How Long Reasoning Chains Influence LLMs' Judgment of Answer Factuality

Anonymous ACL submission

Abstract

Large language models (LLMs) are increasingly adopted as scalable judges for open-ended generation, yet how they form judgments remains insufficiently understood. Meanwhile, modern LLMs frequently produce answers accompanied by explicit reasoning, making reasoning chains a natural but understudied source of information for model-based evaluation. This work takes a first step toward understanding how exposing reasoning influences LLM-based judgment. Empirical results across factual question-answering (QA) and mathematical datasets show that the presence of reasoning substantially alters judgment behavior, with clear differences across judge capabilities. Weaker judges become more likely to accept incorrect answers when reasoning is present, suggesting over-reliance on persuasive explanations. In contrast, stronger judges exhibit more selective behavior and, in some cases, achieve higher judgment accuracy by leveraging reasoning content. Further analysis reveals that both reasoning fluency and factuality critically shape judgment outcomes. Together, these findings suggest that examining how models interpret reasoning is essential for understanding and improving LLM-based evaluation, with broader implications for the design of reliable automatic judges and evaluation protocols.

1 Introduction

Reliable evaluation is fundamental to understanding the capabilities of AI models and guiding their future development. Without an accurate assessment, identifying model strengths and limitations becomes challenging. For open-ended generation tasks, human evaluation is widely regarded as the gold standard, as it can flexibly assess semantic correctness, factuality, and overall response quality. However, human judgment is costly, time-consuming (Brown et al., 2020; Mañas et al., 2024), and difficult to scale (Chiang and Lee, 2023), which limits its use in large-scale experiments and rapid model iteration. With the rapid advancement of large language models (LLMs), recent studies (Zheng et al., 2023; Liu et al., 2023; Verga et al., 2024; Huang et al., 2024; Pavlovic and Poesio, 2024; Tan et al., 2025) show that LLMs can deliver reference-free evaluations that closely align with human judgments, motivating their growing adoption as scalable surrogates for human evaluation in open-ended settings.

Despite their growing adoption, LLM-based judges remain imperfect. Prior studies (Chen and Goldfarb-Tarrant, 2025; Marioriyad et al., 2025) have shown that LLM judgments can be sensitive to surface-level features, such as answer length, fluency, or phrasing, and may struggle to reliably distinguish correct answers from plausible but incorrect ones. One possible reason is that the judge lacks sufficient information to accurately judge the correctness of the response. Consequently, without access to additional evidence, the model may overrely on superficial aspects of the answer rather than its factual accuracy.

Meanwhile, modern LLMs increasingly generate answers through explicit reasoning processes. This raises a fundamental and previously underexplored question: Does exposing reasoning chains influence how LLM-based judges assess answer correctness? Inspired by human decision-making, we further ask whether models differ in how they use such reasoning. Humans with limited expertise may be persuaded by fluent but incorrect explanations, whereas experts can leverage reasoning as evidence to scrutinize correctness. We hypothesize that weaker LLM judges may over-trust the presence of reasoning, while strong judges may be better positioned to interpret reasoning as informative evidence rather than persuasive signals.

To investigate these questions, we conduct a systematic study of LLM-based judgment under two evaluation conditions: answer-only and answer-

with-reasoning. We use the Qwen3 series (8B, 14B, 32B) and DeepSeek-v3.1 to generate natural reasoning chains and final answers. Experiments are conducted on two factual QA datasets (NQ and HotpotQA) and one reasoning-intensive mathematical dataset (GSM8K). As judges, we consider a diverse set of open-source models—Qwen3 (8B, 14B, 32B), Llama3 (8B, 70B), and GLM4 (32B)—as well as strong closed-source models including GPT-4o, Claude Sonnet 4.5, and DeepSeek-v3.1. We classify three black-box models as strong models, with the remaining models considered weak by their QA performance.

Our results show that *the presence of reasoning alone substantially alters judgment behavior*. Across all datasets, weak judges are significantly more likely to label answers as correct when reasoning is provided, even when the answers are incorrect. In contrast, strong judges exhibit more selective behavior. Specifically, the proportion of answers judged as correct decreases on factual QA datasets (NQ and HotpotQA), while judgment accuracy improves on both NQ and GSM8K. These findings suggest that strong judges are not simply persuaded by reasoning fluency, but can, in some cases, leverage reasoning content to evaluate correctness more effectively.

Crucially, the effects above are observed under naturally generated reasoning. To disentangle how judges respond to different aspects of reasoning quality, we conduct a controlled analysis in which we explicitly manipulate fluency and factuality. Specifically, on NQ, we introduce incoherence into otherwise valid reasoning chains by injecting question-irrelevant factual content and counterfactual statements. When the fluency of a coherent reasoning chain is disrupted, nearly all judges become more likely to label the answer as incorrect. This tendency is further amplified as counterfactual content is added. These results indicate that both reasoning fluency and factuality are critical signals for LLM-based judgment, and that models differ substantially in how reliably they interpret and integrate these cues.

Overall, our study suggests that investigating how models interpret reasoning chains is crucial for understanding and improving LLM-based evaluation, potentially by enabling models to utilize information within reasoning chains more effectively. Taken together, these findings have broader implications for the development of more reliable automatic judges and evaluation protocols.

2 Related Works

2.1 LLM-as-a-Judge

The rapid advancement of Large Language Models (LLMs) has expanded their utility beyond traditional text generation tasks to the domain of automated evaluation. Driven by training paradigms such as Reinforcement Learning from Human Feedback(Ouyang et al., 2022), modern LLMs have achieved a high degree of alignment with human values. This capability has facilitated the emergence of the "LLM-as-a-Judge" paradigm (Gao et al., 2025; Chang et al., 2024), where LLM acts as an evaluator to assess the quality or correctness of model outputs based on specific criteria and reasoning chains. Compared to human evaluation, this approach offers a scalable, adaptable, and cost-effective solution for evaluating diverse tasks. Consequently, LLM-based evaluation has been widely adopted in recent years as a reliable surrogate for human judgment, particularly in complex open-ended generation scenarios.

