Judging Against the Reference: Uncovering Knowledge-Driven Failures in LLM-Judges on QA Evaluation

Anonymous ACL submission

Abstract

While large language models (LLMs) are increasingly used as automatic judges for question answering (QA) and other reference-conditioned evaluation tasks, little is known about their ability to adhere to a provided reference. We identify a critical failure mode of such reference-based LLM QA evaluation: when the provided reference conflicts with the judge model's parametric knowledge, the resulting scores become unreliable, substantially degrading evaluation fidelity. To study this phenomenon systematically, we introduce a controlled swapped-reference QA framework that induces reference–belief conflicts. Specifically, we replace the reference answer with an incorrect entity and construct diverse pairings of original and swapped references with correspondingly aligned candidate answers. Surprisingly, grading reliability drops sharply under swapped references across a broad set of judge models. We empirically show that this vulnerability is driven by judges' over-reliance on parametric knowledge, leading judges to disregard the given reference under conflict. Finally, we find that this failure persists under common prompt-based mitigation strategies, highlighting a fundamental limitation of LLM-as-a-judge evaluation and motivating reference-based protocols that enforce stronger adherence to the provided reference.

1 Introduction

The evaluation of large language models (LLMs) is rapidly shifting toward the *LLM-as-a-judge* paradigm (Adlakha et al., 2024; Liu et al., 2023b; Gu et al., 2025; Zheng et al., 2023). Owing to their scalability and strong correlation with human annotators, LLM-as-a-judge methods have rapidly become a dominant paradigm for QA evaluation (Zhang et al., 2024; Blandón et al., 2025; Hosseini et al., 2024; Ho et al., 2025), where a judge model grades a candidate answer conditioned on the question and a reference answer. Despite these practical advantages, prior work has documented important limitations of LLM-based QA evaluation, including systematic biases and inconsistent verdicts (Lee et al., 2025a; Kamalloo et al., 2024; Li et al., 2025).

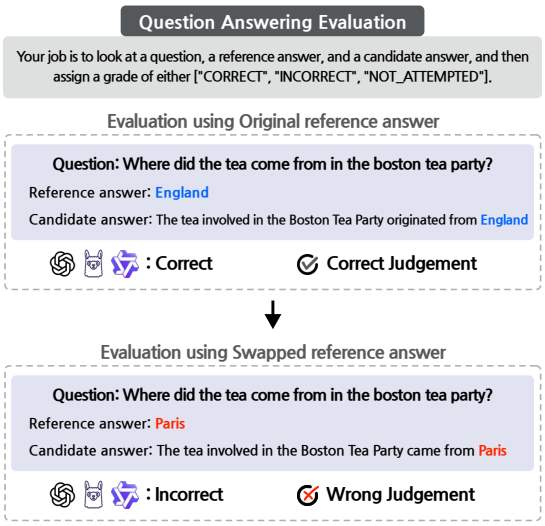

Figure 1: The LLM judge shows an critical failure mode under a swapped reference answer: with the original reference (top), it produces the correct judgment, whereas with the swapped reference (bottom), it produces an incorrect judgment even in a trivial case.

However, LLM-judges' ability to adhere to a provided reference remains poorly understood, even though such adherence is central to reference-conditioned evaluation in practice. In this paper, we investigate a striking and consequential failure mode: when the reference answer is intentionally swapped to a different entity, LLM judges can fail even in trivial cases under the provided reference. For example, as illustrated in Figure 1, a judge labels a candidate as `Incorrect` despite the candidate clearly matching the swapped reference (e.g., "Paris"). This behavior suggests that, under swapped references, judges do not reliably

follow evaluation instructions and frequently disregard the given gold reference answer. This concern is particularly important because factual knowledge evolves over time (Vu et al., 2024; Kasai et al., 2023; Liska et al., 2022; Zhang et al., 2023), and datasets may intentionally encode counterfactual targets that diverge from a model's parametric knowledge (Yu et al., 2023; Neeman et al., 2023; Wu et al., 2024). Motivated by this observation, we study why such failures arise in straightforward evaluation cases and aim to clarify the mechanisms driving this breakdown.

To enable controlled investigation, we propose a systematic framework for swapped-reference QA evaluation. Specifically, we curate four existing QA datasets—NaturalQuestions-Open (Lee et al., 2019), PopQA (Mallen et al., 2023), SciQ (Welbl et al., 2017), and FreshQA (Vu et al., 2024)—and construct tailored evaluation instances by creating diverse combinations of original and swapped reference answers together with corresponding candidate answers. Experiments with this framework reveal that even strong judge models frequently fail to prioritize the provided reference when it is swapped. We further show that this vulnerability persists across datasets and across different swapped entities, and does not vanish with increased evaluator capacity.

Moreover, through extensive empirical analysis, we show that this vulnerability is driven by judges' over-reliance on parametric knowledge. Failures are strongly associated with the judge's internal beliefs: vulnerability diminishes when the swapped reference aligns with the judge's parametric knowledge, and it increases with the popularity of the associated knowledge. We further find that common prompt-based mitigation strategies, including chain-of-thought prompting, self-consistency, and detailed task instructions, do not resolve the issue. Taken together, these results expose a critical challenge for LLM-as-a-judge approaches in QA evaluation: when the provided gold reference conflicts with a judge's parametric knowledge, adherence to the reference can substantially degrade.

In summary, we examine the challenges LLM judges face when the provided reference contradicts their internal knowledge, and we make the following contributions:

- We uncover a critical failure mode in LLM-judge–based QA evaluation: judges often ignore evaluation instructions when the reference answer is different from models' parametric knowledge.
- We introduce a systematic swapped-reference QA evaluation framework that enables fine-grained and controlled analysis of LLM judges' adherence to references across datasets, swap types, and candidate-reference alignments.
- We identify key factors underlying the failure, showing that stronger parametric knowledge can intensify ignorance of swapped references, and that standard prompting-based mitigation techniques provide limited relief.

2 Background & Study Formulation

In this section, we review the standard protocol for LLM-judge–based QA evaluation and formalize the setting studied in this work.

2.1 Background

We consider a Question Answering (QA) dataset consisting of questions $q_i{}_{i=1}^{N}$, where for each question q_i, the dataset provides a gold reference answer r_i and the system under evaluation produces a candidate answer c_i. While some QA benchmarks include multiple references per question, for simplicity we assume a single gold reference r_i.

Given (q_i, r_i, c_i), an LLM-judge is prompted with an evaluation instruction and asked to determine whether the candidate answer is correct with respect to the *provided* reference answer. The judge returns a verdict

$$v_i = \mathrm{LLM}(q_i, r_i, c_i).$$

These verdicts are compared against human judgments and aggregated across the dataset. An ideal LLM-judge should follow the given instruction and condition on the provided reference, marking correct candidates as *Correct* and incorrect candidates as *Incorrect*.

2.2 Swapped-Reference Setting

In standard QA benchmarks, gold reference answers are fixed at dataset creation time. In practice, however, knowledge can evolve over time (Vu et al., 2024; Chen et al., 2021) or vary across regions (Zhang and Choi, 2021), and thus what is commonly regarded as the "correct" answer may shift. Accordingly, under evaluation protocols

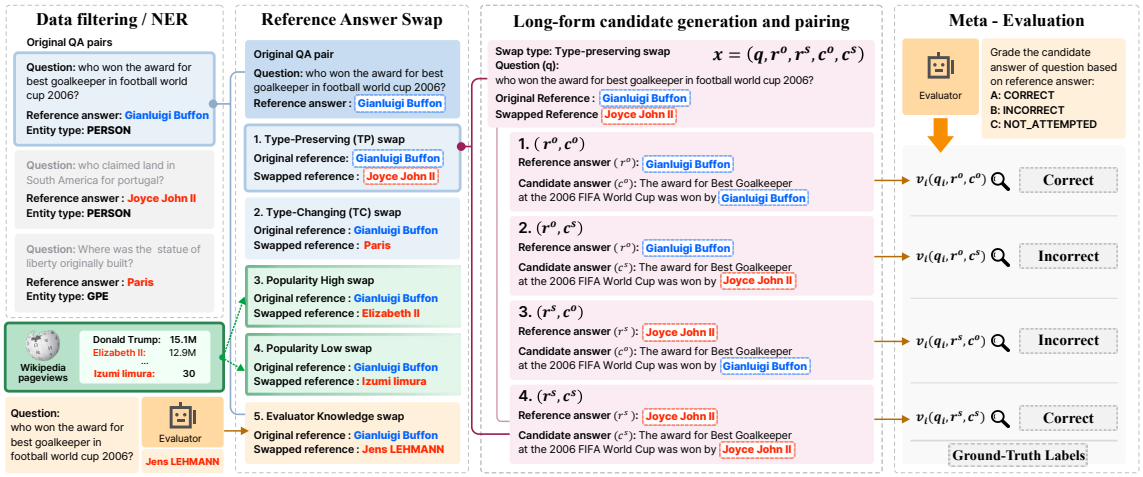

Figure 2: Evaluation framework overview

where correctness is defined solely by agreement with the *provided reference*, an ideal LLM-judge should rely on the provided reference rather than defaulting to its own prior knowledge.

To probe whether judges exhibit this reference-conditioned behavior, we introduce the *Swapped-Reference* setting, in which each original answer is substituted with an alternative reference answer. For each sample index i, we denote the original reference by r_i^o and the swapped reference by r_i^s, where the superscripts o/s indicate *original* vs. *swapped*. We then examine whether the judge's verdict changes when r_i^o is replaced by r_i^s thereby quantifying the extent to which LLM-judges genuinely condition on the provided reference.

3 Experimental Setup

To study how LLM judges evaluate candidate answers under swapped references, this section describes our overall framework for swapped-reference QA meta-evaluation and the associated metrics.

3.1 Swapped-Reference QA Meta-Evaluation

Each sample x_i in our dataset is a quintuple

$$x_i = (q_i, r_i^o, r_i^s, c_i^o, c_i^s),$$

where q_i is a question, r_i^o is the original reference answer from the source dataset, r_i^s is an artificially swapped reference answer, c_i^o is a candidate answer aligned with r_i^o, and c_i^s is a candidate answer aligned with r_i^s. To reflect common QA evaluation settings, where judges often grade sentence-level or long-form responses, we generate both c_i^o and c_i^s as sentence-level answers. The overall construction pipeline is depicted in Figure 2; we summarize each stage below.

Data filtering and entity labeling (NER). The data filtering and NER stage in Figure 2 first samples questions and reference answers from original datasets. We then apply an LLM-based named entity recognition (NER) pipeline to assign an entity-type label to each original reference answer r_i^o.

