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

Anonymous ACL submission

Abstract

Evaluating Large Language Models (LLMs) free-form generated responses remains a challenge due to their diverse and open-ended nature. Traditional 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 disagreement. This selective arbitration mechanism prioritizes evaluation reliability while reducing unnecessary computational demands. DAFE combines 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 (Laskar et al., 2024). In particular, free-form model responses are difficult to evaluate because their correctness depends on understanding the broader context and underlying meaning (Si et al., 2021). Many benchmarks, such as MMLU (Hendrycks et al., 2021), often simplify evaluation by focusing on structured formats (e.g., multiple-choice questions) (Chen et al., 2024). Although effective for certain tasks, such methods 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 (Thakur et al., 2024). This structured approach fails to accommodate the complexity of free-form responses, where multiple valid answers exist (Chang et al., 2024). The rigid, predefined options in such evaluations not only limit the scope of assessment but also overlook the diversity of potential correct responses in free-form tasks (Li et al., 2023; Zhang et al., 2024).

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 (Papineni et al., 2002; Lin, 2004) primarily assess surface-level similarity and often fail to capture semantic equivalence, particularly when lexical or structural diversity conveys the same underlying meaning (Zhu et al., 2023; Chen et al., 2021; Zhang et al., 2020). Neural-based metrics like BERTScore (Zhang et al., 2020) address such limitations by leveraging contextual embeddings to evaluate semantic similarity. However, BERTScore depends on reference quality (Liu et al., 2024) and struggles with domain adaptation and length variations (Zhu et al., 2023). Furthermore, continuous score provider metrics are difficult to interpret (Xu et al., 2023). The limitations in automatic metrics become particularly evident when evaluating instruction-tuned chat models (Doostmohammadi et al., 2024), which tend to produce verbose and diverse responses (Saito et al., 2023; Wang et al., 2024b).

Contrary to automatic metrics, human evalua-

tion provides a more transparent assessment (Chiang and Lee, 2023). 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 (Mañas et al., 2024). Furthermore, the reliability of human evaluation is additionally challenged by variations in rater expertise and inherent subjectivity that affect reproducibility (Clark et al., 2021; Chiang and Lee, 2023).

Recently, a paradigm shift has emerged where LLMs are utilized to judge the candidate model generations for given tasks (Zheng et al., 2024). This model-based method leverages the instruction-following capabilities of LLMs through evaluation prompts or, in some cases, fine-tuned versions of LLMs that are specifically optimized for evaluation. In this new line of work, research primarily focuses on pairwise comparison (Zheng et al., 2024; Wang et al., 2023; Vu et al., 2024), such as instructing an LLM to judge "which assistant response is better", and single-answer scoring (Verga et al., 2024) like evaluating summarization task based on predefined criteria (e.g., likability, relevance, etc.) (Chiang and Lee, 2023; Hu et al., 2024; Liu et al., 2023; Chan et al., 2024; Chu et al., 2024).

Inspired by a recent study on self-correction where external feedback helps models identify and correct their mistakes (Gou et al., 2024a), 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 (Jung et al., 2024; Adlakha et al., 2024; Verga et al., 2024), 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.

2 Methodology

Our methodology employs multiple judge models to evaluate outputs generated by the candidate LLMs. In the case of disagreement among the judges, our method employs an additional LLM as an arbitrator. In the following, we describe our methodology in detail.

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., $x \rightarrow \mathcal{A}_{\text{llm}}$) 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,

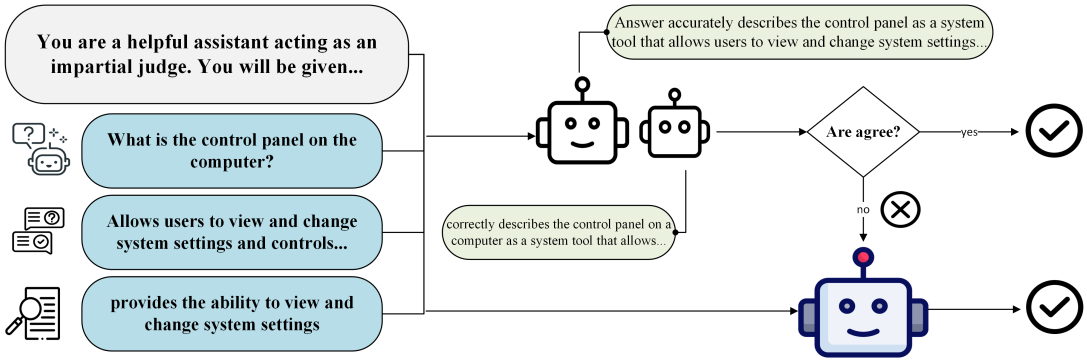

Figure 1: **Our proposed Dynamic Arbitration Framework for Evaluation (DAFE).** Two primary judges, J_1 and J_2, first provide verdicts V_{i_1} and V_{i_2} for an instance i. If agree, that consensus V_i is the final decision D_i. If disagree, a tiebreaker model J_t independently produces a verdict V_t.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

where $\mathcal{J}_{\text{llm}}$ receives no prior examples, to few-shot, which includes several related examples, or a chain of thought, encouraging $\mathcal{J}_{\text{llm}}$ to reason stepwise through the problem.