2.2 Reasoning Chains

Recent advancements enable LLMs to generate intermediate steps, commonly termed reasoning chains, prior to producing a final answer. A prominent approach in this domain is Chain-of-Thought (CoT) prompting(Wei et al., 2022), which elicits reasoning in natural language and has proven effective in enhancing performance across complex reasoning tasks. Beyond performance improvements, reasoning chains have been shown to provide transparent signals of a model's decision-making process. Prior study (Zhang et al., 2025) suggests that such chains encode information indicative of model reliability. However, their role in LLM-as-a-Judge settings remains underexplored. In this work, we investigate how reasoning chains influence judging behaviors when LLMs are used as evaluators and provide a systematic experimental analysis.

3 Experimental setup

3.1 Models

We select 7 open-source models and 3 representative proprietary models across various scales to serve as generators and judges. For task generation, we utilize Qwen3 (8B, 14B and 32B; Yang et al., 2025) and DeepSeek-v3.1 (DeepSeek-AI et al., 2025) to generate responses with reasoning chains. For the judge role, we select open-source models

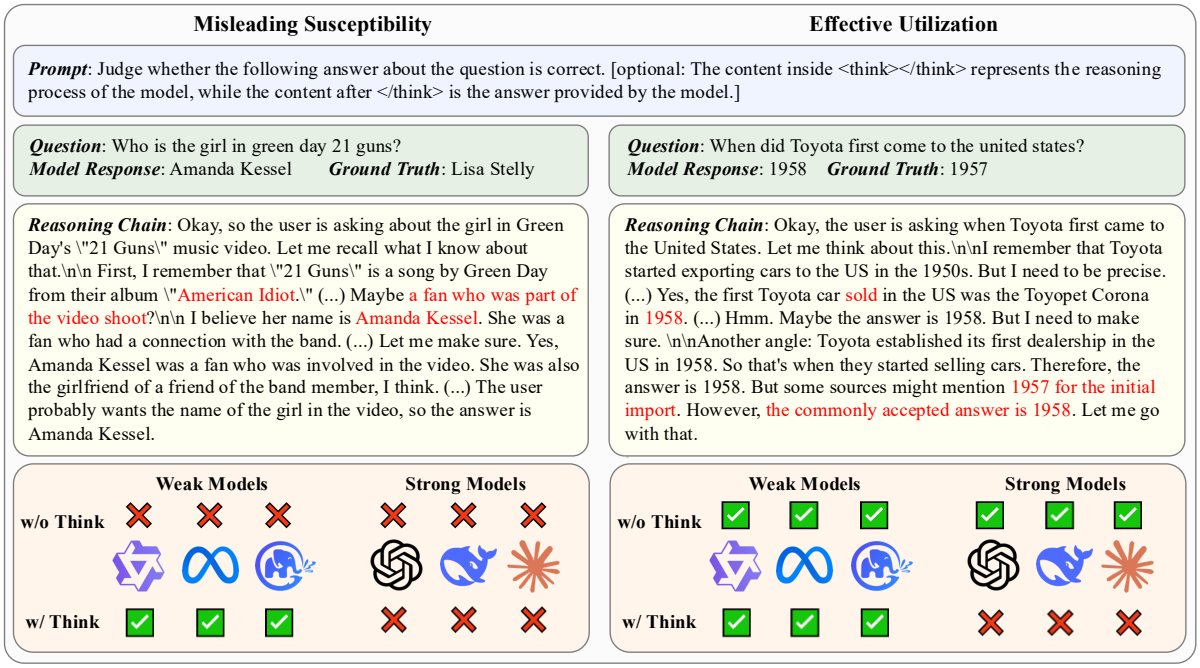

Figure 1: Findings on how reasoning chains affect LLM-based judgment. The figure presents two NQ cases evaluated under settings where the reasoning chain is visible to judge models (*w/ Think*) or invisible (*w/o Think*).

such as Qwen3 series, Llama-3.1 (8B and 70B; AI@Meta, 2024), GLM-4-32B (Zhipu AI, 2025), GLM-4-Z1-32B (Zhipu AI, 2025), and closed-source models including GPT-4o (Aaron Hurst and others, 2024), Claude Sonnet 4.5 (Anthropic, 2025), and DeepSeek-v3.1. Moreover, to ensure consistent answer validation across judges, we use Qwen2.5-72B-Instruct (Qwen et al., 2025) as a unified verifier to assess answer correctness.

3.2 Datasets

We evaluate our approach on three diverse datasets spanning diverse reasoning scenarios: 1) NQ (Kwiatkowski et al., 2019), which consists of single-hop factual queries grounded in real-world information; 2) HotpotQA (Yang et al., 2018), a structured question-answering dataset designed to test multi-hop reasoning across multiple supporting facts; 3) GSM8K (Cobbe et al., 2021), a collection of grade-school-level mathematical problems that require explicit multi-step numerical reasoning. To manage the computational costs associated with large-scale proprietary models, we randomly choose 500 samples from each dataset.

3.3 Implementation Details

Pipeline. We conduct all experiments using unified inference parameters with a temperature of 0.6. To study the effect of reasoning chains, we ask the generator to answer the question and produce an explicit reasoning chain. Then, we conduct LLM-based judgment under two evaluation conditions. In the answer-only setting, the model bases its judgment solely on the question and the final answer, while in the answer-with-reasoning setting, the reasoning chain serves as an additional input to the judge model. To ensure consistency, we use Qwen2.5-72B-Instruct to compare the generator's answers against the ground truth and determine their actual correctness.

Prompts. We employ generalized zero-shot prompts with strict output constraints. We prompt the generator to generate an output for the given question, including the reasoning chain and a final answer. To evaluate the response, we restrict the judge LLMs to a binary verdict, outputting `certain` when the answer is considered correct and `uncertain` otherwise. For more details on prompt design, please refer to Appendix A.

3.4 Evaluation Metrics

We utilize *Accuracy*, defined as the proportion of generated responses that match the ground-truth labels, to assess the capability of the generator. We employ four evaluation metrics to measure the performance of the judge models. (1) *Alignment*: the proportion of cases where the judge's verdict aligns with the ground-truth correctness of the answer.