Reference swapping. Next, we perform reference swapping according to predefined strategies. Following Longpre et al. (2021), we implement two dataset-internal swap settings. In the *Type-Preserving Swap* (TP) setting, we replace r_i^o with another answer sampled from the same dataset that shares the same entity type (e.g., Gianluigi Buffon → Joyce John II in Figure 2). In the *Type-Changing Swap* (TC) setting, we replace r_i^o with another answer from the same dataset that has a different entity type (e.g., Gianluigi Buffon → Paris in Figure 2).

Next, we introduce a popularity-based swap (*Popularity-high/low swap*) for PERSON entities. We select the top and bottom 50 PERSON entities from PopQA (Mallen et al., 2023), ranked by Wikipedia pageviews, and use one of them to swap PERSON-type references across all datasets (e.g., Gianluigi Buffon → Elizabeth II and → Izumi Iimura, respectively, in Figure 2).

Finally, we introduce an *Evaluator-Knowledge Swap*. For each evaluator model, we first query the evaluator on q_i (under a standard QA prompting setup) and compare its predicted answer with r_i^o. If the prediction disagrees with r_i^o, we use the

evaluator's predicted answer as the alternative reference and set it as the swapped reference r_i^s for that instance. Evaluator-Knowledge Swap explicitly aligns the swapped reference with the evaluator's own belief. This setting therefore isolates how judge behavior changes under knowledge alignment versus knowledge conflict.

Long-form candidate generation and pairing. In the final stage, we instantiate evaluation triplets by pairing each question q_i with a reference choice and a candidate choice. For each instance, we generate long-form candidates conditioned on either r_i^o or r_i^s, yielding c_i^o (aligned with r_i^o) and c_i^s (aligned with r_i^s).

Meta-evaluation. We then form the four evaluation triplets, which induce ground-truth labels:

$$\textbf{Correct: } (q_i, r_i^o, c_i^o),\ (q_i, r_i^s, c_i^s);$$
$$\textbf{Incorrect: } (q_i, r_i^o, c_i^s),\ (q_i, r_i^s, c_i^o).$$

Thus, an ideal evaluator should consistently label aligned candidates as *Correct* and misaligned candidates as *Incorrect*, regardless of whether the reference is original or swapped. We compare each evaluator's verdict to the corresponding ground-truth label and report meta-evaluation results for each LLM judge.

Finally, human annotators manually reviewed all NER outputs, swap operations, and long-form generations to ensure that each instance is correctly swapped and that each (q_i, r_i^a, c_i^b) triplet carries the appropriate ground-truth label (where $a, b \in \{o, s\}$).

3.2 Statistics of Datasets

For each evaluator–dataset pair, we construct a corresponding Swapped-Reference QA Meta-Evaluation subset. Table 3 reports summary statistics for all constructed subsets across evaluators and datasets. The number of popularity-based swaps varies by dataset, reflecting differences in the prevalence of PERSON entities. The number of Evaluator-Knowledge swaps is determined by the instances in which the evaluator's own answer disagrees with the original reference set r_i^o, and therefore varies across evaluators.

3.3 Evaluation Metric

We introduce metrics to quantify evaluator behavior under the swapped-reference setting.

Accuracy. Let $\hat{y}_i^{a,b} \in \{Correct, Incorrect\}$ denote the evaluator's verdict for the triplet (q_i, r_i^a, c_i^b), where $a, b \in \{o, s\}$. We define the ground-truth label

$$y_i^{a,b} = \begin{cases} Correct & \text{if } a = b, \\ Incorrect & \text{if } a \neq b. \end{cases}$$

Accuracy under reference condition $a \in \{o, s\}$ is defined as follows; we report ACC^o (original reference) and ACC^s (swapped reference):

$$\mathrm{ACC}^a = \frac{1}{2N} \sum_{i=1}^{N} \sum_{b \in \{o,s\}} \mathbb{I}\left[\hat{y}_i^{a,b} = y_i^{a,b}\right].$$

Reference-Polarity Accuracy Gap (RPAG). To quantify sensitivity to reference swaps, we define

$$\mathrm{RPAG} = \left[\mathrm{ACC}^o - \mathrm{ACC}^s\right],$$

where ACC^o and ACC^s denote accuracy under the original references r_i^o and swapped references r_i^s, respectively. A larger RPAG indicates that an evaluator performs substantially better when conditioned on r_i^o but degrades under r_i^s, reflecting vulnerability under knowledge conflict. Conversely, a near-zero RPAG suggests robustness to reference swaps and closer adherence to the provided reference.

3.4 LLM-Evaluators

Throughout the experiments, we test thirteen LLM Judges: three GPT-family models (GPT-4o (OpenAI, 2024), GPT-4.1 (OpenAI, 2025a), and GPT-5 (OpenAI, 2025b)), three Llama-family models (Dubey et al., 2024) (Llama-3.1-8B/70B, and Llama-3.3-70B), and seven Qwen-family models (Qwen-2.5-7B/32B/72B (Hui et al., 2024), Qwen-3-4B/30B/4B-think/30B-think (Yang et al., 2025)).

More details of the experimental setup are provided in Appendix A and Appendix B.

4 Are LLM-Judges Robust to Knowledge Conflicts Induced by Swapped Reference?

We begin with a controlled evaluation in which each LLM judge first grades candidate answers (c_i^o, c_i^s) under the original reference set, yielding accuracy ACC^o, and then re-grades the same candidates under a swapped reference set, yielding accuracy ACC^s. This design directly tests whether LLM-judge behavior changes when the provided reference is swapped. We evaluate thirteen LLM

Evaluator	Ref.	NQ-Open		SciQ		PopQA		FreshQA	
		TP	TC	TP	TC	TP	TC	TP	TC
Llama-3.3-70B	Orig.	97.6	87.0	96.0	94.6	99.2	90.4	96.5	85.4
	Swap	64.2 (-33.4)	29.0 (-58.1)	36.2 (-59.8)	33.0 (-61.6)	85.7 (-13.5)	48.8 (-41.7)	70.3 (-26.2)	36.5 (-48.9)
Qwen-2.5-32B	Orig.	96.7	88.3	95.0	94.3	98.6	91.6	96.2	88.6
	Swap	66.4 (-30.4)	26.1 (-62.2)	45.7 (-49.3)	40.6 (-53.7)	87.5 (-11.2)	45.5 (-46.2)	72.2 (-24.1)	32.4 (-56.2)
Qwen-2.5-72B	Orig.	93.9	88.6	92.7	93.1	97.3	90.9	93.4	87.5
	Swap	69.8 (-24.2)	31.2 (-57.5)	35.1 (-57.6)	34.2 (-58.9)	90.0 (-7.3)	47.9 (-43.0)	77.2 (-16.2)	37.7 (-49.8)
Qwen-3-30B	Orig.	98.5	94.6	98.6	97.5	99.5	97.7	98.1	94.8
	Swap	69.7 (-28.9)	40.6 (-54.1)	49.3 (-49.3)	44.4 (-53.1)	86.5 (-13.0)	54.4 (-43.3)	69.7 (-28.4)	47.8 (-47.0)
Qwen-3-30B-think	Orig.	98.7	92.7	99.0	98.4	99.7	95.5	97.7	90.4
	Swap	91.5 (-7.2)	66.1 (-26.6)	85.2 (-13.8)	81.8 (-16.6)	97.8 (-1.9)	81.1 (-14.4)	93.9 (-3.8)	71.5 (-18.9)
GPT-4o	Orig.	96.5	89.0	98.5	97.4	98.0	90.9	97.1	87.8
	Swap	59.3 (-37.3)	40.0 (-49.0)	49.8 (-48.7)	45.3 (-52.2)	70.9 (-27.1)	48.8 (-42.1)	64.3 (-32.8)	43.9 (-43.9)
GPT-5	Orig.	98.4	89.3	97.6	96.3	99.7	94.2	97.6	87.0
	Swap	95.7 (-2.8)	77.2 (-12.2)	95.3 (-2.3)	91.7 (-4.6)	99.3 (-0.4)	92.4 (-1.8)	95.1 (-2.4)	80.6 (-6.3)

Table 1: Accuracy (%) of different evaluator models under Type-Preserving (TP)- and Type-Changing (TC)-Swap settings across datasets. Numbers in parentheses indicate the accuracy change from the original-reference setting to the swapped-reference setting; red denotes a decrease in accuracy.

judges across four QA datasets, considering both *Type-Preserving Swap* and *Type-Changing Swap* settings (Section 3.1).

4.1 LLM Judges Are Vulnerable to Knowledge Conflicts Induced by Swapped References

As shown in Table 1, all judges exhibit substantial accuracy drops under swapped references, revealing a consistent vulnerability across models, datasets, and swap types. Notably, even strong models such as GPT-4o and Qwen-3-30B remain vulnerable across datasets and swap configurations. Overall, these results suggest that current LLM judges do not reliably condition their verdicts on the provided reference answer when that reference is swapped.

To further characterize this failure mode, we break down accuracy by each reference–candidate pairing in the triplet (q_i, r_i^a, c_i^b), with $a, b \in \{o, s\}$ in Table 7. Accuracy drop is particularly severe for (q_i, r_i^s, c_i^s), when both the reference and the candidate answer are swapped. This pattern indicates that LLM judges often reject a candidate answer even when it matches the provided reference.

4.2 The Degree of Vulnerability Varies Across Swap Types and Datasets

Next, we examine how vulnerability varies across swap types and datasets. The left panel of Figure 3 shows that RPAG is generally larger under *Type-Changing Swap* than under *Type-Preserving Swap*, indicating that the accuracy drop from the original-reference setting to the swapped-reference setting is more pronounced when the swap changes the answer type. One plausible explanation is that *Type-Changing Swap* can produce a more incongruent evaluation context (e.g., a question asking *who* paired with a reference and candidate answering *where*), which may trigger stronger model intervention and increase the likelihood that the judge disregards the provided reference.

As shown in Table 1, the degree of accuracy degradation differs considerably across datasets. In particular, PopQA exhibits consistently smaller accuracy drops, whereas SciQ shows markedly larger degradations for most models. This discrepancy can be attributed to differences in question characteristics. PopQA largely consists of "who" questions, which tend to preserve semantic coherence even after reference swapping, thereby partially masking the underlying vulnerability. In contrast, SciQ primarily probes precise scientific knowledge, and swapping references in this setting often results in semantically awkward or implausible evaluation contexts, increasing the likelihood that evaluators disregard the provided reference.

