual review, leaving gaps in understanding the full spectrum of evaluation inaccuracies. From our ablation experiments, we found SAGE performs worst in reference-free settings. Thus, in the future, we aim to explore SAGE with LLM agents for automatic and reference-free evaluation, instead of relying on human-annotated dataset-specific reference answers.

References

Vaibhav Adlakha, Parishad BehnamGhader, Xing Han Lu, Nicholas Meade, and Siva Reddy. 2024. Evaluating correctness and faithfulness of instruction-following models for question answering. *Transactions of the Association for Computational Linguistics*, 12:681–699.

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

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, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2024. Chateval: Towards better LLM-based evaluators through multi-agent debate. In *The Twelfth International Conference on Learning Representations*.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45.

Yihan Chen, Benfeng Xu, Quan Wang, Yi Liu, and Zhendong Mao. 2024. Benchmarking large language models on controllable generation under diversified instructions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17808–17816.

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.

Zhumin Chu, Qingyao Ai, Yiteng Tu, Haitao Li, and Yiqun Liu. 2024. Pre: A peer review based large language model evaluator.

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.

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021. All that's 'human' is not gold: Evaluating human evaluation of generated text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7282–7296, Online. Association for Computational Linguistics.

Ehsan Doostmohammadi, Oskar Holmström, and Marco Kuhlmann. 2024. How reliable are automatic evaluation methods for instruction-tuned llms? *arXiv preprint arXiv:2402.10770*.

Owain Evans, Owen Cotton-Barratt, Lukas Finnveden, Adam Bales, Avital Balwit, Peter Wills, Luca Righetti, and William Saunders. 2021. Truthful ai: Developing and governing ai that does not lie.

Shangbin Feng, Wenxuan Ding, Alisa Liu, Zifeng Wang, Weijia Shi, Yike Wang, Zejiang Shen, Xiaochuang Han, Hunter Lang, Chen-Yu Lee, Tomas Pfister, Yejin Choi, and Yulia Tsvetkov. 2025. When one llm drools, multi-llm collaboration rules.

Joseph L Fleiss and Jacob Cohen. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement*, 33(3):613–619.

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

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

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?

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding.

Xinyu Hu, Mingqi Gao, Sen Hu, Yang Zhang, Yicheng Chen, Teng Xu, and Xiaojun Wan. 2024. Are llm-based evaluators confusing nlg quality criteria?

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880, Online. Association for Computational Linguistics.

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.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts.

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

Jaehun Jung, Faeze Brahman, and Yejin Choi. 2024. Trust or escalate: Llm judges with provable guarantees for human agreement.

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.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models.

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.

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.

Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, Xin Zhou, Enzhi Wang, and Xiaohang Dong. 2024. Better zero-shot reasoning with role-play prompting.

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.

Md Tahmid Rahman Laskar, Sawsan Alqahtani, M Saiful Bari, Mizanur Rahman, Mohammad Abdullah Matin Khan, Haidar Khan, Israt Jahan, Amran Bhuiyan, Chee Wei Tan, Md Rizwan Parvez, Enamul Hoque, Shafiq Joty, and Jimmy Huang. 2024. A systematic survey and critical review on evaluating large language models: Challenges, limitations, and recommendations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13785–13816, Miami, Florida, USA. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. 2024a. From generation to judgment: Opportunities and challenges of llm-as-a-judge.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024b. Llms-as-judges: A comprehensive survey on llm-based evaluation methods.

Junlong Li, Shichao Sun, Weizhe Yuan, Run-Ze Fan, Hai Zhao, and Pengfei Liu. 2023. Generative judge for evaluating alignment.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2024. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17889–17904, Miami, Florida, USA. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

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, Singapore. Association for Computational Linguistics.

Yuxuan Liu, Tianchi Yang, Shaohan Huang, Zihan Zhang, Haizhen Huang, Furu Wei, Weiwei Deng, Feng Sun, and Qi Zhang. 2024. Calibrating LLM-based evaluator. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2638–2656, Torino, Italia. ELRA and ICCL.

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.

Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282.

Meta AI. 2024. Introducing meta llama 3: The most capable openly available llm to date. Meta AI Blog. Accessed: 2024-07-25, 12:14:31 p.m.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Keita Saito, Akifumi Wachi, Koki Wataoka, and Youhei Akimoto. 2023. Verbosity bias in preference labeling by large language models.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Chenglei Si, Chen Zhao, and Jordan Boyd-Graber. 2021. What's in a name? answer equivalence for open-domain question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9623–9629, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Guangzhi Sun, Anmol Kagrecha, Potsawee Manakul, Phil Woodland, and Mark Gales. 2024. Skillaggregation: Reference-free llm-dependent aggregation. *arXiv preprint arXiv:2410.10215*.

Tianxiang Sun, Junliang He, Xipeng Qiu, and Xuanjing Huang. 2022. BERTScore is unfair: On social bias in language model-based metrics for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3726–3739, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

DeepSeek-AI Team. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.

OpenAI Team. 2023. Gpt-4 technical report.

OpenAI Team. 2024. Gpt-4 technical report.

Aman Singh Thakur, Kartik Choudhary, Venkat Srinik Ramayapally, Sankaran Vaidyanathan, and Dieuwke Hupkes. 2024. Judging the judges: Evaluating alignment and vulnerabilities in llms-as-judges.

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.

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.

Tu Vu, Kalpesh Krishna, Salaheddin Alzubi, Chris Tar, Manaal Faruqui, and Yun-Hsuan Sung. 2024. Foundational autoraters: Taming large language models for better automatic evaluation.

Cunxiang Wang, Sirui Cheng, Qipeng Guo, Yuanhao Yue, Bowen Ding, Zhikun Xu, Yidong Wang, Xiangkun Hu, Zheng Zhang, and Yue Zhang. 2024a. Evaluating open-qa evaluation. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023a. Large language models are not fair evaluators.

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.

Yuqi Wang, Lyuhao Chen, Songcheng Cai, Zhijian Xu, and Yilun Zhao. 2024b. Revisiting automated evaluation for long-form table question answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14696–14706, Miami, Florida, USA. Association for Computational Linguistics.

Wenda Xu, Danqing Wang, Liangming Pan, Zhenqiao Song, Markus Freitag, William Wang, and Lei Li. 2023. INSTRUCTSCORE: Towards explainable text generation evaluation with automatic feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5967–5994, Singapore. Association for Computational Linguistics.

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.

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.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023. Siren's song in the ai ocean: A survey on hallucination in large language models.

Yue Zhang, Ming Zhang, Haipeng Yuan, Shichun Liu, Yongyao Shi, Tao Gui, Qi Zhang, and Xuanjing Huang. 2024. Llmeval: A preliminary study on how to evaluate large language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):19615–19622.

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*.

A Free-form Question-Answering

In our experiments, we include AmbigQA (Min et al., 2020), FreshQA (Vu et al., 2023), HotpotQA (Yang et al., 2018), Natural Questions (Kwiatkowski et al., 2019), and TriviaQA (Joshi et al., 2017).

- **AmbigQA**: Focuses on 14K ambiguous questions derived from NQ, requiring systems to identify multiple valid interpretations and generate disambiguated questions alongside corresponding answers.
- **FreshQA**: A QA benchmark containing 600 questions that consist of a diverse range of types, including those requiring fast-changing world knowledge and questions with false premises that need debunking. It is regularly updated to reflect current information and is designed to evaluate the factual accuracy of LLMs in handling up-to-date and evolving knowledge.
- **HotpotQA**: Contains 113K questions based on Wikipedia. It is designed to test multi-hop reasoning, requiring connections across multiple paragraphs, and includes annotated supporting facts for evaluation.
- **Natural Questions (NQ)**: Consists of real user queries from Google Search, paired with

Wikipedia articles. The dataset includes 307K training examples annotated with both long (paragraph) and short (entity-level) answers.