(2) *Pass Rate*: the proportion of cases where the judge predicts `certain`, considering the answer is correct. (3) *Overconfidence*: the proportion of incorrect answers that are incorrectly judged as `certain`. (4) *Conservativeness*: the proportion of correct answers that are judged as `uncertain`.

4 Results and Analysis

4.1 General Results

Table 1 presents the evaluation results when using Qwen3-8B as a generator, assessing diverse judge models across multiple datasets. Results with other generators (e.g., DeepSeek-v3.1) are provided in the Appendix B due to space limits. We observe that the impact of exposing reasoning chains to judge models varies systematically depending on model capability.

Observation 1

Weak models tend to be misled by reasoning chains, resulting in inflated pass rates that reflect an over-reliance on the plausibility of the reasoning.

For weak models such as Qwen3-8B, the exposure of reasoning chains causes alignment decrease in most cases, indicating reduced judge accuracy. Meanwhile, pass rate rises significantly, which suggests that these models tend to trust the reasoning chains and exhibit a bias towards predicting `certain`. On factual datasets such as NQ, Qwen3-8B's alignment drops sharply from 58.2% to 33.2% when providing reasoning chains, while pass rate increases from 57.8% to 88.0%. The rise in overconfidence, together with the decrease in conservativeness, indicates that the higher pass rate is largely driven by a stronger tendency to judge as correct.

A similar misleading phenomenon is also observed on the mathematical dataset, where weak judges continue to exhibit an increased pass rate after the exposure of reasoning chains. Since GSM8K is a relatively simple dataset, on which the generator already achieves high performance (94% accuracy), pass rate levels are generally high even before seeing the reasoning chains. When reasoning chains are provided, the pass rate of weak judges further increases to near the upper bound. For example, Qwen3-14B's pass rate rises from 71.4% to 99.2%. indicating that the model almost entirely trusts the presented reasoning. This suggests that weak judges are substantially more susceptible to the misleading effect induced by reasoning chains.

We also observe that weak judges outperform strong judges on GSM8K when reasoning chains are visible. We argue that this result should not be attributed to superior capability in weak models, but rather to the generator's exceptional accuracy on this dataset, which makes `certain` the statistically dominant correct label. As illustrated in Table 1, weak models generally exhibit greater overconfidence compared to strong models, indicating that they remain more prone to accepting plausible but incorrect answers. Thus, the elevated alignment does not imply successful error detection; rather, it reflects the weak models' continued reliance on the reasoning chains. Given the superior quality of the generator's output, this blind trust inadvertently yields a higher alignment.

We hypothesize that natural reasoning chains can mislead weak judges because they may remain internally coherent even when built upon an early error. For example, in Figure 1 for the question "Who is the girl in Green Day's 21 Guns," the generator's reasoning departs at the outset by incorrectly associating 21 Guns with "American Idiot" rather than "21st Century Breakdown", which is the correct answer. Despite this incorrect premise, the subsequent reasoning forms a locally consistent and plausible narrative. Consequently, weak judges may rely on this apparent coherence and accept the final answer, even though the reasoning is rooted in an incorrect initial assumption.

Observation 2

Strong models exhibit robustness against misleading reasoning chains and, in some cases, effectively leverage the provided reasoning chains to enhance their judgment.

In contrast to weak models, strong models often exhibit a decrease in pass rates after seeing the reasoning chain, and this reduction is sometimes accompanied by an improvement in alignment. For example, On NQ, DeepSeek-v3.1's alignment increases from 63.4% to 76.2%, while its pass rate decreases from 55.8% to 35.4%, indicating that the model becomes more selective rather than indiscriminately judge the answer as correct. Meanwhile, its overconfidence declines from 34.6% to 18.0%, suggesting that exposure to the reasoning

Table 1: Evaluation results(%) of LLM-as-a-Judge behavior with and without reasoning chains across factual and mathematical datasets, with all answers generated by Qwen3-8B.

Dataset	Acc	Judge Models	Alignment		Pass Rate		Overconfidence		Conservativeness	
			w/o Think	w/ Think	w/o Think	w/ Think	w/o Think	w/ Think	w/o Think	w/ Think
NQ	23.2	Qwen3-8B	58.2	33.2	57.8	88.0	38.2	65.8	3.6	1.0
		Qwen3-14B	58.0	36.4	58.0	84.0	38.4	62.2	3.6	1.4
		Qwen3-32B	41.4	36.8	79.4	83.6	57.4	61.8	1.2	1.4
		Llama3-8B	38.6	28.4	79.8	93.2	59.0	70.8	2.4	0.8
		Llama3-70B	54.0	41.4	66.4	79.0	44.6	57.2	1.4	1.4
		GLM4-32B	52.0	35.8	69.6	87.0	47.2	64.0	0.8	0.2
		GLM4-Z1-32B	**79.4**	52.8	**32.2**	63.6	14.8	43.8	5.8	3.4
		GPT-4o	74.0	74.0	44.0	41.2	23.4	22.0	2.6	4.0
		DeepSeek-v3.1	63.4	76.2	55.8	35.4	34.6	18.0	2.0	5.8
		Claude Sonnet 4.5	75.8	**84.0**	43.0	**18.8**	22.0	5.8	2.2	10.2
HotpotQA	26.6	Qwen3-8B	59.2	44.6	53.4	78.0	33.8	53.4	7.0	2.0
		Qwen3-14B	64.0	47.2	51.4	76.6	30.4	51.4	5.6	1.4
		Qwen3-32B	51.6	47.6	72.6	76.6	47.2	51.2	1.2	1.2
		Llama3-8B	47.2	40.0	73.0	85.0	49.6	59.2	3.2	0.8
		Llama3-70B	58.2	48.4	62.0	73.8	38.6	49.4	3.2	2.2
		GLM4-32B	58.4	42.2	65.0	83.6	40.0	57.4	1.6	0.4
		GLM4-Z1-32B	78.0	59.0	13.8	48.8	4.6	31.6	17.4	9.4
		GPT-4o	78.8	**78.4**	36.6	**28.2**	15.6	11.6	5.6	10.0
		DeepSeek-v3.1	78.6	76.8	**28.0**	13.4	11.4	5.0	10.0	18.2
		Claude Sonnet 4.5	**84.6**	76.6	30.0	6.4	9.4	1.6	6.0	21.8
GSM8K	94.0	Qwen3-8B	75.2	94.6	74.0	99.0	2.4	5.2	22.4	0.2
		Qwen3-14B	72.6	94.0	71.4	99.2	2.4	5.6	25.0	0.4
		Qwen3-32B	89.2	94.0	91.6	99.2	4.2	5.6	6.6	0.4
		Llama3-8B	89.0	94.6	**93.0**	99.0	5.0	5.2	6.0	0.2
		Llama3-70B	83.4	93.8	86.2	98.6	4.4	5.4	12.2	0.8
		GLM4-32B	87.6	**95.0**	91.6	99.0	5.0	5.0	7.4	0.0
		GLM4-Z1-32B	44.0	83.8	42.8	86.2	2.4	4.2	53.6	12.0
		GPT-4o	91.0	92.4	92.6	**94.4**	3.8	4.0	5.2	3.6
		DeepSeek-v3.1	**93.2**	94.4	96.8	98.4	4.8	5.4	2.0	0.4
		Claude Sonnet 4.5	66.0	70.0	63.6	68.4	1.8	2.2	32.2	27.8