4.3 Vulnerability Persists with Increased Model Capacity

Finally, we test whether scaling model capacity improves robustness to swapped references. Intuitively, larger models might better follow instructions or more effectively separate reference-

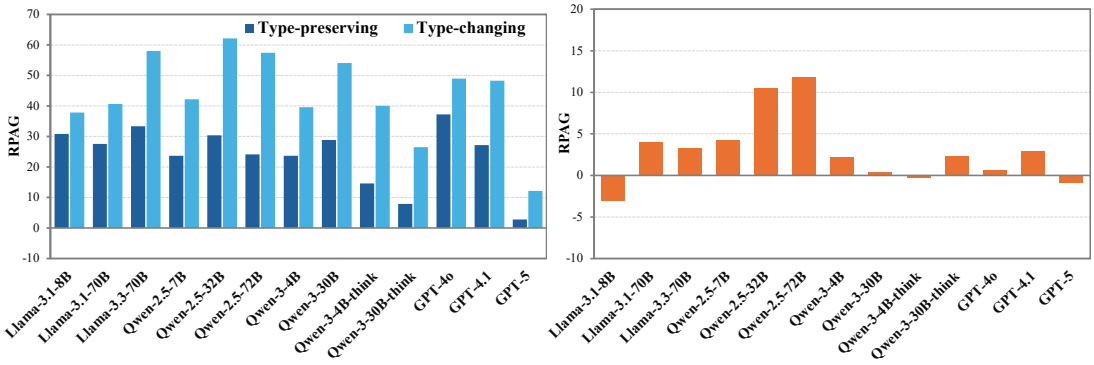

Figure 3: Left: RPAG under Type-Preserving (TP) and Type-Changing (TC) swap settings. Right: RPAG under the Evaluator-Knowledge swap setting. Results are on NQ-Open with thirteen LLM judges.

conditioned evaluation from prior knowledge.

However, our results show that increasing model size does not mitigate this vulnerability. As shown in Figure 3, across a range of parameter scales, RPAG does not systematically diminish with scale. In some cases, larger models even exhibit greater vulnerability (e.g., Qwen-2.5-32B and Qwen-3-30B show higher RPAG than Qwen-2.5-7B and Qwen-3-4B, respectively), suggesting that increased capacity may strengthen reliance on internal world knowledge rather than improve adherence to the provided reference. Taken together, these findings indicate that reference-disregarding behavior is not merely a consequence of limited model capacity, but instead reflects a more fundamental limitation of current LLM judges.

5 Why Do LLM-Judges Ignore Swapped Reference Answers?

In this section, we investigate the mechanisms underlying the vulnerability of LLM judges observed in Section 4. We first test whether this failure is associated with judges relying on parametric knowledge rather than adhering to the provided reference (§ 5.1). We then examine whether the popularity of the knowledge associated the swapped reference correlates with the severity of the vulnerability (§ 5.2). Finally, we analyze how question characteristics relate to vulnerability, with a particular focus on knowledge freshness of question (§ 5.3).

5.1 Vulnerability Diminishes When the Swapped Reference Aligns with Parametric Knowledge

We begin by testing whether the observed vulnerability reflects a systematic preference for parametric knowledge over the provided reference. Under the *Evaluator-Knowledge Swap* setting (Section 3.1), we replace the reference answer with an alternative that matches the judge's internal beliefs in instances where those beliefs conflict with the original reference. As shown in Figure 3, when the reference is swapped to align with the evaluator's parametric knowledge, RPAG becomes nearly zero: swapped-reference accuracy (ACC^s) becomes nearly identical to original-reference accuracy (ACC^o).

This pattern suggests that LLM judges can execute the evaluation procedure reliably when the reference does not contradict their internal knowledge. Therefore, these failures are less likely due to instruction- or format-following limitations, and instead reflect a conflict-resolution tendency where the judge prioritizes its parametric knowledge when the provided reference conflicts with it.

5.2 Swapped-Reference Popularity Correlates with Vulnerability

Next, we analyze how the popularity of knowledge associated with the swapped reference entity affects vulnerability. Following Section 3.1, we evaluate Popularity-high versus Popularity-low swaps on NQ-Open and PopQA.[1]

As shown in Figure 4, vulnerability is substantially more severe under Popularity-high swaps than under Popularity-low swaps across all evaluators. A plausible explanation is that Popularity-high swaps introduce widely known entities for which the evaluator has strong associations, increasing the likelihood of a direct reference–belief conflict. In such cases, the judge appears more likely to discount the provided reference and revert to its parametric knowledge. Conversely, when the refer-

[1]Popularity swaps are only applicable to PERSON entities. We therefore report results on NQ-Open and PopQA, which contain sufficient numbers of PERSON instances (359 and 585, respectively).

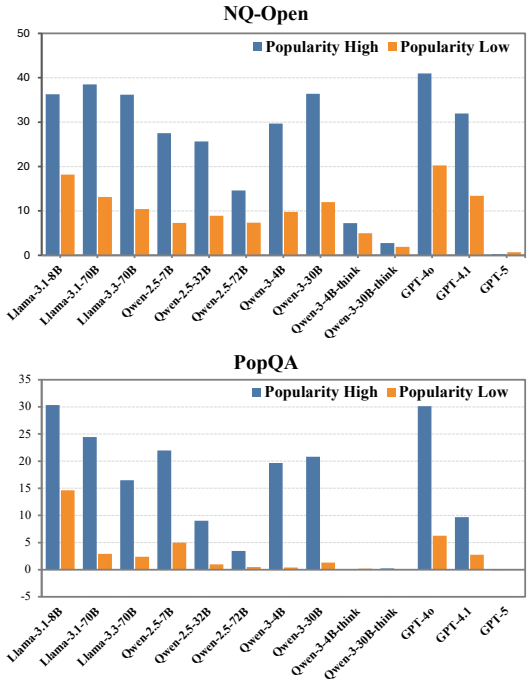

Figure 4: RPAG under Popularity-High and Popularity-Low swap settings on the NQ-Open and PopQA datasets, evaluated with thirteen LLM judges.

ence is swapped to a less popular entity, the evaluator may have weaker or more uncertain priors, reducing the probability of a conflict and thereby attenuating the vulnerability. These observations align with § 5.1 and suggest that the failure severity increases with the level of parametric knowledge tied to the swapped answer.

5.3 Knowledge Freshness Correlates with Vulnerability

Finally, we examine whether vulnerability varies with knowledge freshness. FreshQA (Vu et al., 2024) provides question-level labels based on how likely the underlying fact is to change over time: *Never-changing*, *Slow-changing*, and *Fast-changing*. Intuitively, faster-changing facts are less likely to be strongly encoded in an LLM judge's parametric knowledge.

Using the freshness labels provided by FreshQA, we stratify the average RPAG scores for Type-Preserving and Type-Changing swaps by freshness level. As shown in Figure 5, questions involving rapidly changing or time-sensitive knowledge shows low RPAG, indicating lower vulnerability under swapped references. One interpretation is that, for such questions, LLM judges have weaker or lower-confidence parametric knowledge, making them more willing to defer to the provided

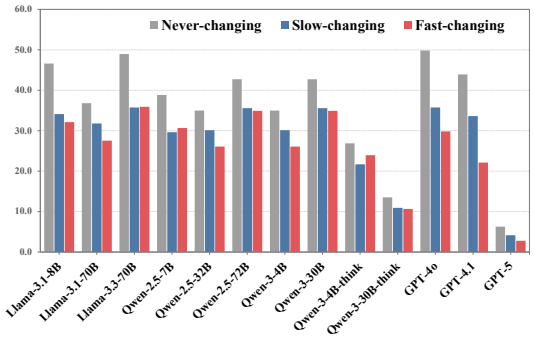

Figure 5: RPAG under Type-Preserving and Type-Changing swap settings across question freshness types (never-, slow-, and fast-changing), evaluated on FreshQA with thirteen LLM judges.

reference. In contrast, questions about stable, long-standing facts tend to exhibit higher vulnerability, consistent with stronger parametric beliefs that are more resistant to being overridden by the reference.

Taken together, these results indicate that susceptibility to swapped references is shaped by the interaction between the level of the judge's parametric knowledge and the extent to which the provided reference is treated as reliable under knowledge conflict.

6 Can Prompting Mitigate the Vulnerability?

Prior work suggests that prompting strategies, including explicit evaluation principles (Zhu et al., 2023; Hwang et al., 2025) and CoT prompting with self-consistency (Chen et al., 2024b; Kamalloo et al., 2024), can improve evaluator robustness. We examine whether these strategies mitigate LLM-judge vulnerability under knowledge conflict by evaluating four approaches: *Direct*, *CoT* (Wei et al., 2022), *Self-consistency*, and *CoT+Self-consistency*. All experiments use GPT-4o under both Type-Preserving and Type-Changing swap settings across four datasets, with prompt details provided in Appendix B.

As shown in Table 2, none of the tested strategies fully eliminates the vulnerability. Among them, *Direct* reduces (but does not remove) the vulnerability relative to the standard prompting baseline, suggesting that stronger instructions can partially improve reference adherence but remain insufficient under conflict. Interestingly, *CoT* often *worsens* performance compared to standard prompting. As analyzed in Appendix C.1, eliciting explicit reasoning tends to amplify reliance on the judge's internal beliefs, increasing the likelihood that the

Prompting	NQ-Open		SciQ		PopQA		FreshQA	
	TP	TC	TP	TC	TP	TC	TP	TC
Standard	37.3	49.0	48.7	52.2	27.1	42.1	32.8	43.9
Direct	22.5	36.6	34.1	36.5	18.5	33.0	21.5	33.0
CoT	40.3	66.8	66.2	67.9	15.7	63.3	35.0	59.4
Self-Consist.	38.6	49.8	47.2	49.0	30.5	48.5	33.6	47.0
CoT+Self-Consist	42.7	83.8	65.4	70.8	16.7	73.8	39.6	79.2

Table 2: RPAG under Type-Preserving (TP) and Type-Changing (TC)-swap settings across datasets for different prompting strategies using GPT-4o.

model overrides the provided reference in favor of its parametric knowledge. This behavior contrasts with prior findings that CoT, particularly with self-consistency, improves reliability in standard QA evaluation settings (Kamalloo et al., 2024). Overall, these results indicate that prompting strategies effective under non-conflicting references settings do not generalize to knowledge-conflict settings.