- **TriviaQA**: Features approximately 650K trivia questions, with evidence sourced from Wikipedia and web searches. These questions often require reasoning across multiple documents for complex answer synthesis.

We utilize the validation splits across multiple datasets: the standard validation split for AmbigQA and Natural Questions, the "distractor" subset's validation split for HotpotQA, and the "unfiltered.nocontext" subset's validation split for TriviaQA. We randomly sampled 300 examples from each dataset using Seed 42.

B Baselines

Exact Match (EM): For our selected datasets and also free-form QA tasks, EM serves as a standard lexical matching metric to evaluate candidate LLM performance (Izacard and Grave, 2021; Lewis et al., 2020; Gou et al., 2024b). Due to the verbose nature of LLM-generated responses, we adapt EM to classify an answer as correct if any golden answer $r_i \in R$ appears within the generated response $\bar{y}$ (i.e., $r_i \subseteq \bar{y}$), rather than requiring complete strict string equality (i.e., $\bar{y} = r_i$).

BERTScore: We use BERTScore (Zhang et al., 2020) which measures similarity by comparing contextualized word embeddings derived from a pre-trained BERT model. This enables the evaluation to focus on semantic correctness rather than exact lexical matches. As BERTScore is based on continuous values between -1 and 1, we set a threshold of $\tau = 0.5$ to convert continuous similarity scores into binary 0 and 1. The purpose of this conversion is to allow direct comparison with other evaluation methods. For our implementation, we use the microsoft/deberta-xlarge-mnli[4] model (He et al., 2021).

Majority voting This uses three fixed LLM judges to independently evaluate each instance. The final decision is determined by a simple majority across the three verdicts. Unlike SAGE, which selectively invokes the third judge only in cases of disagreement, this method uniformly engages all judges, leading to higher computational cost.

[4] https://huggingface.co/microsoft/deberta-xlarge-mnli

Self-consistency For self-consistency (Wang et al., 2023b), we sample five outputs from the same judge model using a temperature of 0.7, and take the majority verdict across these samples as the final decision. This baseline evaluates the extent to which response stability from a single model can approximate consistent evaluation.

Prometheus 2 We tailored Prometheus (Kim et al., 2024) for the reference-based (direct-assessment) settings. We converted 5-point scores to a binary by assigning 0 to ratings 1-3 and 1 to ratings 4-5 so that it is directly comparable to SAGE.

C Human evaluation

This section provides detailed guidelines for human annotators responsible for evaluating the outputs of candidate LLMs. The goal is to ensure consistency and objectivity across all evaluations. These guidelines provide clear instructions for assessing each model's response based on its alignment with the reference answer and contextual relevance.

C.1 Guidelines

Dear Evaluator,
Thank you for your valuable contribution to this evaluation process. These guidelines outline the process for evaluating Large Language Model (LLM) outputs for the given tasks. As annotators, you will receive three components for each evaluation instance: the input question, reference answer(s), and the model's response. Your task is to evaluate the responses independently and score them on a binary scale: '1' for 'True' (correct) and '0' for 'False' (incorrect).
A response warrants a score of '1' when it demonstrates semantic equivalence with the reference answer, even if expressed through alternative phrasing or structure. This includes acceptable variations such as synonym usage and structural variations. Additional contextual information is acceptable as long as it doesn't introduce errors.
Responses receive a score of '0' when they contain factual errors, miss crucial elements from the reference answer, or demonstrate contextual misalignment. Partial answers that omit essential information should be marked incorrect, regardless of the accuracy of included content. When multiple reference answers are provided, a response is correct if it fully aligns with at least one reference.
You are encouraged to use internet resources when needed to verify specific facts, terminology, or po-

LLMs	AmbigQA	FreshQA	HotpotQA	NQ-Open	TriviaQA
DeepSeek	0.975	0.949	0.986	0.889	0.456 (κ paradox)
Llama	0.945	0.962	0.973	0.985	0.935
GPT	0.989	0.973	0.982	0.990	0.948
Mixtral	0.981	0.945	0.996	0.977	0.936
Mistral	0.978	0.932	0.981	0.978	0.975

Table 7: Fleiss' Kappa scores of human annotators across models and tasks.

LLMs	AmbigQA	FreshQA	HotpotQA	NQ-Open	TriviaQA
DeepSeek	99.0%	98.0%	99.7%	92.0%	90.0%
Llama	96.3%	98.0%	98.0%	99.0%	99.0%
GPT	99.3%	99.3%	98.7%	99.3%	99.0%
Mixtral	98.7%	98.0%	99.7%	98.3%	98.3%
Mistral	98.3%	97.0%	98.7%	98.3%	99.0%

Table 8: Human annotators percent agreement scores across candidate models and tasks.

tential synonyms that may affect your evaluation decision. However, the reference answer should remain the primary basis for evaluation. Focus on whether the model's response conveys the same core information as the reference answer. To maintain reliability, document any challenging cases requiring further discussion with other annotators.

C.2 Inter-human annotator agreement

We calculate Fleiss' Kappa (κ) (Fleiss and Cohen, 1973) and percent agreement to assess inter-rater reliability among human annotators.

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 is calculated as:

$$\text{Percent Agreement} = \left(\frac{\text{Agreements}}{\text{Total Annotations}}\right) \times 100$$

Table 7 and 8 show the inter-annotator agreement across models and tasks. The results demonstrate high reliability, with Fleiss' Kappa scores consistently above 0.93 for most tasks. The highest agreement is observed in Mixtral evaluations on HotpotQA ($\kappa = 0.996$), and GPT on NQ-Open ($\kappa = 0.990$). In FreshQA, which shows lower Kappa scores, the agreement among annotators remains high including 99.3% in GPT and 98.0% in Mixtral.

The percent agreement scores in Table 8 further confirm strong inter-annotator consistency. Most models achieve over 98% agreement across AmbigQA, HotpotQA, NQ-Open, and TriviaQA. However, DeepSeek exhibits lower agreement on NQ-Open (92.0%) and TriviaQA (90.0%). This indicates a variance in human ratings for these tasks.

D Additional results

This section provides further results and analysis of conventional metrics and LLM-based evaluators.

Table 9 illustrates the raw performance of Llama obtained through various evaluators. Unlike lexical matching and neural-based metrics, each LLM-as-a-judge shows overall performance close to the human majority. The proposed SAGE method consistently achieves comparable or slightly better alignment with the human majority. Conventional metrics such as EM severely underestimate the candidate LLMs' performance. Contrarily, BERTScore tends to overestimate the performance except in some cases such as when evaluating Llama on AmbigQA and NQ-Open (see Table 9 for additional results).

EM underestimates performance because it requires a candidate's response to exactly match one of the reference answers. This rigid, lexical approach fails to account for valid paraphrases, synonyms, or alternative expressions that convey the same meaning. In free-form QA tasks, where there can be multiple correct answers phrased in various ways, EM's strict criteria often penalize responses

that are semantically accurate but differ slightly in wording. As a result, it underestimates the true capabilities of candidate LLMs, leading to an incomplete assessment of their performance.

BERTScore relies on token-level semantic similarity, which rewards shallow lexical overlap rather than actual factual accuracy. For example, in cases where minor differences in wording (e.g., "The Treaty of Versailles was signed in 1919." versus "The Treaty of Versailles ended in 1919.") lead to opposing factual claims, BERTScore still scores the response high due to its emphasis on matching tokens (e.g., "signed" versus "ended"). Additionally, verbosity bias and threshold instability—where a default threshold (threshold = 0.5) is arbitrarily set—further inflate its raw accuracy. However, when comparing raw accuracy with instance-level agreement metrics like Cohen's kappa, which adjusts for class imbalance and penalizes asymmetric errors, the limitations of BERTScore become apparent.