2.3 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 (Kenton et al., 2024). 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 (Liang et al., 2024; Sun et al., 2024).

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 both candidates and judges in our experiment. These models include Mistral 7B[1] (Jiang et al., 2023), Mixtral 8x7B[2] (Jiang et al., 2024), Llama-3.1 70B[3] (Meta AI, 2024), and GPT-3.5-turbo (Brown et al., 2020). To ensure the reproducibility of our experiments, we set the temperature to 0 for all models under study, as the performance of LLM-based evaluators has been shown to drop when temperature increases (Hada et al.,

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

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

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

2024). 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.

3.2 Datasets

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

3.3 Prompts

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

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

3.4 Baselines

We establish the following baselines.

3.4.1 Exact Math

For our selected datasets and also free-form QA tasks, Exact Match (EM) serves as a standard lexical matching metric to evaluate candidate LLM performance (Izacard and Grave, 2021; Lewis et al., 2020; Gou et al., 2024b). EM classifies an answer as correct if the generated response precisely matches one of the golden answers in the reference set. 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., $\bar{y} = r_i$).

3.4.2 BERTScore

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

3.4.3 Human Evaluation

Human evaluation 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.

We calculate Fleiss' Kappa (κ) (Fleiss and Cohen, 1973) to assess inter-rater reliability among human annotators. Table 1 shows the perfect agreement among annotators across all models and tasks (see Table 4 in Appendix B for detail).

4 Results

Figure 2 illustrates the raw performance of candidate LLMs obtained through various evaluators. Unlike lexical matching and neural-based metrics, each LLM-as-a-judge shows overall perfor-

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

LLMs	AmbigQA	HotpotQA	NQ-Open	TriviaQA
Llama	0.945	0.973	0.985	0.935
GPT	0.989	0.982	0.990	0.948
Mixtral	0.981	0.996	0.977	0.936
Mistral	0.978	0.981	0.978	0.975

Table 1: Human annotators Fleiss' Kappa scores across models and given tasks

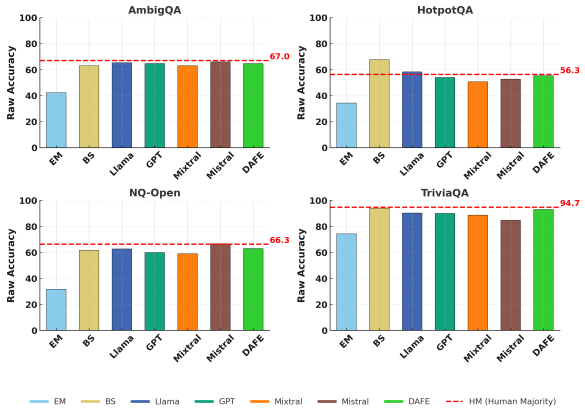

Figure 2: Raw accuracy of candidate LLMs across free-form QA tasks using Exact Match (EM), BERTScore (BS), and model-based evaluation. The Human Majority (HM) serves as the ground truth for all evaluators. See Table 5 in Appendix C for complete results.

mance close to the human majority. The proposed DAFE method consistently achieves comparable or slightly better alignment with the human majority corresponding to individual LLM judges. 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 candidate Llama-3.1-70B on AmbigQA and NQ-Open (see Table 5 in Appendix C for additional results).

4.1 Alignment with human evaluation

We calculate Cohen's kappa (McHugh, 2012) to find the agreement between each evaluator and the human majority (i.e., ground truth) to obtain instance-level comparison. Overall, DAFE is almost perfectly aligned with human judgment than other evaluators (see Table 2). Similarly, individual LLM judges show substantial to a 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"* (Cicchetti and Feinstein, 1990). Therefore, we treat the evaluation as a binary classification task, where we consider each evaluator's predictions against the human majority and report Macro-F1 scores which give equal weight to both classes regardless of their frequency in the selected random samples.

Table 2: 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.