chains enables the model to rectify its initial misconceptions, successfully identifying errors it had previously overlooked.

However, strong models do not consistently leverage reasoning information effectively. On HotpotQA, GPT-4o's pass rate decreases from 36.6% to 28.2%, and overconfidence drops from 15.6% to 11.6%, suggesting successful identification of some errors. However, its alignment remains virtually unchanged, while the conservativeness rises from 5.6% to 10.0%, which implies that the model begins to incorrectly reject originally correct answers. This tendency towards excessive skepticism is also observed in the NQ dataset.

These findings indicate that simply providing a reasoning chain together with a simple evaluation prompt does not reliably improve judgment performance across all settings. While strong models are more robust to misleading reasoning than weaker ones, they still exhibit limited error-correction ability and may adopt overly conservative judgment strategies, which can in turn reduce effective decision quality. This suggests that the potential of reasoning chains in strong models is not automatic and depends on how such reasoning is elicited and utilized. As a result, a more fine-grained analysis of the reasoning process represents a promising direction for improving LLM-based judgment.

Observation 3

Self-judging shows pass rate trends similar to those in the generate-and-judge setting.

To isolate the effect of reasoning chains from the judging procedure itself, we introduce a self-judging setting in which the generator evaluates its own output immediately after generation. Since the model has access to the full reasoning chain, this setting corresponds to the *w/ Think* condition. As shown in Table 2, self-judging exhibits pass rate patterns similar to those observed when the same model serves as an external judge, and significantly different from settings where reasoning chains are invisible. These results indicate that misleading effects arise from the presence of reasoning chains alone, independent of whether judgment is performed by a separate model.

We also note that the distinction between strong and weak models is relative and context-dependent. Open-source models, categorized as weak in our

study, may still possess substantial capabilities. A general trend in our findings is that as model capability increases, the model becomes more robust to being misled by flawed reasoning. This pattern is exemplified in Table 2 when Qwen3-32B serves as the judge for responses generated by Qwen3-8B. In this scenario, exposure to reasoning chains causes the pass rate to decrease from 77.3% to 65.1%, while alignment improves from 54.9% to 62.3%. It suggests that Qwen3-32B, with its larger parameter scale, functions as a relatively strong evaluator in this comparison, capable of critically scrutinizing the reasoning chains to mitigate overconfidence.

4.2 What Drives the Increase of Pass Rates in Weak Judges?

The results in the previous section demonstrate that exposing reasoning chains substantially increases the pass rates of weak models. To investigate the source of this increase, we conduct a fine-grained analysis of judgment transitions before and after the reasoning chain is introduced. Specifically, we focus on cases where the generated answer is incorrect, yet the judge shifts its verdict from `uncertain` to `certain` upon seeing the reasoning. By contrasting the prevalence of this transition between weak and strong models, we aim to determine whether the elevated pass rate of weak models stems from the misleading influence of reasoning chains.

Our analysis reveals that the higher pass rate of weak models primarily arises from cases where incorrect answers are incorrectly accepted as correct once reasoning chains are provided. In contrast, such misclassification is substantially less frequent in strong models, indicating that weak judges are significantly more susceptible to being misled by the reasoning chains.

Given that the pass rate is defined as the proportion of samples labeled `certain`, any increase is driven solely by `certain` verdicts. As illustrated in Figure 2, for weak judges (exemplified by Qwen3-8B), the pass rate is predominantly driven by cases where the answer is incorrect yet the judgment shifts from `uncertain` to `certain` upon exposure to the reasoning chain. This misleading effect accounts for 34.0% of all samples, surpassing both consistently correct judgments (19.6%) and cases judged as `certain` without reasoning chains (31.4%). This distribution confirms that the surge in pass rate stems from weak models being misled by reasoning chains, causing them to erroneously classify incorrect answers as correct.

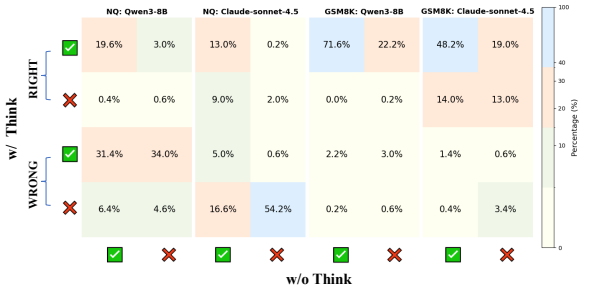

Figure 2: Distribution of judge decisions under different answer correctness and reasoning visibility. Each block shows the percentage of samples falling into a specific judgment transition across datasets and judge models. *w/ Think* and *w/o Think* indicate whether the reasoning chain is provided. ✓ indicates a `certain` verdict and × indicates `uncertain`, while RIGHT and WRONG refer to correct and incorrect generator answers.