D.1 Impact of selective third judge on disagreements

Figure 3 illustrates the impact of SAGE on resolving disagreements between primary judges. SAGE, facilitated by GPT-3.5 as the third judge, consistently improves performance across all tasks, particularly in FreshQA and TriviaQA, where Macro F1 increases by up to 21.5 points. In contrast, tasks like AmbigQA and HotpotQA, where primary judges initially exhibit stronger agreement, show smaller but still meaningful improvements. Notably, evaluations of DeepSeek-v3 show higher disagreement between Llama-3.1 70B and Mistral 7B, particularly in FreshQA (28.3%) and AmbigQA (25.7%). From our analysis, we did not find strong evidence explaining why DeepSeek-v3 leads to higher disagreement between the primary judges.

We observed substantial enhancements in Cohen's Kappa scores across several tasks. For instance, as illustrated in Figure 4, in the AmbigQA Cohen's Kappa increased from 0.881 to 0.911 for Llama. Similarly, in the same task, Cohen's Kappa from 0.467 to 0.773 for candidate DeepSeek. Some Cohen's Kappa scores remain relatively low, particularly in FreshQA and DeepSeek-evaluated outputs. This is partially explained by the Kappa Paradox, where high agreement on extreme cases (e.g., clear correct/incorrect responses) and unbalanced class distributions can artificially lower the Kappa scores. In such cases, even when evaluators mostly agree, Cohen's Kappa can appear lower than expected. Despite this, the SAGE process effectively mitigates inconsistencies, especially in tasks involving evolving knowledge and nuanced interpretations, such as FreshQA.

D.2 Cost analysis

To assess the efficiency of SAGE, we track the number of times the third judge is invoked, which directly corresponds to disagreement between the two primary judges. As shown in Table 11, across 7,500 evaluation instances, the third judge was required only 1,318 times, representing just 17.6% of the total cases. This implies an 82.4% reduction in third-judge usage compared to a full majority-voting setup, where every instance would involve all three models.

Disagreement rates vary across tasks and models. For example, GPT shows only 5.7% disagreement on HotpotQA, while FreshQA exhibits higher disagreement (up to 44.3%) for some judge combinations. This behavior allows SAGE to scale efficiently: it concentrates computational effort only where model uncertainty exists, minimizing redundant inference. In contrast to fixed-cost evaluation schemes, SAGE offers a cost-efficient alternative that maintains high evaluation quality while significantly reducing compute usage.

D.3 DeepSeek as the third judge

To assess the impact of using DeepSeek as the third judge in SAGE, we conducted experiments by replacing GPT-3.5-turbo with DeepSeek-R1 (Team, 2025). We evaluated this setup using different candidate models across multiple tasks. Specifically, we tested GPT-3.5 on TriviaQA, DeepSeek on NQ-Open, and Llama on FreshQA. The primary judges remained Llama and Mistral, and third was invoked only in cases of disagreement. Our findings indicate that DeepSeek as the selective third judge achieves strong performance, with Macro-F1 scores of 91.23 on TriviaQA, 79.11 on NQ-Open, and 0.914 on FreshQA.

D.4 Evaluating with GPT-4o as-a-judge

While a single state-of-the-art evaluator can achieve strong performance in many cases, the dual-LLM framework remains critical for ensuring robustness, particularly in high-stakes or ambiguous scenarios.

To explore the potential of a more powerful single LLM, we evaluated GPT-3.5-turbo on Hot-

LLMs	Tasks	Evaluators							
		EM	**BS**	**HM**	**DeepSeek**	**Llama**	**GPT**	**Mixtral**	**Mistral**
DeepSeek	AmbigQA	56.3	80.0	84.3	86.3	73.7	75.0	62.3	93.3
	FreshQA	31.3	88.0	84.3	84.7	82.7	75.3	58.0	82.3
	HotpotQA	38.6	78.4	57.7	58.0	51.0	51.0	52.7	57.7
	NQ-Open	35.0	78.3	60.3	64.7	63.7	61.3	55.3	68.3
	TriviaQA	77.3	90.7	94.3	90.7	94.0	91.7	81.7	89.7
Llama	AmbigQA	42.3	63.0	67.0	64.0	65.3	64.7	63.0	66.0
	FreshQA	25.6	81.3	77.7	81.3	78.3	72.7	71.0	62.3
	HotpotQA	34.3	67.7	56.3	56.7	58.3	54.0	50.7	52.7
	NQ-Open	31.7	61.7	66.3	62.3	62.7	60.0	59.0	66.7
	TriviaQA	74.3	94.0	94.7	88.0	90.3	90.0	88.7	84.7
GPT	AmbigQA	49.7	78.0	71.7	70.3	70.0	68.0	65.7	71.0
	FreshQA	24.6	89.3	70.7	58.0	51.7	78.7	83.0	83.3
	HotpotQA	33.7	80.0	54.0	50.3	53.0	52.7	51.7	54.0
	NQ-Open	36.3	74.0	65.3	65.3	62.7	59.0	59.0	67.0
	TriviaQA	74.3	95.3	93.0	90.0	89.3	90.7	89.7	86.3
Mixtral	AmbigQA	37.7	70.3	61.7	58.7	57.3	62.0	59.3	61.7
	FreshQA	18.6	89.7	86.0	72.3	67.0	87.0	85.0	77.7
	HotpotQA	25.0	69.7	47.0	46.3	45.3	45.7	44.7	46.0
	NQ-Open	23.7	63.7	56.7	54.0	52.7	47.7	52.3	59.7
	TriviaQA	64.7	91.3	90.7	83.7	86.3	89.7	86.0	85.3
Mistral	AmbigQA	31.0	61.7	49.7	47.7	46.3	47.7	46.3	53.3
	FreshQA	15.6	80.0	81.7	60.7	59.0	83.7	84.0	86.0
	HotpotQA	23.7	64.7	40.0	39.3	39.0	38.0	37.0	39.0
	NQ-Open	22.7	60.0	46.0	41.3	40.0	43.3	41.3	50.0
	TriviaQA	62.0	94.3	83.7	78.0	81.3	81.0	79.7	85.0

Table 9: Raw performance of candidate LLMs across free-form QA tasks evaluated through various methods. HM represents Human Majority and BS denotes BERTScore.

Candidate LLMs	Tasks	Prometheus 2	SAGE
Llama	AmbigQA	0.894	0.955
	HotpotQA	0.891	0.976
	NQ-Open	0.855	0.964
	TriviaQA	0.804	0.842
GPT	AmbigQA	0.937	0.984
	HotpotQA	0.942	0.993
	NQ-Open	0.843	0.978
	TriviaQA	0.796	0.887
Mixtral	AmbigQA	0.939	0.975
	HotpotQA	0.930	0.987
	NQ-Open	0.887	0.956
	TriviaQA	0.801	0.882
Mistral	AmbigQA	0.920	0.977
	HotpotQA	0.931	0.979
	NQ-Open	0.866	0.976
	TriviaQA	0.819	0.927

Table 10: Macro-F1 comparison between the fine-tuned Prometheus 2 evaluator and SAGE in the reference-based setting.

potQA and TriviaQA using GPT-4o as a judge. With this configuration, GPT-4o as the evaluator achieved a Macro-F1 score of 0.946 on HotpotQA, demonstrating its exceptional capability. However, the same GPT-4o judge achieved only 0.784 on TriviaQA, which falls short of SAGE's performance of 0.887. This shows that even the most advanced models show inconsistencies when evaluating free-form QA. This is particularly critical in precision-sensitive domains where minor errors can have outsized consequences.

In such settings, SAGE's ensemble approach acts as a safeguard. When employing SAGE with GPT-3.5-turbo as the selective third judge, we achieved an even higher Macro-F1 of 0.984 on HotpotQA, surpassing the performance of a single GPT-4o. Interestingly, when we experimented with DeepSeek as the third judge in SAGE, performance remained strong at 0.963 Macro-F1, indicating that SAGE's benefits are not solely tied to a specific third judge model.