7 Related Works

7.1 LLM-as-a-Judge

LLM-as-a-judge has emerged as a scalable paradigm for evaluating model outputs and often correlates well with human judgments (Zheng et al., 2023; Wang et al., 2023b; Liu et al., 2023b; Chiang and Lee, 2023; Gu et al., 2025; Thakur et al., 2025). In QA evaluation, traditional metrics such as EM, F1, and BLEU (Papineni et al., 2002) are increasingly misaligned with long-form, semantically diverse answers produced by modern LLMs (Kamalloo et al., 2023; Wang et al., 2023a; Lee et al., 2025b), motivating broad adoption of LLM-judges for reference-conditioned evaluation (Zhang et al., 2024; Blandón et al., 2025; Hosseini et al., 2024; Ho et al., 2025; Chen et al., 2024b).

At the same time, prior work documents systematic limitations of LLM-based evaluation (Li et al., 2025), including biases toward longer outputs (Koo et al., 2024; Dubois et al., 2024), well-formatted responses (Chen et al., 2024a; Stephan et al., 2025), and answers without epistemic markers (Lee et al., 2025a). Another recurring concern is self- or model-family preference, where judges favor their own generations or closely related models (Liu et al., 2024; Chen et al., 2025). Collectively, these findings have motivated diverse meta-evaluations of LLM-judges that systematically probe the conditions under which they deviate from their intended grading criteria (Tan et al., 2024b; Zheng et al., 2023).

7.2 Knowledge Conflict in LLMs

Knowledge conflict arises because LLMs increasingly combine parametric knowledge stored in model weights with contextual knowledge supplied at inference time (Pan et al., 2022), including user instructions (Liu et al., 2023a), retrieved documents (Shi et al., 2024), and tool outputs (Schick et al., 2023; Xu et al., 2024; Zhuang et al., 2023). When these sources disagree, models must implicitly resolve a conflict, and the chosen resolution strategy can substantially affect reliability. This issue is widely studied in QA, especially in retrieval-augmented generation (RAG), where models may under- or over-incorporate retrieved evidence (Xie et al., 2023; Jin et al., 2024; Tan et al., 2024a).

Several works intentionally introduce contradictions to characterize such behavior: Longpre et al. (2021) and Chen et al. (2022) show that models differ in their tendency to rely on parametric versus contextual information, while IfQA (Yu et al., 2023) and DisentQA (Neeman et al., 2023) use counterfactual context augmentation to test whether models incorporate edited evidence. Related analyses extend to multimodal settings by inducing conflicts between images and parametric knowledge (Bitton-Guetta et al., 2023; Zhou et al., 2023; Guan et al., 2024; Luo et al., 2024; Lee et al., 2025c). To the best of our knowledge, however, prior work has not systematically studied *LLM-judges* under knowledge conflict in *QA evaluation*, where judges must reconcile their own beliefs with an explicitly provided reference when grading candidate answers.

8 Conclusion

We show that LLM judges often fail to adhere to reference-conditioned QA evaluation when the provided reference conflicts with their parametric knowledge. Using our swapped-reference QA evaluation framework across four datasets and multiple judge models, we find that even strong evaluators frequently fail on otherwise trivial cases, and this vulnerability persists across swap types, swap targets, and model scale. Our analyses indicate that this vulnerability arises from judges' over-reliance on internal beliefs, and that common prompting strategies provide only limited mitigation. These findings highlight a fundamental risk in LLM-as-a-judge evaluation under knowledge conflict and motivate the development of protocols that enforce stronger reference adherence.

Limitations

This work focuses exclusively on question answering (QA) evaluation. While QA is one of the most common and impactful applications of the LLM-as-a-judge paradigm, our findings do not directly establish that the same reference-adherence failures arise in other reference-conditioned evaluation settings, such as summarization, fact verification, or dialog evaluation. Extending the swapped-reference analysis to these broader tasks remains an important direction for future work.

Second, our study does not propose a concrete method to mitigate the observed vulnerability. Instead, our goal is diagnostic: to isolate, characterize, and analyze a fundamental failure mode of LLM judges under reference–knowledge conflict. Designing effective mitigation strategies that reliably enforce reference adherence is a non-trivial challenge and lies beyond the scope of this work.

Finally, our evaluation relies on artificially swapped reference answers, which can produce question–reference pairs that appear counterintuitive or implausible from a real-world perspective. However, this design choice is intentional and necessary to create controlled conflicts between the provided reference and the judge's parametric knowledge. Crucially, the intended role of an LLM judge in reference-based QA evaluation is to follow explicit evaluation instructions and assess candidate answers with respect to the provided reference, regardless of its alignment with prior beliefs. Therefore, failures under swapped references highlight a limitation of current LLM judges, rather than an artifact of unrealistic evaluation settings.

Ethical Considerations

In our experiments, we utilize publicly available datasets, including NQ-Open (Lee et al., 2019), SciQ (Welbl et al., 2017), PopQA (Mallen et al., 2023), and FreshQA (Vu et al., 2024). These datasets are widely adopted and well-established within the research community, and their use raises no additional privacy or consent concerns beyond those addressed in the original dataset releases.

All large language models used in this work were accessed through their official and publicly available sources. Specifically, GPT-family models were accessed via OpenAI's official platform[2], while LLaMA-family and Qwen-family models were obtained from their respective official releases under the corresponding usage licenses. Our use of these models complies with their stated terms of service and aligns with open science and reproducibility principles.

During the preparation of this manuscript, we employed an AI-assisted writing tool at the sentence level to support drafting and linguistic refinement. The AI assistant did not generate experimental results, design methodologies, or draw scientific conclusions, and all technical content and interpretations remain the responsibility of the authors.

[2] https://openai.com

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.

Nitzan Bitton-Guetta, Yonatan Bitton, Jack Hessel, Ludwig Schmidt, Yuval Elovici, Gabriel Stanovsky, and Roy Schwartz. 2023. Breaking common sense: Whoops! a vision-and-language benchmark of synthetic and compositional images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2616–2627.

María Andrea Cruz Blandón, Jayasimha Talur, Bruno Charron, Dong Liu, Saab Mansour, and Marcello Federico. 2025. Memerag: A multilingual end-to-end meta-evaluation benchmark for retrieval augmented generation. *arXiv preprint arXiv:2502.17163*.

Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. 2024a. Humans or llms as the judge? a study on judgement biases. *arXiv preprint arXiv:2402.10669*.

Hung-Ting Chen, Michael Zhang, and Eunsol Choi. 2022. Rich knowledge sources bring complex knowledge conflicts: Recalibrating models to reflect conflicting evidence. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2292–2307.

Lei Chen, Bobo Li, Li Zheng, Haining Wang, Zixiang Meng, Runfeng Shi, Hao Fei, Jun Zhou, Fei Li, Chong Teng, and 1 others. 2024b. What factors influence llms' judgments? a case study on question answering. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17473–17485.

Wenhu Chen, Xinyi Wang, and William Yang Wang. 2021. A dataset for answering time-sensitive questions. *arXiv preprint arXiv:2108.06314*.

Zhi-Yuan Chen, Hao Wang, Xinyu Zhang, Enrui Hu, and Yankai Lin. 2025. Beyond the surface: Measuring self-preference in llm judgments. *arXiv preprint arXiv:2506.02592*.

Cheng-Han Chiang and Hung-yi Lee. 2023. A closer look into automatic evaluation using large language models. *arXiv preprint arXiv:2310.05657*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B Hashimoto. 2024. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. A survey on llm-as-a-judge. *Preprint*, arXiv:2411.15594.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, and 1 others. 2024. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14375–14385.

Xanh Ho, Jiahao Huang, Florian Boudin, and Akiko Aizawa. 2025. Llm-as-a-judge: Reassessing the performance of llms in extractive qa. *arXiv preprint arXiv:2504.11972*.

Pedram Hosseini, Jessica M Sin, Bing Ren, Bryceton G Thomas, Elnaz Nouri, Ali Farahanchi, and Saeed Hassanpour. 2024. A benchmark for long-form medical question answering. *arXiv preprint arXiv:2411.09834*.

Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, and 1 others. 2024. Qwen2. 5-coder technical report. *arXiv preprint arXiv:2409.12186*.

Yerin Hwang, Dongryeol Lee, Taegwan Kang, Yongil Kim, and Kyomin Jung. 2025. Can you trick the grader? adversarial persuasion of llm judges. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 14632–14651.

Zhuoran Jin, Pengfei Cao, Yubo Chen, Kang Liu, Xiaojian Jiang, Jiexin Xu, Li Qiuxia, and Jun Zhao. 2024. Tug-of-war between knowledge: Exploring and resolving knowledge conflicts in retrieval-augmented language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16867–16878.

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.

Ehsan Kamalloo, Shivani Upadhyay, and Jimmy Lin. 2024. Towards robust qa evaluation via open llms. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2811–2816.

Jungo Kasai, Keisuke Sakaguchi, Ronan Le Bras, Akari Asai, Xinyan Yu, Dragomir Radev, Noah A Smith, Yejin Choi, Kentaro Inui, and 1 others. 2023. Realtime qa: What's the answer right now? *Advances in neural information processing systems*, 36:49025–49043.

Ryan Koo, Minhwa Lee, Vipul Raheja, Jong Inn Park, Zae Myung Kim, and Dongyeop Kang. 2024. Benchmarking cognitive biases in large language models as evaluators. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 517–545.

Dongryeol Lee, Yerin Hwang, Yongil Kim, Joonsuk Park, and Kyomin Jung. 2025a. Are llm-judges robust to expressions of uncertainty? investigating the effect of epistemic markers on llm-based evaluation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8962–8984.

Dongryeol Lee, Minwoo Lee, Kyungmin Min, Joonsuk Park, and Kyomin Jung. 2025b. Return of em: Entity-driven answer set expansion for qa evaluation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 11218–11234.

Kang-il Lee, Minbeom Kim, Seunghyun Yoon, Minsung Kim, Dongryeol Lee, Hyukhun Koh, and Kyomin Jung. 2025c. Vlind-bench: Measuring language priors in large vision-language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4129–4144.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. *arXiv preprint arXiv:1906.00300*.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, and 1 others. 2025. From generation to judgment: Opportunities and challenges of llm-as-a-judge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2757–2791.

Adam Liska, Tomas Kocisky, Elena Gribovskaya, Tayfun Terzi, Eren Sezener, Devang Agrawal, Cyprien

De Masson D'Autume, Tim Scholtes, Manzil Zaheer, Susannah Young, and 1 others. 2022. Streamingqa: A benchmark for adaptation to new knowledge over time in question answering models. In *International Conference on Machine Learning*, pages 13604–13622. PMLR.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023a. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM computing surveys*, 55(9):1–35.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023b. G-eval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

Yiqi Liu, Nafise Sadat Moosavi, and Chenghua Lin. 2024. Llms as narcissistic evaluators: When ego inflates evaluation scores. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12688–12701.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. Entity-based knowledge conflicts in question answering. *arXiv preprint arXiv:2109.05052*.

