Does Context Matter? ContextualJudgeBench for Evaluating LLM-based Judges in Contextual Settings

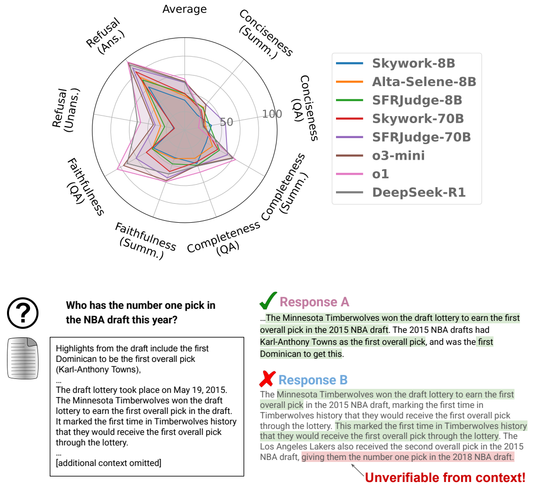
Anonymous ACL submission

Abstract

The large language model (LLM)-as-judge paradigm has been used to meet the demand for a cheap, reliable, and fast evaluation of model outputs during AI system development and post-deployment monitoring. While judge models—LLMs finetuned to specialize in assessing and critiquing model outputs—have been touted as general purpose evaluators, they are typically evaluated only on *non-contextual scenarios*, such as instruction following. The omission of contextual settings—those where external information is used as *context* to generate an output—is surprising given the increasing prevalence of retrieval-augmented generation (RAG) and summarization use cases. Contextual assessment is uniquely challenging, as evaluation often depends on practitioner priorities, leading to conditional evaluation criteria (e.g., comparing responses based on factuality and then considering completeness if they are equally factual). To address the gap, we propose ContextualJudgeBench, a judge benchmark with 2,000 challenging response pairs across eight splits inspired by real-world contextual evaluation scenarios. We build our benchmark with a multi-pronged data construction pipeline that leverages both existing human annotations and model-based perturbations. Our comprehensive study across 11 judge models and 7 general purpose models, reveals that the contextual information and assessment criteria present a significant challenge to even state-of-the-art models. For example, o1, the best-performing model, barely reaches 55% consistent accuracy.

1 Introduction

In the LLM era, timely, affordable, and accurate evaluation of model responses is essential for model development and monitoring. One automated evaluation solution available to practitioners is the *LLM-as-judge* approach, where relatively lightweight *judge models* are trained to evaluate and critique other model responses. Judge models are broadly touted as general-purpose evaluators (e.g., Vu et al. (2024); Alexandru et al. (2025)), capable of being deployed across domains and evaluation settings. However, judges are rarely evaluated on *contextual settings* (Wang et al., 2024c; Saha et al., 2025; Ye et al., 2024), where the evaluated responses are generated from an externally provided context rather than solely from the model's parametric knowledge, like in retrieval-augmented generation (RAG) or summarization.



Who has the number one pick in the NBA draft this year?

Highlights from the draft include the first Dominican to be the first overall pick (Karl-Anthony Towns),
...
The draft lottery took place on May 19, 2015. The Minnesota Timberwolves won the draft lottery to earn the first overall pick in the draft. It marked the first time in Timberwolves history that they would receive the first overall pick through the lottery.
...
[additional context omitted]

✓ **Response A**

...The Minnesota Timberwolves won the draft lottery to earn the first overall pick in the 2015 NBA draft. The 2015 NBA drafts had Karl-Anthony Towns as the first overall pick, and was the first Dominican to get this.

✗ **Response B**

The Minnesota Timberwolves won the draft lottery to earn the first overall pick in the 2015 NBA draft, marking the first time in Timberwolves history that they would receive the first overall pick through the lottery. This marked the first time in Timberwolves history that they would receive the first overall pick through the lottery. The Los Angeles Lakers also received the second overall pick in the 2015 NBA draft, giving them the number one pick in the 2018 NBA draft.

Unverifiable from context!

Figure 1: (Top) An overview of top-performing models on the eight splits of ContextualJudgeBench. (Bottom) A truncated sample from the faithfulness split, where Response A is preferred because all of its content is factually verifiable from the context.

As contextual generation systems gain prominence, specialized generators (Cohere Team, 2024; Contextual AI Team, 2024; Nguyen et al., 2024) have been developed to meet the stringent faithfulness demands of business applications and high-risk fields, like medicine (Xiong et al., 2024) and law (Wiratunga et al., 2024). Reliably evaluating such systems is increasingly important, but presents unique challenges. The presence of contextual information magnifies challenges that exist in non-contextual human evaluation (Liu et al., 2023): Since contextual generation requires responses to be *faithful* to the provided context, humans must first comprehend potentially long, domain-specific contexts before they can evaluate a response. This additional "hallucination detection" step adds another layer of complexity on top of evaluating the substantive quality of responses.

Taken together, contextual settings are the ideal candidate for automatic evaluation: LLMs have strong language understanding across specialized domains (Xie et al., 2023; Ke et al., 2025; Colombo et al., 2024) and have rapidly improving long-context comprehension abilities (Kamradt, 2023). Indeed, many recent benchmarks for contextual generation use prompted (Laban et al., 2024; Jacovi et al., 2025) or finetuned (Friel et al., 2024) LLMs to serve as evaluators due to longer, more complex model outputs. However, to our knowledge, no benchmarks exist to measure the quality of *contextual evaluators*. We bridge this gap by proposing ContextualJudgeBench, which consists of 2,000 challenging pairwise samples across 8 splits that measure different evaluation criteria and settings. Fig. 1 showcases our dataset splits and benchmarking results. Our work *complements* existing contextual generation benchmarks by offering a way to assess contextual evaluators.

The dominant criteria for responses in contextual evaluation center around *faithfulness* and *answer relevancy* (Es et al., 2023; Saad-Falcon et al., 2023; Jacovi et al., 2025; Laban et al., 2024). Such metrics are often assigned independently in a pointwise manner, i.e., a model assigns a faithfulness score and a relevance score to a single response, with each score assigned without considering the other. ContextualJudgeBench, in contrast, proposes a pairwise evaluation setup. This pairwise setup offers utility to practitioners (e.g., evaluation for A/B testing) while eliciting evaluations better aligned with humans judgment from automatic evaluators (Wang et al., 2023; Liu et al., 2024a). However, directly using pointwise scores to do pairwise comparisons can lead to ambiguity: If a response is more relevant but less faithful, is it better?

To remedy this, we propose a principled *conditional* evaluation hierarchy (Sec. 3) that prioritizes refusal accuracy and response faithfulness. First, we evaluate if judges can assess accurate or inaccurate refusals, where a response that refuses to answer due to a perceived lack of evidence is compared against a substantive response. Given two substantive responses, we next assess based on faithfulness: Which response contains more factually supported information? If two responses are equally faithful, then they are evaluated on completeness, with more thorough responses being preferred. Finally, for two equally complete responses, they are evaluated based on conciseness, as responses should not contain extraneous information, even if factual. The splits in ContextualJudgeBench are carefully designed to test judges in each setting that arises in this hierarchy. Concretely, our contributions are:

- With an emphasis on refusals and faithfulness, we propose a hierarchy that provides an "order of operations" for pairwise contextual evaluation.
- We present ContextualJudgeBench, a benchmark for evaluating judge models consisting of 2,000 response pairs across eight splits derived from real-world contextual outcomes.
- We evaluate 11 judge models, ranging in size from 3.8B to 70B parameters on ContextualJudgeBench along with 7 general purpose/reasoning models.

Our findings reveal that contextual assessment remains an open challenge, with o1 and SFRJudge-70B only achieving 55.3 and 51.4 accuracy. Despite the reasoning intense nature of contextual evaluation, our analysis shows that inference-time scaling for judges may lead to performance *degradations*. We will make our code and benchmark available for future research in the evaluation and development of contextual judges.

2 Related work

Our work, rather than evaluating contextual systems, evaluates judge models as contextual *evaluators*. Here, we review current judge benchmarks and contextual evaluation setups.

Evaluation for LLM-as-judges. LLM-as-judge is a generative evaluator paradigm where LLMs are trained to produce an evaluation (natural language explanation and judgment) given the original user input, evaluation protocol (rules and criteria for evaluation), and model responses as input. As the popularity of LLM-as-judges grows, numerous benchmarks have been proposed to evaluate these evaluators. These benchmarks are typically for specific domains, like instruction following (Zeng et al., 2023), fine-grained evaluation (Kim et al., 2023, 2024), bias (Park et al., 2024), reward modeling (Lambert et al., 2024; Frick et al., 2024; Gureja et al., 2024), or reasoning (Tan et al., 2024). While new judge benchmarks are challenging, none focus on contextual evaluation. Of judge benchmarks, a subset of Eval-P (Li et al., 2023a) contains summarization pairs with the winner chosen by aggregating various criteria into an overall score. InstruSum (Liu et al., 2023) has also been used for judge evaluation (Wang et al., 2024b; Alexandru et al., 2025; Liu et al., 2024c). ContextualJudgeBench, in contrast, is dedicated entirely to contextual evaluation, requiring evaluation to be done in under-explored settings like RAG-QA along previously untested criteria such as refusal.