D.5 Majority voting-based evaluation

We conducted additional experiments utilizing a traditional majority voting approach for evaluating candidate LLMs performance. Given n annotators and a binary classification, the majority label is defined as:

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

where y_i represents the label assigned by the ith

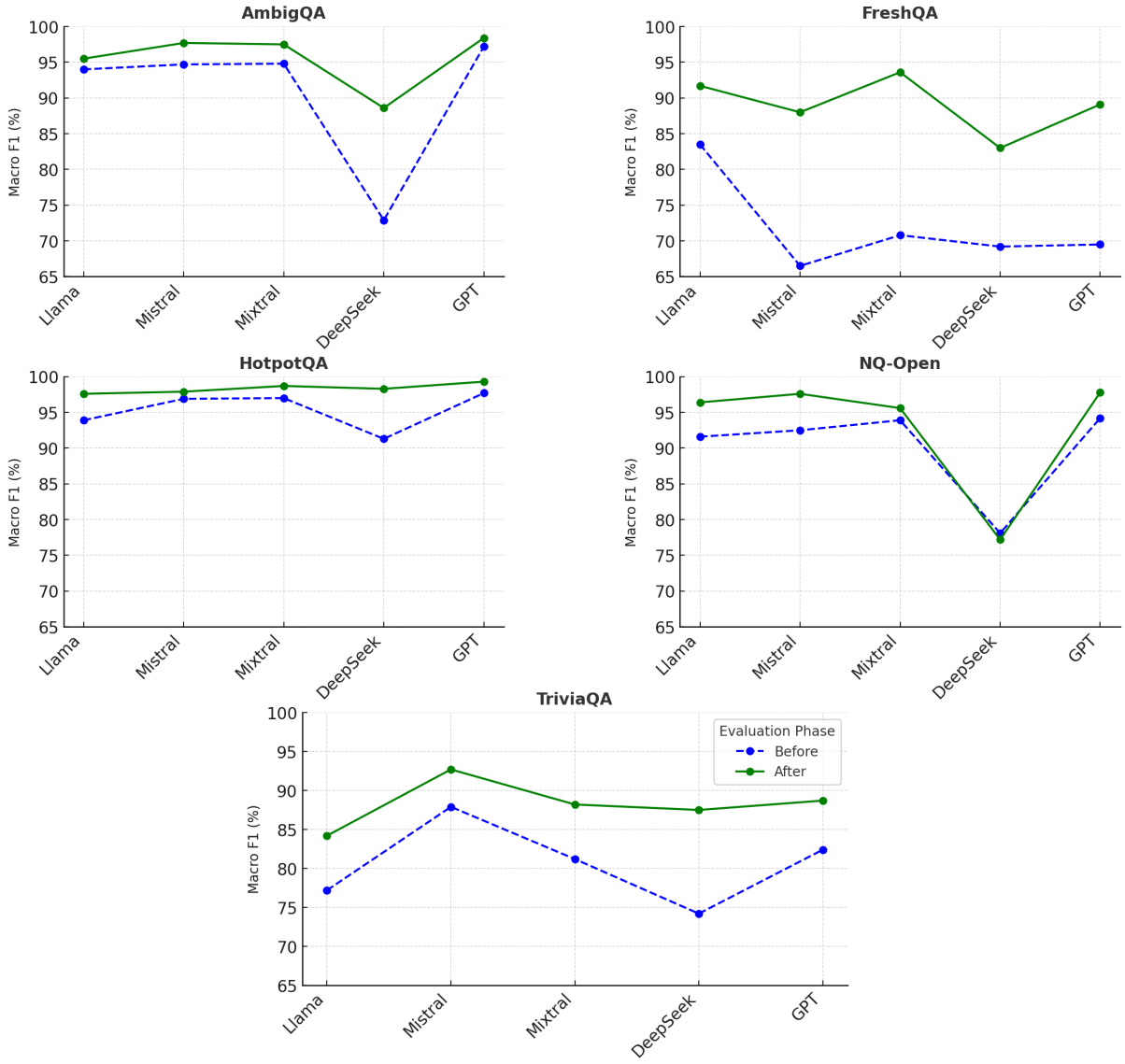

Figure 3: Impact of selective third judge on disagreements between primary judges. Note that we used Llama-3.1-70B and Mistra 7B as primary judges. GPT-3.5-turbo is only utilized when disagreements are found. The models given in the figure are candidate LLMs that generate outputs for the given tasks and are then evaluated through SAGE.

annotator.

In this setup, we employed three LLM judges of equal weight: Llama, GPT-3.5, and Mistral to evaluate candidate models' generated responses. For every evaluation instance, each judge provided an independent binary verdict (True or False). The final decision is determined through a simple majority vote across these three verdicts.

As presented in Table 12, SAGE matches or closely approaches the Macro F1 and Cohen's Kappa scores of the three-judge majority across almost all tasks and candidate LLMs. For example, on HotpotQA, evaluating candidate Llama with SAGE achieves a Macro F1 of 97.6% (compared to 97.6% for majority voting) and a Cohen's Kappa of 0.95, while for GPT-3.5 on AmbigQA, SAGE reaches a Macro F1 of 98.4% (versus 98.3% for majority voting), indicating a negligible performance difference. Even in high-disagreement tasks like TriviaQA, where the primary judges (e.g., Mistral) disagree 20.3% of the time, SAGE retains strong alignment (with a Macro F1 of 92.7 compared to 93.5 for majority voting). Minor deviations, such as the one observed for candidate Mixtral on TriviaQA (SAGE's Macro F1 = 0.88 vs. 0.95 for majority voting), reflect rare instances where both the primary judges and the third judge make errors, yet these outliers are substantially outweighed by the

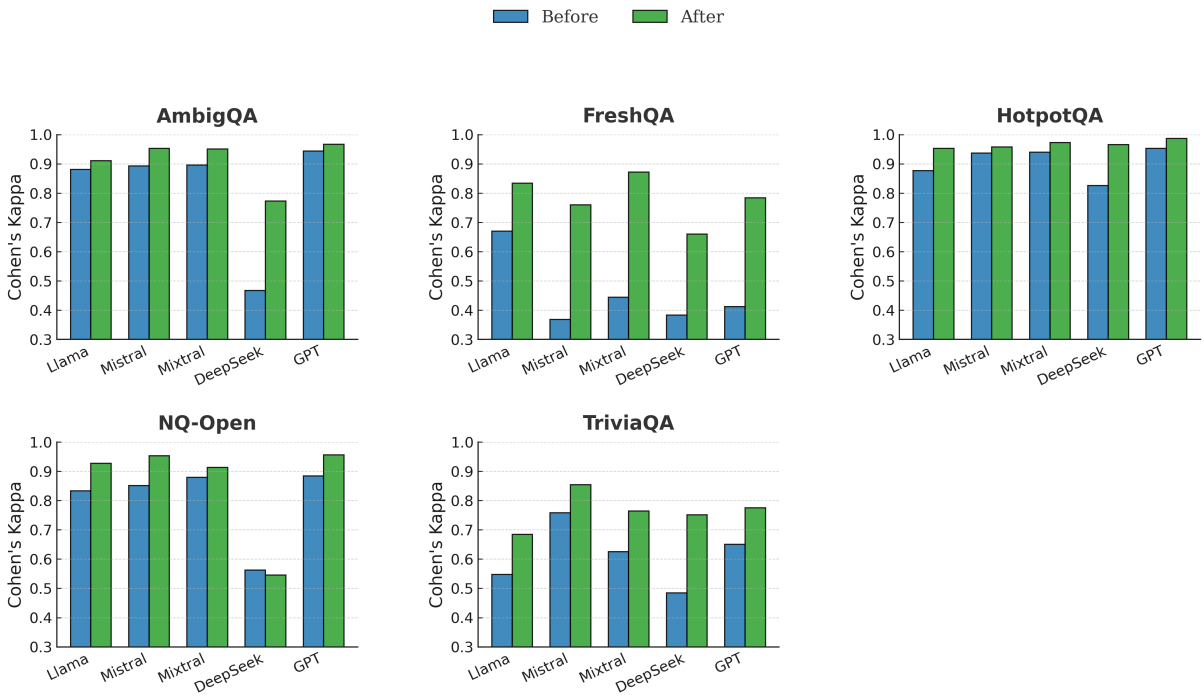

Figure 4: Comparison of Cohen's kappa scores before and after third judge (GPT-3.5-turbo as third judge).

computational savings offered by selective third judge.

