Research Article

DAFE: LLM-Based Evaluation Through Dynamic Arbitration for Free-Form Question-Answering

Sher Badshah[1], **Hassan Sajjad**[1]

1. Faculty of Computer Science, Dalhousie University, Canada

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 handle the variability of open-ended responses, while human evaluation, though reliable, is resource-intensive. Leveraging LLMs as evaluators offers a promising alternative due to their strong language understanding and instruction-following capabilities. Taking advantage of these capabilities, we propose the Dynamic Arbitration Framework for Evaluation (DAFE), which employs two primary LLM-as-judges and engages a third arbitrator only in cases of disagreements. This selective arbitration prioritizes evaluation reliability while reducing unnecessary computational demands compared to conventional majority voting. DAFE utilizes task-specific reference answers with dynamic arbitration to enhance judgment accuracy, resulting in significant improvements in evaluation metrics such as Macro F1 and Cohen's Kappa. Through experiments, including a comprehensive human evaluation, we demonstrate DAFE'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[1]. In particular, free-form model responses are difficult to evaluate because their correctness depends on understanding the broader context and underlying meaning[2]. Many benchmarks, such as MMLU[3], often simplify evaluation by focusing on structured formats (e.g., multiple-choice questions)[4]. Although effective for certain tasks, such methods

 doi.org/10.32388/B69SKY

rely on log probabilities assigned to predefined options, where the model selects the most likely answer, limiting the range of capabilities that can be assessed[5]. This structured approach fails to accommodate the complexity of free-form responses, where multiple valid answers exist[6]. 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[7][8].

Automatic metrics including lexical matching, n-gram, and neural-based have been widely adopted as scalable solutions for the evaluation of free-form model outputs. Lexical matching methods such as Exact Match (EM) evaluate model predictions by assessing strict lexical alignment between generated outputs and reference answers. However, EM fails to account for semantically equivalent variations in phrasing. For instance, despite their equivalence, EM treats "nuclear weapon" and "atomic bomb" as incorrect. Similarly, n-gram-based metrics[9][10] primarily assess surface-level similarity and often fail to capture semantic equivalence, particularly when lexical or structural diversity conveys the same underlying meaning[11][12][13]. Neural-based metrics like BERTScore[13] address such limitations by leveraging contextual embeddings to evaluate semantic similarity. However, BERTScore depends on reference quality[14] and struggles with domain adaptation and length variations[11]. Furthermore, continuous score provider metrics are difficult to interpret[15]. The limitations in automatic metrics become particularly evident when evaluating instruction-tuned chat models[16], which tend to produce verbose and diverse responses[17][18].

Contrary to automatic metrics, human evaluation provides a more transparent assessment[19]. However, despite being the "gold standard", human evaluation is not without its limitations. LLMs' growing complexity and scale have made recruiting and coordinating multiple human raters increasingly resource-intensive and time-consuming[20]. Furthermore, the reliability of human evaluation is additionally challenged by variations in rater expertise and inherent subjectivity that affect reproducibility[21][19].

Recently, a paradigm shift has emerged where LLMs are utilized to judge the candidate model generations for given tasks[22]. 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. In this new line of work, research primarily focuses on pairwise comparison[22][23][24], such as instructing an LLM to judge "which assistant response is better", and single-answer scoring[25] like evaluating summarization task based on predefined criteria (e.g., likability, relevance, etc.)[19][26][27][28][29].

Inspired by a recent study on self-correction where external feedback helps models identify and correct their mistakes[30], we propose to guide LLM-as-a-judge with human-annotated task-specific reference answers in order to explore the potential of LLMs as an alternative to lexical matching (e.g., EM), neural-based (e.g., BERTScore), and human evaluation for automatic evaluation of free-form model responses. Unlike traditional metrics, an LLM judge can leverage its language understanding and instruction-following capabilities to recognize the correctness of open-ended generations.

We propose the Dynamic Arbitration Framework for Evaluation (DAFE), which employs LLM judges to evaluate free-form model responses. Using a single LLM as a judge, while simple, often leads to inconsistent evaluations, undermining trust in the results. On the other hand, the common practice of using large, universally capable models such as GPT-4 as evaluators makes the evaluation process both slow and costly[31][32][25], further limiting its broader applicability. Relying on multiple judges for every evaluation, though more reliable, exacerbates these computational challenges, making such approaches impractical at scale. DAFE offers a middle ground between these approaches by utilizing two complementary primary judges to perform the initial assessment. Only when these judges disagree, is a third independent arbitrator engaged to resolve the conflict. This selective arbitration ensures evaluation reliability and fairness while reducing computational overhead. Our experiments reveal that DAFE achieves significant improvements in metrics such as Macro F1 and Cohen's kappa. Our key contributions include: a detailed analysis of limitations in conventional metrics for free-form QA, an evaluation of LLM judges with insights into their strengths and errors, a comprehensive human evaluation for benchmarking, and the introduction of DAFE—a scalable framework that improves reliability while minimizing the need for additional evaluators through selective arbitration.

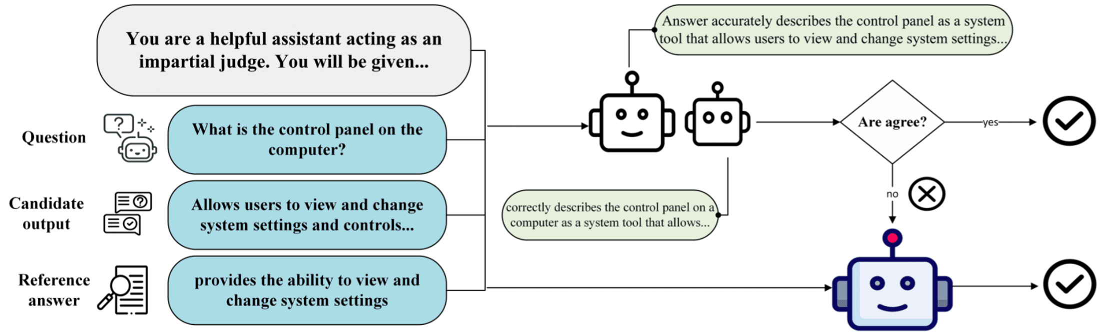

Figure 1. Our proposed Dynamic Arbitration Framework for Evaluation (DAFE). Two primary judges, J_1 and J_2, first provide verdicts V_{i_1} and V_{i_2} for an instance i. If agree, that consensus V_i is the final decision D_i. If disagree, a tiebreaker model J_t independently produces a verdict V_t.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

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. We first utilized candidate LLMs to obtain outputs for the given free-form question-answering tasks.

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., $x \to \mathcal{A}_{\text{llm}}$) and $\bar{y}$. We utilized the reference answer r and prompted 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 this 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

doi.org/10.32388/B69SKY

includes several related examples, or a chain of thought, encouraging $\mathcal{J}_{\text{llm}}$ to reason stepwise through the problem.

2.3. Dynamic Arbitration Framework for Evaluation (DAFE)

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 human arbitration practice, we propose the Dynamic Arbitration Framework for Evaluation (DAFE). Rather than immediately employing a large powerful or a closed-source LLMs-as-a-judge, DAFE adopts a cost-efficient approach by beginning with two complementary open-source models as primary judges based on their past performance[33]. When these judges reach a consensus, no further evaluation is needed. Only in cases of disagreement is the more powerful LLM engaged as an arbitrator, whose decision then creates a majority verdict. This dynamic approach maintains evaluation quality while minimizing reliance on expensive models. The method also accounts for varying skill levels across different LLMs and tasks[34][35].

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 tiebreaker 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.

3. Experiments

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

3.1. Models

We select open and closed-source instruct models to serve as candidates and judges in our experiment. These include DeepSeek-V3[36], Llama-3.1 70B[1][37], GPT-3.5-turbo[38], Mistral 7B[2][39], and Mixtral 8x7B[3][40]. We also utilize GPT-4o[41] and DeepSeek-R1[36] 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[42]. For our proposed DAFE method, we utilized Mistral 7B and Llama 3.1 70B as primary judges with GPT-3.5-turbo as the tiebreaker. In addition, we experiment with other models as tiebreakers in our ablation experiments. In the rest of the paper, we refer both candidate and judge LLMs as: DeepSeek, Llama, GPT, Mistral, and Mixtral.

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[30][43]. In our experiment, we utilize five free-form QA datasets: AmbigQA[44], FreshQA[45], HotpotQA[46], Natural Questions[47], and TriviaQA[48]. See Appendix A for details.

3.3. Prompts

We designed generalized (i.e., with minimum instructions) zero-shot prompts with role-playing[49] 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 D for examples). Binary verdicts explicitly differentiate between correct and incorrect answers, minimize subjective interpretations, and simplify the evaluation process, thus facilitating automatic evaluation. In addition to three key prompt components (i.e., x, $\bar{y}$, r), we define the role of the judge LLMs as "You are a helpful assistant acting as an impartial judge." to mitigate biases in judgments[22]. 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[42][50].

doi.org/10.32388/B69SKY

3.4. Baselines

We establish the following 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[51][52][53]. 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[13] 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[54].

G-Eval: In addition to automatic metrics, we also utilize G-Eval[27], a reference-free framework that uses GPT-4 to assess the quality of the generated text. In this setting, we modify the evaluation prompt by excluding the reference answer r and directly prompted the evaluator model as $P = \{x, \bar{y}\}$ along with instructions.

Human Evaluation: It remains the gold standard for assessing the outputs of candidate LLMs. 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 were 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 contextual relevance. For inter-rater reliability, we compute Fleiss' Kappa (κ)[55] and percent agreement. See Appendix B for details.

4. Results

Figure 2 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 DAFE 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 6 in Appendix C for additional results).