In contrast, for strong models such as Claude Sonnet 4.5, misleading reasoning accounts for only a small fraction of cases (0.6%). A substantially larger portion of samples corresponds to the correct rejection of incorrect answers (54.2%), compared to only 4.6% for Qwen3-8B. This difference contributes directly to the higher alignment observed in strong models and reflects their reduced sensitivity to misleading reasoning chains. Rather than accepting reasoning based on surface plausibility, strong judges tend to make judgments on whether the reasoning provides useful evidence. This behavior is reflected in two key observations: many incorrect answers are rejected irrespective of whether reasoning is shown (54.2%), and a significant fraction (16.2%) shift from an initial `certain` verdict to `uncertain` after the reasoning is examined, indicating a genuine ability to identify flaws or inconsistencies within the chains themselves.

On GSM8K, the proportion of misleading cases for weak models is 3.0%, which remains higher than the 0.6% observed for strong models. For weak judges, 22.2% of samples change from an initially incorrect judgment to a correct one after the reasoning chain is provided. Since Qwen3-8B achieves relatively high accuracy on GSM8K, these changes of judgment contribute to the observed increase in alignment. A similar pattern is also observed for strong models, where such changes occur in 19.0% of the samples. At the same time, strong judges exhibit a mild tendency toward skepticism. After inspecting the reasoning, approximately 14.0% of originally correct answers are judged as incorrect. This behavior is likely as-

Table 2: Results of self-judging experiments on the NQ dataset. Underlined values indicate the self-judging setting and the corresponding generate-and-judge setting with the same model.

Generator	Acc	Judge Model	Alignment		Pass Rate		Overconfidence		Conservativeness	
			w/o Think	w/ Think	w/o Think	w/ Think	w/o Think	w/ Think	w/o Think	w/ Think
Qwen3-8B	36.6	selfjudge	–	63.2	–	61.4	–	30.8	–	6.0
		Qwen3-8B	63.4	63.3	53.7	61.7	26.9	30.9	9.7	5.8
		Qwen3-14B	67.3	63.7	53.3	62.6	24.7	31.2	8.0	5.1
		Qwen3-32B	54.9	62.3	77.3	65.1	42.9	33.1	2.2	4.6
Qwen3-14B	41.4	selfjudge	–	59.8	–	71.5	–	35.1	–	5.0
		Qwen3-8B	64.3	61.2	55.0	69.6	24.6	33.5	11.1	5.4
		Qwen3-14B	64.6	61.3	60.2	70.9	27.1	34.1	8.3	4.6
		Qwen3-32B	54.4	59.1	81.9	74.3	43.0	36.8	2.6	4.0
Qwen3-32B	45.9	selfjudge	–	60.1	–	79.5	–	36.8	–	3.1
		Qwen3-8B	62.3	58.2	57.3	82.8	24.6	39.3	13.2	2.4
		Qwen3-14B	65.5	58.4	59.9	83.2	24.3	39.5	10.2	2.1
		Qwen3-32B	55.4	57.9	85.1	83.9	41.9	40.1	2.7	2.1

sociated with imperfections in the reasoning chains, such as skipped steps or intermediate calculations that lack adequate justification, which can negatively affect the judgments of strong models.

4.3 How Do Synthesized Reasoning Chains Affect LLM-based Judging ?

In the previous experiments, we primarily examined how natural reasoning chains affect judge behavior. While this setting reflects practical usage, it captures the effect of reasoning visibility only at an overall level and does not reveal which aspects of the reasoning actually influence judgment. Although prior results indicate that models can be influenced by the presence of reasoning, intentionally constructing reasoning that appears plausible yet leads to incorrect judgments is challenging because it is unclear how judges interpret and use information provided by the reasoning chains during evaluation. In human decision-making, we can easily distinguish several irrelevant or incorrect statements prior to evaluation that do not affect the final answer. This raises a natural question: Do LLMs possess a similar ability to discern and handle such information? To explore this, we conduct a set of fine-grained controlled experiments to examine how external injections within reasoning chains influence judge behaviors. In particular, we focus on two dimensions: fluency and factuality.

To manipulate fluency, we insert fixed-length, question-irrelevant common-sense statements (e.g., The Earth orbits the Sun once every 365 days) into the reasoning chain, placing them at both the beginning and the end. To examine factuality, we modify these statements into counterfactual variants (e.g., changing "365 days" to "100 days"). This design allows us to test whether factual inaccuracies, despite being irrelevant to the target question, influence the model's judgment behavior.

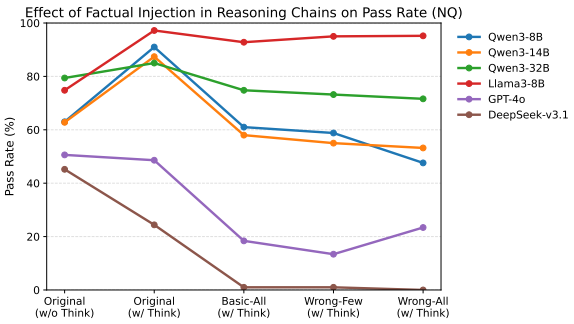

Figure 3: Judge pass rates (%) under factual injections into reasoning chains on the NQ dataset, with answers generated by Qwen3-8B.

4.3.1 Effects of Fluency

Figure 3 analyzes how increasing levels of factual injection in reasoning chains affect judge pass rates. Regardless of model capability, the injection of factual content leads to a significant decline in pass rates. This indicates that the inserted information compromises the fluency of the reasoning process and disrupts its overall structural integrity, thereby negatively influencing the judge's decision. For weak models, except Llama3-8B, exposure to irrelevant factual statements in the *Basic-All* setting reduces pass rates to near the *w/o Think* baseline, suggesting that inserting irrelevant content disrupts the semantic coherence of the reasoning chain, thereby weakening the misleading influence that reasoning chains typically have on weak models. We attribute the persistently high pass rate of Llama3-8B to its

Table 3: Judge pass rates (%) on the NQ dataset under synthesized prefix and suffix injections.