D.6 Impact of prompt variations

The effectiveness and consistency of LLM-based evaluation are significantly influenced by prompt design. Variations in prompt structure, reasoning order, explanation requirements, and task-specific examples can lead to notable differences in model verdicts. To analyze the robustness of the LLM judges in free-form QA, we conducted ablation studies on different prompt variations using Mistral as the candidate model and GPT as the judge.

D.6.1 Consistency in judgment across multiple trials

LLMs generate random text even at a temperature of 0. To assess whether this affects evaluation consistency, we repeated the same evaluation task five times for 100 Mistral-generated responses for HotpotQA.

- **Verdict stability:** GPT produced identical True/False verdicts in 100% of cases. This suggest that its binary decision-making process remains stable even across multiple trials.
- **Explanation variability:** While verdicts remained consistent, the rationales and explanations provided by GPT across trials, often cited different supporting facts for the same judgment.

D.6.2 Few-shot vs. zero-shot prompting

We investigated the impact of few-shot prompting where we included three **task-specific examples** in the prompt to guide the judge's decision-making process. We found that adding few-shot examples resulted in a 2% increase in Macro-F1 scores. However, few-shot prompting introduced rigid decision patterns—the model sometimes over-applied reasoning from the examples rather than adapting flexibly to novel cases. For instance, multi-hop reasoning cases from HotpotQA, the judge model consistently followed the structure of the provided examples, even when the correct reasoning required a different approach.

D.6.3 Explanation requirement: Binary verdict vs. justification-based evaluation

To test whether requiring the model to generate explanations alongside verdicts improves judgment reliability, we compared two settings:

- **Binary verdict-only evaluation:** The model was instructed to provide only a True/False response without any explanation.
- **Justification-based evaluation:** The model was required to explain its reasoning before delivering the final verdict.

Candidate LLMs	Tasks	Samples	Disagreement Rates (%)	Third Judge Usage
DeepSeek	AmbigQA	300	25.7	77
	FreshQA	300	28.3	85
	HotpotQA	300	10.7	32
	NQ-Open	300	12.0	36
	TriviaQA	300	14.3	43
Llama	AmbigQA	300	10.0	30
	FreshQA	300	31.3	94
	HotpotQA	300	13.0	39
	NQ-Open	300	18.0	54
	TriviaQA	300	17.0	51
GPT	AmbigQA	300	7.0	21
	FreshQA	300	44.3	133
	HotpotQA	300	5.7	17
	NQ-Open	300	13.0	39
	TriviaQA	300	15.7	47
Mixtral	AmbigQA	300	9.0	27
	FreshQA	300	37.3	112
	HotpotQA	300	4.7	14
	NQ-Open	300	13.0	39
	TriviaQA	300	17.0	51
Mistral	AmbigQA	300	11.7	35
	FreshQA	300	39.7	119
	HotpotQA	300	6.0	18
	NQ-Open	300	14.7	44
	TriviaQA	300	20.3	61
Total		**7500**		**1318**

Table 11: Cost-efficiency analysis of SAGE: Summary of disagreement rates and third judge usage across candidate models and tasks

We found that:

- **Higher verdict volatility in verdict-only mode:** When explanations were removed, 13% of verdicts changed between repeated evaluations of the same responses.
- **Reduced alignment with human judgment:** Cohen's Kappa agreement with human annotators dropped from 0.95 to 0.72, highlighting that rationale-based prompts lead to more stable and accurate decisions.

D.6.4 Reason-first vs. verdict-first prompting

In the verdict-first approach, the model is instructed to provide a True/False answer before justifying its decision, whereas in the reason-first approach, the model is asked to generate reasoning first and then conclude with a verdict. Experimental results showed no significant difference in accuracy or agreement scores between these two formats.

D.7 G-Eval: reference-free evaluation of free-form question-answering

Existing LLM-based evaluators such as G-Eval (Liu et al., 2023) are designed for reference-free, subjective tasks (e.g., summarization, dialogue), where evaluation criteria (e.g., coherence, fluency) are inherently ambiguous and scored on Likert scales. These frameworks prioritize qualitative judgments rather than binary factual correctness. In contrast, SAGE is explicitly tailored for reference-dependent, objective evaluation in free-form QA, where answers are either factually correct or incorrect based on alignment with explicit ground-truth references.

To validate this distinction, we tailored the G-

Candidate LLM	Task	Majority Voting		Disagreement (%)	SAGE	
		Macro F1	Kappa		Macro F1	Kappa
Llama	AmbigQA	95.5	0.91	10.0	95.5	0.91
	HotpotQA	97.6	0.95	13.0	97.6	0.95
	NQ-Open	96.3	0.93	18.0	96.4	0.92
	TriviaQA	84.1	0.68	17.0	84.2	0.68
GPT	AmbigQA	98.3	0.97	7.0	98.4	0.96
	HotpotQA	99.3	0.99	5.7	99.3	0.98
	NQ-Open	97.8	0.96	13.0	97.8	0.95
	TriviaQA	90.5	0.81	15.7	88.7	0.77
Mixtral	AmbigQA	98.9	0.98	9.0	97.5	0.95
	HotpotQA	98.6	0.97	4.7	98.7	0.97
	NQ-Open	98.3	0.97	13.0	95.6	0.91
	TriviaQA	95.0	0.90	17.0	88.2	0.76
Mistral	AmbigQA	97.6	0.95	11.7	97.7	0.95
	HotpotQA	97.9	0.96	6.0	97.9	0.95
	NQ-Open	97.6	0.95	14.7	97.6	0.95
	TriviaQA	93.5	0.87	20.3	92.7	0.85

Table 12: Comparison between Majority Voting (Llama+GPT-3.5+Mistral) and SAGE (GPT-3.5 as the third judge). For each candidate LLM and task, the table reports Macro F1 and Cohen's Kappa scores under Majority Voting, the disagreement rate (in %), and the corresponding scores using SAGE.

eval (Liu et al., 2023) to investigate the capability of LLM-as-a-judge in reference-free settings. In this setting, we modify the evaluation prompt by excluding the reference answer r and directly prompt the evaluator model as $P = \{x, \bar{y}\}$ along with instructions such as correctness.

The performance of LLM-as-a-judge drastically changes in reference-free settings. Without access to the ground truth references, we observe a stark decline in evaluation capability across all models (see Table 13 and 14 values in blue). This systematic deterioration spans all tasks and model combinations, though its severity varies by context. HotpotQA and NQ-Open, with their demands for complex reasoning, exemplify this challenge most clearly. The substantial gap between reference-based and reference-free evaluation underscores the crucial role of reference answers in reliable assessment.

D.8 SAGE in multi-reference answers

SAGE explicitly accommodates multiple gold reference answers by incorporating all available references into the judge LLM's prompt during evaluation. For datasets like AmbigQA and TriviaQA, where questions often have multiple valid answers (e.g., synonyms, rephrased answers, or alternative factual representations), SAGE aggregates all reference answers into the judge's input prompt (e.g., concatenating them as a comma-separated list).

This design ensures that the judge evaluates the candidate's output against the full spectrum of acceptable answers, mirroring the human evaluation protocol, where annotators are instructed to mark a response as correct if it aligns with any reference answer. However, as presented in our paper, LLM-based judges encounter challenges with multiple reference answers. This confusion is particularly evident in TriviaQA, where multiple reference answers introduce difficulties for the judges to recognize and evaluate a range of correct responses.