Evaluation for contextual responses. RAG generators have been typically evaluated with standard knowledge-based QA tasks, e.g., ContextualBench (Nguyen et al., 2024), or with newer benchmarks that cover scenarios such as faithfulness (Ming et al., 2024; Niu et al., 2024; Tang et al., 2024; Li et al., 2023b; Sadat et al., 2023), diverse domains (Friel et al., 2024), refusals (Peng et al., 2024), and reasoning (Wei et al., 2024; Krishna et al., 2024). Because RAG settings have progressed beyond simple factoid answers, recent benchmarks have deployed carefully prompted frontier LLMs (e.g., Jacovi et al. (2025)) to perform assessment in a pointwise manner, rather than using exact string matching (Nguyen et al., 2024).

Initial evaluation efforts for RAG settings focused on faithfulness, training hallucination detectors (Tang et al., 2024) as both sequence classifiers and generative models (Wang et al., 2024a; Ravi et al., 2024; Ramamurthy et al., 2024). More holistic evaluation systems with multiple metrics have recently been proposed, such as Es et al. (2023); Saad-Falcon et al. (2023). For the most part, these approaches involve specialized prompting (Es et al., 2023), using synthetic data generation to train specialized evaluators (Saad-Falcon et al., 2023).

Summarization evaluation has evolved from n-gram metrics like ROUGE (Lin, 2004) and METEOR (Banerjee and Lavie, 2005) to contextual embedding model scorers (Zhang et al., 2020; Zhao et al., 2019; Yuan et al., 2021). However, these evaluators cannot assess based on multiple criteria and tend to correlate poorly with humans. To evaluate quality, the primary focus has been on model-based factual verification (Laban et al., 2022; Cao and Wang, 2021; Goyal and Durrett, 2021; Kryscinski et al., 2020; Laban et al., 2023). Recent studies have shifted toward human annotations for finer-grained assessment (Song et al., 2024; Lee et al., 2024; Oh et al., 2025), focusing on metrics such as faithfulness and conciseness.

Our proposed work complements these existing benchmarks in summarization and RAG by evaluating contextual *judges*, rather than the generators.

3 ContextualJudgeBench

> *"63 percent of respondents...said that output inaccuracy was the greatest risk they saw in their organizations' use of gen AI."* –McKinsey, 2024 AI Survey

Inaccuracy is the largest reported risk for practitioners using AI systems. 30% of respondents in a Deloitte survey specifically cite trust loss due to hallucinations as a top concern. Hallucinations are especially unacceptable in contextual settings, as the model is expected to generate responses strictly based on the provided context. This grounding context is typically considered a gold-standard source of knowledge. If the relevant information is absent, the model should refrain from responding rather than generate unsupported content. Motivated by real-world concerns, we propose a conditional evaluation workflow (Fig. 2) that prioritizes *answerability* and *faithfulness* before assessing other criteria. Each evaluation step in our workflow requires creating new splits for ContextualJudgeBench.

In developing contextual systems, practitioners often conduct A/B testing between systems with different generator, retriever, pre-processing configurations (Saad-Falcon et al., 2023). ContextualJudgeBench is designed to reflect this pairwise A/B testing setup, containing 2,000 test samples. Each sample includes a user input, a context, and two responses, from which a judge selects the "better" response based on our workflow. The pairwise setting is well-suited for judge-based evaluation as it aligns closely with human preferences (Wang et al., 2023; Liu et al., 2024a). We first describe two methods we use to create the pairwise samples. Then, we present ContextualJudgeBench in four stages (Sec. 3.2 – 3.5), each corresponding to a step in the evaluation workflow (Fig. 2).

3.1 Dataset creation approach

We employ two primary approaches to create ContextualJudgeBench: utilizing existing human annotations and leveraging frontier models for criteria-based response perturbation.

- **Human annotations [H]**: We use existing human annotations (Lee et al., 2024; Wan et al.,

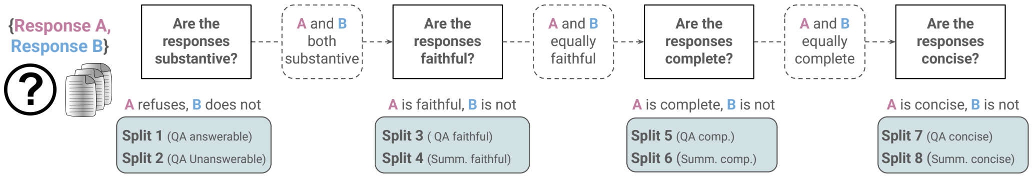


Figure 2: A refusal and faithfulness-first contextual evaluation hierarchy, as assessed by ContextualJudgeBench.

2024; Wu et al., 2023; Liu et al., 2024b) that evaluate multiple model responses for the same context. These assessments include criteria-specific scores or errors, either holistically or sentence-level. We select responses with significant differences based on specific criterion to form pairs, enabling comparative assessments.

- **Model-based perturbations**: In the absence of human labels, we form pairs through criteria-based response perturbation. Specifically, we use frontier LLMs to modify accurate responses based on the context to produce responses that do not align with the intended criteria. We apply this approach in two distinct ways:
  Desired output prompting [M1]: We ask an LLM to directly generate a response based on the context that fits certain output criteria. This includes generating context-based refusals or deliberately unfaithful responses.
  Existing output modification [M2]: We use an LLM to modify an existing response, introducing deviations based on predefined criteria. This can include making the response more verbose or altering its content in specific ways.

See App. A for details on the datasets, prompts, and specific approaches used for each split.

3.2 Step 1: QA refusal validity [splits 1 & 2]

Knowing when to refuse to answer due to lack of information is a critical first step specific to RAG settings[1]. Refusals can be viewed as a form of faithfulness: To remain faithful to the context, the model should refuse to hallucinate an answer if no relevant information is present. Conversely, the model should not refuse if the context is sufficient.

Splits 1 and 2 of ContextualJudgeBench assess if judges can identify appropriate refusals. Each sample consists of a refusal (e.g., "The answer cannot be answered based on the context") and a substantive response. Split 1 contains answerable questions from LFRQA (Han et al., 2024), where the judge should pick the substantive response, whereas split 2 contains unanswerable questions from FaithEval (Ming et al., 2024), making refusal the correct choice. To construct split 1, we use approach **M1** from Sec. 3.1, using an LLM to generate context-based refusals as negative responses to pair up with the provided positive responses. In split 2, we again employ approach **M1** to generate context-based refusal responses to correctly decline the question as positive responses and generate hallucinated (incorrect) responses as negative ones. See App. A for generation prompt.

3.3 Step 2: Faithfulness [splits 3 & 4]

When evaluating two substantive responses, the first criterion is *faithfulness*, as a response cannot be considered accurate if it contains hallucinated content. Faithfulness measures the consistency of the response with the context: all factual statements in a faithful response must be attributable to the context, ensuring there are no hallucinations. Splits 3 and 4 evaluate the judge's ability to select the more faithful response for QA and summarization, respectively. Each pairwise sample is designed to include one substantively more faithful response, allowing the judge to choose the better response based solely on faithfulness.

We construct split 3 by combining multiple QA datasets. For QA-Feedback (Wu et al., 2023) and RAGTruth (Niu et al., 2024), we use the approach **H** to form pairs between RAG responses, annotated with either faithfulness scores or factuality errors. For LFRQA (Han et al., 2024), LFQA (Xu et al., 2023), and short queries from MRQA (Fisch et al., 2019), we treat the provided responses as factually correct (positive) and apply approach **M1** to generate factually inconsistent negative responses based on the context. See App. A for prompt template. We manually reviewed the formed pairs to ensure their reliability. For Split 4, we use approach **H** to create summarization response pairs of different factuality levels. To ensure diversity, we sample contexts from Wan et al. (2024); Lee et al. (2024), which cover both topic-specific and general summarization instructions across diverse domains.

[1] Refusals are uncommon in summarization, as instructions and context are both user provided; In RAG settings, the user has no control over the retrieved context.

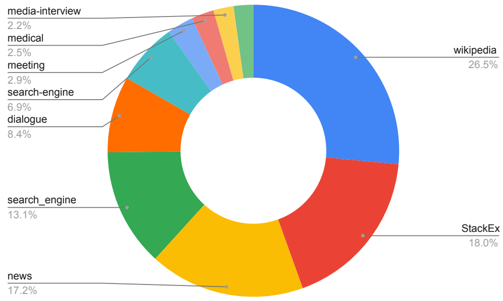


Figure 3: Distributions of context domain as a percent of the total set of preference pairs in the benchmark.

3.4 Step 3: Completeness [splits 5 & 6]

Beyond faithfulness, contextual evaluation must also assess response quality. When comparing two faithful responses, the better one should cover all essential information needed for a thorough and useful answer. As such, we consider *completeness*, i.e., how comprehensive the response is, as the next criteria. Splits 5 and 6 assess the judge's ability to select more complete response when both options are faithful, for QA and summarization tasks, respectively. Each pairwise sample is designed such that one response is more complete than the other while both the responses are faithful.

Judges should first confirm that both responses are faithful and then determine which one is more complete. We construct Split 5 using the LFRQA(Han et al., 2024) and QA-Feedback(Wu et al., 2023) datasets. For LFRQA, we use approach **M2** from Sec. 3.1 to modify a faithful response by omitting lines associated with certain citations while expanding on other citations. This yields a less complete negative response that is still faithful and similar in length to the original (positive) response. See App. A for generation prompt. For QA-Feedback, we use approach **H** to create preference pairs from RAG responses annotated for completeness scores or missing information errors. Similarly, split 6 is created using approach **H** with existing human annotations that assess faithfulness and completeness in summarization responses. To form preference pairs, we first filter unfaithful responses . Then, we form pairs based on completeness, ensuring that one response is significantly more complete (positive) than the other (negative).