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

As evidenced by consistently high Macro F1 scores in Table 3, DAFE maintains a strong alignment with 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 evident in Llama-3.1-70B and GPT-3.5-turbo which achieve higher Macro-F1 scores (0.91-0.98) across AmbigQA, HotpotQA, and NQ-Open compared to smaller models. This reveals an important scaling law in evaluation capability (Kaplan et al., 2020; Zheng et al., 2024; OpenAI et al., 2024). However, we also found that the most advanced models are not always guaranteed to be the best evaluators. We observed slightly comparable performance through 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 5), lexical matching EM typically accomplishes better alignment with human evaluation on instance-level in Table 3 than neural-based BERTScore. EM's strict and conservative nature leads to lower over-

Table 3: 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.

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

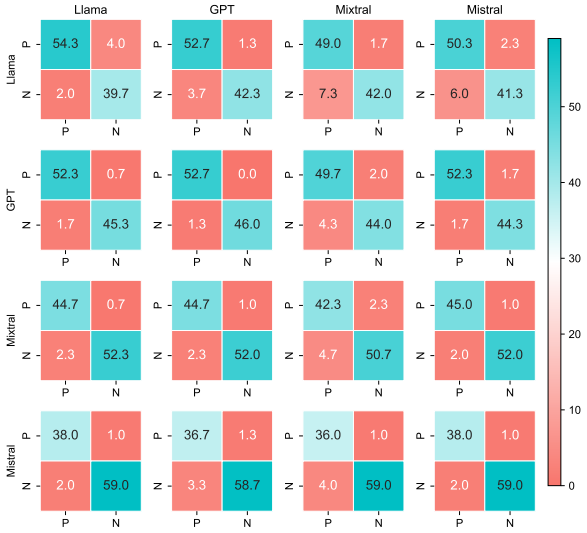

Figure 3: Heatmap illustrating the performance of LLM judges on HotpotQA. Each cell value represents percentages (%). Rows represent predicted outcomes (P: Positive, N: Negative), while columns represent actual outcomes. See Appendix C for full results.

all 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.

4.2 Analysis

In our experiments, candidate LLMs generated 4,800 outputs for the given tasks, with each evaluator producing 4,800 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 10 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 focus 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. Furthermore, LLM-as-judges excel at identifying essential information within responses. Answers that include correct and supplementary details, LLM judges focus on the key components necessary for evaluation and disregard non-critical elements to ensure reliable assessments.

LLMs are prone to hallucination in justification (Zhang et al., 2023), where they fabricate reasoning to support their evaluations, produce detailed but incorrect explanations, or reference 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 (Ye et al., 2024) 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.

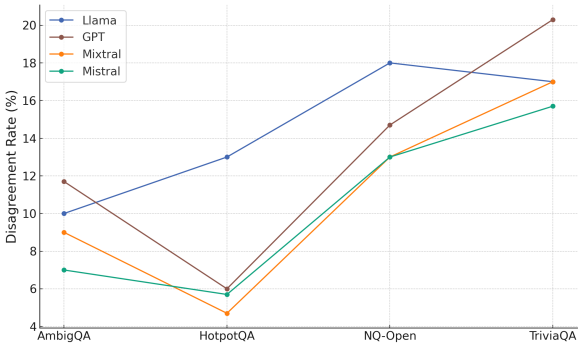

Figure 4: Disagreement rates between the primary judges (Llama-3.1 70B and Mistral 7B) across various candidate LLMs (Llama, GPT, Mixtral, and Mistral) and tasks.

In addition to the stated issues, we found a few temporal limitations in LLM-based evaluators. Although most of our datasets are older and the evaluator models are relatively up-to-date, we still encounter a small number of instances where references to recent events, new terminology, or shifting contexts are misunderstood. This temporal bias underscores the need for evaluation mechanisms that can adapt to or acknowledge evolving information landscapes, ensuring fair and context-sensitive assessments over time.

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 NQ-Open and TriviaQA, with disagreement rates reaching 18.0% and 20.3%, respectively. From the judges' explanations, we interpret that these elevated rates are likely due to the judges' focus on specific reference answers among many possible options and the free-form nature of responses.

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 and 7). Notably, in the TriviaQA task, Macro F1 scores advanced from 77.2% to 84.2%, and Cohen's Kappa increased from 0.547 to 0.684. These substantial improvements highlight the pivotal role of the arbitrator in ensuring reliable and

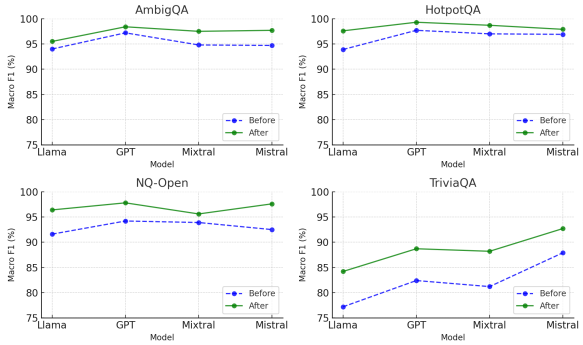

Figure 5: Comparison of Macro F1 scores before and after arbitration.

consistent evaluation outcomes, especially in complex and ambiguous tasks where primary judges are more likely to disagree. By leveraging GPT-3.5-turbo exclusively for contested cases, DAFE effectively maintains high evaluation standards and fosters better accuracy and fairness in the evaluation process (see Appendix C for more results).