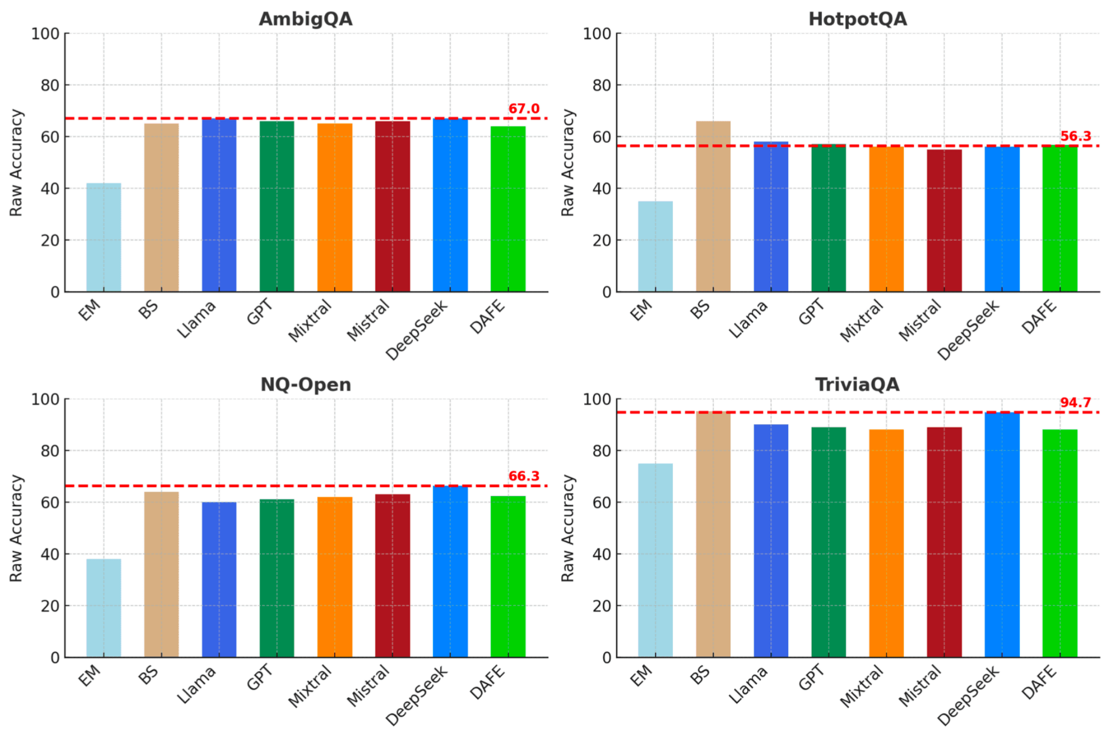

Figure 2. Raw accuracy of candidate Llama across free-form QA tasks using automatic metrics and model-based evaluation. The Human Majority (HM) serves as the ground truth for all evaluators.

4.1. Alignment with human evaluation

We calculate Cohen's kappa[56] to find the agreement between each evaluator and the human majority to obtain instance-level comparison. Overall, DAFE is almost perfectly aligned with human judgment than other evaluators (see Table 1). Similarly, individual LLM judges show more substantial to nearly perfect agreement with human judgments than EM and BERTScore.

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*"[57]. 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.

		Evaluators							
LLMs	**Tasks**	**EM**	**BS**	**DeepSeek**	**Llama**	**GPT**	**Mixtral**	**Mistral**	**DAFE**
Llama	AmbigQA	0.518	0.283	0.897	0.888	0.844	0.824	0.858	**0.911**
	HotpotQA	0.577	0.498	0.885	0.877	0.899	0.820	0.832	**0.953**
	NQ-Open	0.381	0.437	0.797	0.833	0.793	0.816	0.738	**0.927**
	TriviaQA	0.281	0.564	0.460	0.547	0.439	0.396	0.299	**0.684**
GPT	AmbigQA	0.561	0.252	0.951	0.944	0.897	0.861	0.853	**0.967**
	HotpotQA	0.604	0.300	0.807	0.953	0.973	0.873	0.933	**0.987**
	NQ-Open	0.453	0.218	0.809	0.884	0.824	0.824	0.829	**0.956**
	TriviaQA	0.335	0.364	0.594	0.650	0.401	0.580	0.467	**0.775**
Mixtral	AmbigQA	0.546	0.337	0.896	0.896	0.781	0.909	0.887	**0.951**
	HotpotQA	0.546	0.349	0.920	0.940	0.933	0.859	0.940	**0.973**
	NQ-Open	0.371	0.301	0.825	0.879	0.728	0.899	0.815	**0.913**
	TriviaQA	0.317	0.390	0.661	0.625	0.605	0.678	0.436	**0.764**
Mistral	AmbigQA	0.599	0.254	0.893	0.893	0.893	0.893	0.860	**0.953**
	HotpotQA	0.605	0.383	0.903	0.937	0.902	0.895	0.937	**0.958**
	NQ-Open	0.484	0.291	0.797	0.851	0.838	0.878	0.840	**0.953**
	TriviaQA	0.467	0.239	0.754	0.758	0.725	0.645	0.470	**0.854**

Table 1. Cohen's Kappa scores displaying the agreement levels of various evaluators with human judgments across candidate models and tasks. Higher scores indicate better agreement with human judgments.

As evidenced by consistently high Macro F1 scores in Table 2, DAFE maintains a strong alignment with

 doi.org/10.32388/B69SKY

human judgment. This represents a substantial improvement over individual model performance, where individual judges generally revealed varying levels of agreement with human evaluation. LLM-as-a-judge approach generally works better with larger more powerful models. This is particularly noticeable in DeepSeek and GPT which achieve higher Macro-F1 scores (0.97-0.98) across AmbigQA, HotpotQA, and NQ-Open compared to smaller models. This reveals an important scaling law in evaluation capability[58][22][59]. However, we also found that the most advanced models are not always guaranteed to be the best evaluators. We observed slightly comparable performance through small open-source Mistral-7B. For instance, when evaluating candidate Mixtral-8x7B on AmbigQA, Mistral-7B as-a-judge outperformed (0.944) judge GPT-3.5-turbo (0.891). Regardless, we observe relatively lower Macro-F1 scores for all LLM judges in TriviaQA.

Interestingly, despite EM's deviation from the human majority (see Figure 2 and Table 6), lexical matching EM typically accomplishes better alignment with human evaluation on 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 judgment. In contrast, BERTScore takes a more lenient approach to semantic matching. Although this leniency produces higher raw scores, it introduces more false positives, consequently reducing instance-level agreement with human judgments. This pattern emerges clearly in many models and tasks such as when evaluating Llama-3.1-70B on AmbigQA, EM shows a raw score of 42.3% but achieves a Macro-F1 of 0.744, while BERTScore indicates a higher raw score of 63.0% but a lower Macro-F1 of 0.641.

		Evaluators							
LLMs	**Tasks**	**EM**	**BS**	**DeepSeek**	**Llama**	**GPT**	**Mixtral**	**Mistral**	**DAFE**
Llama	AmbigQA	0.744	0.641	0.948	0.944	0.922	0.912	0.929	**0.955**
	HotpotQA	0.778	0.745	0.942	0.939	0.949	0.910	0.916	**0.976**
	NQ-Open	0.653	0.718	0.898	0.916	0.896	0.907	0.869	**0.964**
	TriviaQA	0.612	0.782	0.726	0.772	0.717	0.695	0.640	**0.842**
GPT	AmbigQA	0.792	0.622	0.976	0.972	0.949	0.930	0.927	**0.984**
	HotpotQA	0.794	0.623	0.903	0.977	0.987	0.936	0.966	**0.993**
	NQ-Open	0.703	0.606	0.904	0.942	0.911	0.911	0.914	**0.978**
	TriviaQA	0.646	0.681	0.796	0.824	0.700	0.789	0.730	**0.887**
Mixtral	AmbigQA	0.760	0.666	0.948	0.948	0.891	0.955	0.944	**0.975**
	HotpotQA	0.761	0.657	0.960	0.970	0.966	0.930	0.970	**0.987**
	NQ-Open	0.650	0.649	0.912	0.939	0.863	0.950	0.908	**0.956**
	TriviaQA	0.625	0.695	0.829	0.812	0.803	0.838	0.716	**0.882**
Mistral	AmbigQA	0.792	0.622	0.947	0.947	0.947	0.947	0.930	**0.977**
	HotpotQA	0.796	0.673	0.951	0.969	0.951	0.947	0.969	**0.979**
	NQ-Open	0.726	0.639	0.898	0.925	0.919	0.939	0.920	**0.976**
	TriviaQA	0.718	0.608	0.925	0.879	0.863	0.822	0.735	**0.927**

Table 2. Macro-F1 scores of various evaluators applied to different candidate LLMs and associated tasks. Higher scores indicate better performance. DAFE consistently achieves the highest Macro-F1 across all evaluated settings.

doi.org/10.32388/B69SKY

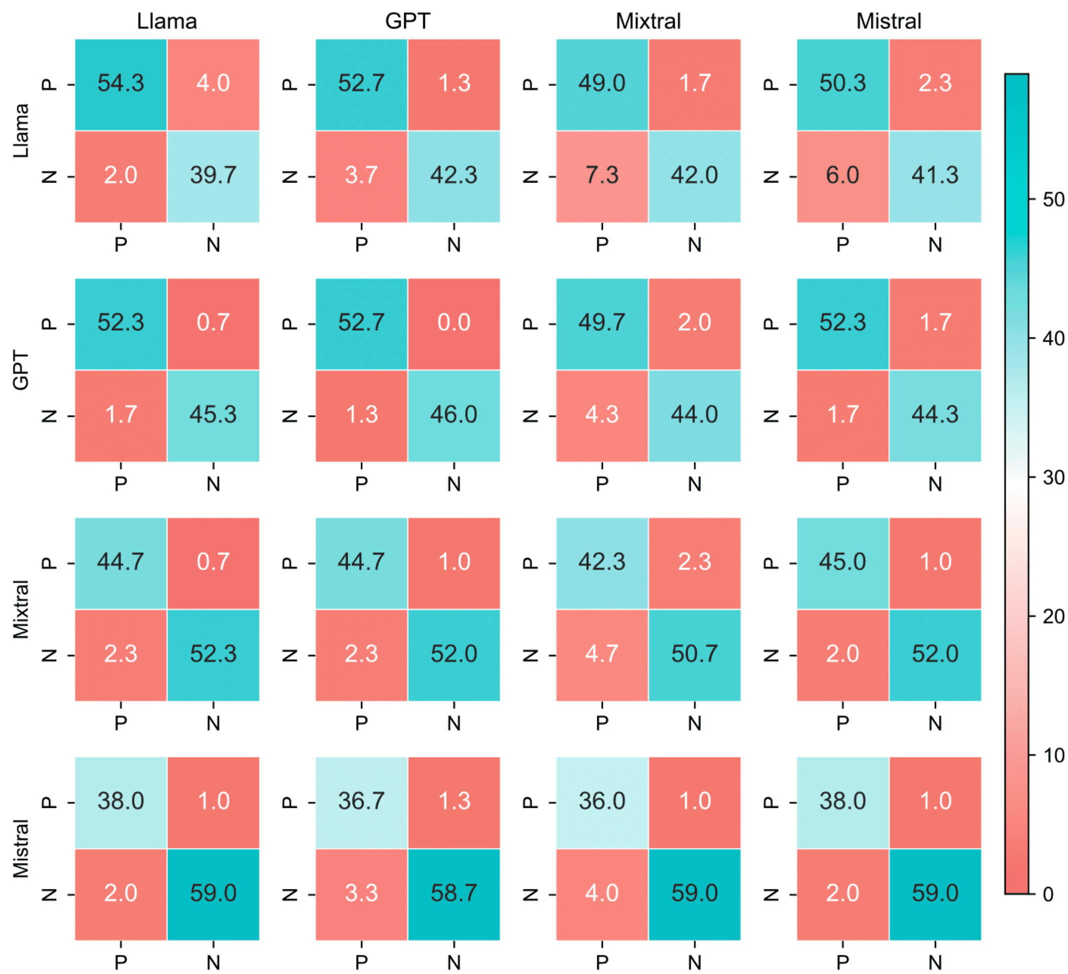

Figure 3. Heatmap illustrating the performance of four individual LLM judges on HotpotQA. Each cell value represents percentages (%). Rows represent predicted outcomes (P: Positive, N: Negative), while columns represent actual outcomes. See Appendix C for full results.

