Reasoning about Uncertainty: Do Reasoning Models Know When They Don't Know?

Zhiting Mei[*], **Christina Zhang**, **Tenny Yin**,
Justin Lidard, **Ola Shorinwa**[*], **Anirudha Majumdar**
Princeton University

Abstract

Reasoning language models have set state-of-the-art (SOTA) records on many challenging benchmarks, enabled by multi-step reasoning induced by reinforcement learning. However, reasoning models are prone to generating confident, plausible responses that are incorrect (hallucinations). Knowing when and how much to trust these models is critical for safe deployment in real-world applications. To this end, we explore uncertainty quantification (UQ) of reasoning models in the form of self-verbalized confidence in this work. We ask three fundamental questions: First, *are reasoning models well-calibrated?* Second, *does deeper reasoning improve model calibration?* Finally, inspired by humans' innate ability to double-check their thought processes to verify the validity of their answers and their confidence, we ask: *can reasoning models improve their calibration by explicitly reasoning about their chain-of-thought traces?* We introduce *introspective uncertainty quantification* (IUQ) to explore this direction. In extensive evaluations on SOTA reasoning models across a broad range of benchmarks focused on knowledge-intensive tasks, we find that reasoning models: (i) are typically overconfident, (ii) become even more overconfident with deeper reasoning, and (iii) can become better calibrated through introspection (e.g., o3-Mini and DeepSeek R1) but not uniformly (e.g., Claude 3.7 Sonnet becomes more poorly calibrated). We conclude with important research directions to design necessary UQ benchmarks and improve the calibration of reasoning models.

1 Introduction

Recent breakthroughs in large language models (LLMs) and vision-language models (VLMs) have largely been driven by *reasoning* language models—language models that are trained to perform multi-step reasoning through reinforcement learning, e.g., Claude 3.7 (Anthropic, 2024), o3-Mini (OpenAI, 2024b), Gemini 2 Flash Thinking (Team et al., 2024), and DeepSeek R1 (Guo et al., 2025). By breaking down complex tasks into subtasks and analyzing and synthesizing responses from the resulting subtasks, reasoning models have achieved new state-of-the-art (SOTA) performance on a broad range of challenging real-world tasks, such as programming, e.g., Codeforces (Mikhail Mirzayanov and Codeforces Team, 2025); math, e.g., AIME (Art of Problem Solving, 2025); engineering and science, e.g., GPQA (Rein et al., 2023); and arts, e.g., MMLU (Hendrycks et al., 2020), outperforming domain experts on many benchmarks.

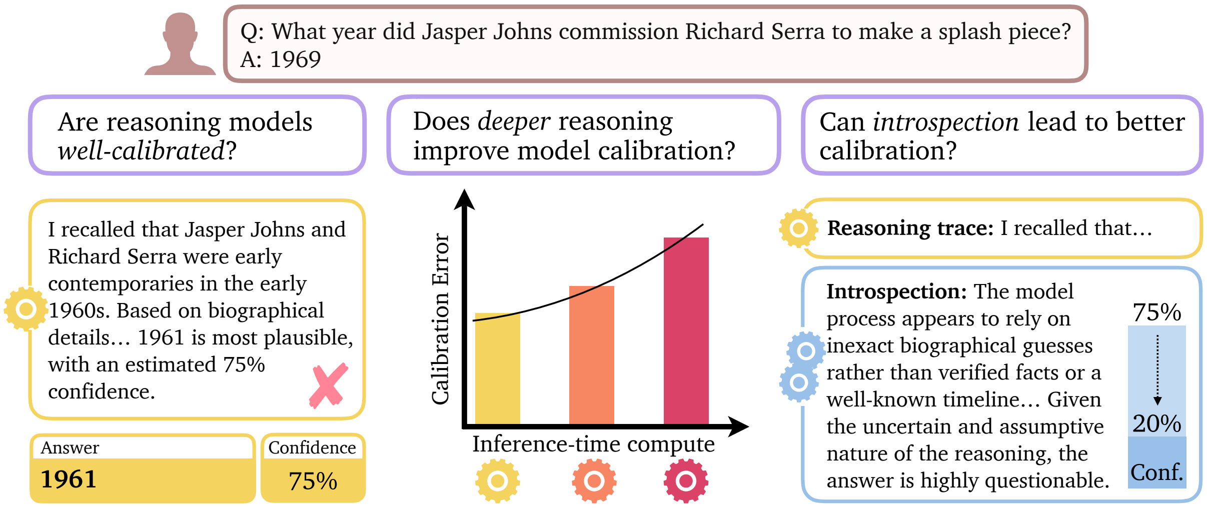
Despite their remarkable capabilities, SOTA reasoning language models often generate confident but factually-incorrect responses to questions, known as *hallucinations* (Bai et al., 2024; Liu et al., 2024), similar to their non-reasoning counterparts. Consequently, the safe integration of reasoning models into real-world tasks requires knowing when and how much to trust these models, i.e., uncertainty quantification (UQ) of these models. Prior work has examined UQ of non-reasoning models, showing that non-reasoning models are poorly calibrated in general (Xiong et al., 2023) but can become better calibrated through chain-of-thought prompting (Liang et al., 2024). However, UQ of reasoning models remains an unexplored frontier. To address this knowledge gap, we examine uncertainty quantification of reasoning models to better evaluate their trustworthiness. Considering that SOTA reasoning models are often closed-source, making access to their internal logits difficult if not impossible, we leverage self-verbalized confidence as the primary method for uncertainty quantification, following prior works (Lin et al., 2022; Shorinwa et al., 2025).

Specifically, in this work, we ask three critical questions (illustrated in Figure 1). First, *are reasoning models well-calibrated?* We evalu-

---

[*] Equal Contribution.

Findings of the Association for Computational Linguistics: EACL 2026, pages 3408–3458
March 24-29, 2026 ©2026 Association for Computational Linguistics


Q: What year did Jasper Johns commission Richard Serra to make a splash piece?
A: 1969

Are reasoning models *well-calibrated*?

Does *deeper* reasoning improve model calibration?

Can *introspection* lead to better calibration?

I recalled that Jasper Johns and Richard Serra were early contemporaries in the early 1960s. Based on biographical details… 1961 is most plausible, with an estimated 75% confidence.

Reasoning trace: I recalled that…

Introspection: The model process appears to rely on inexact biographical guesses rather than verified facts or a well-known timeline… Given the uncertain and assumptive nature of the reasoning, the answer is highly questionable.

Figure 1: To examine if reasoning models know when they don't know, we ask three fundamental questions: (i) are reasoning models well-calibrated? (ii) does deeper reasoning improve model calibration? and (iii) can introspection improve calibration? We find that reasoning models are typically overconfident and become more overconfident with deeper reasoning. However, through introspection, reasoning models can provide better calibrated confidence estimates.

ate the calibration of reasoning models across a broad range of benchmarks to examine underconfidence/overconfidence in these models across different problem domains. Our findings reveal that *reasoning models are typically overconfident,* often expressing confidence estimates greater than 85% (in increments of 5%) even when they are incorrect. Given that overconfidence is often masked by high accuracy, we find that reasoning models are well-calibrated in knowledge-retrieval tasks and poorly calibrated in tasks requiring domain expertise, e.g., graduate-level research domains. Moreover, our results highlight that prompt engineering is significantly less effective with reasoning models, unlike non-reasoning models.

Second, *does deeper reasoning improve model calibration?* Prior work (OpenAI, 2024b) has shown that reasoning models can achieve higher accuracy through inference-time scaling of reasoning effort, measured by the number of available reasoning tokens. Here, we explore the effects of deeper reasoning on model calibration across low, medium, and high amounts of reasoning. Our evaluations reveal that *reasoning models become more overconfident with deeper reasoning.* Specifically on wrongly-answered questions, reasoning models become significantly more overconfident as we scale the reasoning effort.