D.9 Analysis of automatic metrics

Figures 5, 6, 7, and 8 illustrate the fundamental trade-offs in automatic metrics. In TriviaQA, where multiple normalized reference answers exist, EM achieves impressive true positives (61.7-74.3%) compared to HotpotQA (23.0-34.3%) which contains single reference answers. EM's near-zero false positives across tasks (0-0.7%) stem from its strict string matching – it only flags matches when answers are identical to references. Our er-

Candidate LLMs	Tasks	Evaluators						
		EM	BERTScore	Human Majority	Llama-3.1-70B	GPT-3.5-turbo	Mixtral-8x7B	Mistral-7B
Llama-3.1-70B	AmbigQA	42.3	63.0	67.0	65.3 [83.3]	64.7 [84.7]	63.0 [76.0]	66.0 [80.3]
	HotpotQA	34.3	67.7	56.3	58.3 [81.0]	54.0 [81.0]	50.7 [67.3]	52.7 [69.3]
	NQ-Open	31.7	61.7	66.3	62.7 [89.0]	60.0 [89.3]	59.0 [81.0]	66.7 [81.0]
	TriviaQA	74.3	94.0	94.7	90.3 [90.3]	90.0 [90.3]	88.7 [89.0]	84.7 [84.0]
GPT-3.5	AmbigQA	49.7	78.0	71.7	70.0 [79.0]	68.0 [81.0]	65.7 [79.0]	71.0 [84.3]
	HotpotQA	33.7	80.0	54.0	53.0 [85.3]	52.7 [85.7]	51.7 [82.3]	54.0 [86.3]
	NQ-Open	36.3	74.0	65.3	62.7 [83.7]	59.0 [90.7]	59.0 [87.0]	67.0 [89.7]
	TriviaQA	74.3	95.3	93.0	89.3 [89.0]	90.7 [88.7]	89.7 [90.3]	86.3 [84.3]
Mixtral-8x7B	AmbigQA	37.7	70.3	61.7	57.3 [74.7]	62.0 [82.3]	59.3 [79.7]	61.7 [80.7]
	HotpotQA	25.0	69.7	47.0	45.3 [80.0]	45.7 [84.7]	44.7 [72.0]	46.0 [78.0]
	NQ-Open	23.7	63.7	56.7	52.7 [81.7]	47.7 [90.3]	52.3 [85.7]	59.7 [89.7]
	TriviaQA	64.7	91.3	90.7	86.3 [85.7]	89.7 [89.0]	86.0 [86.7]	85.3 [86.0]
Mistral-7B	AmbigQA	31.0	61.7	49.7	46.3 [61.0]	47.7 [78.7]	46.3 [74.7]	53.3 [85.0]
	HotpotQA	23.7	64.7	40.0	39.0 [64.3]	38.0 [83.3]	37.0 [62.0]	39.0 [77.0]
	NQ-Open	22.7	60.0	46.0	40.0 [72.3]	43.3 [85.7]	41.3 [78.0]	50.0 [92.3]
	TriviaQA	62.0	94.3	83.7	81.3 [80.7]	81.0 [81.0]	79.7 [80.7]	85.0 [84.7]

Table 13: Overall performance (Raw Accuracy) of candidate LLMs across free-form QA tasks. Values [in blue] represent LLM-as-a-judge in the reference-free mood.

Candidate LLMs	Tasks	Evaluators						
		EM	BERTScore	Llama-3.1-70B	GPT-3.5-turbo	Mixtral-8x7B	Mistral-7B	SAGE
Llama-3.1-70B	AmbigQA	0.744	0.641	0.944 [0.629]	0.922 [0.604]	0.912 [0.669]	0.929 [0.631]	0.955 [0.637]
	HotpotQA	0.778	0.745	0.939 [0.628]	0.949 [0.574]	0.910 [0.665]	0.916 [0.640]	0.976 [0.623]
	NQ-Open	0.653	0.718	0.916 [0.606]	0.896 [0.560]	0.907 [0.639]	0.869 [0.622]	0.964 [0.610]
	TriviaQA	0.612	0.782	0.772 [0.772]	0.717 [0.628]	0.695 [0.678]	0.640 [0.633]	0.842 [0.747]
GPT-3.5	AmbigQA	0.792	0.622	0.972 [0.686]	0.949 [0.603]	0.930 [0.596]	0.927 [0.553]	0.984 [0.607]
	HotpotQA	0.794	0.623	0.977 [0.566]	0.987 [0.521]	0.936 [0.543]	0.966 [0.494]	0.993 [0.522]
	NQ-Open	0.703	0.606	0.942 [0.671]	0.911 [0.544]	0.911 [0.601]	0.914 [0.536]	0.978 [0.575]
	TriviaQA	0.646	0.681	0.824 [0.817]	0.700 [0.690]	0.789 [0.760]	0.730 [0.701]	0.887 [0.882]
Mixtral-8x7B	AmbigQA	0.760	0.666	0.948 [0.704]	0.891 [0.636]	0.955 [0.654]	0.944 [0.622]	0.975 [0.650]
	HotpotQA	0.761	0.657	0.970 [0.587]	0.966 [0.470]	0.930 [0.582]	0.970 [0.577]	0.987 [0.536]
	NQ-Open	0.650	0.649	0.939 [0.652]	0.863 [0.517]	0.950 [0.590]	0.908 [0.529]	0.956 [0.563]
	TriviaQA	0.625	0.695	0.812 [0.800]	0.803 [0.754]	0.838 [0.818]	0.716 [0.725]	0.882 [0.858]
Mistral-7B	AmbigQA	0.792	0.622	0.947 [0.730]	0.947 [0.627]	0.947 [0.628]	0.930 [0.523]	0.977 [0.647]
	HotpotQA	0.796	0.673	0.969 [0.649]	0.951 [0.478]	0.947 [0.680]	0.969 [0.578]	0.979 [0.673]
	NQ-Open	0.726	0.639	0.925 [0.652]	0.919 [0.515]	0.939 [0.597]	0.920 [0.433]	0.976 [0.527]
	TriviaQA	0.718	0.608	0.879 [0.881]	0.863 [0.840]	0.822 [0.846]	0.735 [0.744]	0.927 [0.913]

Table 14: Performance (Macro F1) of various evaluators across candidate LLMs and tasks. Values [in blue] represent the reference-free mode.

ror analysis found three primary causes of such rare false positives including preprocessing errors, where character normalization removes crucial distinctions, and reference ambiguities, where incomplete or ambiguous references lead to incorrect matches. Additionally, a semantic mismatch occurs when the EM incorrectly labels a prediction as true by matching text without considering its context. For instance, despite their different contextual meanings, EM wrongly marks a match between a model prediction of "1944" (describing the start of a war) and a reference answer containing "1944" (representing the end of the war).

EM string-matching guarantees high precision and makes EM particularly effective when exact wording is crucial, such as mathematical problems. However, its rigid criteria also result in substantial false negatives (17.0-34.7%). These false negatives primarily occur when the candidate LLM generates semantically correct responses that differ from references in format or expression. Common cases include synonym usage and paraphrases, structural variations in phrasing (e.g., "School of Medicine at Harvard" vs. "Harvard Medical School"), granularity discrepancies where answers differ in levels of detail from references (e.g., answering "British writer" instead of "William Shakespeare"), and partial matches that contain valid information but don't exactly mirror the reference.