4.2. Analysis

In our experiments, candidate LLMs generated 7,500 outputs for the given tasks, with each evaluator producing 7,500 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 space constraints, we moved the detailed analysis of EM and BERTScore to Appendix C 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[60], where they fabricate reasoning to support their evaluations, produce detailed but incorrect explanations, or reference non-existent criteria or standards. In LLM judges, false positives and negatives (e.g., see Figure 3) 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.

Verbosity[61] emerges as a subtle source of bias, where more elaborate answers are sometimes overrated simply due to their detail and fluency, while concise yet correct responses are undervalued. This misplaced emphasis leads to irrelevant judgment criteria, such as praising the presence of irrelevant information or penalizing perfectly valid but succinct answers. We also found that LLM-based judges encounter challenges in multiple reference answers and more open-ended questions. This confusion is especially pronounced in the TriviaQA where the diversity and flexibility of valid responses present challenges for the judges' ability to consistently recognize and evaluate a range of correct answers.

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[45], being recent, serves as a valuable testbed for assessing these temporal deficiencies. As shown in Table 3, 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 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.*"

	Evaluators					
LLMs	**DeepSeek**	**Llama**	**GPT**	**Mixtral**	**Mistral**	**DAFE**
DeepSeek	0.714	0.692	0.715	0.614	0.724	**0.830**
Llama	0.801	0.835	0.737	0.817	0.730	**0.917**
GPT	0.659	0.695	0.824	0.780	0.746	**0.891**
Mixtral	0.732	0.708	0.779	0.738	0.703	**0.936**
Mistral	0.687	0.665	0.802	0.818	0.723	**0.880**

Table 3. Performance (in Macro F1) of LLM judges on FreshQA.

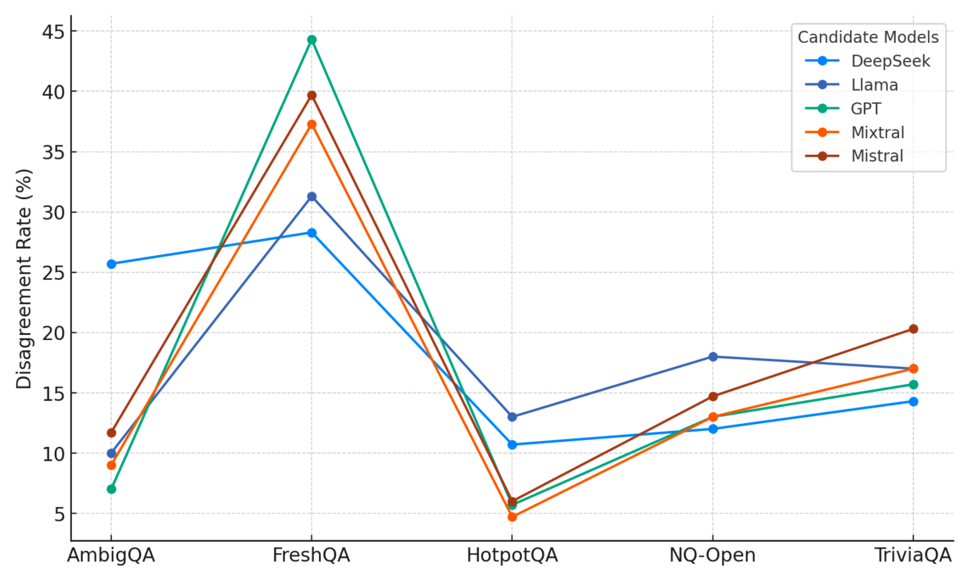

Figure 4. Disagreement rates between primary judges (Llama+Mistral) across candidate LLMs and tasks.

4.3. Disagreements between primary judges

Figure 4 shows that disagreements between our primary judges, Llama-3.1 70B and Mistral 7B, mainly occur in the TriviaQA and FreshQA, with disagreement rates reaching 20.3% and 44.3%, respectively. Interestingly, higher disagreement rates between primary judges create a greater opportunity for DAFE

doi.org/10.32388/B69SKY

to refine evaluations. As depicted in Figure 4, FreshQA (31.3% for Llama-70B, 39.7% for Mistral-7B) demonstrates the highest disagreement, allowing DAFE to improve Macro F1 scores (see Table 3).

4.4. Impact of arbitration

Our proposed arbitration approach significantly enhanced evaluation performance by resolving disputes through an independent judge, GPT-3.5-turbo (see Figure 5). Notably, in the AmbigQA, Macro F1 scores advanced from 72.9% to 86.6%, and Cohen's Kappa increased from 0.467 to 0.773 (see Figure 7). These improvements highlight the pivotal role of the arbitrator in ensuring reliable and consistent evaluation outcomes.

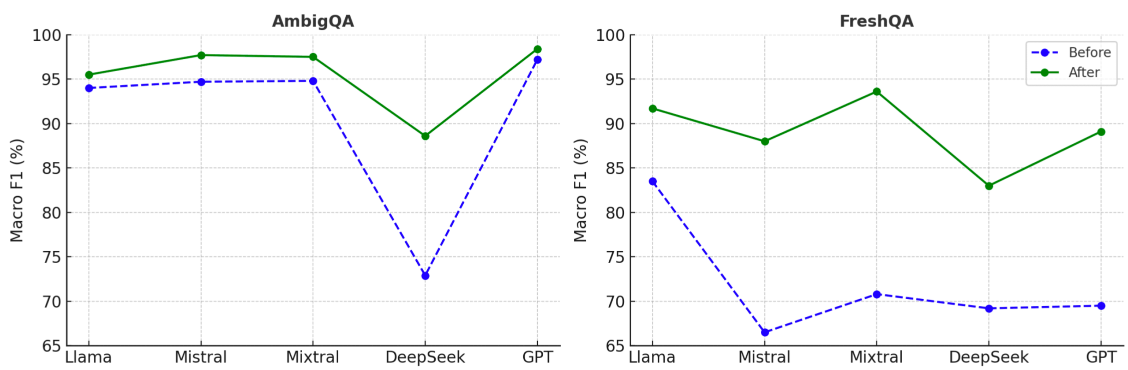

Figure 5. Comparison of Macro F1 scores before and after arbitration (see Appendix C for more results).

5. Related work

Evaluation of natural language generation has traditionally relied on supervised signal-based metrics such as EM which evaluates the exact lexical match between generated outputs and reference answers. Despite its simplicity and efficiency, EM overlooks semantically equivalent variations, often penalizing accurate responses that use different phrasing[62][63]. Other commonly used metrics including BLEU[9] and ROUGE[10] primarily focus on n-gram overlap with human written reference texts. Despite their widespread use, these metrics have significant limitations in capturing semantic subtleties and contextual relevance[13]. To address the limitations of conventional metrics, various model-based methods such as BERTScore[13] offer semantically informed evaluation. However, even BERTScore and similar embedding-based methods struggle to effectively evaluate open-ended generation[22][64].

 doi.org/10.32388/B69SKY

Recent advances in LLMs have unlocked new opportunities for automatic and context-aware evaluation[65][19][22]. A key strength of LLM-based evaluators lies in their ability to operate in reference-free settings, where evaluation does not rely on pre-defined 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[66]. For instance, LLMs are frequently used in subjective evaluations such as pairwise comparison ("Which response is better?") or single-response scoring ("How good is this response based on criteria X?"[25][28]. LLM-based evaluators are specifically effective for tasks like summarization, where subjective criteria are central to evaluation[27]. However, they are less effective for fact-based tasks such as free-form question-answering, where responses are either correct or incorrect and require explicit verification against reference answers.

Furthermore, LLM-based evaluators face several challenges, particularly in ensuring consistency and fairness[61][67]. In reference-free settings, the absence of a definitive ground truth increases the risk of bias in evaluations[61][68][69]. Common biases include positional bias, where LLMs may favor responses based on their order[22][67], verbosity bias, which favors longer or more detailed responses[70], and self-enhancement bias, where models may disproportionately prefer their own outputs[22]. These biases can distort evaluations and undermine the reliability of the results.

6. Conclusion

We present DAFE, 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 closer alignment with human evaluations. However, relying solely on individual judges poses challenges including inherent biases and prompt sensitivity, which can affect evaluation performance. DAFE addresses these challenges through a dynamic arbitration mechanism. This design achieves near-perfect agreement with human evaluations, establishing DAFE as a trustworthy and reliable framework for evaluating open-ended language generation tasks. In the future, we aim to explore DAFE by excluding reference answers and integrating LLM agents with tools-interacting capabilities for evaluation.

7. 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. Inconsistent or ambiguous references could affect evaluation outcomes. Similarly, this study primarily uses binary verdicts which might 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. Finally, our study focuses exclusively on English, and the applicability of our approach to other languages, particularly morphologically rich or resource-scarce ones, remains unexplored.

Appendix A. Free-form Question-Answering

In our experiments, we include AmbigQA[44], FreshQA[45], HotpotQA[46], Natural Questions[47], and TriviaQA[48].

- **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.

Appendix B. 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.

B.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 potential 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

doi.org/10.32388/B69SKY

information as the reference answer. To maintain reliability, document any challenging cases requiring further discussion with other annotators.

B.2. Inter human annotator agreement

We calculate Fleiss' Kappa (κ)[55] to assess inter-rater reliability among human annotators. Table 4 and 5 show the inter-annotator agreement across models and tasks.

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 4. Fleiss' Kappa scores of human annotators 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 5 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.

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 5. Human annotators percent agreement scores across candidate models and tasks.

Appendix C. Additional results

This section provides further results and analysis of conventional metrics and LLM-based evaluators. Table 6 illustrates the overall performance of candidate LLMs obtained through various evaluators. Unlike lexical matching and neural-based metrics, each LLM-as-a-judge indicates overall performance close to the human majority. Automatic metrics like EM severely underestimate the candidate LLMs' performance. On the other hand, BERTScore tends to overestimate the performance.

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.

doi.org/10.32388/B69SKY

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

LLMs	Tasks	Evaluators							
		EM	BS	HM	DeepSeek	Llama	GPT	Mixtral	Mistral
	TriviaQA	62.0	94.3	83.7	78.0	81.3	81.0	79.7	85.0

Table 6. Raw performance of candidate LLMs across free-form QA tasks evaluated through various methods. HM represents Human Majority and BS denotes BERTScore.