3.5 Step 4: Conciseness [splits 7 & 8]

Our final criterion is *conciseness*: does the response avoid including more than what was asked? Our hierarchy intentionally places conciseness after completeness, as an answer should not sacrifice relevant content for the sake of brevity. However, complete responses may not be *minimally* complete: They may contain faithful yet extraneous information, repeated content, or unnecessary stylistic details. In splits 7 and 8, each pairwise sample has one response that is more concise while maintaining the same faithfulness and completeness. Judges should first verify both responses are faithful and complete, then choose the more concise one.

For Split 7, we use LFRQA (Han et al., 2024) and QA-Feedback (Wu et al., 2023). For LFRQA, we apply approach **M2**, tasking the model to insert direct quotations from the context without modifying the substance of provided responses. See App. A for generation prompt. For QA-Feedback, we use approach **H** to create pairs from responses annotated along conciseness, redundancy, and irrelevance. Preference pairs are formed by pairing faithful and complete responses by conciseness. For split 8, we again use approach **H**, using human annotations (Lee et al., 2024; Liu et al., 2024b) that assess summarization faithfulness, completeness, and conciseness.

3.6 Overall dataset statistics

ContextualJudgeBench is constructed based on our evaluation workflow (Fig. 2), resulting in 8 splits across 4 evaluation criteria, covering two common use cases of contextual generation: RAG-QA (5 splits) and Summarization (3 splits). We present the domain distribution in Fig. 3 and dataset statistics in Tab. 1. Overall, ContextualJudgeBench consists of 2,000 preference pairs, balanced across all splits, with over 1,500 unique contexts to minimize duplication. We include a wide range of context lengths, from a few tokens to nearly 10K tokens, with summarization contexts typically longer than QA ones. Response lengths range from brief answers to summaries over 1,000 tokens. To account for length bias in judges (Zeng et al., 2023; Park et al., 2024), we ensure minimal length differences between positive and negative responses across all splits; however, conciseness correlates with response length, resulting in longer positive responses.

4 Evaluation and analysis

4.1 Evaluation setup and baselines

Because the order of responses influences judge decisions (Wang et al., 2023), we adopt a consistency evaluation setup, like Tan et al. (2024); Li et al. (2023a). We run evaluation for each test sample

Split	# Pairs	# Context	L_c	L_r	L_{pos}	L_{neg}
Refusal (Ans.)	250	250	1,444	102	108	95
Refusal (Unans.)	250	250	418	64	64	63
Faithfulness (QA)	250	213	414	100	99	101
Faithfulness (Summ.)	250	192	1,754	94	97	91
Completeness (QA)	250	250	658	106	98	113
Completeness (Summ.)	251	171	1,066	91	93	89
Conciseness (QA)	255	254	1,086	199	116	281
Conciseness (Summ.)	244	117	1,557	98	77	118
Total	2,000	1,537	1,048	94	119	107

Table 1: ContextualJudgeBench statistics. # Context denotes unique contexts across all pairsL_c and L_r represent the mean context and response lengths, while L_{pos} and L_{neg} denote the mean positive and negative response lengths per split.

Model		# Params	Expl.	Context len.
GLIDER	(Deshpande et al., 2024)	3.8B	✓	128K
Prometheus-2	(Kim et al., 2024)	7,8x7B	✓	16K
OffsetBias	(Park et al., 2024)	8B	✗	8K
Atla-Selene	(Alexandru et al., 2025)	8B	✓	128K
Skywork-Critic	(Shiwen et al., 2024)	8,70B	✗	128K
SFRJudge	(Wang et al., 2024b)	8,12,70B	✓	128K
STEval.	(Wang et al., 2024c)	70B	✓	128K
Llama-3.1	(Dubey et al., 2024)	8,70B	✓	128K
GPT-4o,4o-mini	(Hurst et al., 2024)	?	✓	128K
GPT-o1,o3-mini	(Jaech et al., 2024)	?	✓	128K
DeepSeek-R1	(Guo et al., 2025)	671B	✓	128K

Table 2: Judge (top) and general (bottom) models evaluated. Expl. denotes if model outputs explanations.

twice, swapping the order of responses for the second run, and measure *consistent accuracy*: A judge output is considered correct if the judge selects the correct response for both runs. Under this setup, randomly choosing responses achieves a consistent accuracy of 25%. We also measure *consistency* , the fraction of times a judge selects the same response in both runs, regardless of correctness.

We evaluate 11 competitive LLM-as-judge models, ranging in size from 3.8B to 70B parameters: Prometheus (Kim et al., 2024), OffsetBias (Park et al., 2024), SFRJudge (Wang et al., 2024b), Skywork-Critic (Shiwen et al., 2024), Self-taugh-evaluator (Wang et al., 2024c), GLIDER (Deshpande et al., 2024), and Atla-Selene (Alexandru et al., 2025). See Tab. 2 for an overview of judges and App. B.1 for a more detailed description of each evaluated judge. For each judge, we retain the original prompt template while modifying evaluation instructions to align with our proposed workflow. Please see App. B.2 for prompt samples. In addition to specialized judges, we use Llama-3.1-8B & 70B (instruct versions) and GPT-4o, GPT-4o-mini, o3-mini, o1, and Deepseek-R1 as prompted judge model baselines. For all non-reasoning model-based judges, we generate with greedy sampling.

As a reference point, we also run RAGAS (Es et al., 2023), a pointwise RAG evaluator that leverages both prompted frontier models and embedding models, as well as MiniCheck (Tang et al., 2024), a hallucination detector. We apply these two methods to benchmark splits covered by their respective metrics: refusal and faithfulness for both, and completeness for RAGAS. For RAGAS, we score each response pointwise and derive corresponding pairwise outcomes in line with our hierarchy (e.g., for the completeness split, two responses must be considered equally faithful). For MiniCheck, we directly compare the classifier probabilities of each response to determine the pairwise winner.

4.2 Judge model evaluation

The results presented in Tab. 3 highlight the challenges of contextual evaluation. Overall, the best models on ContextualJudgeBench are o1 (55.3), o3-mini (52.6) and DeepSeek-R1 (51.9), two large-scale *reasoning* models. The best-performing judge, SFRJudge-70B (51.4), nearly matches DeepSeek-R1. Judge model performance generally increases with model size, with the best-performing judges exceeding their similarly-sized API counterparts (e.g., SFRJudge-8B at 39.3 and GPT-4o-mini at 38.8). The scaling trend, along with the strong performance of reasoning models, suggests that contextual evaluation is a reasoning-intensive task. However, we show that two inference-time reasoning techniques, self-consistency and better chain-of-thought prompting, do not boost judge performance in Sec. 4.5 and App. C.1.

Generative judge models tend to lag specialized evaluators. MiniCheck naturally excels for faithfulness, while RAGAS offers more balanced, yet still competitive performance across refusal and faithfulness splits. However, most judges outperform the embedding-based RAGAS completeness score, showing an advantage of generative evaluation.

Models tend to struggle with conciseness and unanswerable refusals. The difficulty with conciseness may be exacerbated by length bias (Zeng et al., 2023), as selecting shorter concise responses conflicts with the tendency of judge models to prefer longer ones. Likewise, struggling to select accurate refusals may be a special case of concreteness bias (Park et al., 2024), as judges are biased towards substantive responses. Further analysis in App. C.2 reveal that poor accurate refusal performance may be an unintended result of judge finetuning.

	Model	Refusal (Ans.)	Refusal (Unans.)	Faithfulness (QA)	Faithfulness (Summ.)	Completeness (QA)	Completeness (Summ.)	Conciseness (QA)	Conciseness (Summ.)	Average
Small Judge	Glider-3.8B	12.0	8.8	45.6	9.2	20.8	28.7	5.1	4.1	16.8
	Promtheus-2-7b	12.4	44.0	27.2	32.0	24.0	42.6	6.7	29.5	27.3
	Llama-3-OffsetBias-8B	64.8	11.2	34.0	26.4	33.2	21.1	46.3	23.0	32.6
	Skywork-8B	60.8	12.0	38.8	31.6	38.4	26.7	29.4	21.3	32.4
	Alta-Selene-8B	74.4	26.4	40.8	32.8	32.4	34.7	23.1	32.0	37.1
	SFRJudge-8B	70.8	22.0	40.4	38.8	40.4	43.4	27.5	31.1	**39.3**
	SFRJudge-12B	68.4	28.4	45.2	43.6	28.0	51.0	16.1	29.5	38.8
Large Judge	Promtheus-2-8x7b	22.0	29.6	22.4	29.6	20.4	39.8	10.2	18.4	24.1
	Skywork-70B	82.4	11.2	48.0	47.6	36.8	41.4	21.6	27.9	39.6
	ST-Eval-70B	50.0	42.0	51.2	45.6	40.8	39.4	36.1	29.9	41.9
	SFRJudge-70B	87.6	32.4	60.8	54.8	40.8	53.4	44.7	36.1	**51.4**
Instruct + Reasoning	Llama-3.1-8B	28.0	43.2	34.8	34.8	23.2	41.0	11.4	21.3	29.7
	Llama-3.1-70B	59.6	48.0	58.0	48.4	38.0	51.8	15.7	27.5	43.4
	GPT-4o-mini	71.2	22.8	45.6	42.4	33.2	54.2	11.8	29.5	38.8
	GPT-4o	64.0	52.0	68.0	50.8	39.6	56.2	12.9	22.5	45.8
	o3-mini	95.2	34.4	76.4	58.0	40.4	59.8	20.8	35.7	52.6
	o1	96.0	48.4	84.4	59.2	48.4	63.7	15.3	27.0	55.3
	DeepSeek-R1	92.0	52.0	72.0	50.4	41.2	60.6	20.4	26.2	51.9
Other	RAGAS	62.4	60.0	78.8	54.4	22.4	23.1	–	–	–
	Minicheck-7B	93.6	20.4	83.2	70.4	–	–	–	–	–

Table 3: Consistent accuracy for judge models, open-source instruct models, and API models on ContextualJudgeBench.