5 Related work

Evaluation of natural language generation has traditionally relied on 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 (Wang et al., 2024a; Kamalloo et al., 2023). Other commonly used metrics including BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) primarily focus on n-gram overlap with human written reference texts. Despite their widespread use, these metrics have significant limitations in capturing semantic subtleties and contextual relevance (Zhang et al., 2020). To address the limitations of conventional metrics, various model-based methods such as BERTScore (Zhang et al., 2020) offer semantically informed evaluation. However, even BERTScore and similar embedding-based methods struggle to effectively evaluate open-ended generation (Zheng et al., 2024; Sun et al., 2022).

Recent advances in LLMs have unlocked new opportunities for automatic and context-aware evaluation (Li et al., 2024b; Chiang and Lee, 2023; Zheng et al., 2024). A key strength of LLM-based evaluators lies in their ability to operate in reference-free settings, where evaluation does not rely on 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 (Li et al., 2024a). 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?" (Verga et al., 2024; Chan et al., 2024). LLM-based evaluators are specifically effective for tasks like summarization, where subjective criteria are central to evaluation (Liu et al., 2023). However, 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 (Ye et al., 2024; Khan et al., 2024). In reference-free settings, the absence of a definitive ground truth increases the risk of bias in evaluations (Ye et al., 2024; Kim et al., 2024; Huang et al., 2024a). Common biases include positional bias, where LLMs may favor responses based on their order (Zheng et al., 2024; Khan et al., 2024), verbosity bias, which favors longer or more detailed responses (Huang et al., 2024b), and self-enhancement bias, where models may disproportionately prefer their own outputs (Zheng et al., 2024). 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.

Another limitation is the sensitivity of LLM judges to prompt design which can lead to different results as developing more robust prompts or standardizing prompt templates may help improve judges' performance. Additionally, we employed two primary judges — one small Mistral 7B and one large Llama 3.1-70 —but did not explore configurations involving two smaller models, with arbitration only invoked during disagreements. This could be an avenue for future work to reduce computational costs while maintaining evaluation reliability. Furthermore, we did not analyze the resource usage and cost-benefit trade-offs of our framework, which are important considerations for practical deployment. The high computational demand for running multiple LLMs may also limit the practicality of our method in resource-constrained settings (Badshah and Sajjad, 2024).

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.

References

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

Sher Badshah and Hassan Sajjad. 2024. Quantifying the capabilities of llms across scale and precision.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

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

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

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. 2021. Evaluating large language models trained on code.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Hui Huang, Yingqi Qu, Xingyuan Bu, Hongli Zhou, Jing Liu, Muyun Yang, Bing Xu, and Tiejun Zhao. 2024a. An empirical study of llm-as-a-judge for llm evaluation: Fine-tuned judge model is not a general substitute for gpt-4.

Hui Huang, Yingqi Qu, Hongli Zhou, Jing Liu, Muyun Yang, Bing Xu, and Tiejun Zhao. 2024b. On the limitations of fine-tuned judge models for llm evaluation.

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

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

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

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

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

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

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

Zachary Kenton, Noah Y. Siegel, János Kramár, Jonah Brown-Cohen, Samuel Albanie, Jannis Bulian, Rishabh Agarwal, David Lindner, Yunhao Tang, Noah D. Goodman, and Rohin Shah. 2024. On scalable oversight with weak llms judging strong llms.

Akbir Khan, John Hughes, Dan Valentine, Laura Ruis, Kshitij Sachan, Ansh Radhakrishnan, Edward Grefenstette, Samuel R. Bowman, Tim Rocktäschel, and Ethan Perez. 2024. Debating with more persuasive llms leads to more truthful answers.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button,

Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

A Free-form Question-Answering

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

- **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.
- **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.
- **AmbigQA**: Focuses on 14K ambiguous questions derived from NQ, requiring systems to identify multiple valid interpretations and generate disambiguated questions alongside corresponding answers.