C.1. Impact of arbitration on dispute resolution

Figure 6 illustrates the impact of arbitration on resolving disagreements between primary judges. Arbitration, facilitated by GPT-3.5 as the tiebreaker, 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. This highlights the critical role of arbitration in enhancing agreement and achieving closer alignment with ground truth, especially in cases of significant disagreement among primary judges.

Notably, evaluations of DeepSeek-v3 exhibit 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.

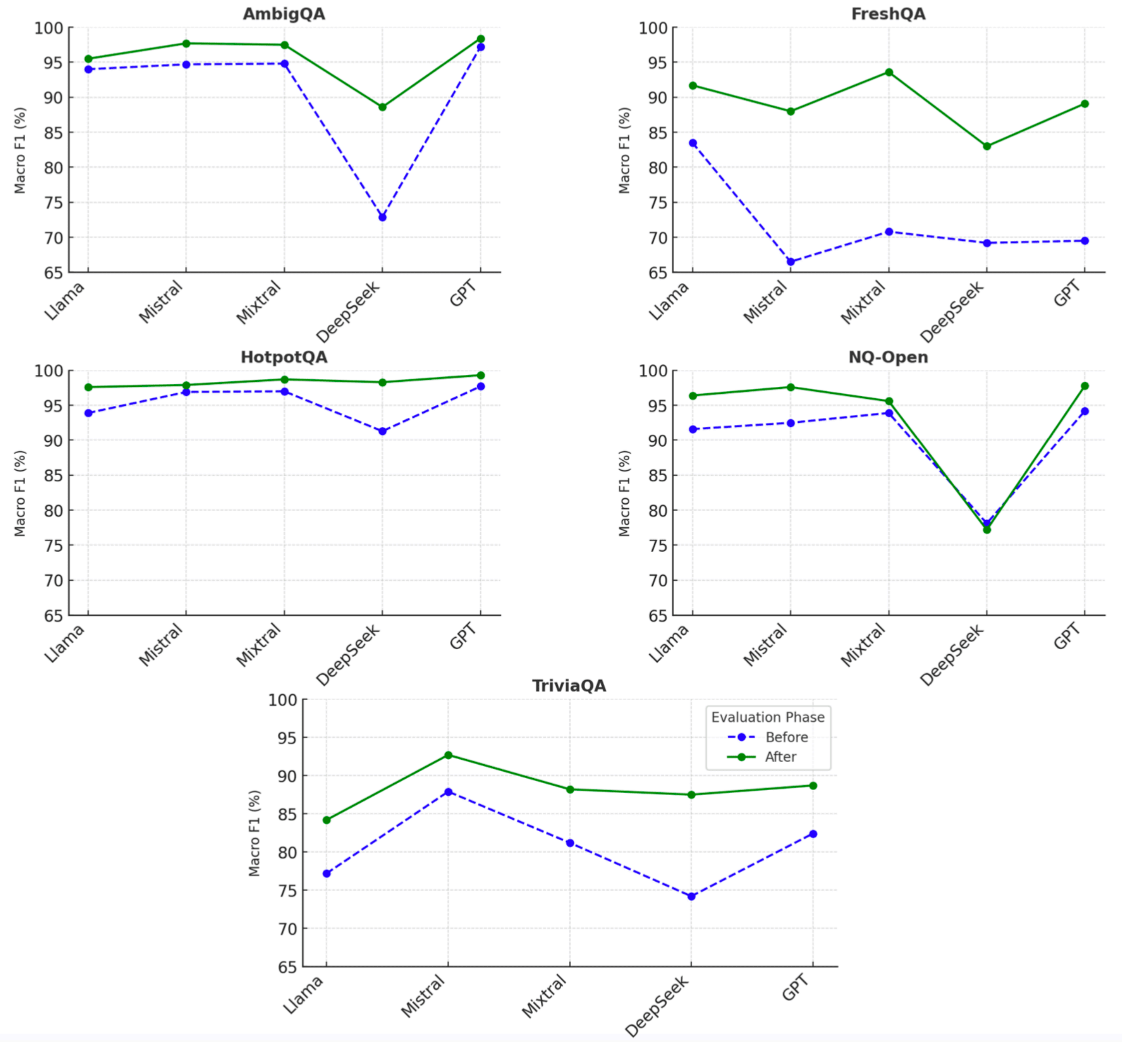

Figure 6. Impact of arbitration on disagreements between primary judges. Note that we used Llama-3.1-70B and Mistra 7B as primary judges. GPT-3.5-turbo is only utilized when disagreements are found. The models given in the figure are candidate LLMs which generate outputs for the given tasks and are then evaluated through DAFE.

We observed substantial enhancements in Cohen's Kappa scores across several tasks. For instance, as illustrated in Figure 7, 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. These improvements demonstrate that the arbitration mechanism effectively enhances the reliability and consistency of evaluations, particularly in complex and ambiguous tasks where primary judges are more likely to disagree.

doi.org/10.32388/B69SKY

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 arbitration process effectively mitigates inconsistencies, especially in tasks involving evolving knowledge and nuanced interpretations, such as FreshQA.

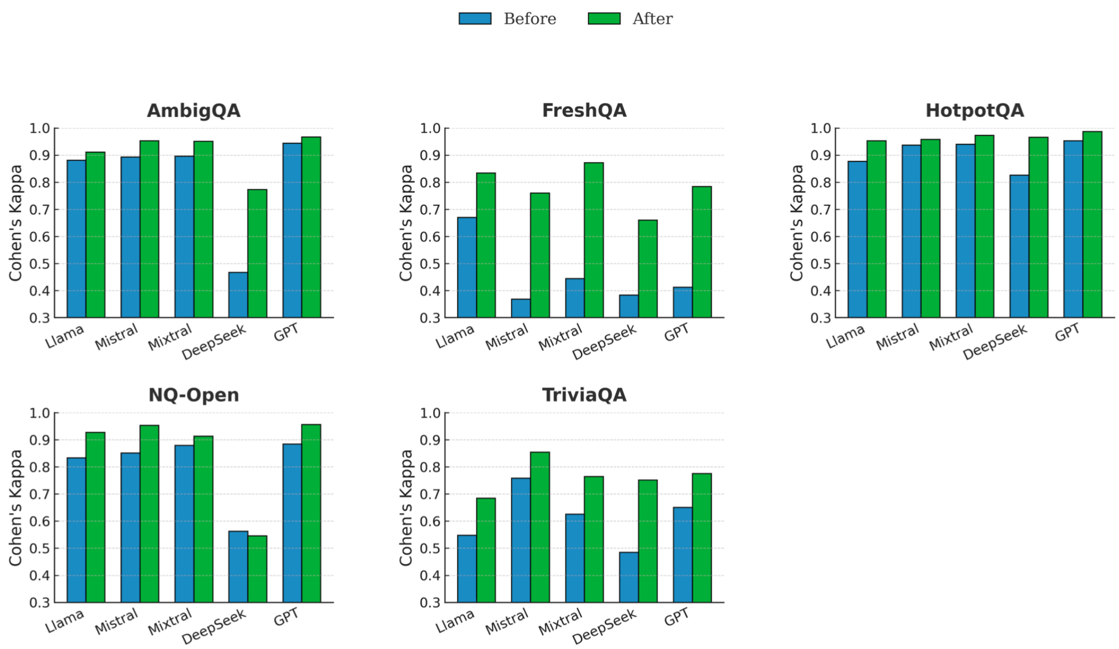

Figure 7. Comparison of Cohen's kappa scores before and after arbitration (GPT-3.5-turbo as arbitrator). The performance is illustrated across candidate LLMs and tasks.

C.2. Cost analysis

Human evaluation is the gold standard for assessing LLM-generated responses, but it is expensive and time-consuming. In our setup, we employed three human annotators who volunteered their efforts. However, if these annotators were compensated based on standard annotation rates, the cost of evaluating such outputs would be significantly higher. On the other hand, GPT-3.5-turbo, acting as an arbitrator in DAFE, incurs a cost that depends on the number of arbitration cases. In our evaluations, GPT-3.5-turbo was invoked 1,318 times, with an estimated total cost of $0.59, which increases to $5.40 if a 2048 max token setting is used (see Table 7). Since GPT-3.5 is only invoked when primary judges disagree, this selective arbitration substantially reduces overall evaluation expenses while maintaining high reliability in assessments. Rather than relying on a single model for evaluation, this multi-model arbitration approach enhances trust by mitigating biases and weaknesses inherent in any individual model.

By invoking the arbitrator only when disagreements occur (rather than evaluating all responses), DAFE reduces arbitration usage by 82–94% compared to a majority-voting system. This leads to:

- Over 90% fewer third-judge inferences, drastically lowering computational demand.
- Up to 95% cost savings by avoiding redundant model evaluations.
- Better scalability, making it practical for large-scale deployments

doi.org/10.32388/B69SKY

Candidate LLMs	Tasks	Samples	Disagreement Rates (%)	Tiebreaker 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

Candidate LLMs	Tasks	Samples	Disagreement Rates (%)	Tiebreaker Usage
Total		**7500**		**1318**

Table 7. Cost-efficiency analysis of DAFE: Summary of disagreement rates and tiebreaker usage across candidate models and tasks

C.3. DeepSeek as the arbitrator

To assess the impact of using DeepSeek as the arbitrator in DAFE, we conducted experiments by replacing GPT-3.5-turbo with DeepSeek. 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 arbitration was invoked only in cases of disagreement. Our findings indicate that DeepSeek as the arbitrator achieves strong performance, with Macro-F1 scores of 91.23 on TriviaQA, 79.11 on NQ-Open, and 0.914 on FreshQA.

C.4. Evaluating with one strong LLM-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 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 DAFE'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, DAFE's ensemble approach acts as a safeguard. When employing DAFE with GPT-3.5-turbo as the arbitrator, 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 arbitrator in DAFE, performance remained strong at 0.963 Macro-F1, indicating that DAFE's benefits are not solely tied to a specific arbitrator model.

doi.org/10.32388/B69SKY

C.5. Majority voting-based evaluation

We conducted additional experiments utilizing a traditional majority voting approach for evaluating candidate LLM performance. In this setup, we employed three LLM judges of equal weight: Llama, GPT-3.5, and Mistral to evaluate candidate models generated response. 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 8, DAFE 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 DAFE 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, DAFE 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, DAFE 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 (DAFE's Macro F1 = 0.88 vs. 0.95 for majority voting), reflect rare instances where both the primary judges and the arbitrator make errors, yet these outliers are substantially outweighed by the computational savings offered by selective arbitration.

doi.org/10.32388/B69SKY

Candidate LLM	Task	Majority Voting		Disagreement (%)	DAFE	
		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 8. Comparison between Majority Voting (Llama+GPT-3.5+Mistral) and DAFE (GPT-3.5 as arbitrator). 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 DAFE.

C.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

doi.org/10.32388/B69SKY

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.

C.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.

C.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.

C.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.