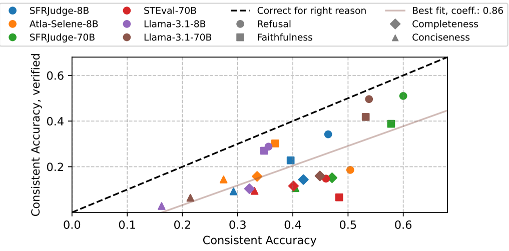


Figure 4: Accuracy vs. verified accuracy for six models, aggregated by criteria. The larger the drop from the dashed black line, the larger fraction of correct outcomes used incorrect criteria, as assessed by GPT-4o.

4.3 How does context impact positional bias?

Past studies have noted that judges are not robust to the order of the response pairs (Wang et al., 2023; Li et al., 2023a). This *positional bias* may be further exacerbated by the inclusion of context. As we show in Fig. 5 (top), consistency decreases as the context and response lengths increase, with evaluations of very long (>11K) inputs 33% less consistent than very short (<1K) input tokens. Inconsistency leads to performance variations, as shown in Fig. 5 (bottom). Here, we visualize the accuracy of each individual consistency run (Run 1 and Run 2 accuracy) and *optimistic accuracy*, where the judgment is considered correct if the judge identifies the better response in either of the two consistency runs, irrespective of the consistency. The inter-run performance gap tends to be small for stronger models, reflecting more consistent judgments. Weaker models exhibit higher levels of positional bias, but the favored position is model-dependent. For example, Prometheus-7B prefers the first response while OffsetBias prefers the second. The optimistic accuracy shows that judges are not wrong in a consistent manner, but often flip-flop based on position. Notably, optimistic accuracy of finetuned judge models is generally *higher* than that of prompted judges (e.g., 73.1 for OffsetBias vs. 68.3 for o1), revealing that judge finetuning may raise the upper bound of evaluation. Additional results can be found in App. C.3.

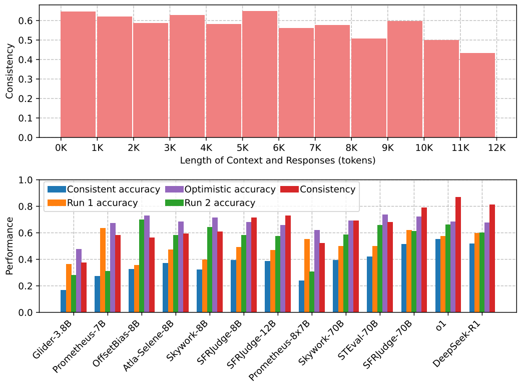


Figure 5: (Top) Consistency computed across all models and splits is inversely correlated with input length. . (Bottom) Four accuracy measures showing performance variations due to inconsistency, averaged across all splits for each model.

4.4 How do judges handle criteria?

Our analysis thus far has been outcome driven: We have not verified that judges make correct judgments based on the specified criteria. Here, we conduct model-assisted verification on a subset of judge models that generate explanations: SFRJudge-8B,70B, Atla-Selene-8B, Self-taught-evaluator, and the two Llama models. For all judgments with the correct outcome, we prompt GPT-4o to determine from the judge explanation if the judgment was decided by the correct criteria (Full prompt in App. B.3). From this, we compute a *verified consistent accuracy*. In Fig. 4, we plot the verified accuracies of each judge against its original accuracies, with the black dashed-line indi-

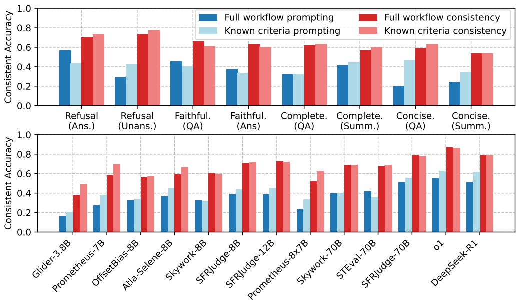


Figure 6: Judge performance changes are minor when given exact criteria vs. full workflow, indicating challenges in contextual evaluation beyond criteria. Per-split metrics averaged across all models, per-judge metrics averaged across all splits.

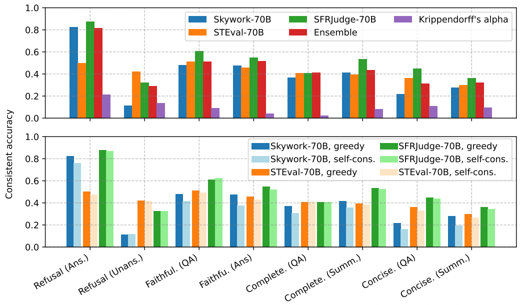


Figure 7: (Top) Ensembling judges into juries rarely outperforms the strongest judge in the jury due to weak judge agreement. (Bottom) Self-consistency rarely improves judge performance.

cating the upper bound, where all correct responses use the right criteria. On average, verified accuracies tend to be 20 absolute percent lower than outcome-based accuracy, revealing that judges are using incorrect reasoning to reach correct outcomes. Refusals and faithfulness are generally determined for the correct reasons, whereas completeness and conciseness are not, further highlighting the challenges of evaluation in the contextual settings.

While judges struggle to use the correct criteria when evaluating based on the contextual hierarchy, they are slightly more capable when given the correct criteria to use, as shown in Fig. 6. For each split, we prompt the judge with only the split criteria, omitting any mention of the contextual hierarchy. We compare judge performance against prompting with the full hierarchy. Conciseness and unanswerable refusals receive the greatest benefit, showing that length bias and concreteness bias can be mitigated to a degree with specific prompting. However, performance gains are relatively muted across judges due to little change in judge consistency between the two settings. Judge inconsistency, even after abstracting away the hierarchical structure, suggests that contextual evaluation poses challenges beyond applying the correct criteria.

4.5 Can scaling inference-time compute help?

Inspired by recent efforts in inference-time scaling (Jaech et al., 2024; Snell et al., 2024), we investigate the impact of two test-time scaling techniques: LLM-as-jury (Verga et al., 2024) and self-consistency (Wang et al., 2022). We experiment with three 70B judges, and for both settings, aggregate judgments via majority vote[2]. In Fig. 7, we present our results for LLM-as-jury (top) using responses from different three judges and self-consistency (bottom) using 10 responses per judge (using a temperature of 0.7) . LLM-as-jury rarely outperforms the strongest judge in the jury, while using self-consistency similarly has little impact on judge performance. Similar trends hold for smaller judge models, as shown in App. C.5.

These trends may be surprising given the strong performance of reasoning models like o1 and DeepSeek-R1. Lack of improvement from self-consistency likely results from the fact that contextual assessment is largely unseen in judge training. As a result, better judgments cannot be extracted via random sampling. The lack of jury success stems from the fact that judges do not exhibit *structured* agreement. We use all judge outputs to compute Krippendorff's alpha coefficient (Krippendorff, 2011), which measures inter-annotator agreement on a range from -1 to 1, with 0 indicating random chance. As shown in Fig. 7, judge agreement is extremely random: Even on the best-performing split, the alpha coefficient barely exceeds 0.2.

5 Conclusion

We introduce ContextualJudgeBench, a benchmark designed to evaluate LLM-judges in contextual settings. Building on a principled contextual evaluation hierarchy, we construct eight benchmark splits that assess refusals, faithfulness, completeness, and conciseness. This benchmark presents a significant challenge for state-of-the-art judge and reasoning models, with SFRJudge-70B and o1 achieving consistent accuracies of 51.4% and 55.3%, respectively. Additionally, we conduct a thorough analysis of reasoning correctness and examine the impact of common methods for scaling test-time compute, result of which further validate the unique challenges of contextual evaluation.

[2] We treat inconsistent judgments as ties. For a sample, if the aggregated judgments do not have a clear winner, e.g., (A, Tie, B) or (Tie, Tie, Tie), then we consider it incorrect.

Limitations

Our evaluations center around generative evaluators, as they are the most flexible in terms of incorporating context and indicating different evaluation criteria. However, reward models (RMs) are a common class of evaluators that may be applicable to this setting. However, to our knowledge, no contextual reward models exist. While in practice, one can embed the context in the input, it is unclear how to derive criteria-specific rewards from current models. A fruitful direction of future work is developing and benchmarking classifier based RMs for contextual settings.

As we repurposed existing annotated datasets – particularly for faithfulness and completeness – we are constrained by their coverage. This limitation may prevent us from making observations that generalize beyond their original distribution. Furthermore, ContextualJudgeBench is constructed primarily from English sources, a language abundant with context, model responses, and corresponding annotations. Further research should aim to rigorously assess contextual assessment in low-resource languages, where contextual content and corresponding annotations may be more scarce.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Andrei Alexandru, Antonia Calvi, Henry Broomfield, Jackson Golden, Kyle Dai, Mathias Leys, Maurice Burger, Max Bartolo, Roman Engeler, Sashank Pisupati, et al. 2025. Atla selene mini: A general purpose evaluation model. *arXiv preprint arXiv:2501.17195*.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Shuyang Cao and Lu Wang. 2021. CLIFF: Contrastive learning for improving faithfulness and factuality in abstractive summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6633–6649, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Cohere Team. 2024. Command r: Retrieval-augmented generation at production scale.