Lastly, humans often double-check their problem-solving approach to verify their answers and estimate their confidence in the validity of the answers (Hewitt et al., 2016). In line with this observation, we ask the question: *can reasoning models improve their calibration by explicitly reasoning about their chain-of-thought traces?* We introduce *introspective uncertainty quantification* to explore this question. Introspective UQ utilizes a multi-stage reasoning paradigm to estimate the confidence in the accuracy of a given response: a model first reasons about a given question, and then a second model reasons about the chain of thought and final answer produced by the first model in order to assess its confidence in the answer. We find that *more critical introspection improves the calibration of reasoning models, e.g., o3-Mini and DeepSeek R1, although the calibration of Claude 3.7 Sonnet worsens.*

2 Related Work

Reasoning in LLMs. Prior work (Wei et al., 2022; Kojima et al., 2022; Wang et al., 2022) has demonstrated that language models can achieve significantly higher accuracy by generating intermediate reasoning steps, known as a chain of thought (CoT). Chain-of-thought prompting enables remarkable zero-shot or few-shot learning in LLMs, enabling generalist LLMs to outperform task-specific, fine-tuned models (Wei et al., 2022). However, the complexity of the few-shot CoT examples significantly influences the induced accuracy of the models (Fu et al., 2022), spurring research on prompt engineering for effective chain-of-thought reasoning (Zhang et al., 2022; Huang and Chang, 2022; Hao et al., 2024). Subsequent work introduces multi-path reasoning strategies to enable backtracking on incorrect reasoning paths: e.g., Tree-of-Thought prompting (Yao et al., 2023) and Graph-of-Thoughts prompting (Besta et al., 2024), which construct parallel reasoning traces; least-to-most prompting (Zhou et al., 2022), which solves a series of subproblems in sequence; and reasoning via planning (Hao et al., 2023), which performs Monte Carlo Tree Search on a reasoning tree to find the optimal reasoning path. However, existing approaches require domain expertise to construct effective chain-of-thought prompts/examples, posing a challenge. By training language models to reason using reinforcement learning (RL), recent work (Jaech et al., 2024; Guo et al., 2025; Anthropic, 2024; OpenAI, 2024b; Team et al., 2024; Team, 2025) has shown that reasoning language models can solve complex tasks more accurately than non-reasoning LLMs. Through inference-time scaling, reasoning models can be made to think deeper to break down complex problems into multiple steps, achieving SOTA performance on many benchmarks (Anthropic, 2024; OpenAI, 2024b; Team et al., 2024; Guo et al., 2025). However, the effects of reasoning on the calibration of confidence estimates of these models remains unknown, motivating uncertainty quantification.

Uncertainty Quantification of LLMs. Uncertainty quantification of LLMs has gained notable prominence given the tendency of LLMs to hallucinate. In general, existing UQ methods for LLMs eschew traditional techniques such as Bayesian networks (Jospin et al., 2022) due to their computational cost, given that LLMs often have billions to trillions of parameters. UQ methods for LLMs can be grouped into two broad classes: *white-box* and *black-box* methods (Shorinwa et al., 2025). White-box methods utilize the internal outputs of LLMs (Geva et al., 2020; Azaria and Mitchell, 2023), such as tokens for entropy/perplexity-based UQ (Manakul et al., 2023; Fadeeva et al., 2024; Ling et al., 2024; Xiao and Wang, 2021; Kuhn et al., 2023) or neural activations for mechanistic interpretability (MI) (Belinkov, 2022; Dunefsky et al., 2024; Yun et al., 2021; Gao et al., 2024; Tamkin et al., 2023; Ferrando et al., 2024). While these UQ metrics can be well calibrated for pre-trained models (Kadavath et al., 2022), reinforcement learning from human feedback (RLHF) leads to poorly calibrated token-level UQ (OpenAI, 2024a) since RL-training objectives are not proper scoring rules (Gneiting and Raftery, 2007). In contrast to white-box methods, black-box methods estimate uncertainty without access to the model's internal outputs which might not be available, given the notable shift from fully open-source language models to open-weight or closed-source models, e.g., (Hurst et al., 2024; Anthropic, 2024). In fact, state-of-the-art closed-source reasoning models generally do not provide access to the models' logits, which is required for token-based, semantic similarity-based, or MI-based UQ. Consequently, self-verbalized UQ methods (Mielke et al., 2022; Smith et al., 2020; Tanneru et al., 2024; Aichberger et al., 2024; Zhang et al., 2024), which directly prompt an LLM for its confidence in natural language, have become increasingly important. In contrast to prior work which has focused exclusively on non-reasoning models, we assess the uncertainty of reasoning models without finetuning, using self-verbalized UQ methods given the lack of access to the internal outputs.

3 Method

As noted in the preceding discussion, widely used state-of-the-art reasoning models do not provide native access to the models' logits or other internal outputs, motivating us to leverage self-verbalized uncertainty quantification methods as a broadly applicable approach without relying on the internal features of the reasoning model. We seek to maximize the practical effectiveness of our work, given the widespread use of reasoning models, especially black-box reasoning models like GPT, Claude, etc. However, a naïve implementations of self-verbalized UQ are prone to subjective bias arising from the different cultural influences. We specifically guard against this bias in our evaluations, by only asking for **numerical** estimates of the model's confidence to avoid all forms of language-related bias, e.g., in tone or word choice.

This challenge motivated the use of self-verbalized uncertainty quantification methods as a widely-applicable approach without relying on the internal features of the reasoning model. We agree that a naïve implementation of self-verbalized UQ is prone to subjective bias, given that different cultures interpret confidence in different ways. We specifically guard against this bias in our evaluations, by only asking for **numerical** estimates of the model's confidence. By so doing, we avoid all language-related bias covering tone, word choice, etc. In order to effectively explore uncertainty quantification in reasoning language models, we consider a broad range of prompting strategies to examine the ability of reasoning models to accurately express their confidence. We draw insights from existing work on prompt engineering to identify potentially effective prompt designs for reasoning models. Beyond the established prompt strategies, we examine the effects of *introspection* on a reasoning model's ability to quantify its uncertainty. We call this approach *introspective uncertainty quantification*. We discuss these strategies subsequently.

3.1 Prompt Strategies

We design prompts for self-verbalized uncertainty quantification of reasoning models following prompt templates for LLMs introduced in prior work. Specifically, we consider zero-shot basic prompting, chain-of-thought prompting (Wei et al., 2022; Kojima et al., 2022), multi-step prompting (Xiong et al., 2023), and Top-K prompting (Tian et al., 2023). We emphasize that the LLM is prompted zero-shot, i.e., without any examples. We provide the full prompts in Appendix D.

Basic Prompt Strategy. In the basic prompt strategy, we provide the question directly to the LLM and ask for its answer to the question. For example, we ask the LLM to provide its answer and confidence without any description of the problem-solving approach.

Chain-of-Thought Prompt Strategy. Prior work (Wei et al., 2022; Kojima et al., 2022; Zhang et al., 2022) has shown that LLMs are highly capable zero-shot reasoners when asked to *think step by step*. Although prior work has demonstrated the effectiveness of chain-of-thought prompting in improving accuracy, the effects of chain-of-thought prompting on calibration of self-verbalized UQ estimates remain relatively unexplored. We explore this research direction in this paper. Specifically, we ask the LLM to *analyze* the question *step by step* to induce intermediate reasoning and ask for its confidence.