We found that:

doi.org/10.32388/B69SKY

- **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.

C.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.

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

Existing LLM-based evaluators such as G-Eval[27] 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, DAFE 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 G-eval based method 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 prompted 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 9 and 10 values in blue). This systematic deterioration spans all tasks and model combinations, though its severity varies by context. HotpotQA, with its demands for complex reasoning, exemplifies this challenge most clearly. The substantial gap between reference-based and reference-free evaluation underscores the crucial role of reference answers in reliable assessment.

Candidate LLMs	Tasks	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 9. Overall performance (Accuracy) of candidate LLMs across free-form QA tasks. Values [in blue] represent LLM-as-a-judge in the reference-free mood.

doi.org/10.32388/B69SKY

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

Table 10. Performance (Macro F1) of various evaluators across candidate LLMs and tasks. Values [in blue] represent LLM-as-a-judge in the reference-free mode.

C.8. DAFE in multi-reference answers

DAFE 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), DAFE 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.

C.9. Analysis of automatic metrics

Figures 8, 9, 10, and 11 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 error 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 require multi-hop reasoning, true positives reach 36.0-50.3%, with an increase in false

doi.org/10.32388/B69SKY

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.

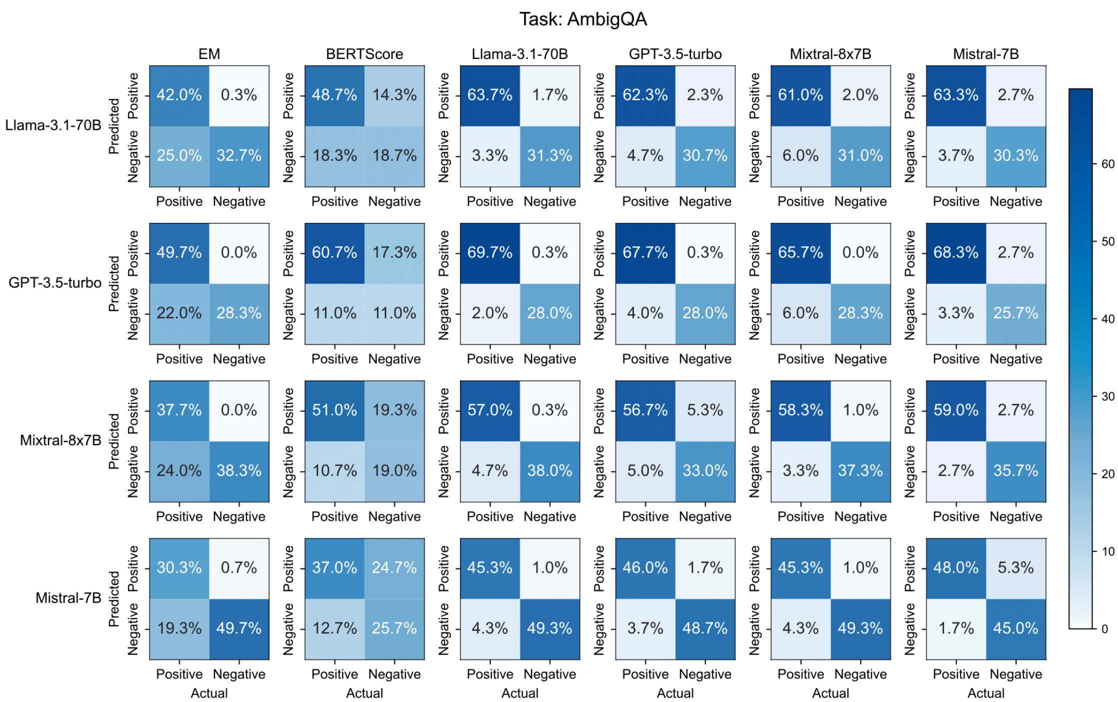

Figure 8. Confusion matrices comparing the performance of automatic metrics (EM, BERTScore) and individual LLM judges on AmbigQA.

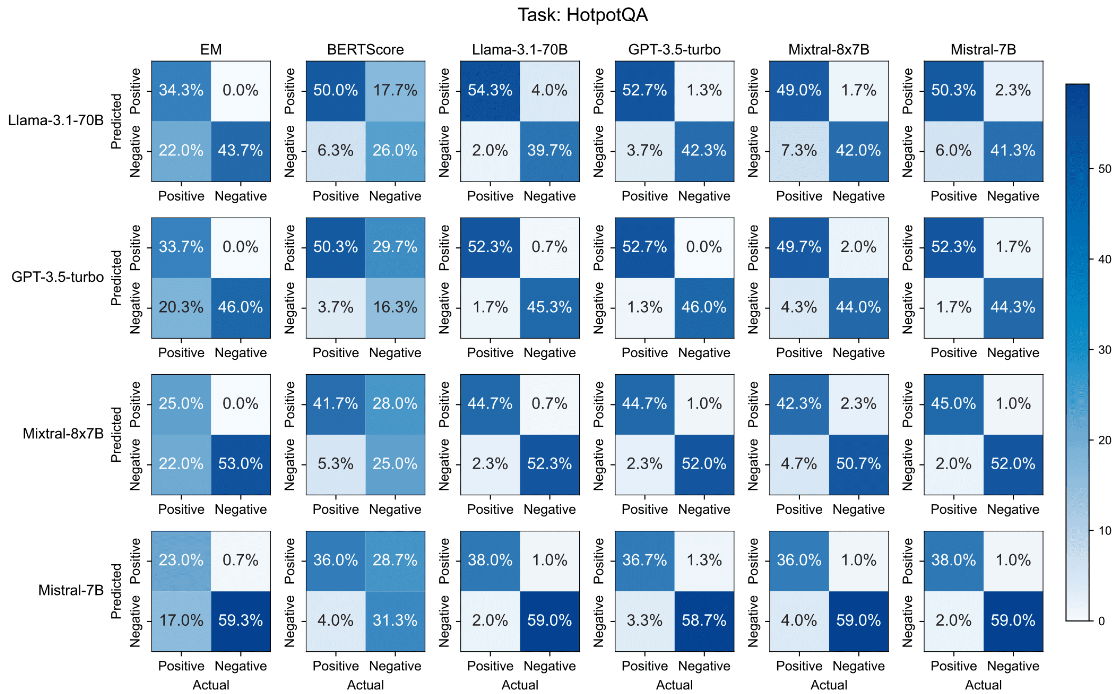

Figure 9. Confusion matrices comparing the performance of automatic metrics (EM, BERTScore) and individual LLM judges on HotpotQA.

doi.org/10.32388/B69SKY

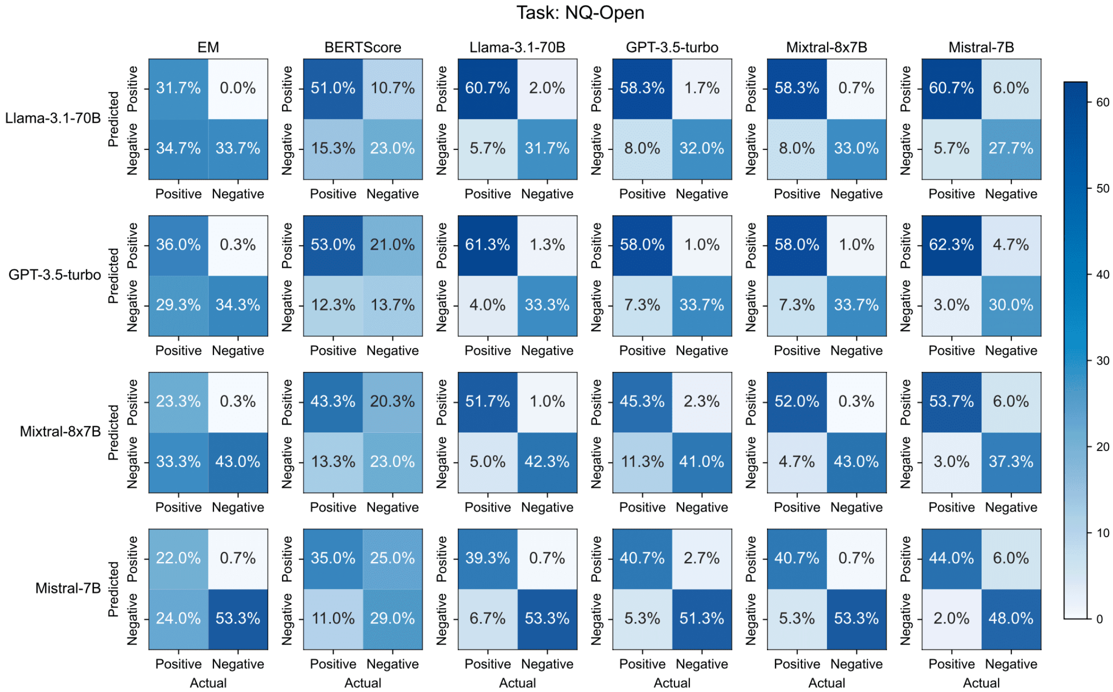

Figure 10. Confusion matrices comparing the performance of automatic metrics (EM, BERTScore) and individual LLM judges on NQ-Open.

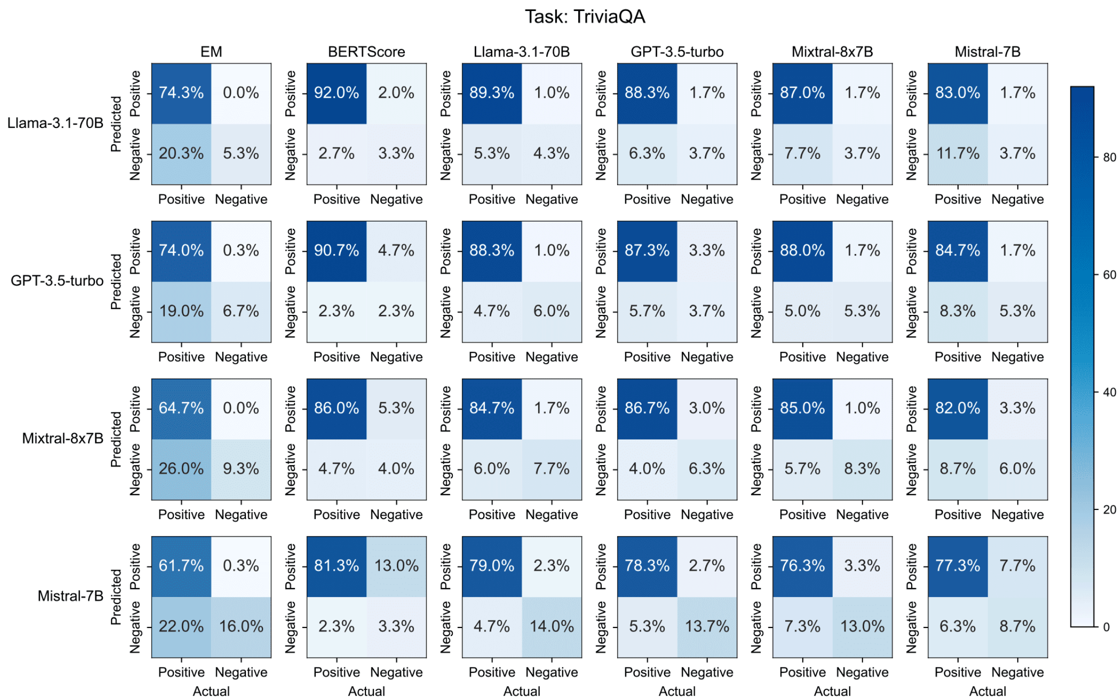

Figure 11. Confusion matrices comparing the performance of automatic metrics (EM, BERTScore) and individual LLM judges on TriviaQA.