Unlike EM, BERTScore offers advantages in capturing semantic similarities. In TriviaQA, it gains high true positive rates (81.3-92.0%) with relatively low false positives (2.0-13.0%). BERTScore's performance varies significantly across tasks and is influenced by its sensitivity to the threshold setting. In HotpotQA, where answers

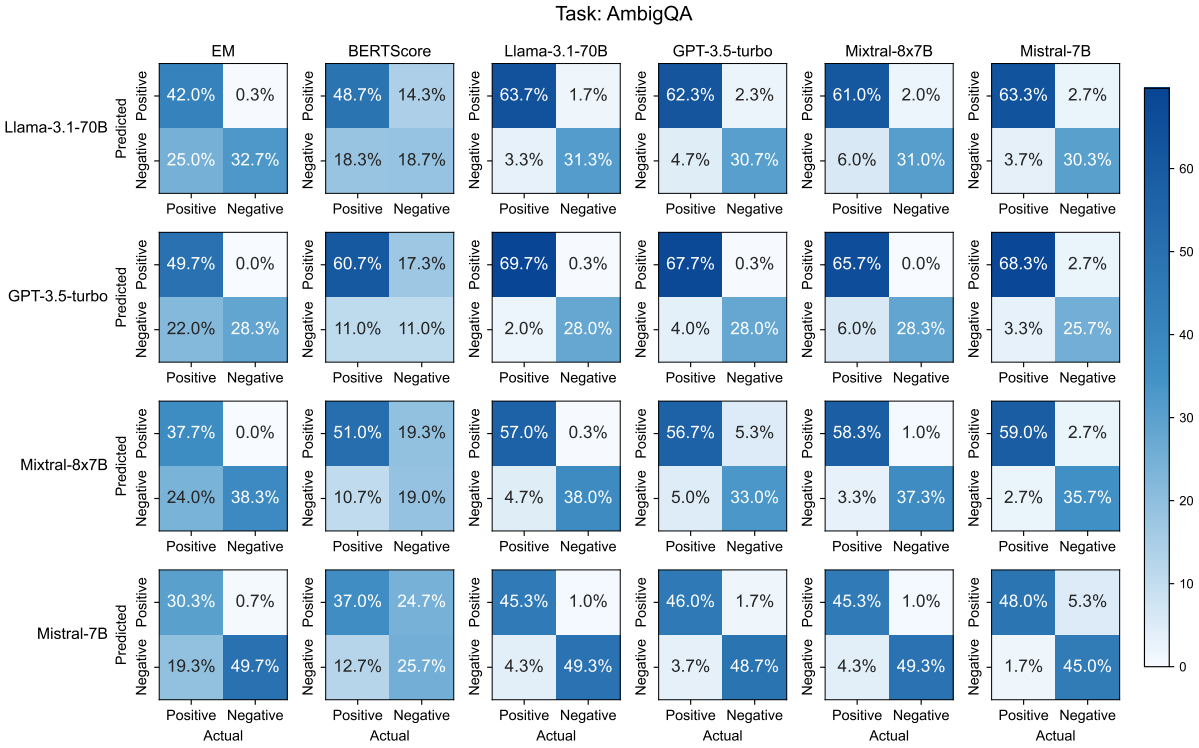

Figure 5: Confusion matrices comparing the performance of automatic metrics (EM, BERTScore) and individual LLM judges on AmbigQA.

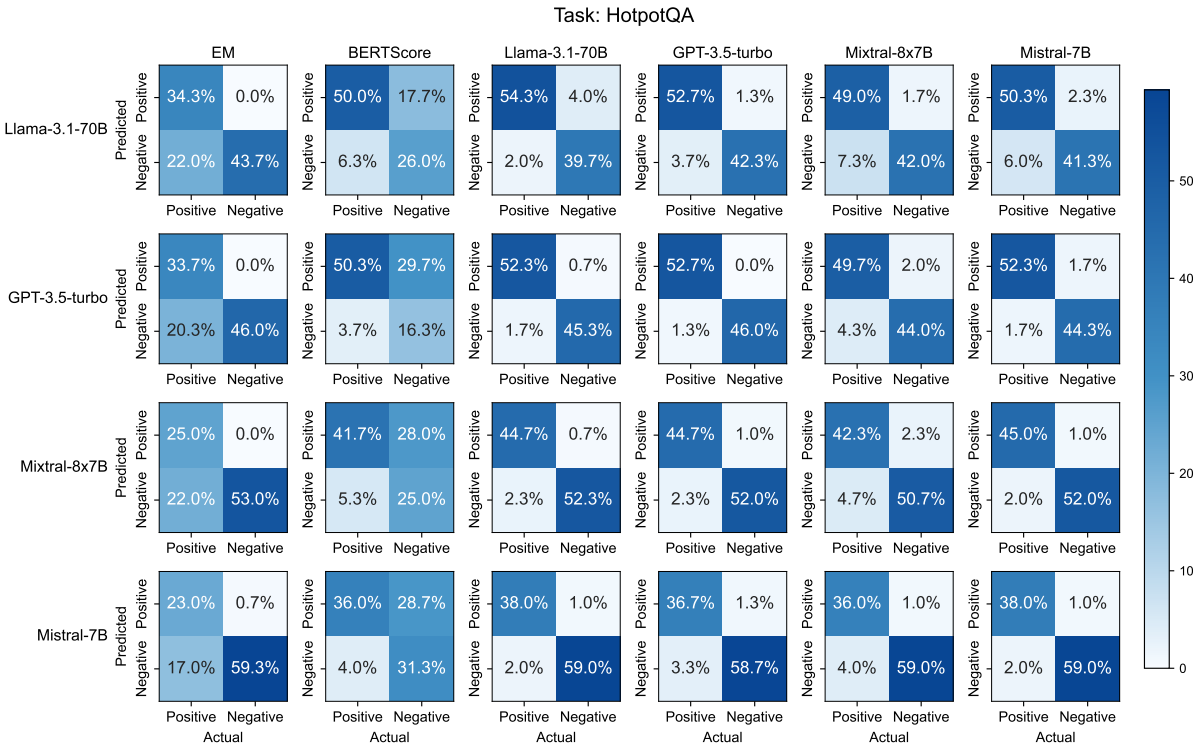

Figure 6: Confusion matrices comparing the performance of automatic metrics (EM, BERTScore) and individual LLM judges on HotpotQA.