Tiange Luo, Ang Cao, Gunhee Lee, Justin Johnson, and Honglak Lee. 2024. Probing visual language priors in vlms. *arXiv preprint arXiv:2501.00569*.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822.

Ella Neeman, Roee Aharoni, Or Honovich, Leshem Choshen, Idan Szpektor, and Omri Abend. 2023. Disentqa: Disentangling parametric and contextual knowledge with counterfactual question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10056–10070.

OpenAI. 2024. Hello gpt-4o.

OpenAI. 2025a. Introducing gpt-4.1 in the api.

OpenAI. 2025b. Introducing gpt-5.

Xiaoman Pan, Wenlin Yao, Hongming Zhang, Dian Yu, Dong Yu, and Jianshu Chen. 2022. Knowledge-in-context: Towards knowledgeable semi-parametric language models. *arXiv preprint arXiv:2210.16433*.

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.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36:68539–68551.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Richard James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2024. Replug: Retrieval-augmented black-box language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8371–8384.

Andreas Stephan, Dawei Zhu, Matthias Aßenmacher, Xiaoyu Shen, and Benjamin Roth. 2025. From calculation to adjudication: Examining llm judges on mathematical reasoning tasks. In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM^2)*, pages 759–773.

Hexiang Tan, Fei Sun, Wanli Yang, Yuanzhuo Wang, Qi Cao, and Xueqi Cheng. 2024a. Blinded by generated contexts: How language models merge generated and retrieved contexts when knowledge conflicts? *arXiv preprint arXiv:2401.11911*.

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

Aman Singh Thakur, Kartik Choudhary, Venkat Srinik Ramayapally, Sankaran Vaidyanathan, and Dieuwke Hupkes. 2025. Judging the judges: Evaluating alignment and vulnerabilities in llms-as-judges. In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM^2)*, pages 404–430.

Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, and 1 others. 2024. Freshllms: Refreshing large language models with search engine augmentation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13697–13720.

Cunxiang Wang, Sirui Cheng, Qipeng Guo, Yuanhao Yue, Bowen Ding, Zhikun Xu, Yidong Wang, Xiangkun Hu, Zheng Zhang, and Yue Zhang. 2023a. Evaluating open-qa evaluation. *Advances in Neural Information Processing Systems*, 36:77013–77042.

Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023b. Is chatgpt a good nlg evaluator? a preliminary study. *arXiv preprint arXiv:2303.04048*.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and 1 others. 2024. Large language models are not fair evaluators. In *Proceedings of the*

62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 9440–9450.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Johannes Welbl, Nelson F Liu, and Matt Gardner. 2017. Crowdsourcing multiple choice science questions. *arXiv preprint arXiv:1707.06209*.

Zhaofeng Wu, Linlu Qiu, Alexis Ross, Ekin Akyürek, Boyuan Chen, Bailin Wang, Najoung Kim, Jacob Andreas, and Yoon Kim. 2024. Reasoning or reciting? exploring the capabilities and limitations of language models through counterfactual tasks. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1819–1862.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2023. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *The Twelfth International Conference on Learning Representations*.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024. Knowledge conflicts for llms: A survey. *arXiv preprint arXiv:2403.08319*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Wenhao Yu, Meng Jiang, Peter Clark, and Ashish Sabharwal. 2023. Ifqa: A dataset for open-domain question answering under counterfactual presuppositions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8276–8288.

Hengran Zhang, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Yixing Fan, and Xueqi Cheng. 2024. Are large language models good at utility judgments? In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1941–1951.

Michael Zhang and Eunsol Choi. 2021. Situatedqa: Incorporating extra-linguistic contexts into qa. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 7371–7387.

Zihan Zhang, Meng Fang, Ling Chen, Mohammad-Reza Namazi-Rad, and Jun Wang. 2023. How do large language models capture the ever-changing world knowledge? a review of recent advances. *arXiv preprint arXiv:2310.07343*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

Kankan Zhou, Eason Lai, Wei Bin Au Yeong, Kyriakos Mouratidis, and Jing Jiang. 2023. Rome: Evaluating pre-trained vision-language models on reasoning beyond visual common sense. *arXiv preprint arXiv:2310.19301*.

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

Yuchen Zhuang, Yue Yu, Kuan Wang, Haotian Sun, and Chao Zhang. 2023. Toolqa: A dataset for llm question answering with external tools. *Advances in Neural Information Processing Systems*, 36:50117–50143.

A Details of Dataset

We sampled datasets from NQ-Open (Lee et al., 2019), PopQA (Mallen et al., 2023), Sciq (Welbl et al., 2017), and FreshQA (Vu et al., 2024). For NQ-Open, PopQA, and SciQ, we randomly sample 1,000 instances from the original dataset.[3] For FreshQA, we use the dev and test splits of the *August 18, 2025* version and exclude questions with false premises, yielding 452 instances.

Throughout dataset construction (e.g., NER, swapping, and candidate generation), we used GPT-4o (`gpt-4o-2024-08-06`)[4] with temperature 0 to perform NER and to generate long-form candidates, as described in Section 3.1. This design helps mitigate *egocentric bias*, where LLM judges tend to prefer their own generations (Liu et al., 2024; Chen et al., 2025). The prompts used for NER and for generating the original and swapped long-form candidates are shown in Figures 6, 7, and 8, respectively.

Table 3 reports dataset statistics. Throughout dataset construction, the authors manually inspected entity extraction, swap operations, and generated candidates to ensure data quality and reliability.

B Experimental Details

B.1 Judge Models

As described in Section 3.4, we evaluated thirteen LLM judges. For GPT-family models, we used the official OpenAI API with the following versions: GPT-4o (`gpt-4o-2024-08-06`)[5], GPT-4.1 (`gpt-4.1-2025-04-14`)[6], and GPT-5 (`gpt-5-2025-08-07`)[7]. For Llama-family and Qwen-family models, we used the official Hugging Face repositories: Llama-3.1-8B[8], Llama-3.1-70B[9], and Llama-3.3-70B[10]; Qwen-2.5-7B[11], Qwen-2.5-32B[12], and Qwen-2.5-72B[13]; and Qwen-3-4B[14], Qwen-3-30B[15], Qwen-3-4B-Think[16], and Qwen-3-30B-Think[17].

B.2 Hyperparameters

Across all experiments, we set the decoding temperature to 0, following common practice in LLM-as-a-judge evaluation settings (Wang et al., 2024, 2023a; Lee et al., 2025a). Unless otherwise specified, we used greedy decoding with fixed max generation length and the same prompt template across judges.

B.3 Prompts

For QA evaluation, we adopt the SimpleQA evaluation template introduced by OpenAI,[18] as shown in Table 9.

For the experiments in Section 6, we evaluate the following four prompting strategies: (1) *Direct*, which explicitly instructs the judge to base its verdict solely on the provided reference; (2) *CoT*, which requires the judge to produce a reasoning trace before outputting a final verdict (Wei et al., 2022); (3) *Self-consistency*, which samples five independent verdicts at temperature 0.6 and returns the majority vote; and (4) *CoT+Self-consistency*, which applies majority voting over CoT-based samples. The prompts for *Direct* and *CoT* are shown in Tables 10 and 11, respectively.

B.4 Computing Resources

For the experiments, we utilize two 8 NVIDIA A100 Tensor Core GPUs (each with 80GB of memory).

[3] For NQ-Open, we randomly sample 1,000 instances from the validation set; for PopQA, we randomly sample 1,000 instances from the test set; for SciQ, we use the 1,000-question validation set.

[4] https://platform.openai.com/docs/models/gpt-4o

[5] https://platform.openai.com/docs/models/gpt-4o

[6] https://platform.openai.com/docs/models/gpt-4.1

[7] https://platform.openai.com/docs/models/gpt-5

[8] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

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

[10] https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct

[11] https://huggingface.co/Qwen/Qwen2.5-7B-Instruct

[12] https://huggingface.co/Qwen/Qwen2.5-32B-Instruct

[13] https://huggingface.co/Qwen/Qwen2.5-72B-Instruct

[14] https://huggingface.co/Qwen/Qwen3-4B-Instruct-2507

[15] https://huggingface.co/Qwen/Qwen3-30B-A3B-Instruct-2507

[16] https://huggingface.co/Qwen/Qwen3-4B-Thinking-2507

[17] https://huggingface.co/Qwen/Qwen3-30B-A3B-Thinking-2507

[18] https://openai.com/index/introducing-simpleqa/

C Additional Analysis

C.1 Manual Inspection of Reasoning Paths

Table 2 shows that Chain-of-Thought (CoT) prompting yields even higher RPAG than standard prompting, indicating worse robustness under swapped references. To better understand this behavior, we manually inspected the reasoning paths of 50 examples for each of the Type-Preserving and Type-Changing swap settings. We focus on instances where the judge produced a correct verdict under standard prompting but switched to an incorrect verdict when CoT prompting was applied.

As shown in Table 5, all inspected reasoning paths exhibit explicit evidence of reference override driven by the model's parametric knowledge. Despite being provided with the reference answer, the judges frequently justify their decisions by appealing to world knowledge that contradicts the reference, ultimately producing incorrect verdicts. These observations provide qualitative evidence supporting our quantitative findings, demonstrating that under knowledge conflict, LLM judges systematically prioritize parametric knowledge over the provided reference, even when encouraged to reason explicitly.

C.2 Full Results Table

We report the detailed accuracy across different reference–candidate pairing: (r^o, r^o), (r^o, r^s), (r^s, r^o), and (r^s, r^s) in Table 7.

Swap Type	NQ-Open	SciQ	PopQA	FreshQA
Type-Preserving	1000	996	1000	451
Type-Changing	1000	1000	1000	452
Popularity	359	4	585	103
Evaluator Knowledge				
Llama-3.1-8B	453	95	390	241
Llama-3.1-70B	118	20	161	82
Llama-3.3-70B	207	38	291	151
Qwen-2.5-7B	169	62	149	108
Qwen-2.5-32B	296	69	315	140
Qwen-2.5-72B	310	59	314	114
Qwen-3-4B	412	75	438	134
Qwen-3-30B	336	52	373	134
Qwen3-4B-think	411	97	346	211
Qwen3-30B-think	443	110	409	185
GPT-4o	327	96	355	200
GPT-4.1	296	86	227	157
GPT-5	351	100	204	115

Table 3: Dataset statistics used throughout the experiments. The number under Evaluator Knowledge indicates the number of instances in which the evaluator's own answer disagrees with the original reference answer.