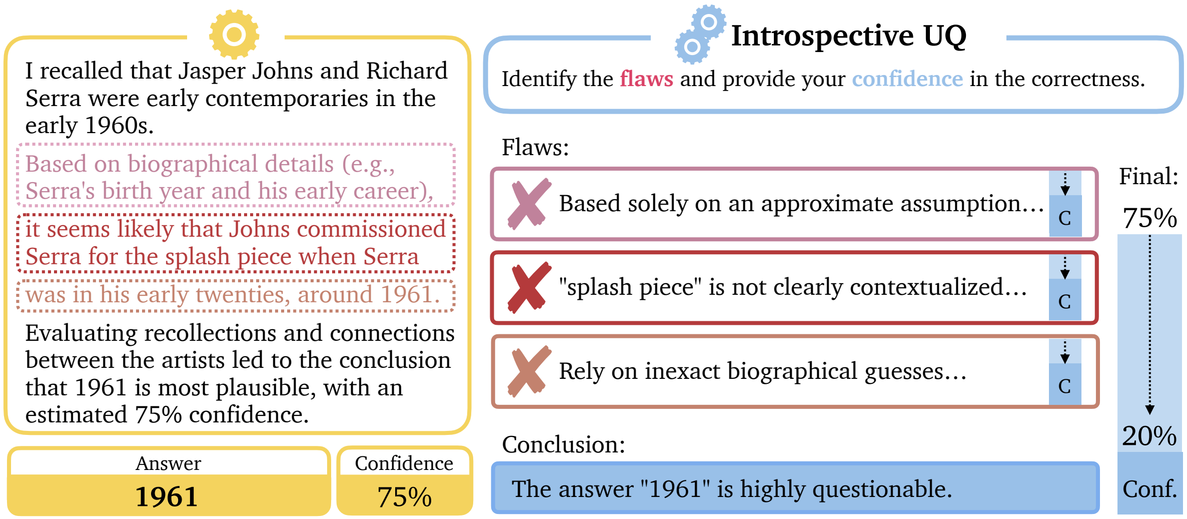
Multi-Step Prompt Strategy. The multi-step prompt strategy takes chain-of-thought prompting a step further. In the multi-step prompt strategy, we ask the LLM not only to reason about the problem, but to also provide its confidence at each step. By doing so, we hope to induce the LLM to reason about its confidence to improve the calibration of its self-verbalized confidence estimates. We ask the LLM to *break down* the problem into *multiple steps* and provide its answer and confidence in each step.

Top-K Prompt Strategy. Identifying candidate answers to a given question (and in particular, how likely these answers are to be correct) can result in better estimation of the model's confidence, as demonstrated in prior work (Tian et al., 2023), where the LLM is asked to provide its top-K guesses for the answer to a question and the associated confidence. We explore this prompt strategy for reasoning models in our work.

3.2 Introspective Uncertainty Quantification

We explore the effect of reasoning on the calibration of uncertainty quantification of reasoning models (an open question in LLM research). Drawing inspiration from how humans double-check their thought processes to ascertain their confidence in a given response, we introduce *introspective uncertainty quantification*: a self-verbalized UQ method which utilizes a multi-stage reasoning architecture to reason about and update the confidence of reasoning models. Figure 2 illustrates the two-stage introspective UQ procedure using the question shown in Figure 1, where a reasoning model is asked to identify flaws in the prior reasoning trace and to provide its confidence on the correctness of the first answer. The model identifies weaknesses in the chain-of-thought traces and correctly reduces the confidence in the original response, ultimately improving its calibration.

At the initial stage of introspective UQ, we ask a reasoning model to answer a question and provide its confidence in the answer. In subsequent stages, we provide the reasoning model with the reasoning traces at earlier stages and ask the model to analyze these traces to identify flaws and to ultimately provide a confidence estimate in the answer provided in the initial stage. Specifically in the second stage, we consider three kinds of prompts for introspection defined by varying conservativeness — (i) *IUQ-Low*: a neutral prompt asking the model to reason about its uncertainty given earlier reasoning traces and associated confidence estimates (without any hints on possible flawed reasoning), (ii) *IUQ-Medium*: a more *conservative* prompt explicitly asking the model to find flaws in the input reasoning traces including the prior confi-



I recalled that Jasper Johns and Richard Serra were early contemporaries in the early 1960s. Based on biographical details (e.g., Serra's birth year and his early career), it seems likely that Johns commissioned Serra for the splash piece when Serra was in his early twenties, around 1961. Evaluating recollections and connections between the artists led to the conclusion that 1961 is most plausible, with an estimated 75% confidence.

Identify the flaws and provide your confidence in the correctness.

Based solely on an approximate assumption…

"splash piece" is not clearly contextualized…

Rely on inexact biographical guesses…

The answer "1961" is highly questionable.

Figure 2: **Introspective UQ.** Although the reasoning model is initially highly overconfident and wrong (as shown in Figure 1), through introspection, the model can identify flaws in its chain-of-thought traces to provide better calibrated confidence estimates, mitigating overconfidence.

dence estimates, and (iii) *IUQ-High*: the most *conservative* prompt which utilizes the same prompt as *IUQ-Medium* without the prior confidence estimates. We provide a template of these prompts in Appendix Section C. Although introspective UQ can utilize many stages, we limit the number of stages to two in our work and leave an exhaustive examination of the optimal number of stages to future work.

We restrict the second stage of introspective UQ to *reasoning about uncertainty* and do not ask the model to update the answer to the original question for two key reasons: (i) to encourage focused reflection on prior reasoning traces, and (ii) to guard against hallucinations, especially in problems where the model is not confident. Prior work (Liu et al., 2024) has shown that language models tend to hallucinate or vacillate between different choices when prompted with a question that they are not confident about, e.g., in knowledge-retrieval problems. Such vacillations present unnecessary distractions, undermining the effectiveness of introspective UQ.

With introspective UQ, we seek to elicit human-like internal reflections in reasoning models for calibrated uncertainty quantification, which could be important in downstream applications such as hallucination detection and mitigation. However, we note that humans sometimes become unjustifiably more confident in their responses after analyzing their reasoning process (Shynkaruk and Thompson, 2006), a challenge that could also arise with reasoning models. We explore these questions in the experiments discussed in Section 4.

4 Experiments

We evaluate the calibration and accuracy of reasoning models on benchmark datasets and assess the confidence estimates of reasoning models using the self-verbalized UQ methods presented in Section 3. Specifically, we seek to answer the following questions: (i) Are reasoning models calibrated, systemically underconfident, or overconfident? (ii) Is accuracy correlated with calibration (positively or negatively, strongly or weakly)? (iii) Does deeper reasoning (e.g., greater number of reasoning steps) improve calibration? (iv) Can reasoning models reason about their uncertainty for better calibration? Lastly, we explore the effects of different prompt strategies on the calibration and accuracy of reasoning models in Appendix J.

4.1 Evaluation Setup

We summarize the experiment setup, including the reasoning models, datasets, prompt strategies, and evaluation metrics.

Models. We evaluate SOTA reasoning models that are publicly accessible via an API, namely: (i) Claude 3.7 Sonnet (*claude-3-7-sonnet-20250219*, released 02/19/2025), (ii) DeepSeek R1 (*deepseek-reasoner*, released 01/20/2025), (iii) Gemini (*gemini-2.0-flash-thinking-exp*, released 12/19/2024), and (iv) o3-mini (*o3-mini*, released 01/31/2025). We do not benchmark Grok 3 Beta, since we were not able to get access to the model. With these models, we examine calibration across different model sizes.

Datasets. We consider five standard question-answering benchmark datasets, namely: (i) *ARC-Challenge* Set (Clark et al., 2018), (ii) Graduate-Level Google-Proof Q&A (*GPQA*) Benchmark (Rein et al., 2023), (iii) Measuring Massive Multitask Language Understanding (*MMLU*) dataset (Hendrycks et al., 2020),(iv) *StrategyQA* (Geva et al., 2021), and (v) *SimpleQA* (Wei et al., 2024). All these datasets are open-sources; hence, our results can be easily reproduced.

Prompt Strategies and Metrics. We utilize the prompt strategies discussed in Section 3.1, with $K = 5$. We report the calibration of the self-verbalized confidence estimates from the model using the expected calibration error (ECE) and maximum calibration error (MCE) (see Section A for a brief review). Further, we provide the accuracy, successful query-completion rate (completion) to show how well the models follow instructions, and reliability diagram to show the deviation between the bin-wise confidence and accuracy as a measure of underconfidence or overconfidence.