Pierre Colombo, Telmo Pires, Malik Boudiaf, Rui Filipe Coimbra Pereira de Melo, Gabriel Hautreux, Etienne Malaboeuf, Johanne Charpentier, Dominic Culver, and Michael Desa. 2024. Saullm-54b & saullm-141b: Scaling up domain adaptation for the legal domain. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Contextual AI Team. 2024. Introducing rag 2.0.

Darshan Deshpande, Selvan Sunitha Ravi, Sky CH-Wang, Bartosz Mielczarek, Anand Kannappan, and Rebecca Qian. 2024. Glider: Grading llm interactions and decisions using explainable ranking. *arXiv preprint arXiv:2412.14140*.

Karel D'Oosterlinck, Winnie Xu, Chris Develder, Thomas Demeester, Amanpreet Singh, Christopher Potts, Douwe Kiela, and Shikib Mehri. 2024. Anchored preference optimization and contrastive revisions: Addressing underspecification in alignment. *arXiv preprint arXiv:2408.06266*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. 2023. Ragas: Automated evaluation of retrieval augmented generation. *arXiv preprint arXiv:2309.15217*.

Adam Fisch, Alon Talmor, Robin Jia, Minjoon Seo, Eunsol Choi, and Danqi Chen. 2019. MRQA 2019 shared task: Evaluating generalization in reading comprehension. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 1–13, Hong Kong, China. Association for Computational Linguistics.

Evan Frick, Tianle Li, Connor Chen, Wei-Lin Chiang, Anastasios N Angelopoulos, Jiantao Jiao, Banghua Zhu, Joseph E Gonzalez, and Ion Stoica. 2024. How to evaluate reward models for rlhf. *arXiv preprint arXiv:2410.14872*.

Robert Friel, Masha Belyi, and Atindriyo Sanyal. 2024. Ragbench: Explainable benchmark for retrieval-augmented generation systems. *arXiv preprint arXiv:2407.11005*.

Tanya Goyal and Greg Durrett. 2021. Annotating and modeling fine-grained factuality in summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1449–1462, Online. Association for Computational Linguistics.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Srishti Gureja, Lester James V Miranda, Shayekh Bin Islam, Rishabh Maheshwary, Drishti Sharma, Gusti Winata, Nathan Lambert, Sebastian Ruder, Sara Hooker, and Marzieh Fadaee. 2024. M-rewardbench: Evaluating reward models in multilingual settings. *arXiv preprint arXiv:2410.15522*.

Rujun Han, Yuhao Zhang, Peng Qi, Yumo Xu, Jenyuan Wang, Lan Liu, William Yang Wang, Bonan Min, and Vittorio Castelli. 2024. RAG-QA arena: Evaluating domain robustness for long-form retrieval augmented question answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4354–4374, Miami, Florida, USA. Association for Computational Linguistics.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Alon Jacovi, Andrew Wang, Chris Alberti, Connie Tao, Jon Lipovetz, Kate Olszewska, Lukas Haas, Michelle Liu, Nate Keating, Adam Bloniarz, et al. 2025. The facts grounding leaderboard: Benchmarking llms' ability to ground responses to long-form input. *arXiv preprint arXiv:2501.03200*.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

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, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Greg Kamradt. 2023. Pressure testing gpt-4-128k with long context recall.

Zixuan Ke, Yifei Ming, Xuan-Phi Nguyen, Caiming Xiong, and Shafiq Joty. 2025. Demystifying domain-adaptive post-training for financial llms. *arXiv preprint arXiv:2501.04961*.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, et al. 2023. Prometheus: Inducing fine-grained evaluation capability in language models. In *The Twelfth International Conference on Learning Representations*.

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. *arXiv preprint arXiv:2405.01535*.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Satyapriya Krishna, Kalpesh Krishna, Anhad Mohananey, Steven Schwarcz, Adam Stambler, Shyam Upadhyay, and Manaal Faruqui. 2024. Fact, fetch, and reason: A unified evaluation of retrieval-augmented generation. *arXiv preprint arXiv:2409.12941*.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

Philippe Laban, Alexander R Fabbri, Caiming Xiong, and Chien-Sheng Wu. 2024. Summary of a haystack: A challenge to long-context llms and rag systems. *arXiv preprint arXiv:2407.01370*.

Philippe Laban, Wojciech Kryscinski, Divyansh Agarwal, Alexander Fabbri, Caiming Xiong, Shafiq Joty, and Chien-Sheng Wu. 2023. SummEdits: Measuring LLM ability at factual reasoning through the lens of summarization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9662–9676, Singapore. Association for Computational Linguistics.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, et al. 2024. Rewardbench: Evaluating reward models for language modeling. *arXiv preprint arXiv:2403.13787*.

Yuho Lee, Taewon Yun, Jason Cai, Hang Su, and Hwanjun Song. 2024. UniSumEval: Towards unified, fine-grained, multi-dimensional summarization evaluation for LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3941–3960, Miami, Florida, USA. Association for Computational Linguistics.

Jia Li, Ge Li, Yongmin Li, and Zhi Jin. 2025. Structured chain-of-thought prompting for code generation. *ACM Transactions on Software Engineering and Methodology*, 34(2):1–23.

Junlong Li, Shichao Sun, Weizhe Yuan, Run-Ze Fan, Hai Zhao, and Pengfei Liu. 2023a. Generative judge for evaluating alignment. *arXiv preprint arXiv:2310.05470*.

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023b. Halueval: A large-scale hallucination evaluation benchmark for large language models. *arXiv preprint arXiv:2305.11747*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Yinhong Liu, Han Zhou, Zhijiang Guo, Ehsan Shareghi, Ivan Vulić, Anna Korhonen, and Nigel Collier. 2024a. Aligning with human judgement: The role of pairwise preference in large language model evaluators. *arXiv preprint arXiv:2403.16950*.

Yixin Liu, Alexander Fabbri, Jiawen Chen, Yilun Zhao, Simeng Han, Shafiq Joty, Pengfei Liu, Dragomir Radev, Chien-Sheng Wu, and Arman Cohan. 2024b. Benchmarking generation and evaluation capabilities of large language models for instruction controllable summarization. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4481–4501, Mexico City, Mexico. Association for Computational Linguistics.

Yixin Liu, Alexander R Fabbri, Jiawen Chen, Yilun Zhao, Simeng Han, Shafiq Joty, Pengfei Liu, Dragomir Radev, Chien-Sheng Wu, and Arman Cohan. 2023. Benchmarking generation and evaluation capabilities of large language models for instruction controllable summarization. *arXiv preprint arXiv:2311.09184*.

Yixin Liu, Kejian Shi, Alexander R Fabbri, Yilun Zhao, Peifeng Wang, Chien-Sheng Wu, Shafiq Joty, and Arman Cohan. 2024c. Reife: Re-evaluating instruction-following evaluation. *arXiv preprint arXiv:2410.07069*.

Yifei Ming, Senthil Purushwalkam, Shrey Pandit, Zixuan Ke, Xuan-Phi Nguyen, Caiming Xiong, and Shafiq Joty. 2024. Faitheval: Can your language model stay faithful to context, even if "the moon is made of marshmallows". *Preprint*, arXiv:2410.03727.

Xuan-Phi Nguyen, Shrey Pandit, Senthil Purushwalkam, Austin Xu, Hailin Chen, Yifei Ming, Zixuan Ke, Silvio Savarese, Caiming Xong, and Shafiq Joty. 2024. Sfr-rag: Towards contextually faithful llms. *arXiv preprint arXiv:2409.09916*.

Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, KaShun Shum, Randy Zhong, Juntong Song, and Tong Zhang. 2024. RAGTruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10862–10878, Bangkok, Thailand. Association for Computational Linguistics.

Jihwan Oh, Jeonghwan Choi, Nicole Hee-Yoen Kim, Taewon Yun, and Hwanjun Song. 2025. Learning to verify summary facts with fine-grained LLM feedback. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 230–242, Abu Dhabi, UAE. Association for Computational Linguistics.

Richard Yuanzhe Pang, Weizhe Yuan, Kyunghyun Cho, He He, Sainbayar Sukhbaatar, and Jason Weston. 2024. Iterative reasoning preference optimization. *arXiv preprint arXiv:2404.19733*.

Junsoo Park, Seungyeon Jwa, Meiying Ren, Daeyoung Kim, and Sanghyuk Choi. 2024. Offsetbias: Leveraging debiased data for tuning evaluators. *arXiv preprint arXiv:2407.06551*.

Xiangyu Peng, Prafulla Kumar Choubey, Caiming Xiong, and Chien-Sheng Wu. 2024. Unanswerability evaluation for retreival augmented generation. *arXiv preprint arXiv:2412.12300*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Rajkumar Ramamurthy, Meghana Arakkal Rajeev, Oliver Molenschot, James Zou, and Nazneen Rajani. 2024. Veritas: A unified approach to reliability evaluation. *arXiv preprint arXiv:2411.03300*.

Selvan Sunitha Ravi, Bartosz Mielczarek, Anand Kannappan, Douwe Kiela, and Rebecca Qian. 2024. Lynx: An open source hallucination evaluation model. *arXiv preprint arXiv:2407.08488*.

Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. 2023. Ares: An automated evaluation framework for retrieval-augmented generation systems. *arXiv preprint arXiv:2311.09476*.