Appendix D. Prompting

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

D.1. Prompting Candidate LLMs

We prompted candidate LLMs (see Figure 12) 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 13 shows the candidate GPT-3.5-turbo response at zero temperature for the input given in Figure 12.

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

Figure 12. 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 13. Candidate GPT-3.5-turbo response to the input given in Figure 12.

D.2. Prompting LLM Judges

We prompted LLMs-as-judges to perform the evaluation (see Figure 14). In Figure 15, judge Llama-3.1-70B evaluating 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. So the answer 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 14. 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 15. Llama-3.1-70B-Judge verdict on the candidate GPT-3.5-turbo output.

Footnotes

[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

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

References

1. ^Laskar MTR, Alqahtani S, Bari MS, Rahman M, Khan MAM, Khan H, Jahan I, Bhuiyan A, Tan CW, Parvez MR, Hoque E, Joty S, Huang J (2024). "A systematic survey and critical review on evaluating large language models: Challenges, limitations, and recommendations". Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. Miami, Florida, USA: Association for Computational Linguistics. pp. 13785–13816. doi:10.18653/v1/2024.emnlp-main.764.
2. ^Si C, Zhao C, Boyd-Graber J (2021). "What's in a name? answer equivalence for open-domain question answering". Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 9623–9629. doi:10.18653/v1/2021.emnlp-main.757.
3. ^Hendrycks D, Burns C, Basart S, Zou A, Mazeika M, Song D, Steinhardt J (2021). "Measuring Massive Multitask Language Understanding". arXiv. arXiv:2009.03300.
4. ^Chen Y, Xu B, Wang Q, Liu Y, Mao Z (2024). "Benchmarking large language models on controllable generation under diversified instructions". Proceedings of the AAAI Conference on Artificial Intelligence. **38**: 17808–17816. doi:10.1609/aaai.v38i16.29734.
5. ^Thakur AS, Choudhary K, Ramayapally VS, Vaidyanathan S, Hupkes D (2024). "Judging the Judges: Evaluating Alignment and Vulnerabilities in LLMs-as-Judges". arXiv. arXiv:2406.12624.
6. ^Chang Y, Wang X, Wang J, Wu Y, Yang L, Zhu K, Chen H, Yi X, Wang C, Wang Y, et al. (2024). "A survey on evaluation of large language models". ACM Transactions on Intelligent Systems and Technology. **15** (3): 1–45. doi:10.1145/3641289.
7. ^Li J, Sun S, Yuan W, Fan RZ, Zhao H, Liu P (2023). "Generative Judge for Evaluating Alignment". arXiv. arXiv:2310.05470 [cs.CL].
8. ^Zhang Y, Zhang M, Yuan H, Liu S, Shi Y, Gui T, Zhang Q, Huang X (2024). "LLMEval: A Preliminary Study on How to Evaluate Large Language Models". Proceedings of the AAAI Conference on Artificial Intelligence.

38 (17): 19615–19622. doi:10.1609/aaai.v38i17.29934.

9. [a], [b] *Papineni K, Roukos S, Ward T, Zhu WJ (2002). "Bleu: a Method for Automatic Evaluation of Machine Translation". In: Isabelle P, Charniak E, Lin D, editors. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics. p. 311-318. doi:10.3115/1073083.1073135. Available from: https://aclanthology.org/P02-1040.*
10. [a], [b] *Lin CY (2004). "ROUGE: A Package for Automatic Evaluation of Summaries". In: Text Summarization Branches Out. Barcelona, Spain: Association for Computational Linguistics; p. 74-81.*
11. [a], [b] *Zhu L, Wang X, Wang X (2023). "Judgelm: Fine-tuned large language models are scalable judges". arXiv preprint arXiv:2310.17631. doi:10.48550/arXiv.2310.17631.*
12. [^] *Chen M, Tworek J, Jun H, Yuan Q, Ponde de Oliveira Pinto H, Kaplan J, Edwards H, Burda Y, Joseph N, Brockman G, Ray A, Puri R, Krueger G, Petrov M, Khlaaf H, Sastry G, Mishkin P, Chan B, Gray S, Ryder N, Pavlov M, Power A, Kaiser L, Bavarian M, Winter C, Tillet P, Petroski Such F, Cummings D, Plappert M, Chantzis F, Barnes E, Herbert-Voss A, Hebgen Guss W, Nichol A, Paino A, Tezak N, Tang J, Babuschkin I, Balaji S, Jain S, Saunders W, Hesse C, Carr AN, Leike J, Achiam J, Misra V, Morikawa E, Radford A, Knight M, Brundage M, Murati M, Mayer K, Welinder P, McGrew B, Amodei D, McCandlish S, Sutskever I, Zaremba W (2021). "Evaluating Large Language Models Trained on Code". arXiv. Available from: https://arxiv.org/abs/2107.03374.*
13. [a], [b], [c], [d], [e] *Zhang T, Kishore V, Wu F, Weinberger KQ, Artzi Y (2020). "BERTScore: Evaluating Text Generation with BERT". 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net. BibTeX.*
14. [^] *Liu Y, Yang T, Huang S, Zhang Z, Huang H, Wei F, Deng W, Sun F, Zhang Q (2024). "Calibrating LLM-Based Evaluator". In: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). Torino, Italia: ELRA and ICCL. p. 2638–2656. Available from: https://aclanthology.org/2024.lrec-main.237.*
15. [^] *Xu W, Wang D, Pan L, Song Z, Freitag M, Wang W, Li L (2023). "INSTRUCTSCORE: Towards explainable text generation evaluation with automatic feedback". Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Singapore: Association for Computational Linguistics. pp. 5967–5994. doi:10.18653/v1/2023.emnlp-main.365. Available from: https://aclanthology.org/2023.emnlp-main.365.*
16. [^] *Doostmohammadi E, Holmstr6m O, Kuhlmann M (2024). "How Reliable Are Automatic Evaluation Methods for Instruction-Tuned LLMs?" arXiv preprint arXiv:2402.10770.*
17. [^] *Saito K, Wachi A, Wataoka K, Akimoto Y (2023). "Verbosity Bias in Preference Labeling by Large Language Models". arXiv. arXiv:2310.10076 [cs.CL].*

18. ^Wang Y, Chen L, Cai S, Xu Z, Zhao Y (2024). "Revisiting automated evaluation for long-form table question answering". Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. Miami, Florida, USA: Association for Computational Linguistics. pp. 14696–14706. doi:10.18653/v1/2024.emnlp-main.815. Available from: https://aclanthology.org/2024.emnlp-main.815.

19. a, b, c, d Chiang C-H, Lee H (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), Toronto, Canada: Association for Computational Linguistics; p. 15607–15631. doi:10.18653/v1/2023.acl-long.870. Available from: https://aclanthology.org/2023.acl-long.870.

20. ^Mañas O, Krojer B, Agrawal A (2024). "Improving automatic vqa evaluation using large language models". Proceedings of the AAAI Conference on Artificial Intelligence. **38**: 4171–4179. Available from: https://arxiv.org/pdf/2310.02567v2.

21. ^Clark E, August T, Serrano S, Haduong N, Gururangan S, Smith NA (2021). "All that's 'human' is not gold: Evaluating human evaluation of generated text". 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. doi:10.18653/v1/2021.acl-long.565. Link.

22. a, b, c, d, e, f, g, h Zheng L, Chiang W-L, Sheng Y, Zhuang S, Wu Z, Zhuang Y, Lin Z, Li Z, Li D, Xing EP, Zhang H, Gonzalez JE, Stoica I (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. Available from: https://proceedings.neurips.cc/paper_files/paper/2023/file/91f18a1287b398d378ef22505bf41832-Paper-Datasets_and_Benchmarks.pdf.

23. ^Wang P, Li L, Chen L, Cai Z, Zhu D, Lin B, Cao Y, Liu Q, Liu T, Sui Z (2023). "Large Language Models are not Fair Evaluators". arXiv. arXiv:2305.17926 [cs.CL].

24. ^Vu T, Krishna K, Alzubi S, Tar C, Faruqui M, Sung YH (2024). "Foundational Autoraters: Taming Large Language Models for Better Automatic Evaluation". arXiv. arXiv:2407.10817 [cs.CL].

25. a, b, c Verga P, Hofstatter S, Althammer S, Su Y, Piktus A, Arkhangorodsky A, Xu M, White N, Lewis P (2024). "Replacing Judges with Juries: Evaluating LLM Generations with a Panel of Diverse Models". arXiv. **cs.CL**: 2404.18796.

26. ^Hu X, Gao M, Hu S, Zhang Y, Chen Y, Xu T, Wan X (2024). "Are LLM-based Evaluators Confusing NLG Quality Criteria?" arXiv. arXiv:2402.12055.

27. a, b, c, d Liu Y, Iter D, Xu Y, Wang S, Xu R, Zhu C (2023). "G-Eval: NLG Evaluation using Gpt-4 with Better Human Alignment". Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing.

Singapore: Association for Computational Linguistics. pp. 2511–2522. doi:10.18653/v1/2023.emnlp-main.153. Available from: https://aclanthology.org/2023.emnlp-main.153.

28. a, b Chan CM, Chen W, Su Y, Yu J, Xue W, Zhang S, Fu J, Liu Z (2024). "ChatEval: Towards Better LLM-based Evaluators through Multi-Agent Debate". The Twelfth International Conference on Learning Representations. Available from: https://openreview.net/forum?id=FQepisCUWu.

29. ^ Chu Z, Ai Q, Tu Y, Li H, Liu Y (2024). "PRE: A Peer Review Based Large Language Model Evaluator". arXiv:2401.15641 [cs.IR].

30. a, b Gou Z, Shao Z, Gong Y, Shen Y, Yang Y, Duan N, Chen W (2024). "CRITIC: Large Language Models Can Self-Correct with Tool-Interactive Critiquing". arXiv. **cs.CL**: 2305.11738.

31. ^ Jung J, Brahman F, Choi Y (2024). "Trust or Escalate: LLM Judges with Provable Guarantees for Human Agreement". arXiv **2407.18370** [cs.LG].