4.2 Calibration of Reasoning Models

Here, we assess the uncertainty of reasoning models to characterize their calibration across the aforementioned benchmarks. Figure 3 summarizes the ECE, MCE, accuracy, and completion rate of each model across these problem domains, showing the average performance of each model over the four prompt strategies. We provide each model's individual performance for each prompt strategy in Appendix F. In general, we find that reasoning models are not well-calibrated, particularly in challenging problem domains (e.g., SimpleQA). For example, the original GPT-4 without RLHF was better calibrated with an ECE of 0.7 on a subset of the MMLU dataset (OpenAI, 2024a), which is significantly lower than the ECE of all the models in Figure 3. Specifically, reasoning models tend to be overconfident, although this issue is less obvious on benchmarks where models achieve near-perfect accuracy (e.g., ARC-Challenge). Next, we provide a detailed discussion of these results.

Model Calibration on Benchmarks. From Figure 3, reasoning models are better calibrated on ARC-Challenge and MMLU; however, calibration of these models degrades significantly on StrategyQA, GPQA, and SimpleQA. These results suggest that reasoning models are better calibrated on older benchmarks, e.g., ARC-Challenge (released in 2018) and MMLU (released in 2020), compared to more recent benchmarks where the models are poorly calibrated, e.g., StrategyQA (released in 2021), GPQA (released in 2023), and SimpleQA (released in 2024). We observe that new benchmarks are often created due to performance saturation on older ones. We hypothesize that the difference in calibration of these models can be explained by the difference in accuracy across these benchmarks and explore these connections in Section 4.3. The results highlight that calibration on a particular dataset is not always indicative of calibration on another dataset, especially in cases with near-perfect model accuracy where overconfidence is difficult to detect. This finding raises a common, important challenge in benchmarking LLMs: *existing benchmarks become relatively uninformative for UQ as new reasoning models are trained to beat these benchmarks, necessitating the introduction of newer benchmarks.* Moreover, from the reliability diagrams, we find that reasoning models tend to be overconfident and observe that Claude 3.7 Sonnet is the most calibration model with o3-Mini being the least calibrated (see Appendix F).

4.3 Correlation Between Accuracy and Calibration of Reasoning Models

The preceding results showed that reasoning models are poorly calibrated in datasets where these models have lower accuracies and better calibrated otherwise. Here, we investigate the relationship between accuracy and calibration of reasoning models by examining the correlation between these two metrics. We compute the average accuracy and ECE over the four prompt strategies per model, aggregating statistics among 20 data points.

In Figure 4, we observe a strong negative correlation between accuracy and ECE across all models, which in the case of Gemini, is a perfect negative correlation, i.e., $r = -1$. Similarly, DeepSeek and o3-Mini exhibit near-perfect correlation. This finding underscores that reasoning models are poorly calibrated. As shown in Figure 7, reasoning models are overconfident, which explains the strong correlation between accuracy and calibration. In general, if an overconfident model consistently outputs high confidence values, then its ECE will decrease as its accuracy increases. The confidence of a well-calibrated model should reflect its accuracy,

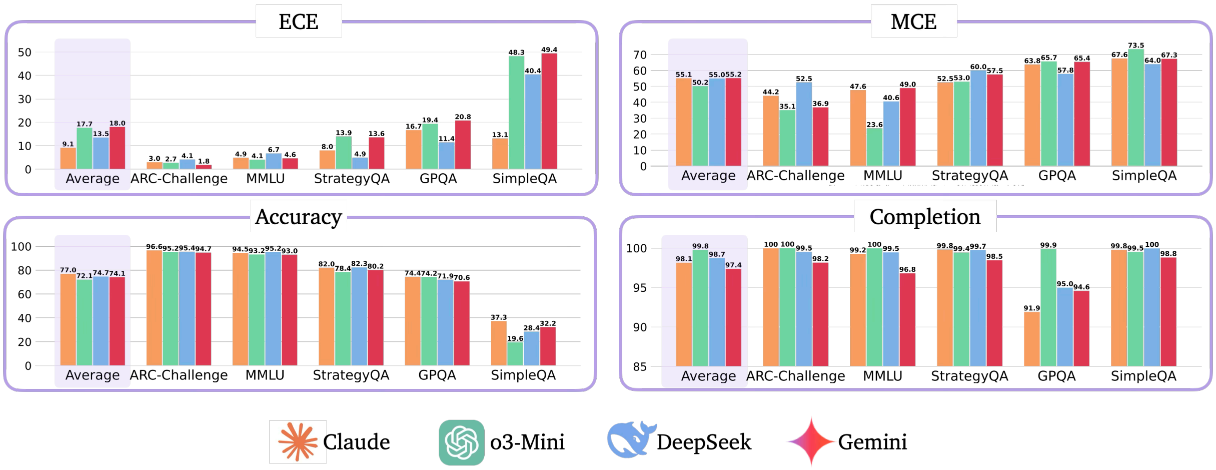


Figure 3: **Are reasoning models calibrated?** Reasoning models are well-calibrated on the *ARC-Challenge* and *MMLU* benchmarks and poorly calibrated on *StrategyQA*, *GPQA*, and *SimpleQA*, suggesting better model calibration on older benchmarks and worse calibration on more recent benchmarks.

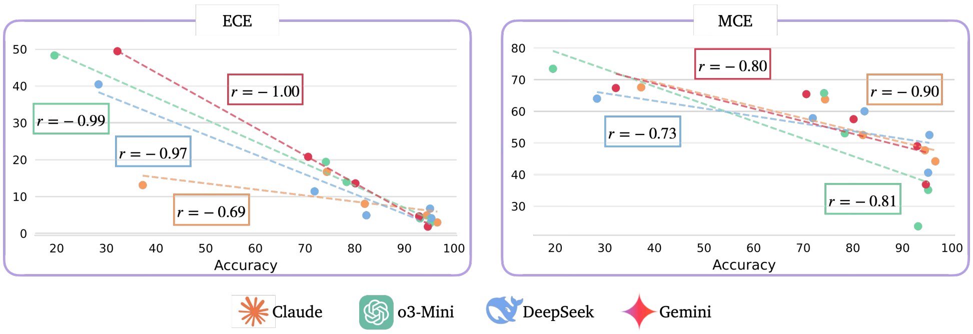


Figure 4: **Is accuracy correlated with calibration in reasoning models?** The ECE is strongly correlated with accuracy in state-of-the-art reasoning models. In essence, the self-verbalized confidence of reasoning models is generally not trustworthy in problems where they achieve low accuracy. Remarkably, Claude remains well-calibrated even in these low-accuracy settings. We observe the same trends in the correlation between accuracy and the MCE.

represented by a near-zero correlation coefficient between accuracy and ECE. However, unlike the other models, Claude has a much weaker correlation with $r = -0.69$ (28% smaller than that of the second-best-performing model DeepSeek), showing its relatively better calibration.

Further, these results suggest that self-verbalized confidence estimates from reasoning models should be taken with great care, especially with poorly calibrated models. Lastly, as with the ECE, accuracy and MCE have a strong negative correlation, with DeepSeek showing the weakest correlation. Training reasoning models to not only be accurate but also be well-calibrated remains a fundamental research challenge, as highlighted by these results. We discuss the correlation between calibration error and accuracy for each prompt in Appendix G. We explore open research problems and potential future research directions in Section 6.

4.4 Reasoning Depth vs. Calibration

One of the most appealing properties of reasoning models is *inference-time scaling*: the ability to boost the accuracy of responses via *deeper reasoning* (albeit at the expense of increased inference time and energy costs). However, the impact of deeper reasoning on calibration has not been explored in prior work. Here, we examine the coupled effects of reasoning depth on the accuracy and calibration of reasoning models, highlighting some notable findings. We summarize our results in Figure 5, evaluating Claude 3.7 Sonnet and o3-Mini since DeepSeek-R1 and Gemini do not sufficiently support/utilize the specification of a reasoning token budget. We provide detailed results for each model in Appendix H. At the time of the experiments, finer control over other open-source reasoning models was not yet supported. We consider three reasoning levels: *RE-Low*, *RE-Medium*, and

RE-High, implemented by the OpenAI API. We estimate the number of thinking tokens corresponding to these levels, and specify these parameters as a thinking budget to Claude. We compute the average of each metric across the two reasoning models.