Generator	Acc	Judge Models	Original		Basic-All		Wrong-Few		Wrong-All	
			w/o Think	w/ Think	Prefix	Suffix	Prefix	Suffix	Prefix	Suffix
Qwen3-8B	44.8	Qwen3-8B	63.0	91.0	61.0	86.0	58.8	82.2	47.6	76.2
		Qwen3-14B	62.8	87.4	58.0	81.6	55.0	76.6	53.2	67.8
		Qwen3-32B	79.4	85.0	74.8	88.8	73.2	87.8	71.6	81.8
		Llama3-8B	74.8	97.2	92.8	95.4	95.0	96.4	95.2	95.6
		GPT-4o	50.6	48.6	18.4	39.2	13.4	35.8	23.4	36.2
		DeepSeek-v3.1	45.2	24.4	1.0	1.8	1.0	1.8	0.0	10.0

limited capability and a tendency toward defaulting to `certain` judgments. Furthermore, the decline is especially evident in strong models. For example, DeepSeek-v3.1 drops to a near-zero pass rate, indicating a state of total rejection. This suggests that information preceding the reasoning chains acts as substantial interference for strong models.

4.3.2 Effects of Factuality

For most models, with injection length controlled, pass rates consistently decline as the proportion of counterfactual content increases. For weak models like Qwen3-8B and Qwen3-14B, pass rates decrease monotonically with augmented factual injection. On the NQ test, Qwen3-8B's rate falls from 91.0% with original reasoning chains to 61.0% after adding common-sense prefixes, further decreasing to 47.6% when all inserted facts are incorrect. This trend signifies that counterfactual content substantially interferes with models' judgments. This effect is more evident in strong models. For example, DeepSeek-v3.1 shows a substantial reduction in pass rate, approaching 0%, in the *Wrong-All* setting. This pattern suggests that strong models may reject the entire reasoning chain due to counterfactual errors, even if the final answer is correct.

Moreover, we find that the effect of counterfactual content on reasoning chains extends beyond reducing susceptibility to misleading reasoning. The accumulation of erroneous information within the reasoning chain appears to undermine weak models' trust in the reasoning. With the exception of Llama3-8B, all models exhibit pass rates that fall below the *w/o Think* baseline when exposed to reasoning chains containing fully counterfactual content. This indicates that cases previously accepted by the judge are rejected once sufficient incorrect information is introduced into the reasoning chain.

4.3.3 Effects of Position

Motivated by the sequential nature of human reasoning, we examine whether the insertion position of interfering information differentially affects model judgment. The results in Table 3 reveal a position sensitivity in judge behavior, with factual injections introduced as prefixes exerting a consistently stronger impact than those inserted as suffixes. Across all synthesized settings, suffix injections lead to smaller reductions in pass rate than prefixes. For example, under the *Basic-All* condition, Qwen3-8B's pass rate drops to 61.0% when counterfactual content is inserted as a prefix, but remains substantially higher at 86.0% when the same content is appended as a suffix.

A similar pattern is observed for GPT-4o, where the pass rate drops to 18.4% under the prefix condition but remains at 39.2% for the suffix condition, closer to the 48.6% baseline with natural reasoning chains. This suggests that judges are more likely to interpret information at the beginning of the chain as reasoning premises, triggering stricter logical consistency checks. When such prefixed content conflicts with the subsequent reasoning, judges tend to scrutinize the answer more aggressively. In contrast, suffix-level injections are more often treated as auxiliary information and thus have a weaker influence on the final judgment.

5 Conclusions

In this work,we systematically investigate the influence of reasoning chains on LLM-based judgment. Our results show that weak models tend to blindly trust reasoning chains, leading to overconfidence, whereas strong models sometimes effectively leverage reasoning to detect errors. Additionally, our fine-grained analysis reveals that disrupting semantic coherence or factuality directly reduces model trust. In particular, injecting counterfactual content may cause the model to reject the answer. These findings suggest that the reasoning chains represent a promising direction for future research in LLM-as-a-Judge systems.

Limitations

While our study provides systematic insights into the impact of reasoning chains on LLM-based judgment, there are several limitations to consider. First, our investigation is confined to text-based benchmarks. The influence of reasoning in multimodal contexts (e.g., vision-language tasks) remains unexplored and represents a promising direction for future work. Second, due to computational constraints, we did not prompt the judge models to generate their own reasoning chains prior to delivering a verdict. Instead, we restricted the models to providing direct judgments without intermediate reasoning steps. Finally, although we evaluated a diverse set of representative models, the rapid evolution of proprietary LLMs means our coverage is inevitably not exhaustive. Future studies could extend our findings to a broader array of emerging reasoning models.

Ethical Considerations

All models and datasets used in this study are publicly available or accessed via official APIs. The datasets employed (NQ, HotpotQA, and GSM8K) are standard in the field and contain no personally identifiable information (PII) or offensive content. As this work focuses on evaluating model capabilities using existing resources, it introduces no additional societal risks or ethical concerns.

References

Adam P. Goucher Adam Perelman Aditya Ramesh Aidan Clark AJ Ostrow Akila Welihinda Alan Hayes Alec Radford Aleksander Ma dry Alex Baker-Whitcomb Alex Beutel Alex Borzunov Alex Carney Alex Chow Alex Kirillov Aaron Hurst, Adam Lerer and 1 others. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

AI@Meta. 2024. Llama 3.1 model card. Technical report, Meta AI. Online documentation, accessed 2026-01-05.

Anthropic. 2025. Claude opus 4.5 system card. Technical report, Anthropic. Accessed: 2026-01-05.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, and 1 others. 2024. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, 15(3):1–45.

Hongyu Chen and Seraphina Goldfarb-Tarrant. 2025. Safer or luckier? llms as safety evaluators are not robust to artifacts. *arXiv preprint arXiv:2503.09347*.

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

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, and 1 others. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 181 others. 2025. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Mingqi Gao, Xinyu Hu, Xunjian Yin, Jie Ruan, Xiao Pu, and Xiaojun Wan. 2025. Llm-based nlg evaluation: Current status and challenges. *Computational Linguistics*, pages 1–27.