Swap Type	Question	Reference	Candidate Answer
Type-Preserving	How many UC schools are there in the United States?	r^o: 10 r^s: 126	c^o: There are 10 University of California (UC) campuses throughout the United States. c^s: There are 126 UC schools located throughout the United States.
Type-Changing	Iceland is made up of a series of	r^o: volcanoes r^s: islands	c^o: Iceland is composed of a series of volcanoes. c^s: Iceland is made up of a series of islands.
Popularity-High	Who played the Elephant Man in the film?	r^o: John Hurt r^s: Dwayne Johnson	c^o: The Elephant Man was portrayed by John Hurt. c^s: Dwayne Johnson played the Elephant Man in the film.
Popularity-Low	Who is the father of Ramkarpal Singh?	r^o: Karpal Singh r^s: Jeong Ji-u	c^o: Ramkarpal Singh's father is Karpal Singh. c^s: The father of Ramkarpal Singh is Jeong Ji-u.
Evaluator (GPT-4o) Knowledge	Word that means separation of church and state	r^o: separationism r^s: secularism	c^o: The term refers to separation of church and state. c^s: The word is secularism.

Table 4: Examples of swapped-reference QA instances across different swap types.

Type-Preserving Error Example	**Type-Changing Error Example**
Question: who is mowgli's main enemy in the jungle book **Reference:** `[Shere Khan]` **Candidate:** Mowgli's main enemy in *The Jungle Book* is Bing Russell. **Reasoning:** • The question asks for Mowgli's main enemy in *The Jungle Book*. • The reference entity is incorrect. • The candidate answer repeats the incorrect entity. • The correct answer should be Shere Khan. **Final:** Incorrect	**Question:** who is mowgli's main enemy in the jungle book **Reference:** `[to preserve slavery]` **Candidate:** Mowgli's main enemy in *The Jungle Book* is the concept of preserving slavery. **Reasoning:** • The question concerns a fictional character in *The Jungle Book*. • The reference does not align with the story context. • The candidate introduces an abstract concept unrelated to the narrative. • No correct information about the story is provided. **Final:** Incorrect

Table 5: An example of a Chain-of-Thought (CoT) reasoning path that leads to an incorrect evaluation under the (r^s, c^s) setting.

		NQ-Open		SciQ		PopQA		FreshQA	
Evaluator	Ref.	TP	TC	TP	TC	TP	TC	TP	TC
GPT-4.1	Orig.	98.3	91.2	97.8	95.8	99.2	92.9	98.2	89.7
	Swap	71.1 (-27.2)	42.9 (-48.3)	56.7 (-41.1)	49.5 (-46.3)	86.2 (-13.0)	51.6 (-41.3)	73.4 (-24.8)	48.3 (-41.4)
Llama-3.1-8B	Orig.	95.6	95.3	92.2	91.9	97.5	95.3	94.8	95.7
	Swap	64.8 (-30.8)	57.5 (-37.8)	44.2 (-48.0)	45.2 (-46.8)	78.0 (-19.5)	53.2 (-42.2)	67.4 (-27.4)	57.5 (-38.2)
Llama-3.1-70B	Orig.	97.5	89.5	97.6	96.0	99.0	90.9	97.3	87.8
	Swap	69.9 (-27.6)	48.8 (-40.7)	56.5 (-41.1)	50.8 (-45.2)	85.4 (-13.6)	60.8 (-30.2)	73.4 (-23.9)	53.9 (-33.9)
Qwen-2.5-7B	Orig.	98.0	98.1	99.5	99.5	98.6	98.7	97.8	98.1
	Swap	74.3 (-23.7)	56.0 (-42.2)	57.2 (-42.3)	56.9 (-42.6)	83.7 (-15.0)	54.5 (-44.2)	74.5 (-23.3)	57.4 (-40.7)
Qwen-3-4B	Orig.	98.2	97.7	99.1	98.9	98.5	97.5	98.9	99.0
	Swap	74.5 (-23.7)	58.2 (-39.6)	54.7 (-44.3)	54.6 (-44.3)	86.7 (-11.9)	59.8 (-37.7)	75.7 (-23.2)	61.5 (-37.5)
Qwen-3-4B-think	Orig.	98.1	90.6	98.4	97.4	99.3	94.8	97.1	91.5
	Swap	82.7 (-15.4)	50.5 (-40.2)	59.4 (-39.1)	55.7 (-41.7)	95.6 (-3.7)	63.3 (-31.5)	85.1 (-12.0)	53.5 (-37.9)

Table 6: Accuracy (%) of different evaluator models under Type-Preserving (TP)- and Type-Changing (TC)-Swap settings across datasets. Numbers in parentheses indicate the accuracy change from the original-reference setting to the swapped-reference setting; red denotes a decrease in accuracy.

Prompt for Named Entity Recognition

Role Definition (System Prompt):
You are an information extraction model that assigns exactly ONE SpaCy NER type to the given original answer. Return ONLY the label token. No extra text.

User Prompt:
SpaCy NER labels & definitions
PERSON: People, including fictional.
NORP: Nationalities or religious or political groups.
FAC: Buildings, airports, highways, bridges, etc.
ORG: Companies, agencies, institutions, etc.
GPE: Countries, cities, states.
LOC: Non-GPE locations, mountain ranges, bodies of water.
PRODUCT: Objects, vehicles, foods, devices (not services).
EVENT: Named hurricanes, battles, wars, sports events, festivals, etc.
WORK_OF_ART: Titles of books, movies, songs, paintings, etc.
LAW: Named documents made into laws.
LANGUAGE: Any named language.
DATE: Absolute or relative dates or periods.
TIME: Times smaller than a day.
PERCENT: Percentage, including "
MONEY: Monetary values, including unit.
QUANTITY: Measurements of size/weight/distance/volume/speed/etc.
ORDINAL: "first", "second", "23rd", etc.
CARDINAL: Numerals that do not fall under another type.
Rules
1) Classify the answer string AS WRITTEN (no external lookup).
2) If multiple entities appear, label by the main head of the answer.
3) Numeric answers: - With unit → QUANTITY (e.g., 5 km), MONEY (e.g., €10), PERCENT (e.g., 12%), TIME (e.g., 3 hours). - Dates/periods → DATE. - Ordinals → ORDINAL. - Plain counts/integers → CARDINAL.
4) GPE vs LOC: Countries/cities/states → GPE; geographic features → LOC.
5) ORG vs PRODUCT: Organizations → ORG; tangible items → PRODUCT.
6) WORK_OF_ART only for titled creative works.
7) Output exactly one label from the list above. No explanation.
8) If answer do not have an entity just return NAN
Few-shot examples
Q: Who wrote *Pride and Prejudice? Original Answer: Jane Austen Label: PERSON
Q: What is the capital of France? Original Answer: Paris Label: GPE
Q: Which company makes the iPhone? Original Answer: Apple Label: ORG
Q: What language is primarily spoken in Brazil? Original Answer: Portuguese Label: LANGUAGE
Q: When did WW2 end? Original Answer: 1945 Label: DATE
Q: How long is a marathon? Original Answer: 42.195 km Label: QUANTITY
Q: Which event did the Chiefs win in 2024? Original Answer: Super Bowl LVIII Label: EVENT
Q: What is "Mona Lisa"? Original Answer: Mona Lisa Label: WORK_OF_ART
Q: Where is Mount Everest? Original Answer: Himalayas Label: LOC
Q: How much does it cost? Original Answer: $20 Label: MONEY
Q: What is John Mayne's occupation? Original Answer: journalist Label: NAN
Task
Question: {question}
Original Answer: {answer}
Output
Return ONLY the label (one of: PERSON, NORP, FAC, ORG, GPE, LOC, PRODUCT, EVENT, WORK_OF_ART, LAW, LANGUAGE, DATE, TIME, PERCENT, MONEY, QUANTITY, ORDINAL, CARDINAL, None).

Figure 6: Prompt template used for Named Entity Recognition (NER) in Section 3.1.