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

B 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 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) to assess inter-rater reliability among human annotators. The results demonstrate exceptionally high reliability, with Fleiss' Kappa scores consistently above 0.93 and perfect agreement rates exceeding 96%. The highest agreement is observed in GPT-3.5 evaluations on NQ-Open (κ = 0.990, 99.3% perfect agreement) and Mixtral-8x7B on HotpotQA (κ = 0.996, 99.7% perfect agreement). Even for traditionally challenging tasks like TriviaQA, annotators maintain strong consensus with κ values between 0.935-0.975 and perfect agreement rates of 98.3-99.0%, indicating robust and reliable human evaluation across all experimental conditions.

LLMs	AmbigQA	HotpotQA	NQ-Open	TriviaQA
Llama	96.3%	98.0%	99.0%	99.0%
GPT	99.3%	98.7%	99.3%	99.0%
Mixtral	98.7%	99.7%	98.3%	98.3%
Mistral	98.3%	98.7%	98.3%	99.0%

Table 4: Human annotators percent agreement scores across candidate models and tasks.

C Additional results

This section provides further results and analysis of conventional metrics and LLM-based evaluators. Table 5 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.

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 on TriviaQA and NQ-Open, where improvements of up to 7.0% are observed. For tasks like AmbigQA and HotpotQA, where initial performance was already high, arbitration yields smaller but still notable gains. 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.

We observed substantial enhancements in Cohen's Kappa scores across several tasks. For instance, in the AmbigQA Cohen's Kappa increased from 0.881 to 0.911. Similarly, the NQ-Open Cohen's Kappa from 0.833 to 0.927. In the TriviaQA, the scores increased from 0.547 to 0.684. These improvements demonstrate that the arbitration mechanism effectively enhances the reliability and consis-

LLMs	Tasks	Evaluators							
		EM	BS	HM	Llama-3.1-70B	GPT-3.5	Mixtral-8x7B	Mistral-7B	DAFE
Llama-3.1-70B	AmbigQA	42.3	63.0	67.0	65.3	64.7	63.0	66.0	64.7
	HotpotQA	34.3	67.7	56.3	58.3	54.0	50.7	52.7	55.3
	NQ-Open	31.7	61.7	66.3	62.7	60.0	59.0	66.7	63.0
	TriviaQA	74.3	94.0	94.7	90.3	90.0	88.7	84.7	93.0
GPT-3.5	AmbigQA	49.7	78.0	71.7	70.0	68.0	65.7	71.0	71.0
	HotpotQA	33.7	80.0	54.0	53.0	52.7	51.7	54.0	53.3
	NQ-Open	36.3	74.0	65.3	62.7	59.0	59.0	67.0	63.3
	TriviaQA	74.3	95.3	93.0	89.3	90.7	89.7	86.3	92.7
Mixtral-8x7B	AmbigQA	37.7	70.3	61.7	57.3	62.0	59.3	61.7	60.7
	HotpotQA	25.0	69.7	47.0	45.3	45.7	44.7	46.0	45.7
	NQ-Open	23.7	63.7	56.7	52.7	47.7	52.3	59.7	52.3
	TriviaQA	64.7	91.3	90.7	86.3	89.7	86.0	85.3	90.7
Mistral-7B	AmbigQA	31.0	61.7	49.7	46.3	47.7	46.3	53.3	48.7
	HotpotQA	23.7	64.7	40.0	39.0	38.0	37.0	39.0	38.0
	NQ-Open	22.7	60.0	46.0	40.0	43.3	41.3	50.0	43.7
	TriviaQA	62.0	94.3	83.7	81.3	81.0	79.7	85.0	83.7

Table 5: Raw performance of candidate LLMs across free-form QA tasks evaluated through various methods. HM represents Human Majority and BS denotes BERTScore.

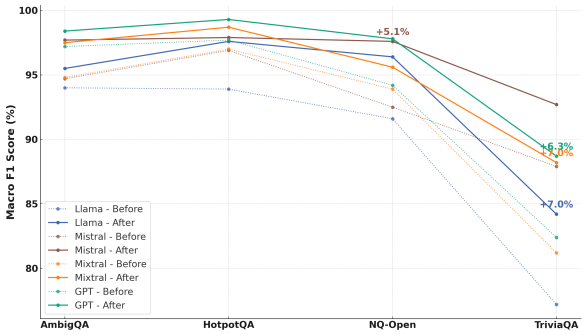

Figure 6: Impact of arbitration on disagreements between primary judges. Note that we used Llama-3.1-70B and Mistra 7B as primary judges. GPT-3.5-turbo is only utilized when disagreements are found. The models given in the figure are candidate LLMs which generate outputs for the given tasks and are then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