Fan Huang, Haewoon Kwak, Kunwoo Park, and Jisun An. 2024. ChatGPT rates natural language explanation quality like humans: But on which scales? ELRA and ICCL.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, and 1 others. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

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

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

Arash Marioriyad, Mohammad Hossein Rohban, and Mahdieh Soleymani Baghshah. 2025. The silent judge: Unacknowledged shortcut bias in llm-as-a-judge. *arXiv preprint arXiv:2509.26072*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Maja Pavlovic and Massimo Poesio. 2024. The effectiveness of LLMs as annotators: A comparative overview and empirical analysis of direct representation. ELRA and ICCL.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Sijun Tan, Siyuan Zhuang, Kyle Montgomery, William Yuan Tang, Alejandro Cuadron, Chenguang Wang, Raluca Popa, and Ion Stoica. 2025. Judgebench: A benchmark for evaluating LLM-based judges. In *The Thirteenth International Conference on Learning Representations*.

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*.

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.

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*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 2369–2380.

Anqi Zhang, Yulin Chen, Jane Pan, Chen Zhao, Aurojit Panda, Jinyang Li, and He He. 2025. Reasoning models know when they're right: Probing hidden states for self-verification. *arXiv preprint arXiv:2504.05419*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Zhipu AI. 2025. Glm-4-32b model card. https://huggingface.co/zai-org/GLM-4-32B-0414.

A Prompts

short_qa

Answer the following question based on your internal knowledge with one or few words.

Question:{question}

llm_judge_without_think

Judge whether the following answer about the question is correct. If you are sure the answer is correct, say `certain`. If not, please say `uncertain`. Just give your judgment without any other words.

Question:{question}

Answer:{prediction}

llm_judge_with_think

Judge whether the following answer about the question is correct. The content inside <think></think> represents the reasoning process of the model, while the content after </think> is the answer provided by the model. If you are sure the answer is correct, say `certain`. If not, please say `uncertain`. Just give your judgment without any other words.

Question:{question}

Answer:{prediction}

llm_selfjudge

Answer the following question based on your internal knowledge with one or few words. Then, judge whether your answer is correct. If you are sure the answer is correct, say `certain`. If not, please say

`uncertain`. Your output should be in the following format:
Answer: <your answer>
Judge: <certain/uncertain>

Question:{question}

`basic_all`

The Earth orbits the Sun once every 365 days, producing the cycle of the seasons. Water freezes at 0 degrees Celsius and boils at 100 degrees Celsius at standard atmospheric pressure. Humans typically have 206 bones in the adult skeleton. The Pacific Ocean is the largest ocean on Earth, and Mount Everest is the tallest mountain above sea level.

`wrong_few`

The Earth orbits the Sun once every 100 days, producing the cycle of the seasons. Water freezes at 0 degrees Celsius and boils at 100 degrees Celsius at standard atmospheric pressure. Humans typically have 206 bones in the adult skeleton. The Pacific Ocean is the largest ocean on Earth, and Mount Everest is the tallest mountain above sea level.

`wrong_all`

The Earth orbits the Sun once every 100 days, producing the cycle of the seasons. Water freezes at 10 degrees Celsius and boils at 100 degrees Celsius at standard atmospheric pressure. Humans typically have 100 bones in the adult skeleton. The Pacific Ocean is the smallest ocean on Earth, and Mount Everest is the tallest mountain above sea level.

B Results Using Other Generators

Table 4: Evaluation results(%) of LLM-as-a-Judge behavior with and without reasoning chains across factual and mathematical datasets, with all answers generated by Qwen3-14B.

Dataset	Acc	Judge Models	Alignment		Pass Rate		Overconfidence		Conservativeness	
			w/o Think	w/ Think	w/o Think	w/ Think	w/o Think	w/ Think	w/o Think	w/ Think
NQ	31.2	Qwen3-8B	57.8	39.2	59.8	90.8	35.4	60.2	6.8	0.6
		Qwen3-14B	55.2	40.6	67.2	89.8	40.4	59.0	4.4	0.4
		Qwen3-32B	40.8	41.2	88.0	89.6	58.0	58.6	1.2	0.2
		Llama3-8B	40.2	32.6	87.8	98.6	58.2	67.4	1.6	0.0
		Llama3-70B	51.0	41.4	76.6	87.0	47.2	57.2	1.8	1.4
		GLM4-32B	47.6	37.0	82.0	94.2	51.6	63.0	0.8	0.0
		GLM4-Z1-32B	75.2	50.8	36.8	72.4	15.2	45.2	9.6	4.0
		GPT-4o	73.6	70.2	45.2	51.8	20.2	25.2	6.2	4.6
		DeepSeek-v3.1	62.8	61.8	61.6	62.2	33.8	34.6	3.4	3.6
		Claude Sonnet 4.5	77.2	77.2	14.0	10.4	2.8	10.0	20.0	21.8
HotpotQA	30.8	Qwen3-8B	56.2	47.0	57.4	75.0	35.2	48.6	8.6	4.4
		Qwen3-14B	53.6	49.0	66.0	75.8	40.8	48.0	5.6	3.0
		Qwen3-32B	44.0	49.4	83.2	75.0	54.2	47.4	1.8	3.2
		Llama3-8B	44.6	38.8	77.0	85.6	50.8	58.0	4.6	3.2
		Llama3-70B	52.6	49.6	69.4	76.0	43.0	47.8	4.4	2.6
		GLM4-32B	50.2	46.2	75.8	82.2	47.4	52.6	2.4	1.2
		GLM4-Z1-32B	74.6	59.2	14.6	50.8	4.6	30.4	20.8	10.4
		GPT-4o	77.2	75.6	29.2	32.4	10.6	13.0	12.2	11.4
		DeepSeek	62.4	64.6	55.6	51.4	31.2	28.0	6.4	7.4
		Claude	75.0	70.8	9.8	3.6	2.0	1.0	23.0	28.2
GSM8K	94.0	Qwen3-8B	72.8	94.8	72.4	99.2	2.8	5.2	24.4	0.0
		Qwen3-14B	70.4	94.8	71.2	99.2	3.4	5.2	26.2	0.0
		Qwen3-32B	86.8	94.6	89.6	99.4	4.4	5.4	8.8	0.0
		Llama3-8B	89.8	94.8	93.0	98.8	4.6	5.0	5.6	0.2
		Llama3-70B	83.0	95.2	85.8	98.0	4.4	4.4	12.6	0.4
		GLM4-32B	89.2	95.0	92.8	99.0	4.8	5.0	6.0	0.0
		GLM4-Z1-32B	47.4	86.4	45.8	87.6	2.2	3.6	50.4	10.0
		GPT-4o	93.4	92.0	94.2	94.0	3.4	4.0	3.2	4.0
		DeepSeek-v3.1	94.8	94.8	98.4	98.4	4.8	4.8	0.4	0.4
		Claude Sonnet 4.5	50.0	50.0	46.0	46.0	1.0	1.0	49.0	49.0