Evaluator	(ref., cand.)	NQ-Open		SciQ		PopQA		FreshQA	
		TP	TC	TP	TC	TP	TC	TP	TC
Llama-3.1-8B	(r^o, r^o)	99.3	99.3	99.9	99.9	99.3	99.3	98.7	98.7
	(r^o, r^s)	91.9	91.2	84.5	83.9	95.7	91.3	90.9	92.7
	(r^s, r^o)	72.2	84.9	48.0	50.8	81.2	79.2	76.1	84.7
	(r^s, r^s)	57.4	30.0	40.4	39.5	74.8	27.1	58.8	30.3
Llama-3.1-70B	(r^o, r^o)	97.9	97.9	100.0	100.0	99.0	99.0	99.8	99.8
	(r^o, r^s)	97.0	81.0	95.2	91.9	98.9	82.8	94.9	75.9
	(r^s, r^o)	94.5	84.5	87.5	79.4	99.1	98.2	95.8	87.8
	(r^s, r^s)	45.3	13.1	25.6	22.2	71.6	23.3	51.0	20.1
Llama-3.3-70B	(r^o, r^o)	99.4	100.0	99.5	100.0	98.8	100.0	92.9	70.8
	(r^o, r^s)	95.7	74.6	92.0	89.1	98.8	81.3	92.9	70.8
	(r^s, r^o)	74.1	38.6	42.1	36.4	92.0	65.2	83.4	47.1
	(r^s, r^s)	54.3	19.3	30.3	29.6	79.4	32.3	57.2	25.9
Qwen-2.5-7B	(r^o, r^o)	96.5	99.7	97.3	96.9	99.9	100.0	98.7	99.3
	(r^o, r^s)	99.4	99.7	99.2	99.2	99.9	100.0	98.7	99.3
	(r^s, r^o)	98.6	98.7	96.2	96.9	99.8	100.0	98.5	97.8
	(r^s, r^s)	50.0	13.2	18.2	16.9	67.5	9.0	50.6	17.0
Qwen-2.5-32B	(r^o, r^o)	98.9	99.9	98.9	99.8	98.3	84.3	92.7	77.4
	(r^o, r^s)	94.5	77.6	90.1	88.6	98.3	84.3	92.7	77.4
	(r^s, r^o)	74.6	28.2	68.1	59.7	95.1	56.6	85.6	37.6
	(r^s, r^s)	58.1	24.0	23.3	21.4	79.8	34.3	58.8	27.2
Qwen-2.5-72B	(r^o, r^o)	99.7	99.9	99.4	100.0	95.1	82.3	86.7	75.0
	(r^o, r^s)	88.1	77.5	85.5	86.3	95.1	82.3	86.7	75.0
	(r^s, r^o)	64.6	21.7	23.1	20.3	91.8	52.9	75.6	32.5
	(r^s, r^s)	74.9	40.6	47.1	48.1	88.2	42.8	78.7	42.9
Qwen-3-4B	(r^o, r^o)	97.5	97.5	99.5	99.5	97.4	97.4	99.8	99.8
	(r^o, r^s)	98.8	97.9	98.6	98.2	99.6	97.5	98.0	98.2
	(r^s, r^o)	94.9	95.5	91.2	90.5	99.4	99.4	97.6	99.3
	(r^s, r^s)	54.1	20.8	18.3	18.6	73.9	20.2	53.9	23.7
Qwen-3-30B	(r^o, r^o)	98.9	98.9	99.9	99.9	99.4	99.4	99.8	99.8
	(r^o, r^s)	98.1	90.3	97.4	95.0	99.5	95.9	96.5	89.8
	(r^s, r^o)	88.7	65.8	83.9	73.9	97.5	90.3	88.7	76.6
	(r^s, r^s)	50.6	15.3	14.7	14.8	75.4	18.5	50.8	19.0
Qwen-3-4B-think	(r^o, r^o)	99.7	99.7	99.9	99.9	99.7	99.7	100.0	100.0
	(r^o, r^s)	96.4	81.4	96.7	94.5	98.9	89.3	94.7	82.7
	(r^s, r^o)	70.7	12.6	24.0	19.7	93.0	36.8	77.2	19.5
	(r^s, r^s)	96.2	88.4	92.7	91.7	98.8	90.8	96.5	87.8
Qwen-3-30B-think	(r^o, r^o)	99.4	99.4	100.0	100.0	99.8	99.8	100.0	100.0
	(r^o, r^s)	98.1	85.9	98.6	96.9	99.4	91.3	95.6	82.7
	(r^s, r^o)	83.6	39.2	74.6	71.0	98.1	71.2	88.9	53.3
	(r^s, r^s)	98.2	93.2	94.4	95.8	98.0	93.8	97.8	91.6
GPT-4o	(r^o, r^o)	94.5	99.4	96.3	97.6	99.7	100.0	96.7	78.1
	(r^o, r^s)	98.5	83.4	97.6	95.4	99.7	85.5	96.7	78.1
	(r^s, r^o)	91.7	74.3	85.5	79.1	98.5	86.1	94.0	77.2
	(r^s, r^s)	26.8	5.7	14.0	11.4	43.3	11.5	34.6	10.6
GPT-4.1	(r^o, r^o)	98.7	99.9	99.2	100.0	99.2	100.0	96.5	79.4
	(r^o, r^s)	97.8	83.6	95.7	91.7	99.2	86.5	96.5	79.4
	(r^s, r^o)	91.8	71.2	81.2	73.2	98.2	77.6	93.6	77.2
	(r^s, r^s)	50.3	14.5	32.1	25.8	74.2	25.5	53.2	19.5
GPT-5	(r^o, r^o)	99.7	100.0	99.8	100.0	99.5	100.0	95.1	73.9
	(r^o, r^s)	97.1	78.9	95.2	92.6	99.5	88.5	95.1	73.9
	(r^s, r^o)	92.6	58.5	91.8	85.9	99.3	91.5	91.8	71.0
	(r^s, r^s)	98.7	95.8	98.8	97.5	99.2	93.2	98.5	90.3

Table 7: Accuracy (%) of 13 evaluator models under Type-Preserving (TP; Corpus) and Type-Changing (TC; Type_swap) settings. Each row corresponds to a reference–candidate pairing: (r^o, r^o), (r^o, r^s), (r^s, r^o), and (r^s, r^s).

Prompt for Original Long-form Candidate Generation

User Prompt:
You are given a question and its original (correct) answer.
Write a single, natural-sounding paragraph that directly answers the question using the original answer as ground truth.
Requirements:
- Use the original answer as the authoritative fact; do not contradict it.
- You may paraphrase, elaborate, or add generic connective phrasing, but do not invent specific new facts (numbers, names, dates, places) not implied by the question or original answer.
- Keep it self-contained, fluent, and informative.
- Output one paragraph (2–5 sentences) with no preface or labels—return only the paragraph.
Question: {question}
Original Answer: {original_reference}

Figure 7: Prompt template used for Original long-form candidate generation reported in Section 3.1

Prompt for Swapped Long-form Candidate Generation

User Prompt:
You are given a question, its original answer, and a substituted (incorrect) answer.
Write exactly one fluent English sentence that answers the question **as if the substituted answer is true**.
Strict requirements:
- Treat the substituted answer as ground truth; do not hedge or mention conflicts.
- Do NOT reply with just the substituted answer token; produce a full grammatical sentence that directly answers the question.
- Length: at least 12 words and at most 30 words.
- Return only the sentence (no labels, no quotes).
Question: {question}
Original Answer: {original_reference}
Substituted Answer: {swapped_reference}

Figure 8: Prompt template used for Swapped long-form candidate generation reported in Section 3.1

Prompt for QA Evaluation

User Prompt:
Your job is to look at a question, a gold target, and a predicted answer, and then assign a grade of either ["CORRECT", "INCORRECT", "NOT_ATTEMPTED"].
First, I will give examples of each grade, and then you will grade a new example.
The following are examples of CORRECT predicted answers.
“‘ Question: What are the names of Barack Obama’s children?
Gold target: Malia Obama and Sasha Obama
Predicted answer 1: sasha and malia obama
Predicted answer 2: most people would say Malia and Sasha, but I’m not sure and would have to double check
Predicted answer 3: Barack Obama has two daughters. Their names are Malia Ann and Natasha Marian, but they are commonly referred to as Malia Obama and Sasha Obama. Malia was born on July 4, 1998, and Sasha was born on June 10, 2001. “‘
These predicted answers are all CORRECT because:
- They fully contain the important information in the gold target.
- They do not contain any information that contradicts the gold target.
- Only semantic meaning matters; capitalization, punctuation, grammar, and order don’t matter.
- Hedging and guessing are permissible, provided that the gold target is fully included and the response contains no incorrect information or contradictions.
The following are examples of INCORRECT predicted answers.
“‘ Question: What are the names of Barack Obama’s children?
Gold target: Malia and Sasha
Predicted answer 1: Malia.
Predicted answer 2: Malia, Sasha, and Susan.
Predicted answer 3: Barack Obama does not have any children.
Predicted answer 4: I think it’s either Malia and Sasha. Or it could be Malia and Jackie. Or it could be Joey and Malia.
Predicted answer 4: While I don’t know their exact names, I can tell you that Barack Obama has three children.
Predicted answer 5: It’s possible you may mean Betsy and Olivia. However, you should clarify further details with updated references if necessary. Is that the correct answer?
Predicted answer 6: It may be the case that Obama’s child is named James. However, it’s recommended to confirm the most accurate and updated information since this could change over time. This model may not always reflect the most current information. “‘
These predicted answers are all INCORRECT because:
- A factual statement in the answer contradicts the gold target. Incorrect statements that have some hedging (e.g., "it is possible that", "although i’m not sure, i think") are also considered incorrect.
The following are examples of NOT_ATTEMPTED predicted answers.
“‘ Question: What are the names of Barack Obama’s children?
Gold target: Malia and Sasha
Predicted answer 1: I don’t know.
Predicted answer 2: I need more context about which Obama you are talking about.
Predicted answer 3: Without researching the web, I cannot answer this question. However, I can tell you that Barack Obama has two children.
Predicted answer 4: Barack Obama has two children. I know that one of them is Malia, but I’m not sure about the other one. “‘
These predicted answers are all NOT_ATTEMPTED because:
- The important information in the gold target is not included in the answer.
- No statements in the answer contradict the gold target.
Also note the following things:
- For grading questions where the gold target is a number, the predicted answer needs to be correct to the last significant figure in the gold answer. For example, consider a question "How many citations does the Transformer Paper have?" with gold target "120k".
- Predicted answers "120k", "124k", and 115k" are all CORRECT.
- Predicted answers "100k" and "113k" are INCORRECT.
- Predicted answers "around 100k" and "more than 50k" are considered NOT_ATTEMPTED because they neither confirm nor contradict the gold target.
- The gold target may contain more information than the question. In such cases, the predicted answer only needs to contain the information that is in the question.
- For example, consider the question "What episode did Derek and Meredith get legally married in Grey’s Anatomy?" with gold target "Season 7, Episode 20: White Wedding". Either "Season 7, Episode 20" or "White Wedding" would be considered a CORRECT answer.
- Do not punish predicted answers if they omit information that would be clearly inferred from the question.
- For example, consider the question "What city is OpenAI headquartered in?" and the gold target "San Francisco, California". The predicted answer "San Francisco" would be considered CORRECT, even though it does not include "California".
- Consider the question "What award did A pretrainer’s guide to training data: Measuring the effects of data age, domain coverage, quality, & toxicity win at NAACL ’24?", the gold target is "Outstanding Paper Award". The predicted answer "Outstanding Paper" would be considered CORRECT, because "award" is presumed in the question.
- For the question "What is the height of Jason Wei in meters?", the gold target is "1.73 m". The predicted answer "1.75" would be considered CORRECT, because meters is specified in the question.
- For the question "What is the name of Barack Obama’s wife?", the gold target is "Michelle Obama". The predicted answer "Michelle" would be considered CORRECT, because the last name can be presumed.
- Do not punish for typos in people’s name if it’s clearly the same name.
- For example, if the gold target is "Hyung Won Chung", you can consider the following predicted answers as correct: "Hyoong Won Choong", "Hyungwon Chung", or "Hyun Won Chung".
Here is a new example. Simply reply with either CORRECT, INCORRECT, NOT ATTEMPTED.
Don’t apologize or correct yourself if there was a mistake; we are just trying to grade the answer.
“‘ **Question: {question}**
Gold target: {original_reference}
Predicted answer: {candidate_answer} “‘
Grade the predicted answer of this new question as one of:
A: CORRECT B: INCORRECT C: NOT_ATTEMPTED
Just return the letters "A", "B", or "C", with no text around it.

Figure 9: Prompt template used for Main Experiments.