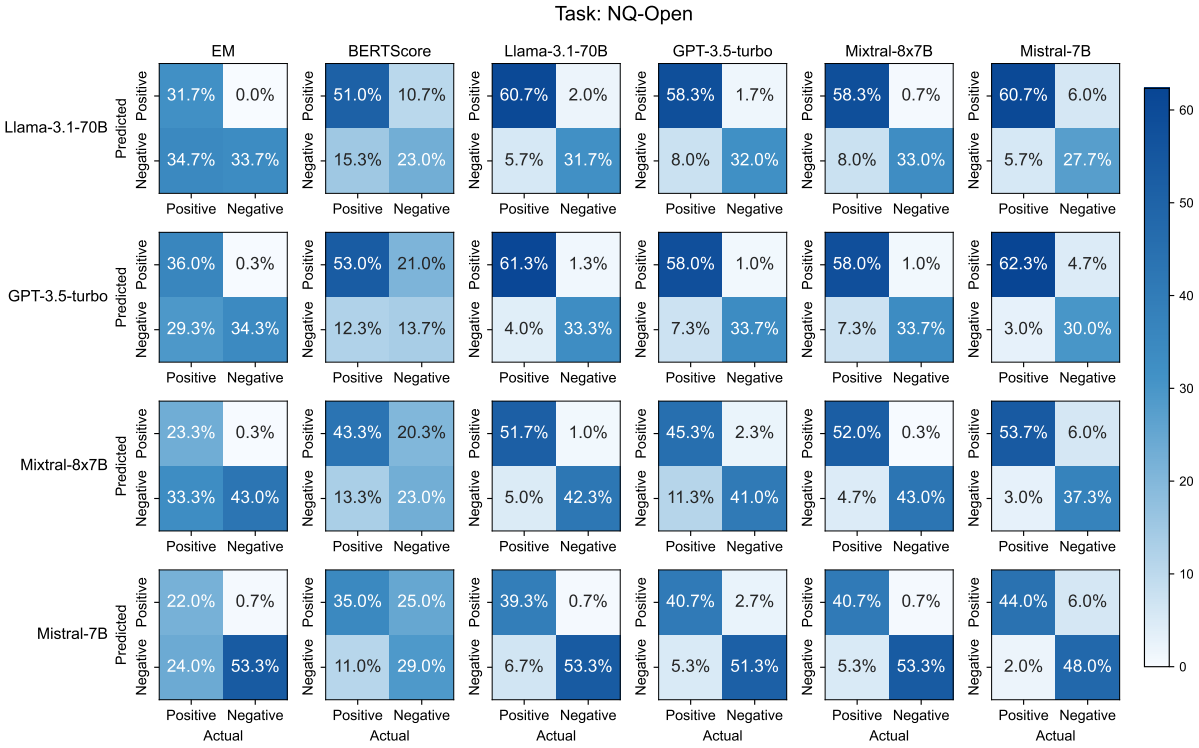

Figure 7: Confusion matrices comparing the performance of automatic metrics (EM, BERTScore) and individual LLM judges on NQ-Open.

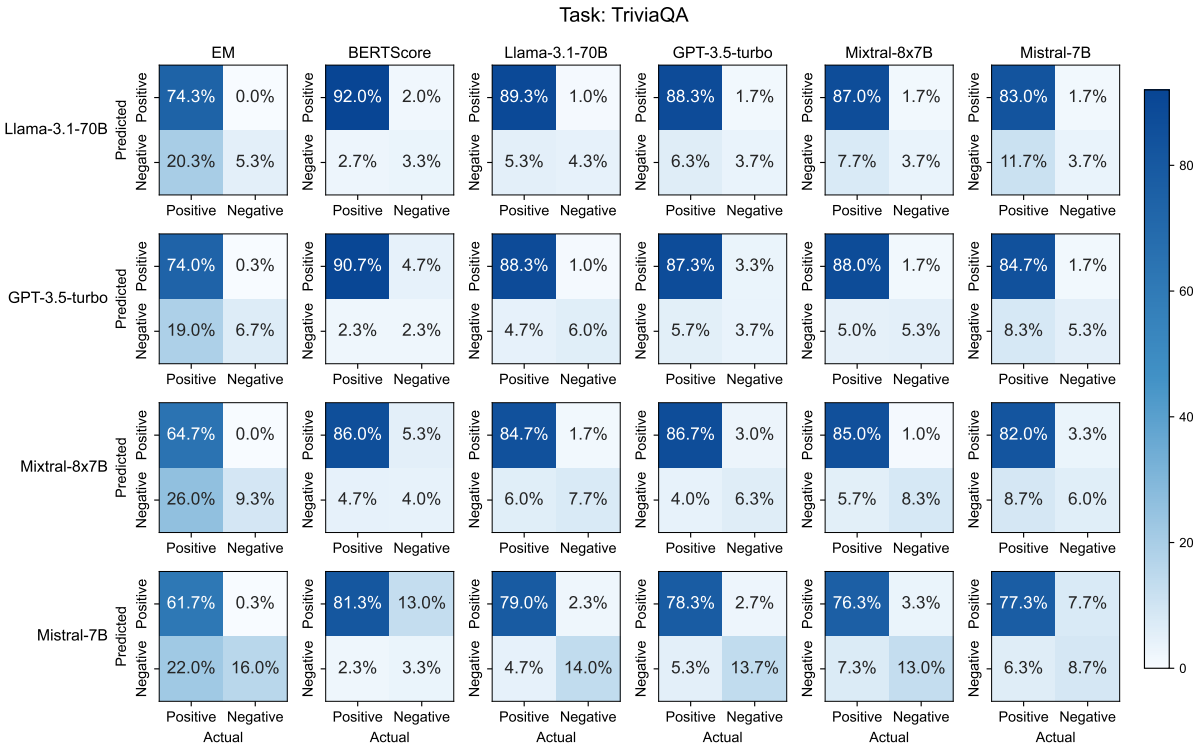

Figure 8: Confusion matrices comparing the performance of automatic metrics (EM, BERTScore) and individual LLM judges on TriviaQA.

require multi-hop reasoning, true positives reach 36.0-50.3%, with an increase in false positives (17.7-29.7%). A similar pattern appears in NQ-Open, with true positives of 43.3-53.0% and false positives of 10.7-21.0%. Its tendency toward false positives indicates that relying solely on embedding similarity often accepts answers that are contextually related but factually incorrect. The false positives emerge through semantic drift (where similar embeddings yield false matches), contextual misalignment (where word meanings shift based on context), and threshold instability (where similarity cutoffs fail to distinguish subtle semantic differences). Additionally, false positives emerge due to the verbose responses where additional content artificially increases similarity scores.

E Prompting

In our main experiment, we performed zero-shot prompting in the following two stages.

E.1 Prompting Candidate LLMs

We prompted candidate LLMs (see Figure 9) to record generations for each task. We set the same role and prompt structure for each candidate model to ensure the reproducibility of our results. Figure 10 shows the candidate GPT-3.5-turbo response at zero temperature for the input given in Figure 9.

> You are a helpful assistant. Which comic book was also written by the writer of Crossed?

Figure 9: Prompting candidate GPT-3.5-turbo to elicit outputs for HotpotQA.

> Crossed was written by Garth Ennis. The Boys is a comic book series also written by Garth Ennis. So the comic book that was also written by the writer of Crossed is The Boys.

Figure 10: Candidate GPT-3.5-turbo response to the input given in Figure 9.

E.2 Prompting LLM Judges

We prompted LLMs-as-judges to perform the evaluation (see Figure 11). In Figure 12, judge Llama-3.1-70B evaluates candidate GPT-3.5-turbo.

> You are a helpful assistant acting as an impartial judge. You will be given a Question and a Proposed Answer. Your task is to judge whether the Proposed Answer is correct by comparing it to the Reference Answer. If the Proposed Answer is correct, choose 'True', otherwise, choose 'False'. Provide a brief explanation for your decision.
> Question: Which comic book was also written by the writer of Crossed?
> Provided Answer: Crossed was written by Garth Ennis. The Boys is a comic book series also written by Garth Ennis. So the comic book that was also written by the writer of Crossed is The Boys.
> Reference Answer: the Vertigo series "Preacher"
> Evaluation:
> Provide your response in the following format:
> Decision: [True/False]
> Explanation: [Your brief explanation]

Figure 11: Prompting judge Llama-3.1-70B for evaluating candidate GPT-3.5-turbo.

> **Decision:** False
> Explanation: The Proposed Answer incorrectly identifies "The Boys" as the comic book written by the writer of Crossed. While it is true that Garth Ennis wrote Crossed, the Proposed Answer fails to consider other works written by Garth Ennis. The Reference Answer, "Preacher", is indeed another comic book series written by Garth Ennis, specifically for the Vertigo imprint.

Figure 12: Llama-3.1-70B-Judge verdict on the candidate GPT-3.5-turbo output.

E.3 Prompting Prometheus 2

As Prometheus 2 is based on subjective criteria, we tailored it for our setting (see below Figure 13).

Criteria: Is the model's answer correct and adequately addresses the question based on the reference answer?

Score 1 ("Completely incorrect") — The answer is completely incorrect or irrelevant to the question.
Score 2 ("Minor relevance") — The answer has minor relevance but contains significant inaccuracies.
Score 3 ("Partially correct") — The answer is partially correct and addresses some aspects of the question.
Score 4 ("Mostly correct") — The answer is mostly correct with minor omissions or imprecisions.
Score 5 ("Fully correct") — The answer is fully correct, comprehensive, and aligns perfectly with the reference answer.

Figure 13: Prompt rubric used by Prometheus 2 in the reference-based setting.