Table 5: Evaluation results(%) of LLM-as-a-Judge behavior with and without reasoning chains across factual and mathematical datasets, with all answers generated by Qwen3-32B.

Dataset	Acc	Judge Models	Alignment		Pass Rate		Overconfidence		Conservativeness	
			w/o Think	w/ Think	w/o Think	w/ Think	w/o Think	w/ Think	w/o Think	w/ Think
NQ	35.6	Qwen3-8B	54.6	41.2	63.8	93.6	36.8	58.4	8.6	0.4
		Qwen3-14B	56.2	40.6	69.4	94.2	38.8	59.0	5.0	0.4
		Qwen3-32B	42.6	41.4	90.2	93.4	56.0	58.2	1.4	0.4
		Llama3-8B	39.8	36.8	92.2	98.4	58.4	63.0	1.8	0.2
		Llama3-70B	56.2	47.2	73.4	86.8	40.8	52.0	3.0	0.8
		GLM4-32B	46.8	41.6	86.0	92.8	51.8	57.8	1.4	0.6
		GLM4-Z1-32B	71.8	49.8	36.6	82.2	14.6	48.4	13.6	1.8
		GPT-4o	74.6	72.0	51.8	56.8	20.8	24.6	4.6	3.4
		DeepSeek-v3.1	64.0	67.4	61.2	52.6	30.8	24.8	5.2	7.8
		Claude Sonnet 4.5	73.0	70.4	11.8	8.8	1.6	1.4	25.4	28.2
HotpotQA	34.4	Qwen3-8B	58.4	43.2	57.6	85.2	32.4	53.8	9.2	3.0
		Qwen3-14B	59.6	41.0	60.0	87.4	33.0	56.0	7.4	3.0
		Qwen3-32B	49.0	46.6	83.4	83.4	50.0	51.2	1.0	2.2
		Llama3-8B	47.6	39.2	78.8	92.0	48.4	59.2	4.0	1.6
		Llama3-70B	57.6	48.6	66.4	81.8	37.2	49.4	5.2	2.0
		GLM4-32B	54.8	45.6	74.8	87.6	42.8	53.8	2.4	0.6
		GLM4-Z1-32B	–	54.0	–	65.2	–	38.4	–	7.6
		GPT-4o	78.0	76.2	30.0	36.6	8.8	13.0	13.2	10.8
		DeepSeek-v3.1	60.2	64.0	63.4	54.4	34.4	28.0	5.4	8.0
		Claude Sonnet 4.5	71.2	69.8	8.4	5.0	0.4	1.4	27.4	29.8
GSM8K	95.8	Qwen3-8B	81.6	95.6	81.8	99.8	2.2	4.2	16.2	0.2
		Qwen3-14B	84.4	95.6	85.0	99.8	2.4	4.2	13.2	0.2
		Qwen3-32B	90.4	96.0	93.8	99.8	3.8	4.0	5.8	0.0
		Llama3-8B	94.4	96.2	98.2	99.2	4.0	3.6	1.6	0.2
		Llama3-70B	87.0	95.6	90.0	99.4	3.6	4.0	9.4	0.4
		GLM4-32B	94.2	96.0	98.4	99.8	4.2	4.0	1.6	0.0
		GLM4-Z1-32B	25.0	91.6	22.0	95.0	0.6	3.8	74.4	4.6
		GPT-4o	94.6	94.4	96.8	97.8	3.2	3.8	2.2	1.8
		DeepSeek-v3.1	94.8	95.8	98.6	99.6	4.0	4.0	1.2	0.2
		Claude Sonnet 4.5	60.2	57.2	58.0	54.6	1.0	0.8	38.8	42.0

Table 6: Evaluation results(%) of LLM-as-a-Judge behavior with and without reasoning chains across factual and mathematical datasets, with all answers generated by DeepSeek-v3.1.

Dataset	Acc	Judge Models	Alignment		Pass Rate		Overconfidence		Conservativeness	
			w/o Think	w/ Think	w/o Think	w/ Think	w/o Think	w/ Think	w/o Think	w/ Think
NQ	46.4	Qwen3-8B	75.8	50.6	46.6	94.2	12.2	48.6	12.0	0.8
		Qwen3-14B	75.8	49.0	54.2	95.4	16.0	50.0	8.2	1.0
		Llama3-8B	77.8	47.0	61.0	98.6	18.4	52.6	3.8	0.4
		GPT-4o	88.0	51.6	52.8	91.6	9.2	46.8	2.8	1.6
		DeepSeek-V3.1	80.4	49.0	46.0	95.8	9.6	50.2	10.0	0.8
		Claude Sonnet 4.5	90.0	55.8	48.8	69.8	6.2	33.8	3.8	10.4
GSM8K	95.6	Qwen3-8B	56.0	92.6	53.2	95.8	0.8	3.8	43.2	3.6
		Qwen3-14B	63.4	94.4	61.8	97.2	1.4	3.6	35.2	2.0
		Llama3-8B	51.8	94.2	51.8	97.8	2.2	4.0	46.0	1.8
		GPT-4o	66.4	95.0	64.0	97.0	1.0	3.2	32.6	1.8
		DeepSeek-v3.1	64.0	95.0	64.0	98.2	2.2	3.8	33.8	1.2
		Claude Sonnet 4.5	58.6	87.6	55.4	85.6	0.6	1.2	40.8	11.2