32. ^ Adlakha V, BehnamGhader P, Lu XH, Meade N, Reddy S (2024). "Evaluating correctness and faithfulness of instruction-following models for question answering". Transactions of the Association for Computational Linguistics. **12**: 681–699. doi:10.1162/tacl_a_00667. Available from: https://aclanthology.org/2024.tacl-1.38.

33. ^ Kenton Z, Siegel NY, Kramár J, Brown-Cohen J, Albanie S, Bulian J, Agarwal R, Lindner D, Tang Y, Goodman ND, Shah R (2024). "On scalable oversight with weak LLMs judging strong LLMs". arXiv. arXiv:2407.04622 [cs.LG].

34. ^ Liang T, He Z, Jiao W, Wang X, Wang Y, Wang R, Yang Y, Shi S, Tu Z (2024). "Encouraging divergent thinking in large language models through multi-agent debate". Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. Miami, Florida, USA: Association for Computational Linguistics. pp. 17889–17904. doi:10.18653/v1/2024.emnlp-main.992. Available from: https://aclanthology.org/2024.emnlp-main.992.

35. ^ Sun G, Kagrecha A, Manakul P, Woodland P, Gales M (2024). "SkillAggregation: Reference-free LLM-Dependent Aggregation". arXiv preprint arXiv:2410.10215.

36. a, b DeepSeek-AI, Guo D, Yang D, Zhang H, Song J, Zhang R, Xu R, Zhu Q, Ma S, Wang P, Bi X, Zhang X, Yu X, Wu Y, Wu ZF, Gou Z, Shao Z, Li Z, Gao Z, Liu A, Xue B, Wang B, Wu B, Feng B, Lu C, Zhao C, Deng C, Zhang C, Ruan C, Dai D, Chen D, Ji D, Li E, Lin F, Dai F, Luo F, Hao G, Chen G, Li G, Zhang H, Bao H, Xu H, Wang H, Ding H, Xin H, Gao H, Qu H, Li H, Guo J, Li J, Wang J, Chen J, Yuan J, Qiu J, Li J, Cai JL, Ni J, Liang J, Chen J, Dong K, Hu K, Gao K, Guan K, Huang K, Yu K, Wang L, Zhang L, Zhao L, Wang L, Zhang L, Xu L, Xia L, Zhang M, Zhang M, Tang M, Li M, Wang M, Li M, Tian N, Huang P, Zhang P, Wang Q, Chen Q, Du Q, Ge R, Zhang R, Pan R, Wang R, Chen RJ, Jin RL, Chen R, Lu S, Zhou S, Chen S, Ye S, Wang S, Yu S, Zhou S, Pan S, Li SS, Zhou S, Wu S, Ye S, Yu

n T, Pei T, Sun T, Wang T, Zeng W, Zhao W, Liu W, Liang W, Gao W, Yu W, Zhang W, Xiao WL, An W, Liu X, Wang X, Chen X, Nie X, Cheng X, Liu X, Xie X, Liu X, Yang X, Li X, Su X, Lin X, Li XQ, Jin X, Shen X, Chen X, Sun X, Wang X, Song X, Zhou X, Wang X, Shan X, Li YK, Wang YQ, Wei YX, Zhang Y, Xu Y, Li Y, Zhao Y, Sun Y, Wang Y, Yu Y, Zhang Y, Shi Y, Xiong Y, He Y, Piao Y, Wang Y, Tan Y, Ma Y, Liu Y, Guo Y, Ou Y, Wang Y, Gong Y, Zou Y, He Y, Xiong Y, Luo Y, You Y, Liu Y, Zhou Y, Zhu YX, Xu Y, Huang Y, Li Y, Zheng Y, Zhu Y, Ma Y, Tang Y, Zha Y, Yan Y, Ren ZZ, Ren Z, Sha Z, Fu Z, Xu Z, Xie Z, Zhang Z, Hao Z, Ma Z, Yan Z, Wu Z, Gu Z, Zhu Z, Liu Z, Li Z, Xie Z, Song Z, Pan Z, Huang Z, Xu Z, Zhang Z, Zhang Z. "DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning". arXiv. 2025. Available from: http://arxiv.org/abs/2501.12948.

37. ^ ***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.*
38. ^ *Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S, Herbert-Voss A, Krueger G, Henighan T, Child R, Ramesh A, Ziegler D, Wu J, Winter C, Hesse C, Chen M, Sigler E, Litwin M, Gray S, Chess B, Clark J, Berner C, McCandlish S, Radford A, Sutskever I, Amodei D (2020). "Language Models are Few-Shot Learners". In: Larochelle H, Ranzato M, Hadsell R, Balcan MF, Lin H, editors. Advances in Neural Information Processing Systems. Curran Associates, Inc.; 2020. p. 1877–1901.*
39. ^ *Jiang AQ, Sablayrolles A, Mensch A, Bamford C, Chaplot DS, de las Casas D, Bressand F, Lengyel G, Lample G, Saulnier L, Lavaud LR, Lachaux MA, Stock P, Le Scao T, Lavril T, Wang T, Lacroix T, El Sayed W (2023). "Mistral 7B". arXiv. 2310.06825.*
40. ^ *Jiang AQ, Sablayrolles A, Roux A, Mensch A, Savary B, Bamford C, Chaplot DS, de las Casas D, Bou Hanna E, Bressand F, Lengyel G, Bour G, Lample G, Lavaud LR, Saulnier L, Lachaux MA, Stock P, Subramanian S, Yang S, Antoniak S, Le Scao T, Gervet T, Lavril T, Wang T, Lacroix T, El Sayed W (2024). "Mixtral of Experts". arXiv. arXiv:2401.04088 [cs.LG].*
41. ^ *OpenAI, Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman FL, Almeida D, Altenschmidt J, Altman S, Anadkat S, Avila R, Babuschkin I, Balaji S, Balcom V, Baltescu P, Bao H, Bavarian M, Belgum J, Bello I, Berdine J, Bernadett-Shapiro G, Berner C, Bogdonoff L, Boiko O, Boyd M, Brakman AL, Brockman G, Brooks T, Brundage M, Button K, Cai T, Campbell R, Cann A, Carey B, Carlson C, Carmichael R, Chan B, Chang C, Chantzis F, Chen D, Chen S, Chen R, Chen J, Chen M, Chess B, Cho C, Chu C, Chung HW, Cummings D, Currier J, Dai Y, Decareaux C, Degry T, Deutsch N, Deville D, Dhar A, Dohan D, Dowling S, Dunning S, Ecoffet A, Eleti A, Eloundou T, Farhi D, Fedus L, Felix N, Fishman SP, Forte J, Fulford I, Gao L, Georges E, Gibson C, Goel V, Gogineni T, Goh G, Gontijo-Lopes R, Gordon J, Grafstein M, Gray S, Greene R, Gross J, Gu SS, Guo Y, Hallacy C, Han J, Harris J, He Y, Heaton M, Heidecke J, Hesse C, Hickey A, Hickey W, Hoeschele P, Houghton B, Hsu K, Hu S, Hu X, Hui*

zinga J, Jain S, Jain S, Jang J, Jiang A, Jiang R, Jin H, Jin D, Jomoto S, Jonn B, Jun H, Kaftan T, Kaiser Ł, Kamali A, Kanitscheider I, Keskar NS, Khan T, Kilpatrick L, Kim JW, Kim C, Kim Y, Kirchner JH, Kiros J, Knight M, Kokotajlo D, Kondraciuk Ł, Kondrich A, Konstantinidis A, Kosic K, Krueger G, Kuo V, Lampe M, Lan I, Lee T, Leike J, Leung J, Levy D, Li CM, Lim R, Lin M, Lin S, Litwin M, Lopez T, Lowe R, Lue P, Makanju A, Malfacini K, Manning S, Markov T, Markovski Y, Martin B, Mayer K, Mayne A, McGrew B, McKinney SM, McLeavey C, McMillan P, McNeil J, Medina D, Mehta A, Menick J, Metz L, Mishchenko A, Mishkin P, Monaco V, Morikawa E, Mossing D, Mu T, Murati M, Murk O, Mély D, Nair A, Nakano R, Nayak R, Neelakantan A, Ngo R, Noh H, Ouyang L, O'Keefe C, Pachocki J, Paino A, Palermo J, Pantuliano A, Parascandolo G, Parish J, Parparita E, Passos A, Pavlov M, Peng A, Perelman A, Peres FdAB, Petrov M, Pinto HPdO, Pokorny M, Pokrass M, Pong VH, Powell T, Power A, Power B, Proehl E, Puri R, Radford A, Rae J, Ramesh A, Raymond C, Real F, Rimbach K, Ross C, Rotsted B, Roussez H, Ryder N, Saltarelli M, Sanders T, Santurkar S, Sastry G, Schmidt H, Schnurr D, Schulman J, Selsam D, Sheppard K, Sherbakov T, Shieh J, Shoker S, Shyam P, Sidor S, Sigler E, Simens M, Sitkin J, Slama K, Sohl I, Sokolowsky B, Song Y, Staudacher N, Such FP, Summers N, Sutskever I, Tang J, Tezak N, Thompson MB, Tillet P, Tootoonchian A, Tseng E, Tuggle P, Turley N, Tworek J, Uribe JFC, Vallone A, Vijayvergiya A, Voss C, Wainwright C, Wang JJ, Wang A, Wang B, Ward J, Wei J, Weinmann CJ, Welihinda A, Welinder P, Weng J, Weng L, Wiethoff M, Willner D, Winter C, Wolrich S, Wong H, Workman L, Wu S, Wu J, Wu M, Xiao K, Xu T, Yoo S, Yu K, Yuan Q, Zaremba W, Zellers R, Zhang C, Zhang M, Zhao S, Zheng T, Zhuang J, Zhuk W, Zoph B. 2023. "GPT-4 Technical Report". arXiv [cs.CL].