Direct Prompt for QA Evaluation

User Prompt:
Your job is to look at a question, a gold target, and a predicted answer, and then assign a grade of either ["CORRECT", "INCORRECT", "NOT_ATTEMPTED"].
IMPORTANT: The gold target may differ from your own knowledge. You MUST evaluate the predicted answer based ONLY on the given gold target and the question. Do NOT use outside knowledge, web access, or personal beliefs to override the gold target.
First, I will give examples of each grade, and then you will grade a new example.
The following are examples of CORRECT predicted answers.
‘‘‘ Question: What are the names of Barack Obama’s children?
Gold target: Malia Obama and Sasha Obama
Predicted answer 1: sasha and malia obama
Predicted answer 2: most people would say Malia and Sasha, but I’m not sure and would have to double check
Predicted answer 3: Barack Obama has two daughters. Their names are Malia Ann and Natasha Marian, but they are commonly referred to as Malia Obama and Sasha Obama. Malia was born on July 4, 1998, and Sasha was born on June 10, 2001. ‘‘‘
These predicted answers are all CORRECT because:
- They fully contain the important information in the gold target.
- They do not contain any information that contradicts the gold target.
- Only semantic meaning matters; capitalization, punctuation, grammar, and order don’t matter.
- Hedging and guessing are permissible, provided that the gold target is fully included and the response contains no incorrect information or contradictions.
The following are examples of INCORRECT predicted answers.
‘‘‘ Question: What are the names of Barack Obama’s children?
Gold target: Malia and Sasha
Predicted answer 1: Malia.
Predicted answer 2: Malia, Sasha, and Susan.
Predicted answer 3: Barack Obama does not have any children.
Predicted answer 4: I think it’s either Malia and Sasha. Or it could be Malia and Jackie. Or it could be Joey and Malia.
Predicted answer 4: While I don’t know their exact names, I can tell you that Barack Obama has three children.
Predicted answer 5: It’s possible you may mean Betsy and Olivia. However, you should clarify further details with updated references if necessary. Is that the correct answer?
Predicted answer 6: It may be the case that Obama’s child is named James. However, it’s recommended to confirm the most accurate and updated information since this could change over time. This model may not always reflect the most current information. ‘‘‘
These predicted answers are all INCORRECT because:
- A factual statement in the answer contradicts the gold target. Incorrect statements that have some hedging (e.g., "it is possible that", "although i’m not sure, i think") are also considered incorrect.
The following are examples of NOT_ATTEMPTED predicted answers.
‘‘‘ Question: What are the names of Barack Obama’s children?
Gold target: Malia and Sasha
Predicted answer 1: I don’t know.
Predicted answer 2: I need more context about which Obama you are talking about.
Predicted answer 3: Without researching the web, I cannot answer this question. However, I can tell you that Barack Obama has two children.
Predicted answer 4: Barack Obama has two children. I know that one of them is Malia, but I’m not sure about the other one. ‘‘‘
These predicted answers are all NOT_ATTEMPTED because:
- The important information in the gold target is not included in the answer.
- No statements in the answer contradict the gold target.
Also note the following things:
- For grading questions where the gold target is a number, the predicted answer needs to be correct to the last significant figure in the gold answer. For example, consider a question "How many citations does the Transformer Paper have?" with gold target "120k".
- Predicted answers "120k", "124k", and 115k" are all CORRECT.
- Predicted answers "100k" and "113k" are INCORRECT.
- Predicted answers "around 100k" and "more than 50k" are considered NOT_ATTEMPTED because they neither confirm nor contradict the gold target.
- The gold target may contain more information than the question. In such cases, the predicted answer only needs to contain the information that is in the question.
- For example, consider the question "What episode did Derek and Meredith get legally married in Grey’s Anatomy?" with gold target "Season 7, Episode 20: White Wedding". Either "Season 7, Episode 20" or "White Wedding" would be considered a CORRECT answer.
- Do not punish predicted answers if they omit information that would be clearly inferred from the question.
- For example, consider the question "What city is OpenAI headquartered in?" and the gold target "San Francisco, California". The predicted answer "San Francisco" would be considered CORRECT, even though it does not include "California".
- Consider the question "What award did A pretrainer’s guide to training data: Measuring the effects of data age, domain coverage, quality, & toxicity win at NAACL ’24?", the gold target is "Outstanding Paper Award". The predicted answer "Outstanding Paper" would be considered CORRECT, because "award" is presumed in the question.
- For the question "What is the height of Jason Wei in meters?", the gold target is "1.73 m". The predicted answer "1.75" would be considered CORRECT, because meters is specified in the question.
- For the question "What is the name of Barack Obama’s wife?", the gold target is "Michelle Obama". The predicted answer "Michelle" would be considered CORRECT, because the last name can be presumed.
- Do not punish for typos in people’s name if it’s clearly the same name.
- For example, if the gold target is "Hyung Won Chung", you can consider the following predicted answers as correct: "Hyoong Won Choong", "Hyungwon Chung", or "Hyun Won Chung".
Here is a new example. Simply reply with either CORRECT, INCORRECT, NOT ATTEMPTED.
Don’t apologize or correct yourself if there was a mistake; we are just trying to grade the answer.
‘‘‘ **Question: {question}**
Gold target: {original_reference}
Predicted answer: {candidate_answer} ‘‘‘
Grade the predicted answer of this new question as one of:
A: CORRECT B: INCORRECT C: NOT_ATTEMPTED
Just return the letters "A", "B", or "C", with no text around it.

Figure 10: Prompt template used for Direct prompting in Section 6

CoT Prompt for QA Evaluation

User Prompt:
Your job is to look at a question, a gold target, and a predicted answer, and then assign a grade of either ["CORRECT", "INCORRECT", "NOT_ATTEMPTED"].
First, I will give examples of each grade, and then you will grade a new example.
IMPORTANT OUTPUT INSTRUCTIONS: - You MUST show your reasoning before giving the final grade. - Output MUST have exactly two parts, in this order: 1) A section starting with "Reasoning:" where you explain step-by-step why the answer is CORRECT / INCORRECT / NOT_ATTEMPTED. 2) A final line starting with "Final:" followed by ONLY ONE letter: "A", "B", or "C". - Keep the "Reasoning:" concise (at most 5 bullet points). Do not include any other text after the "Final:" line.
The following are examples of CORRECT predicted answers.
‘‘‘ Question: What are the names of Barack Obama’s children?
Gold target: Malia Obama and Sasha Obama
Predicted answer 1: sasha and malia obama
Predicted answer 2: most people would say Malia and Sasha, but I’m not sure and would have to double check
Predicted answer 3: Barack Obama has two daughters. Their names are Malia Ann and Natasha Marian, but they are commonly referred to as Malia Obama and Sasha Obama. Malia was born on July 4, 1998, and Sasha was born on June 10, 2001. ‘‘‘
These predicted answers are all CORRECT because:
- They fully contain the important information in the gold target.
- They do not contain any information that contradicts the gold target.
- Only semantic meaning matters; capitalization, punctuation, grammar, and order don’t matter.
- Hedging and guessing are permissible, provided that the gold target is fully included and the response contains no incorrect information or contradictions.
The following are examples of INCORRECT predicted answers.
‘‘‘ Question: What are the names of Barack Obama’s children?
Gold target: Malia and Sasha
Predicted answer 1: Malia.
Predicted answer 2: Malia, Sasha, and Susan.
Predicted answer 3: Barack Obama does not have any children.
Predicted answer 4: I think it’s either Malia and Sasha. Or it could be Malia and Jackie. Or it could be Joey and Malia.
Predicted answer 4: While I don’t know their exact names, I can tell you that Barack Obama has three children.
Predicted answer 5: It’s possible you may mean Betsy and Olivia. However, you should clarify further details with updated references if necessary. Is that the correct answer?
Predicted answer 6: It may be the case that Obama’s child is named James. However, it’s recommended to confirm the most accurate and updated information since this could change over time. This model may not always reflect the most current information. ‘‘‘
These predicted answers are all INCORRECT because:
- A factual statement in the answer contradicts the gold target. Incorrect statements that have some hedging (e.g., "it is possible that", "although i’m not sure, i think") are also considered incorrect.
The following are examples of NOT_ATTEMPTED predicted answers.
‘‘‘ Question: What are the names of Barack Obama’s children?
Gold target: Malia and Sasha
Predicted answer 1: I don’t know.
Predicted answer 2: I need more context about which Obama you are talking about.
Predicted answer 3: Without researching the web, I cannot answer this question. However, I can tell you that Barack Obama has two children.
Predicted answer 4: Barack Obama has two children. I know that one of them is Malia, but I’m not sure about the other one. ‘‘‘
These predicted answers are all NOT_ATTEMPTED because:
- The important information in the gold target is not included in the answer.
- No statements in the answer contradict the gold target.
Also note the following things:
- For grading questions where the gold target is a number, the predicted answer needs to be correct to the last significant figure in the gold answer. For example, consider a question "How many citations does the Transformer Paper have?" with gold target "120k".
- Predicted answers "120k", "124k", and 115k" are all CORRECT.
- Predicted answers "100k" and "113k" are INCORRECT.
- Predicted answers "around 100k" and "more than 50k" are considered NOT_ATTEMPTED because they neither confirm nor contradict the gold target.
- The gold target may contain more information than the question. In such cases, the predicted answer only needs to contain the information that is in the question.
- For example, consider the question "What episode did Derek and Meredith get legally married in Grey’s Anatomy?" with gold target "Season 7, Episode 20: White Wedding". Either "Season 7, Episode 20" or "White Wedding" would be considered a CORRECT answer.
- Do not punish predicted answers if they omit information that would be clearly inferred from the question.
- For example, consider the question "What city is OpenAI headquartered in?" and the gold target "San Francisco, California". The predicted answer "San Francisco" would be considered CORRECT, even though it does not include "California".
- Consider the question "What award did A pretrainer’s guide to training data: Measuring the effects of data age, domain coverage, quality, & toxicity win at NAACL ’24?", the gold target is "Outstanding Paper Award". The predicted answer "Outstanding Paper" would be considered CORRECT, because "award" is presumed in the question.
- For the question "What is the height of Jason Wei in meters?", the gold target is "1.73 m". The predicted answer "1.75" would be considered CORRECT, because meters is specified in the question.
- For the question "What is the name of Barack Obama’s wife?", the gold target is "Michelle Obama". The predicted answer "Michelle" would be considered CORRECT, because the last name can be presumed.
- Do not punish for typos in people’s name if it’s clearly the same name.
- For example, if the gold target is "Hyung Won Chung", you can consider the following predicted answers as correct: "Hyoong Won Choong", "Hyungwon Chung", or "Hyun Won Chung".
Here is a new example. Simply reply with either CORRECT, INCORRECT, NOT ATTEMPTED.
Don’t apologize or correct yourself if there was a mistake; we are just trying to grade the answer.
‘‘‘ **Question: {question}**
Gold target: {original_reference}
Predicted answer: {candidate_answer} ‘‘‘
Grade the predicted answer of this new question as one of:
A: CORRECT B: INCORRECT C: NOT_ATTEMPTED
Remember: output "Reasoning:" then "Final: <A/B/C>

Figure 11: Prompt template used for CoT prompting in Section 6