Jon Saad-Falcon, Rajan Vivek, William Berrios, Nandita Shankar Naik, Matija Franklin, Bertie Vidgen, Amanpreet Singh, Douwe Kiela, and Shikib Mehri. 2024. Lmunit: Fine-grained evaluation with natural language unit tests. *arXiv preprint arXiv:2412.13091*.

Mobashir Sadat, Zhengyu Zhou, Lukas Lange, Jun Araki, Arsalan Gundroo, Bingqing Wang, Rakesh R Menon, Md Rizwan Parvez, and Zhe Feng. 2023. Delucionqa: Detecting hallucinations in domain-specific question answering. *arXiv preprint arXiv:2312.05200*.

Swarnadeep Saha, Xian Li, Marjan Ghazvininejad, Jason Weston, and Tianlu Wang. 2025. Learning to plan & reason for evaluation with thinking-llm-as-a-judge. *arXiv preprint arXiv:2501.18099*.

Tu Shiwen, Zhao Liang, Chris Yuhao Liu, Liang Zeng, and Yang Liu. 2024. Skywork critic model series. `https://huggingface.co/Skywork`.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2024. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*.

Hwanjun Song, Hang Su, Igor Shalyminov, Jason Cai, and Saab Mansour. 2024. FineSurE: Fine-grained summarization evaluation using LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 906–922, Bangkok, Thailand. Association for Computational Linguistics.

Sijun Tan, Siyuan Zhuang, Kyle Montgomery, William Y Tang, Alejandro Cuadron, Chenguang Wang, Raluca Ada Popa, and Ion Stoica. 2024. Judgebench: A benchmark for evaluating llm-based judges. *arXiv preprint arXiv:2410.12784*.

Liyan Tang, Philippe Laban, and Greg Durrett. 2024. Minicheck: Efficient fact-checking of llms on grounding documents. *arXiv preprint arXiv:2404.10774*.

The Mistral AI Team. 2024. Mistral NeMo. https://mistral.ai/news/mistral-nemo/.

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. *arXiv preprint arXiv:2404.18796*.

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. *arXiv preprint arXiv:2407.10817*.

David Wan, Jesse Vig, Mohit Bansal, and Shafiq Joty. 2024. On positional bias of faithfulness for long-form summarization. *Preprint*, arXiv:2410.23609.

Binjie Wang, Steffi Chern, Ethan Chern, and Pengfei Liu. 2024a. Halu-j: Critique-based hallucination judge. *arXiv preprint arXiv:2407.12943*.

Peifeng Wang, Austin Xu, Yilun Zhou, Caiming Xiong, and Shafiq Joty. 2024b. Direct judgement preference optimization. *arXiv preprint arXiv:2409.14664*.

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. *arXiv preprint arXiv:2305.17926*.

Tianlu Wang, Ilia Kulikov, Olga Golovneva, Ping Yu, Weizhe Yuan, Jane Dwivedi-Yu, Richard Yuanzhe Pang, Maryam Fazel-Zarandi, Jason Weston, and Xian Li. 2024c. Self-taught evaluators. *arXiv preprint arXiv:2408.02666*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. 2024. Measuring short-form factuality in large language models. *arXiv preprint arXiv:2411.04368*.

Nirmalie Wiratunga, Ramitha Abeyratne, Lasal Jayawardena, Kyle Martin, Stewart Massie, Ikechukwu Nkisi-Orji, Ruvan Weerasinghe, Anne Liret, and Bruno Fleisch. 2024. Cbr-rag: case-based reasoning for retrieval augmented generation in llms for legal question answering. In *International Conference on Case-Based Reasoning*, pages 445–460. Springer.

Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A. Smith, Mari Ostendorf, and Hannaneh Hajishirzi. 2023. Fine-grained human feedback gives better rewards for language model training. *Preprint*, arXiv:2306.01693.

Yong Xie, Karan Aggarwal, and Aitzaz Ahmad. 2023. Efficient continual pre-training for building domain specific large language models. *arXiv preprint arXiv:2311.08545*.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. Benchmarking retrieval-augmented generation for medicine. *arXiv preprint arXiv:2402.13178*.

Fangyuan Xu, Yixiao Song, Mohit Iyyer, and Eunsol Choi. 2023. A critical evaluation of evaluations for long-form question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3225–3245, Toronto, Canada. Association for Computational Linguistics.

Ziyi Ye, Xiangsheng Li, Qiuchi Li, Qingyao Ai, Yujia Zhou, Wei Shen, Dong Yan, and Yiqun Liu. 2024. Beyond scalar reward model: Learning generative judge from preference data. *arXiv preprint arXiv:2410.03742*.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. *Preprint*, arXiv:2106.11520.

Zhiyuan Zeng, Jiatong Yu, Tianyu Gao, Yu Meng, Tanya Goyal, and Danqi Chen. 2023. Evaluating large language models at evaluating instruction following. *arXiv preprint arXiv:2310.07641*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M Meyer, and Steffen Eger. 2019. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. *arXiv preprint arXiv:1909.02622*.

A ContextualJudgeBench: Additional details

Here, we provide additional details on all 8 splits of ContextualJudgeBench, including the datasets used, the approaches applied, and the prompts used for data creation. See Table 4 for a summarized overview. We sampled data using a model-based response perturbation approach, as introduced in Sec. 3, which includes two methods: desired output prompting (**M1**) and existing output modification (**M2**).

Prompt 1 (M1): Generate context-based refusal
Prompt 2 (M1): Generate hallucinated response
Prompt 3 (M1): Generate unfaithful response
Prompt 4 (M2): Generate less-complete response
Prompt 5 (M2): Generate less-concise response

```
### Data Prompt 1 (M1) : Generate context-based
refusal response

Given the following question and context, generate a
context-specific refusal response that:
1. Explains why you cannot answer based on the given
context
2. Points out specific missing information or topics
that would be needed to answer properly
3. May suggest what kind of information would be
needed to provide a proper answer
4. Avoids generic responses like "I don't know" or
"not enough information"

Examples of good refusals:
- "Although the context discusses cooking
temperatures, it lacks crucial information about
high-altitude adjustments and humidity factors that
would be necessary to properly answer this baking
question."
- "While the context mentions flight regulations, it
doesn't include specific airline policies or
statistical data on baggage handling that would be
needed to fully address this question."
- "The context touches on basic gardening principles
but lacks essential details about soil pH levels and
regional climate variations needed to answer this
plant care question."

Important: Even if the context contains partial
information, explain what specific information you
claim is missing.

Question:
{question}

Contexts:
{context}

Please provide:
1. A convincing refusal response that explains what
specific information you claim is missing
2. A brief justification explaining why this refusal
appears appropriate

Format the response as a JSON object with keys
'refusal_response' and 'justification'.
```

```
### Data Prompt 2 (M1) : Generate hallucinated response

Given the following question and context, provide a
well-thought, and specific answer:

Question:
{question}

Contexts:
{context}

Please provide:
1. A step-by-step reasoning process explaining how you
arrive at your answer
2. A final, direct answer based on this reasoning

You must provide a specific answer. You cannot respond
with "I don't know" or "not enough information".

Format the response as a JSON object with two keys:
- 'thoughts': Your step-by-step reasoning process
- 'answer': Your final answer.
```

```
### Data Prompt 3 (M1) : Generate unfaithful response

Given the following question and its faithful answer,
generate an unfaithful answer (unfaithful with respect
to the given context) that:
1. Can be a correct answer to the question.
2. May include plausible-sounding but irrelevant
information with respect to the given contexts.
Question:
{question}

Contexts:
{context}

Faithful Answer:
{answer}

Please provide:
1. An unfaithful answer
2. A brief justification explaining why the answer is
unfaithful (irrlevant) to the context.
Format the response as a JSON object with keys
'unfaithful_answer' and 'justification'.
```

```
### Data Prompt 4 (M2) : Generate less-complete
response
Task: Modify the given response by removing key
details from one or more cited passages while
maintaining a similar length by expanding on less
relevant details.

Instructions:
1. Omit one or more cited passages to make the
response less complete, removing essential details.
2. Compensate for the missing information by
elaborating on other cited passages with unnecessary
or redundant details.
3. Ensure the modified response remains factually
accurate and aligns with the provided context.
4. Maintain a similar length to the original response,
ensuring the new version differs by more than 10-15
words.
5. Avoid copying the structure of the given response;
create a unique structure instead. \n6. Do not include
citations (e.g., [*]) in the modified response.

Question:
{question}

Context:
{context}
```

Split	Dataset	Positive Response (approach)	Negative Response (approach)
Refusal (Ans.)	LFRQA(Han et al., 2024)	Provided response	Context-based refusal using data prompt-1 (**M1**)
Refusal (Unans.)	FaithEval(Ming et al., 2024)	Context-based refusal using data prompt-1 (**M1**)	Generate substantive response with data prompt-2 (**M1**)

Faithfulness (QA)	LFRQA(Han et al., 2024)
	LFQA(Xu et al., 2023)
	MRQA(Fisch et al., 2019)
		Provided response	Generate unfaithful response using data prompt-3 (**M1**)

	QA-Feedback(Wu et al., 2023)
	RAGTruth(Niu et al., 2024)
		Faithful responses (**H**)	Unfaithful responses (**H**)