42. [a, b]*Hada R, Gumma V, de Wynter A, Diddee H, Ahmed M, Choudhury M, Bali K, Sitaram S (2024). "Are Large Language Model-based Evaluators the Solution to Scaling Up Multilingual Evaluation?" arXiv. arXiv:2309.07462 [cs.CL].*
43. [^]*Evans O, Cotton-Barratt O, Finnveden L, Bales A, Balwit A, Wills P, Righetti L, Saunders W (2021). "Truthful AI: Developing and governing AI that does not lie". arXiv. arXiv:2110.06674 [cs.CY].*
44. [a, b]*Min S, Michael J, Hajishirzi H, Zettlemoyer L. "AmbigQA: Answering ambiguous open-domain questions". Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020:5783-5797. doi:10.18653/v1/2020.emnlp-main.466.*
45. [a, b, c]*Tu Vu, Iyyer M, Wang X, Constant N, Wei J, Wei J, Tar C, Sung YH, Zhou D, Le Q, Luong T (2023). "FreshLLMs: Refreshing Large Language Models with Search Engine Augmentation". arXiv. arXiv:2310.03214.*
46. [a, b]*Yang Z, Qi P, Zhang S, Bengio Y, Cohen WW, Salakhutdinov R, Manning CD (2018). "HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering". arXiv. arXiv:1809.09600 [cs.CL].*

47. ^a, b^ *Kwiatkowski T, Palomaki J, Redfield O, Collins M, Parikh A, Alberti C, Epstein D, Polosukhin I, Devlin J, Lee K, Toutanova K, Jones L, Kelcey M, Chang MW, Dai AM, Uszkoreit J, Le Q, Petrov S (2019). "Natural Questions: A Benchmark for Question Answering Research". Transactions of the Association for Computational Linguistics.* **7***: 452--466. doi:*10.1162/tacl a 00276*.*
48. ^a, b^ *Joshi M, Choi E, Weld DS, Zettlemoyer L (2017). "TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension". arXiv.* **cs.CL***: 1705.03551.*
49. ^ *Kong A, Zhao S, Chen H, Li Q, Qin Y, Sun R, Zhou X, Wang E, Dong X (2024). "Better Zero-Shot Reasoning with Role-Play Prompting". arXiv.* **cs.CL***: 2308.07702.*
50. ^ *Min S, Lyu X, Holtzman A, Artetxe M, Lewis M, Hajishirzi H, Zettlemoyer L. Rethinking the role of demonstrations: What makes in-context learning work? In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics; 2022. p. 11048-11064. doi:10.18653/v1/2022.emnlp-main.759. Available from: https://aclanthology.org/2022.emnlp-main.759.*
51. ^ *Izacard G, Grave E (2021). "Leveraging passage retrieval with generative models for open domain question answering". Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. Online: Association for Computational Linguistics. pp. 874–880. doi:10.18653/v1/2021.eacl-main.74.*
52. ^ *Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, Küttler H, Lewis M, Yih W, Rocktäschel T, Riedel S, Kiela D (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.*
53. ^ *Gou Z, Shao Z, Gong Y, Shen Y, Yang Y, Duan N, Chen W (2024). "CRITIC: Large Language Models Can Self-Correct with Tool-Interactive Critiquing". arXiv [cs.CL].* **2305.11738***.*
54. ^ *He P, Liu X, Gao J, Chen W (2021). "DEBERTA: Decoding-enhanced BERT with disentangled attention". In: International Conference on Learning Representations.*
55. ^a, b^ *Fleiss JL, Cohen J (1973). "The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability". Educational and psychological measurement.* **33** *(3): 613–619.*
56. ^ *McHugh ML (2012). "Interrater reliability: the kappa statistic". Biochemia medica.* **22** *(3): 276–282.*
57. ^ *Cicchetti DV, Feinstein AR (1990). "High agreement but low kappa: II. Resolving the paradoxes". Journal of Clinical Epidemiology.* **43** *(6): 551–558.*

58. ^Kaplan J, McCandlish S, Henighan T, Brown TB, Chess B, Child R, Gray S, Radford A, Wu J, Amodei D (2020). "Scaling Laws for Neural Language Models". arXiv. **2001.08361**.
59. ^OpenAI, Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman FL, Almeida D, Altenschmidt J, Altman S, Anadkat S, Avila R, Babuschkin I, Balaji S, Balcom V, Baltescu P, Bao H, Bavarian M, Belgum J, Bello I, Berdine J, Bernadett-Shapiro G, Berner C, Bogdonoff L, Boiko O, Boyd M, Brakman AL, Brockman G, Brooks T, Brundage M, Button K, Cai T, Campbell R, Cann A, Carey B, Carlson C, Carmichael R, Chan B, Chang C, Chantzis F, Chen D, Chen S, Chen R, Chen J, Chen M, Chess B, Cho C, Chu C, Chung HW, Cummings D, Currier J, Dai Y, Decareaux C, Degry T, Deutsch N, Deville D, Dhar A, Dohan D, Dowling S, Dunning S, Ecoffet A, Eleti A, Eloundou T, Farhi D, Fedus L, Felix N, Fishman SP, Forte J, Fulford I, Gao L, Georges E, Gibson C, Goel V, Gogineni T, Goh G, Gontijo-Lopes R, Gordon J, Grafstein M, Gray S, Greene R, Gross J, Gu SS, Guo Y, Hallacy C, Han J, Harris J, He Y, Heaton M, Heidecke J, Hesse C, Hickey A, Hickey W, Hoeschele P, Houghton B, Hsu K, Hu S, Hu X, Huizinga J, Jain S, Jain S, Jang J, Jiang A, Jiang R, Jin H, Jin D, Jomoto S, Jonn B, Jun H, Kaftan T, Kaiser Ł, Kamali A, Kanitscheider I, Keskar NS, Khan T, Kilpatrick L, Kim JW, Kim C, Kim Y, Kirchner JH, Kiros J, Knight M, Kokotajlo D, Kondraciuk Ł, Kondrich A, Konstantinidis A, Kosic K, Krueger G, Kuo V, Lampe M, Lan I, Lee T, Leike J, Leung J, Levy D, Li CM, Lim R, Lin M, Lin S, Litwin M, Lopez T, Lowe R, Lue P, Makanju A, Malfacini K, Manning S, Markov T, Markovski Y, Martin B, Mayer K, Mayne A, McGrew B, McKinney SM, McLeavey C, McMillan P, McNeil J, Medina D, Mehta A, Menick J, Metz L, Mishchenko A, Mishkin P, Monaco V, Morikawa E, Mossing D, Mu T, Murati M, Murk O, Mély D, Nair A, Nakano R, Nayak R, Neelakantan A, Ngo R, Noh H, Ouyang L, O'Keefe C, Pachocki J, Paino A, Palermo J, Pantuliano A, Parascandolo G, Parish J, Parparita E, Passos A, Pavlov M, Peng A, Perelman A, Peres FdAB, Petrov M, Pinto HPdO, Pokorny M, Pokrass M, Pong VH, Powell T, Power A, Power B, Proehl E, Puri R, Radford A, Rae J, Ramesh A, Raymond C, Real F, Rimbach K, Ross C, Rotsted B, Roussez H, Ryder N, Saltarelli M, Sanders T, Santurkar S, Sastry G, Schmidt H, Schnurr D, Schulman J, Selsam D, Sheppard K, Sherbakov T, Shieh J, Shoker S, Shyam P, Sidor S, Sigler E, Simens M, Sitkin J, Slama K, Sohl I, Sokolowsky B, Song Y, Staudacher N, Such FP, Summers N, Sutskever I, Tang J, Tezak N, Thompson MB, Tillet P, Tootoonchian A, Tseng E, Tuggle P, Turley N, Tworek J, Uribe JFC, Vallone A, Vijayvergiya A, Voss C, Wainwright C, Wang JJ, Wang A, Wang B, Ward J, Wei J, Weinmann CJ, Welihinda A, Welinder P, Weng J, Weng L, Wiethoff M, Willner D, Winter C, Wolrich S, Wong H, Workman L, Wu S, Wu J, Wu M, Xiao K, Xu T, Yoo S, Yu K, Yuan Q, Zaremba W, Zellers R, Zhang C, Zhang M, Zhao S, Zheng T, Zhuang J, Zhuk W, Zoph B. 2024. GPT-4 Technical Report.
60. ^Zhang Y, Li Y, Cui L, Cai D, Liu L, Fu T, Huang X, Zhao E, Zhang Y, Chen Y, Wang L, Luu AT, Bi W, Shi F, Shi S (2023). "Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models". arXiv:2309.012

 doi.org/10.32388/B69SKY

19 [cs.CL].

61. ^a, b, c^ *Ye J, Wang Y, Huang Y, Chen D, Zhang Q, Moniz N, Gao T, Geyer W, Huang C, Chen P-Y, Chawla NV, Zhang X (2024). "Justice or Prejudice? Quantifying Biases in LLM-as-a-Judge". arXiv. arXiv:2410.02736 [cs.CL].*
62. ^ *Wang C, Cheng S, Guo Q, Yue Y, Ding B, Xu Z, Wang Y, Hu X, Zhang Z, Zhang Y (2024). "Evaluating open-QA evaluation." In: Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS '23, New Orleans, LA, USA. Red Hook, NY, USA: Curran Associates Inc. Article No. 3367, 30 pages.*
63. ^ *Kamalloo E, Dziri N, Clarke C, Rafiei D. 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); 2023 Jul; Toronto, Canada. Association for Computational Linguistics. p. 5591-5606. doi:10.18653/v1/2023.acl-long.307. Available from: https://aclanthology.org/2023.acl-long.307.*
64. ^ *Sun T, He J, Qiu X, Huang X (2022). "BERTScore is unfair: On social bias in language model-based metrics for text generation." Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics. pp. 3726–3739. doi:10.18653/v1/2022.emnlp-main.245.*
65. ^ *Li H, Dong Q, Chen J, Su H, Zhou Y, Ai Q, Ye Z, Liu Y (2024). "LLMs-as-Judges: A Comprehensive Survey on LLM-based Evaluation Methods".* ***arXiv*** *[cs.CL]. Available from: https://arxiv.org/abs/2412.05579.*
66. ^ *Li D, Jiang B, Huang L, Beigi A, Zhao C, Tan Z, Bhattacharjee A, Jiang Y, Chen C, Wu T, Shu K, Cheng L, Liu H (2024). "From Generation to Judgment: Opportunities and Challenges of LLM-as-a-judge". arXiv. arXiv:2411.16594 [cs.AI].*
67. ^a, b^ *Khan A, Hughes J, Valentine D, Ruis L, Sachan K, Radhakrishnan A, Grefenstette E, Bowman SR, Rocktäschel T, Perez E (2024). "Debating with More Persuasive LLMs Leads to More Truthful Answers". arXiv. arXiv:2402.06782.*
68. ^ *Kim S, Suk J, Longpre S, Lin BY, Shin J, Welleck S, Neubig G, Lee M, Lee K, Seo M (2024). "Prometheus 2: An Open Source Language Model Specialized in Evaluating Other Language Models". arXiv* ***2405.01535*** *[cs.CL].*
69. ^ *Huang H, Qu Y, Bu X, Zhou H, Liu J, Yang M, Xu B, Zhao T (2024). "An Empirical Study of LLM-as-a-Judge for LLM Evaluation: Fine-tuned Judge Model is not a General Substitute for GPT-4". arXiv. arXiv:2403.02839 [cs.CL].*
70. ^ *Huang H, Qu Y, Zhou H, Liu J, Yang M, Xu B, Zhao T (2024). "On the Limitations of Fine-tuned Judge Models for LLM Evaluation". arXiv. arXiv:2403.02839 [cs.CL].*

Declarations

Funding: No specific funding was received for this work.

Potential competing interests: No potential competing interests to declare.