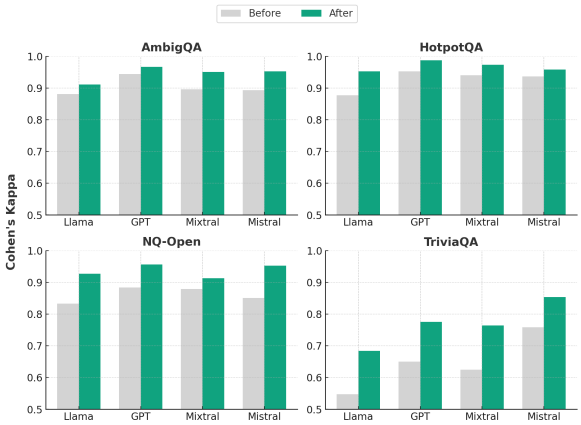

Figure 7: Comparison of Cohen's kappa scores before and after arbitration (GPT-3.5-turbo as arbitrator). The performance is illustrated across candidate LLMs and tasks.

tency of evaluations, particularly in complex and ambiguous tasks where primary judges are more likely to disagree.

C.2 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 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.

D LLM-as-a-judge in reference-free settings

We 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.

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 6 and 7 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.

E Prompting

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

E.1 Prompting Candidate LLMs

We prompted candidate LLMs (see Figure 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.

E.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.

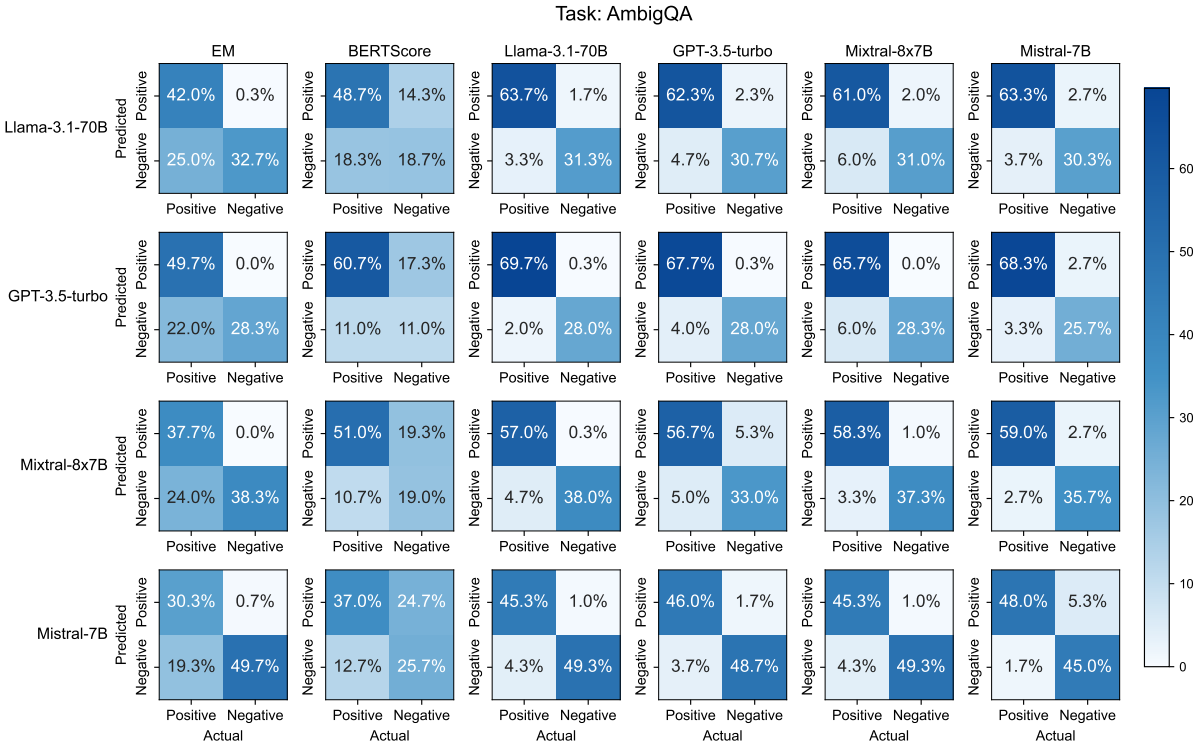

Figure 8: Confusion matrices comparing the performance of automatic metrics (EM, BERTScore) and individual LLM judges on AmbigQA.

