Can You Trust What I Think? Analyzing and Improving Verbalized Uncertainty and Factuality in Reasoning-Based Large Language Models

Tianruo Rose Xu

Cornell University
tx88@cornell.edu

Abstract

Reasoning-based large language models often produce natural-language thinking traces with their answers, but it remains unclear whether the verbalized uncertainties expressed in thinking traces faithfully reflect model's knowledge. We study this question on long-form, knowledge-intensive biography generation. Our pipeline decomposes thinking traces and responses into atomic facts, filters out planning content, labels factual reasoning by certainty, and aligns response facts to their supporting reasoning, enabling plan-based filtering, self-verification, and a classifier that predicts factuality from facts and associated reasoning. Preliminary results suggest that high-certainty reasoning is more likely to be included and correct and that structured use of these signals can improve factuality, though broader validation across models and dataset will be needed.

Introduction

Reasoning-based large language models increasingly "think out loud." Before answering, they produce a natural-language thinking trace that contains intermediate steps, candidate facts, and some expressions of uncertainty (Wei et al. 2022; Guo et al. 2025). These traces are often interpreted as windows into model's internal state and epistemic confidence, and they are used in practice to monitor hallucinations or decide which parts of an answer to trust (Efroni et al. 2025; Devic et al. 2025). At the same time, most reasoning models are trained with reinforcement learning pipelines that reward only the final answer, not faithfulness or calibration of the thinking trace that precedes it (Rafailov et al. 2025; Yao et al. 2025), so we do not yet know whether these verbalized uncertainties actually show what the model "knows," or whether they are mostly stylistic.

This work focuses on long-form, knowledge-intensive generation and asks whether the thinking traces of reasoning models are faithful and can be made useful. The main question we are asking is: Are reasoning LLMs' verbalized uncertainty in reasoning traces faithful indicators of factuality, and can this be operationalized to improve response quality? Breaking down the question, we investigate: (1) whether models follow the plans and certainty expressed in their thinking traces when writing the final answer; (2) whether these verbalized certainties align with factuality; and (3) whether enforcing or otherwise using these certainty labels (through plan-based filtering, self-verification, or an independent classifier) can help improve factuality.

Background

Several recent papers study how reinforcement finetuning and reasoning-style training affect hallucinations and uncertainty: work shows that RL pipelines can make models more willing to produce confident but incorrect content (Rafailov et al. 2025), Yao et al. find that reasoning models can be more prone to hallucination than non-reasoning baselines in knowledge-intensive tasks (Yao et al. 2025), and other work explicitly teaches models to reason for factuality or about their own uncertainty by shaping rewards or objectives (Chen et al. 2025; Damani et al. 2025), reporting that reasoning models may better express confidence than standard LLMs under some conditions (Efroni et al. 2025). A complementary line of work focuses on calibration and confidence signals derived from internal reasoning, for example using agreement across multiple sampled chains as a signal to better calibrate predictions (Xie et al. 2024) or arguing for a broader shift "From Calibration to Collaboration," where uncertainty quantification is evaluated by how it supports human decision-making (Devic et al. 2025). HCI work further shows that surface forms of uncertainty expression strongly influence user reliance and trust (Kim et al. 2024) and that humans can over-interpret coherent but incorrect reasoning narratives (Levy, Elyoseph, and Goldberg 2025).

Approach

Our primary subjects are the DeepSeek distill reasoning models, including Llama models with 8B and 70B parameters and Qwen models with 1.5B, 7B, and 14B parameters, and we plan to extend the analysis to additional reasoning model families over time. For each model, we run long-form biography generation on the FActScore biographies dataset, where each input produces a chain-of-thought-style thinking trace and a corresponding long-form biography response which is evaluated at the level of atomic facts (Min et al. 2023).

To analyze the thinking traces, we deign the following pipeline.

- We first split each trace into sentences and use an LLM to label each as either planning (how to answer, meta-commentary, or surface phrasing) or non-planning factual reasoning, discarding planning sentences and retaining only factual reasoning for further analysis.
- For each remaining reasoning sentence, we then assign a certainty label of speculation, low certainty, or high certainty using an LLM-as-a-judge based on verbalized confidence or hedging expressions in the text.
- We perform atomic fact extraction and factuality evaluation by decomposing both the filtered thinking trace and the final response into atomic facts, evaluating each atomic fact against Wikipedia using the FActScore pipeline.
- Finally, we align response facts to reasoning by asking an LLM judge, for each atomic response fact, to identify possible supporting sentences in the thinking trace.

To operationalize verbalized uncertainty for improving factuality, we design several methods.

- First, we construct plan-based filtering pipelines that remove response facts by dropping facts with no supporting thinking sentence, then dropping facts supported only by speculative reasoning, then removing those supported by low-certainty reasoning, and adding back high-certainty thinking facts that did not appear in the original response.
- Second, we study self-verification with the same generation model by prompting it, for each atomic fact, to classify the fact as true or false under two conditions (one with the full thinking trace in context and one with the fact presented alone) and then comparing performance.
- Finally, we train a small external verifier, Qwen2.5-0.5B-Instruct, to predict factuality from inputs that include the model name, the topic (which may or may not be explicitly given), the atomic fact, and the matched thinking sentences (which may or may not be provided).