On average, we find that deeper reasoning results in higher accuracies and better calibration, highlighted in Figure 5. However, we observe a more nuanced relationship between reasoning depth, accuracy, and calibration upon closer examination. Specifically, our analysis of the results in Figure 5 reveals three key findings. First, model calibration improves with increasing reasoning depth whenever deeper reasoning leads to an increase in accuracy. This observation can be seen in GPQA, where an increase in reasoning depth from low to high leads to a 10.8% increase in accuracy and an associated 5.5% decrease in the ECE. Second, model calibration and accuracy remain relatively unchanged with greater reasoning depth in "easier" benchmarks, i.e., in datasets where the model achieves relatively high accuracy, e.g., in StrategyQA. In these settings, deeper reasoning does not result in an appreciable increase in accuracy. Moreover, model overconfidence becomes difficult to identify, masked by high accuracy. Third, as the accuracy of the model saturates in more challenging datasets, deeper reasoning leads to an *increase* in the calibration errors, which is visible in SimpleQA. As the reasoning depth increases from low to high, we observe a relatively small increase in accuracy and a much larger increase in the ECE. In addition, we find that deeper reasoning makes reasoning models more confident in their incorrect responses by reinforcing their misguided thought process in these problems. We refer readers to Appendix H for more extensive discussion. In summary, for challenging problems, the lack of calibration in the reasoning model typically induces a collapse in the token distribution towards the most likely generation path, generally increasing the confidence of the model without a corresponding increase in accuracy. On the other hand, deeper thinking can also improve accuracy in less challenging task domains, which generally improves calibration, as discussed in Figure 4.

4.5 Reasoning about Confidence

Next, we examine if model calibration can be improved by *reasoning about uncertainty*. Specifically, we utilize a two-stage introspective UQ procedure. We explore three introspective UQ strategies: *IUQ-Low*, *IUQ-Medium*, and *IUQ-High* (see Section 3.2). The second stage uses a new session with a re-initialized context window that does not contain information from the first session. We do not include Gemini in these results due to insufficient support for analysis of the reasoning traces. We limit our scope to inference-time methods, as the data collection and computation costs for training these language models are often out of reach.

We compute the calibration errors of each model across the challenging datasets (GPQA, StrategyQA, and SimpleQA). In Figure 6, when averaging across all models, we observe that introspective UQ leads to a marginal increase in the average ECE in IUQ-Low and a decrease in IUQ-Medium and IUQ-High. The change in the ECE varies significantly with the dataset and the introspective prompt. Unlike the MCE which increases across all datasets, in the least conservative introspection, the ECE decreases in the GPQA and StrategyQA datasets but increases in the SimpleQA dataset. However, with IUQ-Medium and IUQ-High, we observe a decrease in the ECE across all datasets, except in StrategyQA in the case of IUQ-High. Asking the reasoning model to identify flaws in the prior reasoning traces tends to induce more critical introspection, leading to a significant decrease in miscalibration, especially in the most challenging dataset, SimpleQA. When prior confidence estimates are not provided to reasoning models (i.e., in IUQ-High), reasoning models reason more conservatively during introspection, which leads to better calibration in challenging datasets. Conversely, on easier datasets, greater conservativeness from more critical introspection could slightly increase the ECE, e.g., IUQ-High on StrategyQA.

5 Conclusion

In this paper, we examine uncertainty quantification of reasoning models. To this end, we ask three critical questions: First, *are reasoning models well-calibrated?* Second, *does deeper reasoning improve model calibration?* Third, *can reasoning models improve their calibration by explicitly reasoning about their chain-of-thought traces?* Through extensive evaluations, we observe that reasoning models are generally overconfident and become even more overconfident with deeper reasoning, especially in tasks where deeper reasoning does not correspond to higher accuracy. Lastly, the

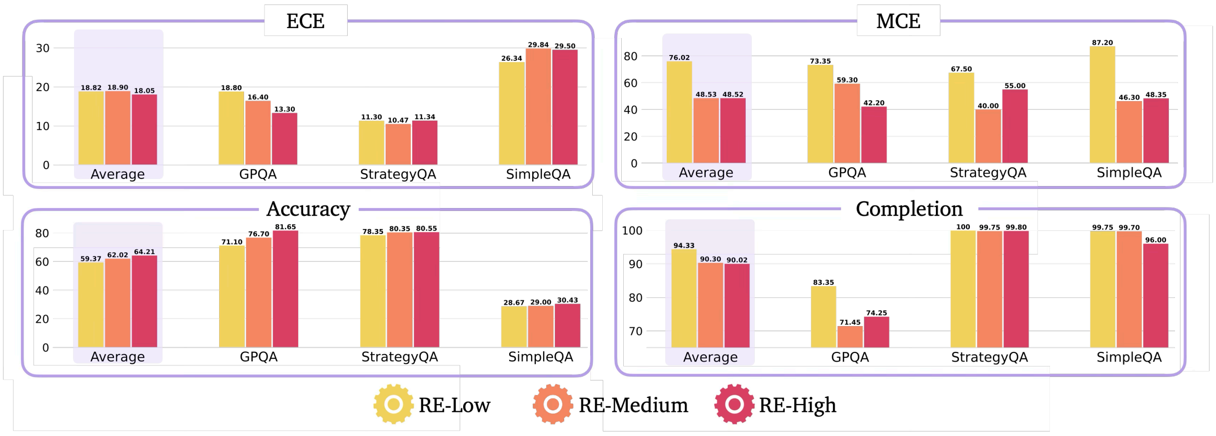


Figure 5: **Does increasing reasoning depth improve calibration?** Deeper reasoning leads to overall higher accuracy; however, as the accuracy of these models saturate, reasoning models become even more overconfident.

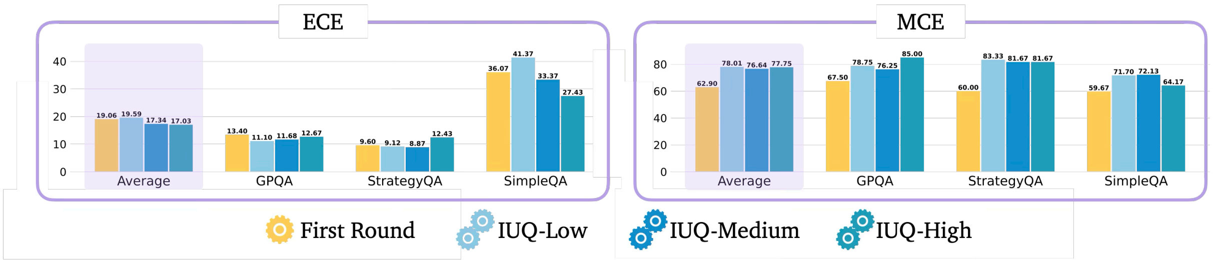


Figure 6: **Introspective UQ.** Calibration of reasoning models improves with introspective uncertainty quantification, particularly in challenging datasets, e.g., SimpleQA, with more conservative prompts, e.g., IUQ-Medium and IUQ-High.

calibration of these model improves with more critical introspection, although introspection increases the overconfidence of Claude.

6 Limitations

The results in our work highlight that in general, reasoning models are poorly calibrated despite their remarkable accuracy, posing a challenge to real-world deployment. We believe that improving the calibration of reasoning models is essential to drive widespread safe adoption of these models. Here, we identify some valuable research directions for future work.

Reducing the Correlation between Accuracy and Calibration. As discussed in Section 4, the accuracy of reasoning models is strongly positively calibrated with their calibration. Stable, well-calibrated models should have a near-zero correlation between accuracy and calibration, i.e., the model should remain well-calibrated in relatively easier problem domains and relatively more challenging problem domains. Training stable, well-calibrated reasoning models remains an open challenge, presenting a valuable direction for future research. Future work will explore techniques that preserve model calibration even when the accuracy of these models is low. For example, reasoning models can be trained to recognize challenging problems and provide more conservative confidence estimates in these knowledge domains.