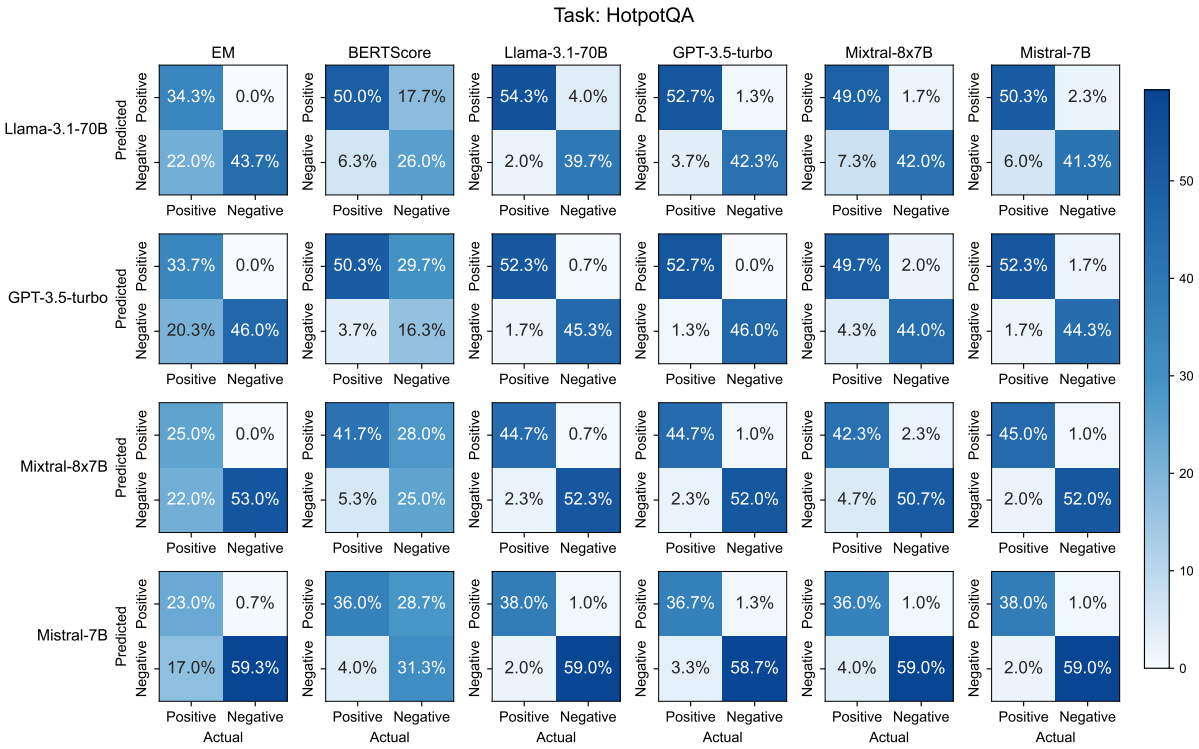

Figure 9: Confusion matrices comparing the performance of automatic metrics (EM, BERTScore) and individual LLM judges on HotpotQA.