Evaluation

We evaluate on the FActScore biography dataset using the DeepSeek reasoner and its distill models. We first assess the thinking-trace labeling by reporting the proportion of sentences classified as planning versus non-planning and as speculation, low certainty, or high certainty, together with FActScore for extracted thinking facts in each sentence before and after matching them to response facts. We then measure plan-following as the fraction of thinking facts in each certainty bucket that are included in the response, compute FActScore separately for matched and unmatched thinking facts, and report FActScore for response facts overall and split by whether they are supported by any thinking sentence, along with the overall matched versus unmatched rate.

To evaluate the effect of our operationalization, we compare the original generations to plan-based filtering baselines (matched-only, no speculation, and high-certainty-only), to self-verification using a cross-verifier with either full-context or fact-only input, and to a small Qwen2.5-0.5B-Instruct classifier evaluated under anonymized versus topic-aware and fact-only versus fact-plus-trace settings. For the verifier and classifier, we track accuracy, F1, and the resulting FActScore at comparable or higher coverage, and treat consistent gains in these metrics as evidence that using AI to analyze and act on verbalized uncertainty improves factuality.

Discussion

On the FActScore dataset and DeepSeek distill models, we find that models do follow the plan expressed in the thinking trace: they include high-certainty facts in the final response at a higher rate than facts associated with low certainty or speculation, although the latter are still included with a bit lower rate. The expressed uncertainty aligns with factual correctness to some degree, but only for larger models with 8B or bigger.

Different strictness levels of plan enforcement, such as dropping facts with speculative or low-certainty support, can improve FActScore, though gains are sensitive to the trade-off between precision and the number of retained facts. We also observe that LLMs can self-verify generations when given just the atomic fact, but are more easily misled when the full thinking trace is present, suggesting that verbalized uncertainty is not always beneficial. Most importantly, the small classifier performs best when given both the fact and the matched thinking sentences, indicating that reasoning traces contain some useful signal, while some characteristics of the fact (e.g. generality) still plays a major role in the decision. If this could be replicated across more datasets and model families, the use of thinking traces could inform how future reasoning models expose their internal uncertainty to users and these can be used to improve factuality and reduce hallucinations.

Conclusion

We study whether verbalized uncertainty in reasoning-based LLMs' thinking traces is a faithful indicator of factuality and how it can be operationalized to improve long-form generation. Focusing on biography generation in the FActScore benchmark, we decompose thinking traces and responses into atomic facts and use LLM-based alignment, certainty labeling, and verification. The evaluation on DeepSeek distill models shows that high-certainty reasoning is more likely to appear in the final answer and be factually correct, and that plan-based filtering, self-verification, and a small independent classifier can each use some information in the thinking traces to improve factuality. Although our experiments center on biographies generation, the same techniques could transfer to other areas: medical information where clinical or patient-facing tools must avoid plausible but unsupported claims, and scientific assistants, where long-form explanations and literature summaries require accurate citations and honest uncertainty. More broadly, aligning verbalized uncertainty with factuality points toward safer, more transparent reasoning models that are better in applications.

References

Chen, X.; Kulikov, I.; Berges, V.-P.; Oğuz, B.; Shao, R.; Ghosh, G.; Weston, J.; and t. Yih, W. 2025. Learning to Reason for Factuality. arXiv:2508.05618.

Damani, M.; Puri, I.; Slocum, S.; Shenfeld, I.; Choshen, L.; Kim, Y.; and Andreas, J. 2025. Beyond Binary Rewards: Training LMs to Reason About Their Uncertainty. arXiv:2507.16806.

Devic, S.; Srinivasan, T.; Thomason, J.; Neiswanger, W.; and Sharan, V. 2025. From Calibration to Collaboration: LLM Uncertainty Quantification Should Be More Human-Centered. arXiv:2506.07461.

Efroni, Y.; Song, Q.; Xia, M.; Schmidt, F.; Allen-Zhu, Z.; Kim, Y.; Li, J. D.; and Le, Q. 2025. Reasoning Models Better Express Their Confidence. arXiv:2505.14489.

Guo, A.; et al. 2025. DeepSeek R1: Incentivizing Reasoning Capability in Large Language Models via Reinforcement Learning. arXiv:2501.12948.

Kim, S. S. Y.; Liao, Q. V.; Vorvoreanu, M.; Ballard, S.; and Vaughan, J. W. 2024. "I'm Not Sure, But...": Examining the Impact of Large Language Models' Uncertainty Expression on User Reliance and Trust. In *FAccT*.

Levy, M.; Elyoseph, Z.; and Goldberg, Y. 2025. Humans Perceive Wrong Narratives from AI Reasoning Texts. arXiv:2508.16599.

Min, S.; Krishna, K.; Lyu, X.; Lewis, M.; Yih, W.; Koh, P. W.; Iyyer, M.; Zettlemoyer, L.; and Hajishirzi, H. 2023. FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation. In *EMNLP*.

Rafailov, R.; Yue, X.; Kim, Y.; Zhou, D.; Hashimoto, T. B.; Li, J. D.; Susskind, J.; and Le, Q. 2025. The Hallucination Tax of Reinforcement Finetuning. arXiv:2505.13988.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E.; Le, Q.; and Zhou, D. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *NeurIPS*.

Xie, Z.; Guo, J.; Yu, T.; and Li, S. 2024. Calibrating Reasoning in Language Models with Internal Consistency. In *NeurIPS*.

Yao, Z.; Ye, X.; Feng, S.; Liu, Z.; Meyers, L. A.; and Durrett, G. 2025. Are Reasoning Models More Prone to Hallucination? arXiv:2505.23646.