Training Reasoning Models to Reason about their Uncertainty. Reinforcement-learning-induced multi-step reasoning has enabled reasoning models to set new performance records on a broad range of benchmarks. By decomposing challenging problems into easier-to-solve subproblems, reasoning models have achieved superior accuracy compared to prior language models. We believe that reasoning models can be trained to perform multi-step reasoning about their *uncertainty* through reinforcement learning, constituting an exciting direction for future work. By training models to explicitly reason about the boundaries of their knowledge, reasoning models can be empowered to better estimate their confidence in the correctness of their responses.

Benchmarks for Uncertainty Quantification. As noted in the preceding discussion in Section 4, the performance of reasoning models on many existing benchmarks has saturated, making these benchmarks essentially uninformative in uncertainty quantification. Future work will seek to introduce new benchmarks that better evaluate the calibration of reasoning models, across a range of disciplines. For example, new benchmarks for evaluating the calibration of reasoning models at intermediate steps in the problem-solving process would shed more light on the calibration of the thought processes of reasoning models.

7 Ethical Considerations

This work identifies that state-of-the-art reasoning models are poorly calibrated, motivating future research on training trustworthy reasoning models. However, we do not introduce any new potential risks.

Acknowledgments

The authors were partially supported by the Office of Naval Research (N00014-23-1-2148) and a Sloan Fellowship.

References

Lukas Aichberger, Kajetan Schweighofer, Mykyta Ielanskyi, and Sepp Hochreiter. 2024. Semantically diverse language generation for uncertainty estimation in language models. *arXiv preprint arXiv:2406.04306*.

Anthropic. 2024. Claude 3.7 sonnet system card. Technical report, Anthropic.

Art of Problem Solving. 2025. American Invitational Mathematics Examination. `https://artofproblemsolving.com/wiki/index.php/American_Invitational_Mathematics_Examination`.

Amos Azaria and Tom Mitchell. 2023. The internal state of an llm knows when it's lying. *arXiv preprint arXiv:2304.13734*.

Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. 2024. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*.

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and 1 others. 2024. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17682–17690.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Jacob Dunefsky, Philippe Chlenski, and Neel Nanda. 2024. Transcoders find interpretable llm feature circuits. *arXiv preprint arXiv:2406.11944*.

Ekaterina Fadeeva, Aleksandr Rubashevskii, Artem Shelmanov, Sergey Petrakov, Haonan Li, Hamdy Mubarak, Evgenii Tsymbalov, Gleb Kuzmin, Alexander Panchenko, Timothy Baldwin, and 1 others. 2024. Fact-checking the output of large language models via token-level uncertainty quantification. *arXiv preprint arXiv:2403.04696*.

Javier Ferrando, Oscar Obeso, Senthooran Rajamanoharan, and Neel Nanda. 2024. Do i know this entity? knowledge awareness and hallucinations in language models. *arXiv preprint arXiv:2411.14257*.

Yao Fu, Hao Peng, Ashish Sabharwal, Peter Clark, and Tushar Khot. 2022. Complexity-based prompting for multi-step reasoning. *arXiv preprint arXiv:2210.00720*.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. 2024. Scaling and evaluating sparse autoencoders. *arXiv preprint arXiv:2406.04093*.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2020. Transformer feed-forward layers are key-value memories. *arXiv preprint arXiv:2012.14913*.

Tilmann Gneiting and Adrian E Raftery. 2007. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Shibo Hao, Yi Gu, Haotian Luo, Tianyang Liu, Xiyan Shao, Xinyuan Wang, Shuhua Xie, Haodi Ma, Adithya Samavedhi, Qiyue Gao, and 1 others. 2024. Llm reasoners: New evaluation, library, and analysis of step-by-step reasoning with large language models. *arXiv preprint arXiv:2404.05221*.

Shibo Hao, Yi Gu, Haodi Ma, Joshua Jiahua Hong, Zhen Wang, Daisy Zhe Wang, and Zhiting Hu. 2023. Reasoning with language model is planning with world model. *arXiv preprint arXiv:2305.14992*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Tanya Hewitt, Samia Chreim, and Alan Forster. 2016. Double checking: a second look. *Journal of evaluation in clinical practice*, 22(2):267–274.

Jie Huang and Kevin Chen-Chuan Chang. 2022. Towards reasoning in large language models: A survey. *arXiv preprint arXiv:2212.10403*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, and 1 others. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Laurent Valentin Jospin, Hamid Laga, Farid Boussaid, Wray Buntine, and Mohammed Bennamoun. 2022. Hands-on bayesian neural networks—a tutorial for deep learning users. *IEEE Computational Intelligence Magazine*, 17(2):29–48.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, and 1 others. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*.

Kaiqu Liang, Zixu Zhang, and Jaime F Fisac. 2024. Introspective planning: Aligning robots' uncertainty with inherent task ambiguity. *Advances in Neural Information Processing Systems*, 37:71998–72031.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Teaching models to express their uncertainty in words. *arXiv preprint arXiv:2205.14334*.

Chen Ling, Xujiang Zhao, Wei Cheng, Yanchi Liu, Yiyou Sun, Xuchao Zhang, Mika Oishi, Takao Osaki, Katsushi Matsuda, Jie Ji, and 1 others. 2024. Uncertainty decomposition and quantification for in-context learning of large language models. *arXiv e-prints*, pages arXiv–2402.

Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. 2024. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*.

Potsawee Manakul, Adian Liusie, and Mark JF Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896*.

Sabrina J Mielke, Arthur Szlam, Emily Dinan, and Y-Lan Boureau. 2022. Reducing conversational agents' overconfidence through linguistic calibration. *Transactions of the Association for Computational Linguistics*, 10:857–872.

Mikhail Mirzayanov and Codeforces Team. 2025. Codeforces. `https://codeforces.com`. Competitive Programming Platform.

Mahdi P. Naeini, Gregory Cooper, and Milos Hauskrecht. 2015. Obtaining well calibrated probabilities using bayesian binning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29.

OpenAI. 2024a. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

OpenAI. 2024b. Openai o3-mini system card. Technical report, OpenAI.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. 2023. Gpqa: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv:2311.12022*.

Ola Shorinwa, Zhiting Mei, Justin Lidard, Allen Z. Ren, and Anirudha Majumdar. 2025. A survey on uncertainty quantification of large language models: Taxonomy, open research challenges, and future directions. *ACM Comput. Surv.*

Jody M Shynkaruk and Valerie A Thompson. 2006. Confidence and accuracy in deductive reasoning. *Memory & cognition*, 34(3):619–632.

Eric Michael Smith, Diana Gonzalez-Rico, Emily Dinan, and Y-Lan Boureau. 2020. Controlling style in generated dialogue. *arXiv preprint arXiv:2009.10855*.

Alex Tamkin, Mohammad Taufeeque, and Noah D Goodman. 2023. Codebook features: Sparse and discrete interpretability for neural networks. *arXiv preprint arXiv:2310.17230*.

Sree Harsha Tanneru, Chirag Agarwal, and Himabindu Lakkaraju. 2024. Quantifying uncertainty in natural language explanations of large language models. In *International Conference on Artificial Intelligence and Statistics*, pages 1072–1080. PMLR.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, and 1 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Qwen Team. 2025. Qwq-32b: Embracing the power of reinforcement learning.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. *arXiv preprint arXiv:2305.14975*.

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

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

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.

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.

Yijun Xiao and William Yang Wang. 2021. On hallucination and predictive uncertainty in conditional language generation. *arXiv preprint arXiv:2103.15025*.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2023. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. *arXiv preprint arXiv:2306.13063*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822.

Zeyu Yun, Yubei Chen, Bruno A Olshausen, and Yann LeCun. 2021. Transformer visualization via dictionary learning: contextualized embedding as a linear superposition of transformer factors. *arXiv preprint arXiv:2103.15949*.