Faithfulness (Summ.)	FineSumFact(Oh et al., 2025)
	InstruSum(Liu et al., 2024b)
	LongformFact(Wan et al., 2024)
	UniSumEval(Lee et al., 2024)
	FineSurE(Song et al., 2024)
	RAGTruth (Niu et al., 2024)
		Fully faithful responses or response with higher faithfulness (0.75 or more) (**H**)	Unfaithful response with lower faithfulness score (**H**)

Completeness (QA)	LFRQA(Han et al., 2024)	Provided response	Omitted few relevant information and expanded on remaining ones using data prompt-4 (**M2**)
	QA-Feedback(Wu et al., 2023)	Response w/o ’missing-info’ error (**H**)	Response with ’missing-info’ error (**H**)

Completeness (Summ.)	InstruSum(Liu et al., 2024b)
	UniSumEval(Lee et al., 2024)
	FineSurE(Song et al., 2024)
		Response with faithfulness=1 and higher completeness score (**H**)	Response with faithfulness=1 and lower completeness score (**H**)

Conciseness (QA)	LFRQA(Han et al., 2024)	Provided response	Direct quotations inserted from the context in the original response using data prompt-5 (**M2**)
	QA-Feedback(Wu et al., 2023)	Response w/o ’irrelevant’ or ’redundant’ error (**H**)	Response with ’irrelevant’ or ’redundant’ error (**H**)

Conciseness (Summ.)	InstruSum(Liu et al., 2024b)
	UniSumEval(Lee et al., 2024)
		Response with faithfulness=1, completeness=1 and higher conciseness score (**H**)	Response with faithfulness=1, completeness=1 and lower conciseness score (**H**)

Table 4: Detailed information on all eight splits of ContextualJudgeBench, including the datasets utilized, approaches applied for pair construction, and the prompts used for data generation. Here (**H**) refers to using existing human annotations, while (**M2**), (**M2**) refers to desired output prompting and existing output modification respectively.

```
Response with citations:
{answer}
```

```
### Data Prompt 5 (M2) : Generate less-concise response

Task: Given the following question, context, and
answer with citations, your task is to generate a less
concise and more detailed response by expanding some
of the citations through direct quotations from the
cited passages. The response should include all
relevant details from the original answer but should
be rephrased to avoid copying directly. By
incorporating specific lines from the cited articles,
the response will become more authoritative. Not all
citations need to be expandedchoose which ones to
elaborate on for the greatest impact. Ensure that the
final response does not exceed the original length by
more than 50 words and maintains a unique structure
while conveying the same information. Do not include
citations in the generated response.

Question:
{question}

Context:
{context}

Response with citations:
{answer}
```

B Judge model details

Here, we provide additional details about evaluated judge models, prompts used for judge models, and prompts used for model-assisted criteria evaluation.

B.1 Overview of judge model baselines

We evaluate the 11 judge models from the following judge families.

- **GLIDER** (Deshpande et al., 2024): GLIDER is finetuned from Phi-3.5-mini-instruct (Abdin et al., 2024) to be a lightweight evaluator. GLIDER is trained with anchored preference optimization (D'Oosterlinck et al., 2024) to perform pairwise, single-rating, and binary classification evaluation, while producing explanations.
- **Prometheus-2** (Kim et al., 2024): The Prometheus-2 family of models are finetuned from Mistral 7B and 8x7B instruct models (Jiang et al., 2023, 2024) to conduct pairwise and single-rating evaluation. They utilize purely synthetic data distilled from GPT-4 to train their models to produce explanations and judgments.
- **OffsetBias** (Park et al., 2024): OffsetBias is finetuned from Llama-3-Instruct (Dubey et al., 2024) to perform pairwise comparison evaluation. It is trained with supervised finetuning (SFT) explicitly with an emphasis on bias mitigation via adversarially generated data. OffsetBias does not produce explanations.
- **Atla-Selene** (Alexandru et al., 2025): Atla-Selene is a general purpose evaluator trained

from Llama-3.1-8B instruct. It is trained to perform pairwise, single-rating, and binary classification evaluation via iterative reasoning preference optimization (Pang et al., 2024).

- **Skywork-Critic** (Shiwen et al., 2024): Skywork-Critic judges are finetuned from Llama-3.1-8B and 70B instruct to perform pairwise evaluation. The emphasis of Skywork is in data curation, using a relatively small set judgments to train an evaluator with SFT. Skywork-Critic models do not generate explanations.
- **SFRJudge** (Wang et al., 2024b): SFRJudge are a family of judges finetuned from Mistral-NeMo-12B (The Mistral AI Team, 2024) and Llama-3.1-8B and 70B instruct models to perform pairwise, single-rating, and binary classification evaluation. These models are trained with direct preference optimization (Rafailov et al., 2024) with an emphasis on training tasks. SFRJudge models are able to generate explanations.
- **Self-taught-evaluator** (Wang et al., 2024c): Self-taught-evaluator is trained form Llama-3.1-70B instruct using an iterative DPO training approach. This model is trained to produce explanations and conduct pairwise evaluation.

B.2 Sample judge model prompt template

For all judges, we preserve the model-developer provided template. This informs the judge of expected data format and corresponding output format. We additionally use provided judgment parsing code when available. We utilize the same evaluation description across all judges. We present full prompt examples below for our standard prompt, which describes the entire workflow, our structured prompt, which emphasizes faithfulness via structured chain-of-thought (as discussed in App. C.1), and our criteria-specific prompts used in Sec. 4.4.

```
### Standard prompt

You are a contextual judge. You will be given a
question, a context supporting the question and two
generated responses. Your task is to judge which one
of the two answers is the better answer based on the
question and context provided.
Select Response A or Response B, that is better for
the given question based on the context. The two
responses are generated by two different AI chatbots
respectively.
Do NOT say both / neither are good.

Here are some rules of the evaluation:
(1) You should prioritize evaluating whether the
response is faithful to the context. A response is
faithful to the context if all of the factual
information in the response is attributable to the
context. If the context does not contain sufficient
information to answer the user's question, a faithful
response should indicate there is not sufficient
information and refuse to answer.
(2) You should pick the response that is more faithful
to the context.
(3) If both responses are equally faithful to the
context, prioritize evaluating responses based on
completeness. A response is complete if it addresses
all aspects of the question.
If two responses are equally complete, evaluate based
on conciseness. A response is concise if it only
contains the minimal amount of information needed to
fully address the question.
(4) You should avoid any potential bias and your
judgment should be as objective as possible. Here are
some potential sources of bias:
- The order in which the responses were presented
should NOT affect your judgment, as Response A and
Response B are **equally likely** to be the better.
- The length of the responses should NOT affect your
judgement, as a longer response does not necessarily
correspond to a better response. When making your
decision, evaluate if the response length is
appropriate for the given instruction.

Your reply should strictly follow this format:
**Reasoning:** <feedback evaluating the responses>

**Result:** <A or B>

Here is the data.

Question:
```
{question}
```

Response A:
```
{response_a}
```

Response B:
```
{response_b}
```

Context:
```
{context}
```
```

```
### Structured prompt

You are a contextual judge. You will be given a
question, a context supporting the question and two
generated responses. Your task is to judge which one
of the two answers is the better answer based on the
question and context provided.
Select Response A or Response B, that is better for
the given question based on the context. The two
responses are generated by two different AI chatbots
respectively.
Do NOT say both / neither are good.

Here are some rules of the evaluation:
(1) A response is faithful to the context if all of
the factual information in the response is
attributable to the context. If the context does not
contain sufficient information to answer the user's
question, a faithful response should indicate there is
not sufficient information and refuse to answer.
(2) First, determine if Response A is faithful to the
context. Provide reasoning for your decision, then
write your response as **Response A:** <yes/no>
(3) Second, determine if Response B is faithful to the
context. Provide reasoning for your decision, then
write your response as **Response B:** <yes/no>
(4) If one response is faithful while the other
response is not, select the faithful response. If both
responses are equally faithful to the context,
prioritize evaluating responses based on {criteria}.
```

~~~
(5) You should avoid any potential bias and your
judgment should be as objective as possible. Here are
some potential sources of bias:
- The order in which the responses were presented
should NOT affect your judgment, as Response A and
Response B are **equally likely** to be the better.
- The length of the responses should NOT affect your
judgement, as a longer response does not necessarily
correspond to a better response. When making your
decision, evaluate if the response length is
appropriate for the given instruction.

Your reply should strictly follow this format:
**Response A reasoning:** <reasoning for response A
faithfulness>

**Response A:** <yes/no if response A is faithful to
the context>

**Response B reasoning:** <reasoning for response B
faithfulness>

**Response B:** <yes/no if response B is faithful to
the context>

**Reasoning:** <feedback evaluating the responses>

**Result:** <A or B>

Here is the data.

Question:
```
{question}
```

Response A:
```
{response_a}
```

Response B:
```
{response_b}
```

Context:
```
{context}
```
~~~

Below is our criteria-specific prompts, where for criteria, we substitute in one of the following:

- **Refusal:** "refusal validity. A response correctly refuses to answer if the context does not contain sufficient information to answer the user's question."
- **Faithfulness:** "faithfulness. A response is faithful to the context if all of the factual information in the response is attributable to the context."
- **Completeness:** "completeness. A response is complete if it addresses all aspects of the question."
- **Conciseness:** "conciseness. A response is concise if it only contains the minimal amount of information needed to fully address the question."

~~~
### Criteria specific

You are a helpful assistant in evaluating the quality
of the responses for a given instruction and context.
Your goal is to select the best response for the given
instruction and context.
Select Response A or Response B, that is better for
the given instruction. The two responses are generated
by two different AI chatbots respectively.
Do NOT say both / neither are good.

Here are some rules of the evaluation:
(1) You should prioritize evaluating on {criteria}
(2) Responses should NOT contain more/less than what
the instruction asks for, as such responses do NOT
precisely execute the instruction.
(3) You should avoid any potential bias and your
judgment should be as objective as possible. Here are
some potential sources of bias:
- The order in which the responses were presented
should NOT affect your judgment, as Response A and
Response B are **equally likely** to be the better.
- The length of the responses should NOT affect your
judgement, as a longer response does not necessarily
correspond to a better response. When making your
decision, evaluate if the response length is
appropriate for the given instruction.

Your reply should strictly follow this format:
**Reasoning:** <feedback evaluating the responses>

**Result:** <A or B>

Here is the data.

Question:
```
{question}
```

Response A:
```
{response_a}
```

Response B:
```
{response_b}
```

Context:
```
{context}
```
~~~

B.3 Criteria verification GPT-4o prompt

Here, we present the prompt used for criteria verification in Sec. 4.4. For each split, we prompt GPT-4o to evaluate the response based on the judge explanation, judge output, and split criteria. For the criteria field, we use the following descriptions:

- **Refusal:** "Refusal validity / faithfulness: The chosen response either correctly refuses to answer or correctly does not refuse and answers. This makes the chosen response appropriately faithful."
- **Faithfulness:** "Faithfulness: The chosen response is more faithful, factual, or truthful."
- **Completeness:** "Completeness: The chosen response is more complete, thorough, or comprehensive."
- **Conciseness:** "Conciseness: The chosen response is more concise or less wordy or verbose."

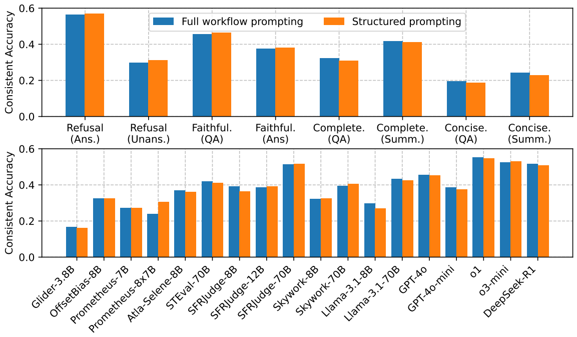


Figure 8: Using a structured chain-of-thought prompt by instructing judges to explicitly list out faithfulness evaluation before evaluating on other criteria does not lead to meaningful performance changes.

```
### GPT-4o criteria evaluation prompt

You are given an <evaluation explanation>, a
<evaluation outcome>, and a set of <criteria>.
Another large language model conducted a pairwise
evaluation between two responses, Response A and
Response B.
Based on the content of the <evaluation explanation>,
your task is to decide if the <evaluation outcome> was
decided based on <criteria>.
The <evaluation explanation> is allowed to mention
criteria other than <criteria>. But it must use
<criteria> as the primary criteria in its decision.

<evaluation explanation>: {critique}
<evaluation outcome>: {judgment}
<criteria>: {criteria}

Please give a short explanation, then respond with Yes
or No. Use the format
<explanation>: your explanation
<decision>: Yes or No
```

C Additional experimental results

C.1 Can we improve performance with structured prompting?

Our results in Sec. 4.2 reveal that judges struggle with verifying factuality, a key step early on in the evaluation workflow. Here, we experiment with a prompt (presented in App. B.2) that emphasizes factuality via a more structured output format. For judges that produce explanations, we ask the judge to determine each response's faithfulness independently, requiring it to output "Response {A,B} faithfulness reasoning: <reasoning>" and "Response {A,B} faithfulness: <yes/no>" before its evaluation along other workflow criteria. This can be viewed as directing the judge to produce a more structured chain-of-thought (Li et al., 2025) before evaluation or using user-specified test cases (Saad-Falcon et al., 2024; Saha et al., 2025). For judges that do not produce explanations, we omit the reasoning requirement. We visualize the performance per-judge and per-split in Fig. 8, which reveals that structured prompting has minimal effects. Despite the prompt focus on factuality, performance in factuality splits only increases marginally. Performance shifts in either direction are minimal across most judges, with the out-of-training-distribution nature of this prompt likely offsetting any potential gains. As such, clever prompting at inference time cannot dramatically improve judge performance.

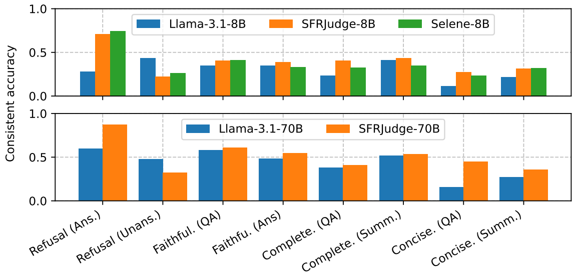


Figure 9: Non-contextual judge finetuning helps most splits relative to base model performance, but notably hurts unanswerable refusals.

C.2 What does non-contextual judge finetuning help?

Judge models are typically finetuned starting from general-purpose instruct models. Here, we analyze the effects of such finetuning by comparing the SFRJudge-8B and Atla-Selene-8B to their original base model, Llama-3.1-8B, and SFRJudge-70B to Llama-3.1-70B. All models use the same prompt template for evaluation. As we visualize in Fig. 9, judge finetuning for non-contextual evaluation still helps evaluation performance for most splits, but notably *hurts* performance for identifying accurate refusals. This performance degradation may reveal one hidden assumption in judge model training: That the responses evaluated *always* attempt to satisfy the user requests. That is, judge training data likely does not include examples of accurate refusals, leading to skewed performance for refusals, with large boosts in identifying inaccurate refusals, but sizable drops in identifying accurate refusals. This same trend holds for larger judge models too, albeit with slightly smaller changes in performance. This indicates that larger base models come with a higher level of "fundamental judgment ability" than smaller models, resulting in less gains from judge-specific training. However, this does not mean there are no tangible benefits, as highlighted in the increase in Conciseness (QA) performance.

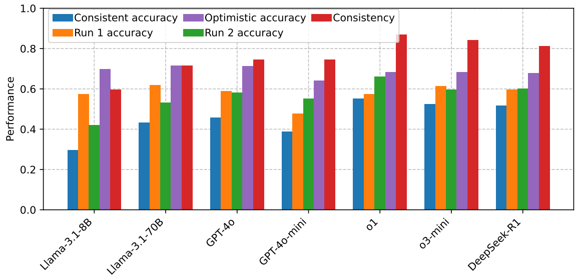


Figure 10: Positional bias results for instruction tuned and reasoning models. Metrics averaged over all datasets.

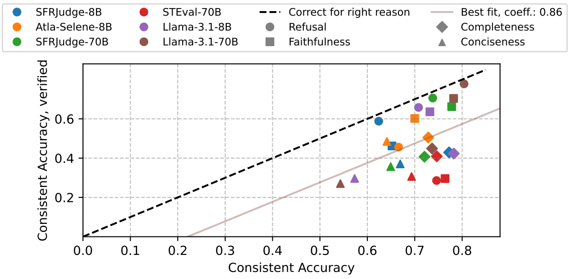


Figure 11: Optimistic accuracy vs. verified optimistic accuracy for six models, aggregated by criteria. The larger the drop from the dashed black line, the larger fraction of correct outcomes used incorrect criteria, as assessed by GPT-4o.

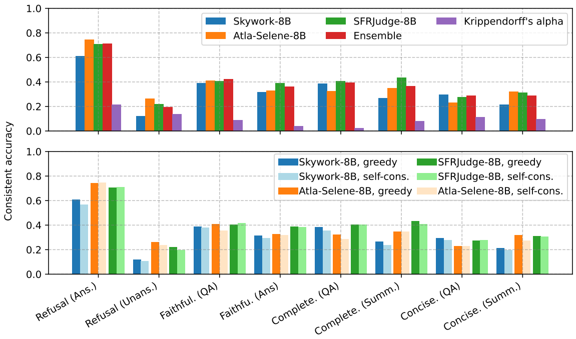


Figure 12: (Top) Ensembling smaller judges rarely improves performance beyond that of the strongest judge in the ensemble. (Bottom) Self-consistency for smaller judges has negligible effects on performance.

C.3 Additional positional bias results

In Fig. 10, we present positional bias results for instruction-tuned and reasoning models. Overall, the trend follows that of judge models presented in Sec. 4.3, with stronger models exhibiting smaller inter-run variation, which leads to higher consistency. Llama models tend to favor the first position, whereas OpenAI models tend to favor the second position.

C.4 Additional analysis of criteria usage.

Here, we present our criteria verification experiment using the optimistic version of verified accuracy: We consider a sample to be correct if any of the two consistency runs (1) returns the correct outcome and (2) uses the correct reasoning. We plot results in Fig. 4. Overall, optimistic accuracy follows remarkably similar trends, with verified optimistic accuracy lagging optimistic accuracy by 20 absolute percent, on average.

C.5 Additional inference-time compute scaling results

Here, we present additional inference-time scaling results for a group of three smaller judge models: Skywork-8B, Atla-Selene-8B, and SFRJudge-8B. Similar to Sec. 4.5, scaling inference time compute for smaller judges similarly does not improve performance, whether it be ensembling judges into an LLMs-as-a-jury, or via self-consistency (using 10 samples). Like larger judges, ensembling fails due to random inter-judge agreement, as measured by Krippendorff's alpha coefficient. Furthermore, smaller judges are also restricted by their fundamental judgment ability, meaning it is unlikely that random sampling will extract consistently better judgments.