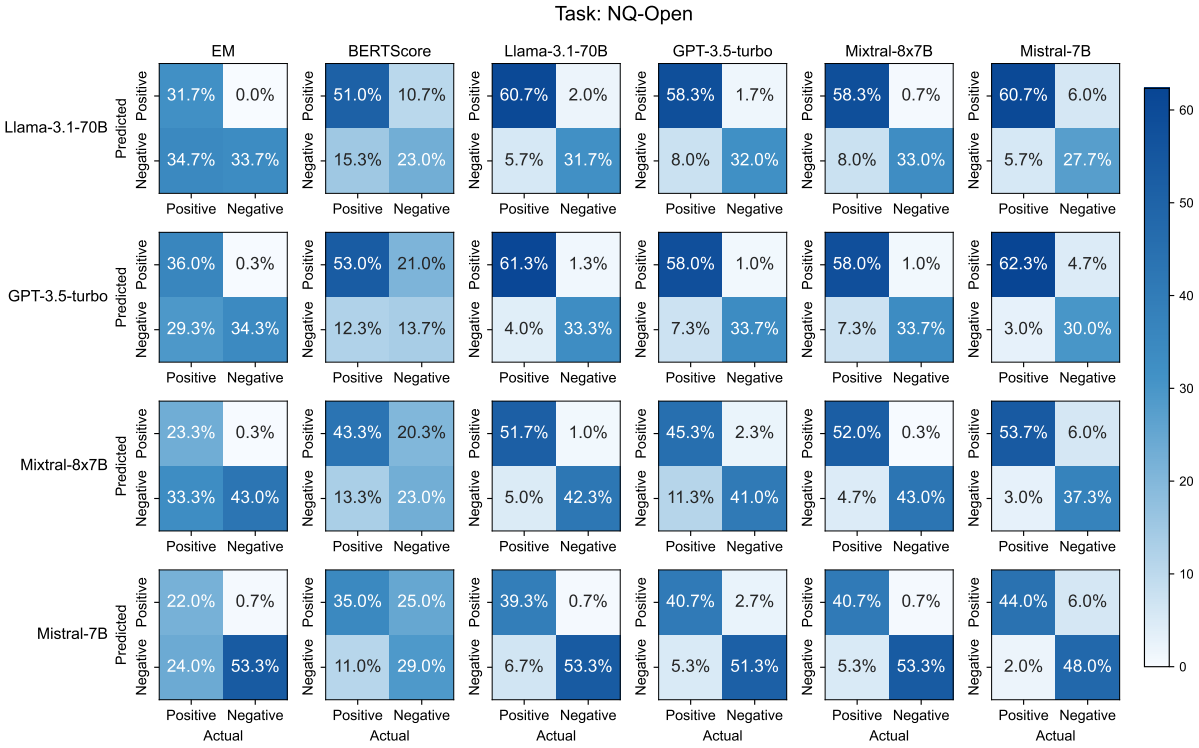

Figure 10: Confusion matrices comparing the performance of automatic metrics (EM, BERTScore) and individual LLM judges on NQ-Open.

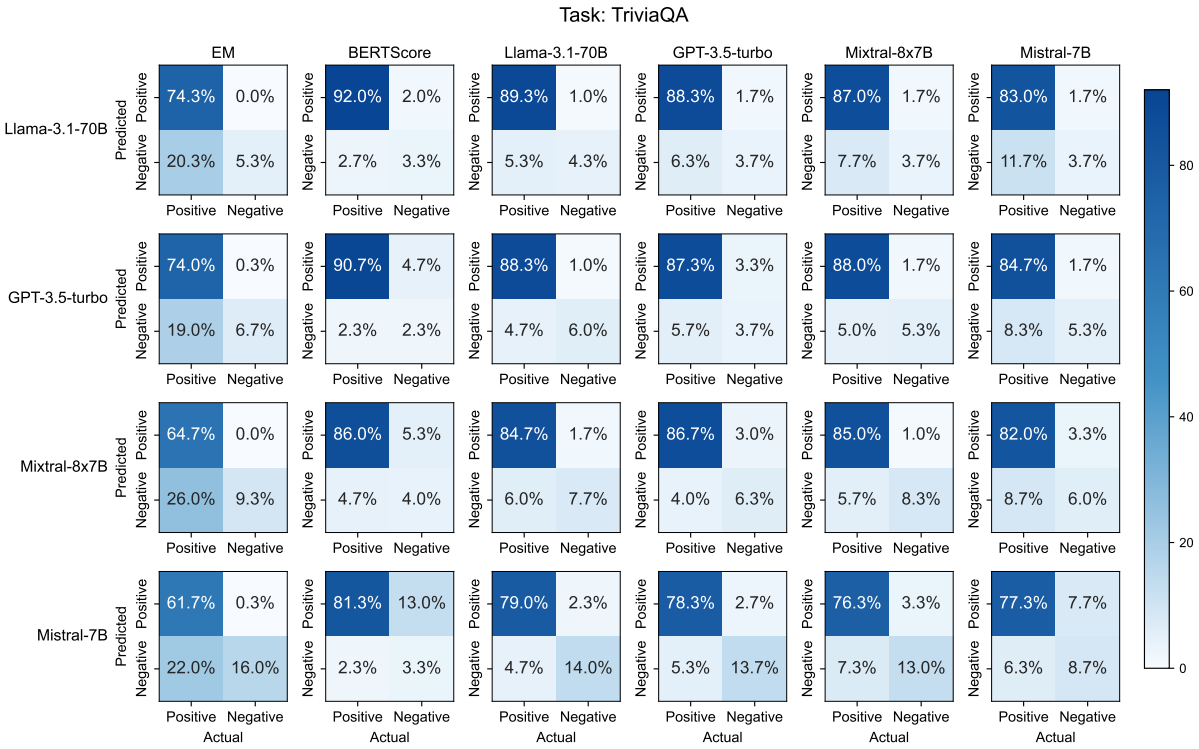

Figure 11: Confusion matrices comparing the performance of automatic metrics (EM, BERTScore) and individual LLM judges on TriviaQA.

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

Table 6: Overall performance of candidate LLMs across free-form QA tasks. Values [in blue] represent LLM-as-a-judge in the reference-free mood.

Candidate LLMs	Tasks	Evaluators					
		EM	BERTScore	Llama-3.1-70B	GPT-3.5-turbo	Mixtral-8x7B	Mistral-7B
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 7: Performance of various evaluators across candidate LLMs and tasks. Values [in blue] represent LLM-as-a-judge in the reference-free mode.

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.

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.