Hanning Zhang, Shizhe Diao, Yong Lin, Yi Fung, Qing Lian, Xingyao Wang, Yangyi Chen, Heng Ji, and Tong Zhang. 2024. R-tuning: Instructing large language models to say 'i don't know'. 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 7106–7132.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2022. Automatic chain of thought prompting in large language models. *arXiv preprint arXiv:2210.03493*.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and 1 others. 2022. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*.

Appendix A Preliminaries

Consider a model $\mathcal{M}$ that takes an input random variable $X \in \mathcal{X}$ and predicts an output $\hat{Y} \in \mathcal{Y}$, along with a confidence score $\hat{P} \in [0, 1]$. We focus on settings with a deterministic ground-truth Y given the input X.[1] In this section, we provide a brief overview of calibration of uncertainty and calibration metrics.

Calibration of Uncertainty. The model $\mathcal{M}$ is defined to be perfectly calibrated if its confidence estimate $\hat{P}$ matches the true probability p that the associated answer $\hat{Y}$ is correct. Formally, perfect calibration (Guo et al., 2017) can be expressed as:

$$\mathbb{P}_{\mathcal{Z}}[\hat{Y} = Y \mid \hat{P} = p] = p, \tag{1}$$

where the probability is over the joint distribution on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. However, achieving perfect calibration is practically impossible, motivating metrics that capture deviation from (1) empirically.

Calibration Metrics. Calibration metrics evaluate how closely a model's predicted confidence aligns with its true accuracy. Widely-used metrics include the expected calibration error (ECE) and maximum calibration error (MCE). The ECE measures the average absolute difference between predicted confidence (right-hand side of (1)) and empirical accuracy (left-hand side of (1)) with:

$$\mathbb{E}_{\hat{P}}\left[\left|\mathbb{P}[Y = \hat{Y} \mid \hat{P} = p] - p\right|\right]. \tag{2}$$

Since this expectation is generally intractable, the ECE is typically approximated by grouping the empirical confidence estimates into M equal-width bins. Given a set of samples B_m whose confidence falls into bin m with the bin average accuracy $\text{acc}(B_m)$ and average confidence $\text{conf}(B_m)$, the ECE is approximated by:

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \left|\text{acc}(B_m) - \text{conf}(B_m)\right|, \tag{3}$$

where n represents the total number of samples across all bins.

On the other hand, the MCE captures the worst-case discrepancy between confidence and accuracy, i.e., the maximum deviation between the left- and right-hand sides of (1):

$$\max_{p \in [0,1]} \left|\mathbb{P}[Y = \hat{Y} \mid \hat{P} = p] - p\right|. \tag{4}$$

Similar to the ECE, we approximate the MCE by taking the largest bin-wise error (Naeini et al., 2015):

$$\text{MCE} = \max_{m \in \{1,\ldots,M\}} \left|\text{acc}(B_m) - \text{conf}(B_m)\right|. \tag{5}$$

Appendix B Dataset Licenses

We use the datasets in Section 4.1 in accordance with their licenses (e.g., CC-BY 4.0), which allow for research use. These datasets, which contain primarily English content, have been used in prior work and have not been flagged as having offensive content.

Appendix C Introspective Uncertainty Quantification

Here, we provide the full system prompts for the three introspective UQ prompt strategies: IUQ-Low, IUQ-Medium, and IUQ-High.

Introspective UQ-Low

You are provided with the reasoning trace of a model asked to answer a question and provide the associated confidence between 0 and 100. Your task is to think about the reasoning trace from the first model and provide your confidence in the correctness of the answer provided by the first model.

Introspective UQ-Medium

You are provided with the reasoning trace of a model asked to answer a question and provide the associated confidence between 0 and 100. Your task is to identify the flaws in the reasoning trace from the first model and provide your confidence in the correctness of the answer provided by the first model.

---

[1] This applies to a wide range of question-and-answering problems, e.g., multiple-choice questions, true-or-false questions, or any question that has an explicit correct answer. The definition of $\hat{Y} = Y$ is not constrained to exact correspondence in form, i.e., $\hat{Y} = Y$ if $\hat{Y}$ and Y have the same meaning. For example, both "true" and "True" are correct for a true-or-false question, if Y=True, while for freeform Q&A, $\hat{Y}$ is correct if it captures the ground truth (Wei et al., 2024). If the ground truth answer Y is inherently stochastic, then it should be expressed as a stochastic distribution described in natural language, for example, a Gaussian distribution with $\mu = 0$ and $\sigma^2 = 1$.

Introspective UQ-High

You are provided with the reasoning trace of a model asked to answer a question. Your task is to identify the flaws in the reasoning trace from the first model and provide your confidence in the correctness of the answer provided by the first model.

Lastly, we provide an example of the full prompt with the question and reasoning trace using IUQ-Medium. We append the text “You must include your reasoning trace in your response," particularly for models that do not natively support access to the internal reasoning trace.

Introspective UQ: Example Full Prompt

System: You are a smart entity. Do not use the internet.
System: You are provided with the reasoning trace of a model asked to answer a question and provide the associated confidence between 0 and 100. Your task is to identify the flaws in the reasoning trace from the first model and provide your confidence in the correctness of the answer provided by the first model.
User: The question is: How many original artworks were shown in the Society of Illustrators' first "Annual Exhibition"?
The given reasoning process and answer is: ...
System: Given the answer and confidence from the first model, what is your confidence in the validity of the first model's answer as a number in the form {rec, Y}? The first element in the curly braces should always be rec. For example, if you are 80 percent confident that the first model is correct, your answer should be {rec, 80}.

Appendix D LLM Prompt Strategies

For completeness, we provide the full prompts for the basic, chain-of-thought, multi-step, and top-K prompt strategies, introduced in prior work.

Standard Prompt

Provide an answer and the confidence in your answer between 0 and 100.

Chain-of-Thought Prompt

Analyze step by step and provide an answer and the confidence in your answer between 0 and 100.

Multi-Step Prompt

Break down the problem into multiple steps, and provide an answer to each step, along with a justification and a confidence in each step. Give your final answer with your confidence in your final answer and the confidence in your answer between 0 and 100.

Top-K Prompt

Give $K = 5$ best guesses and your confidence in each guess (i.e., the probability that each one is correct). Give your final answer with your confidence in your final answer and the confidence in your answer between 0 and 100.

The system prompts are followed with a description of the dataset including the desired answer format, as well as the actual question. We show an example full prompt for the SimpleQA dataset, with the basic prompt strategy.

Example Full Prompt

System: You are a smart entity. Do not use the internet.
System: Analyze step by step and provide an answer and the confidence in your answer between 0 and 100.
System: Each question requires a simple, short fact as the answer: only provide your final answer, which could be a specific name, date, or other fact, and a number for the confidence in the form X, Y. For example, if you think the answer is 1950 and you are 80 percent confident, then you should only say {1950, 80}.
User: Who received the IEEE Frank Rosenblatt Award in 2010?

Appendix E Evaluation Setup

We provide additional details on the evaluation setup.

Datasets. We evaluate the models on the following datasets: (i) AI2 Reasoning Challenge (*ARC-Challenge*) Set (Clark et al., 2018), the challenging subset of the ARC dataset; (ii) Graduate-Level

Google-Proof Q&A (*GPQA*) Benchmark (Rein et al., 2023), consisting of questions determined to be challenging by PhD-level domain experts; (iii) Measuring Massive Multitask Language Understanding (*MMLU*) dataset (Hendrycks et al., 2020), which encompasses a broad range of fields; (iv) *StrategyQA* (Geva et al., 2021), which requires implicit reasoning steps, and (v) *SimpleQA* (Wei et al., 2024), which examines the factuality of LLMs. While SimpleQA consists of open-ended questions, all other datasets use the multiple-choice format. For adequate coverage of a range of tasks in MMLU, we select the abstract algebra, professional accounting, professional medicine, international law, and sociology subsets, representing the more challenging subsets.

Appendix F Calibration of Reasoning Models.

We further discuss the calibration and accuracy of reasoning models, across the basic, chain-of-thought, multi-step, and top-K prompt strategies, in addition to the results in Section 4.

Are Reasoning Models Underconfident or Overconfident? We plot the reliability diagrams of the reasoning models in Figure 7 for the three challenging datasets (StrategyQA, GPQA, and SimpleQA), aggregating the empirical confidence estimates across all prompt strategies. In these more challenging datasets, we see that reasoning models are generally overconfident, with confidence estimates typically greater than 85%. This is seen by examining the density of samples in each confidence bin (i.e., the number of times the model predicts a confidence level within the given bin divided by the total number of queries), denoted by the red crosses in Figure 7. Larger red gaps in the reliability diagram signify greater overconfidence (miscalibration), representing the deviation between the estimated bin-wise confidence and perfect calibration. Figure 7 shows that Gemini and o3-Mini are higly overconfident, with significant concentration of their confidence estimates in the 85% to 100% confidence bins. In particular, the accuracy of the samples within these bins significantly deviates from the estimated bin-wise confidence. DeepSeek also exhibits overconfidence, although to a smaller degree. We observe large miscalibration and high sample densities especially between the 70% and 80% confidence intervals. Claude is the best calibrated model, with significantly smaller miscalibration gaps. However, Claude still appears to be overconfident, particularly in the 60% and 80% confidence range. We provide the reliability diagram for all datasets in Appendix F.

In Figure 8, we plot the reliability diagrams aggregated across all datasets, showing that reasoning models are typically highly confident, with most of their confidence estimates in the 90%-100% confidence bins. However, the high accuracies of the models makes overconfidence detection difficult. For example, in Figure 9, we visualize the reliability diagrams for the easier datasets (ARC-Challenge and MMLU), showing that overconfidence is masked by the models' high accuracy.

Comparison among the Models. We compare the models' calibration, accuracy, and completion capabilities to identify the best-performing model. From Figures 3 and 7, we find that Claude is the most calibrated model by a significant margin, followed by DeepSeek, based on the ECE. This finding appears to be consistent across all datasets, except StrategyQA and GPQA, where DeepSeek achieves the best calibration. In contrast, Gemini is the least calibrated model, although the calibration of o3-Mini is almost the same as that of Gemini in many datasets. In addition, Claude 3.7 Sonnet is about 2% more accurate than DeepSeek (the best-competing method) and over 4% more accurate than o3-Mini (the least accurate model), which is not entirely surprising given that the compact o3-Mini trades off performance for cost efficiency relative to o3. Further, DeepSeek and Gemini achieve almost the same accuracy, with only a 0.6% gap. The superior accuracy of Claude 3.7 Sonnet may be explained by the recency of its release compared to the other models: Claude 3.7 Sonnet is the latest model to be released among the models. Lastly, all models completed more than 97% of the questions, with o3-Mini achieving the highest completion percentage, indicating its ability to interpret the questions and follow the instructions required to provide valid answers.

Model Calibration for Each Prompt Strategy. In Figures 10 to 13, we provide the calibration and accuracies of each model across all datasets for each prompt strategy. In line with the results in Section 4.2, across all the prompt strategies, Claude achieves the lowest ECE by relatively significant margins followed by DeepSeek (except in the top-K prompt strategy). In general, we find that Claude 3.7 Sonnet is the best calibrated and most accurate

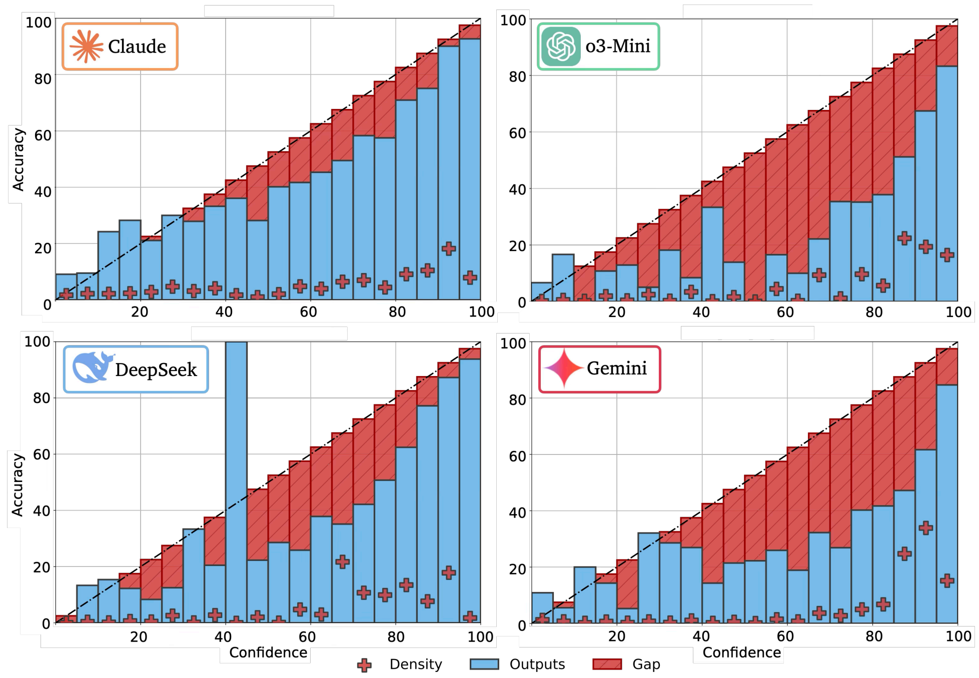


Figure 7: **Are reasoning models consistently underconfident or overconfident?** Reasoning models are systematically overconfident, with Gemini and o3-Mini being the most overconfident followed by DeepSeek, while Claude is the best calibrated one, highlighted by the reliability diagrams for StrategyQA, GPQA, and SimpleQA aggregated over all prompt strategies.

of the reasoning models considered in this work. Further, Gemini and o3-Mini achieve almost the same calibration and accuracy, with Gemini outperforming o3-Mini with basic and top-K prompts by narrow margins.

Appendix G Coupling between Accuracy and Calibration of Reasoning Models

Here, we discuss the accuracy of each model on the benchmarks and examine the correlation between accuracy and calibration.

Model Accuracy on Benchmarks. We find that the accuracy of reasoning models varies significantly across the benchmarks. In Figure 3, all reasoning models achieve over 90% accuracy on the ARC-Challenge and MMLU datasets, with about a $10 - 20\%$ absolute drop on the StrategyQA and GPQA datasets. The drop in accuracy can be explained by the relative difficulty of the more recent benchmarks StrategyQA and GPQA, which emphasize analytical problem-solving over memorization, compared to ARC-Challenge and MMLU. Prior work has shown that LLMs struggle more with analytical tasks compared to information-retrieval tasks (which require recitation) (Wu et al., 2024), which is in line with our results. Likewise, the results indicate that the accuracy of the reasoning models decreases with the recency of the benchmark, which is not surprising, given that these models are trained to perform well on these benchmarks. Notably, all reasoning models achieve less than 40% accuracy on the SimpleQA dataset, which was was constructed adversarially from GPT-4's responses (i.e., a question was added to SimpleQA, if at least one GPT-4 response was incorrect). We observe that SimpleQA remains notably challenging for reasoning models, making it useful for uncertainty quantification.

Correlation between Accuracy and Calibration. In Section 4.3, we show that calibration and accuracy are strongly correlated (Figure 4), with the results averaged over the four prompt strategies. Here, we show the correlation between calibration errors and accuracy for each prompt strategy in Figures 14 through 17. For all prompt strategies, we observe a strong negative correlation between ECE and accuracy, as well as between MCE and accuracy. Similar to results discussed in Section 4.3, among all models, Claude's calibration is least strongly correlated accuracy, while Gemini almost always exhibits perfect negative correlation. For MCE and accuracy, the trend is weaker, affected by many outliers, but o3-Mini and Gemini generally

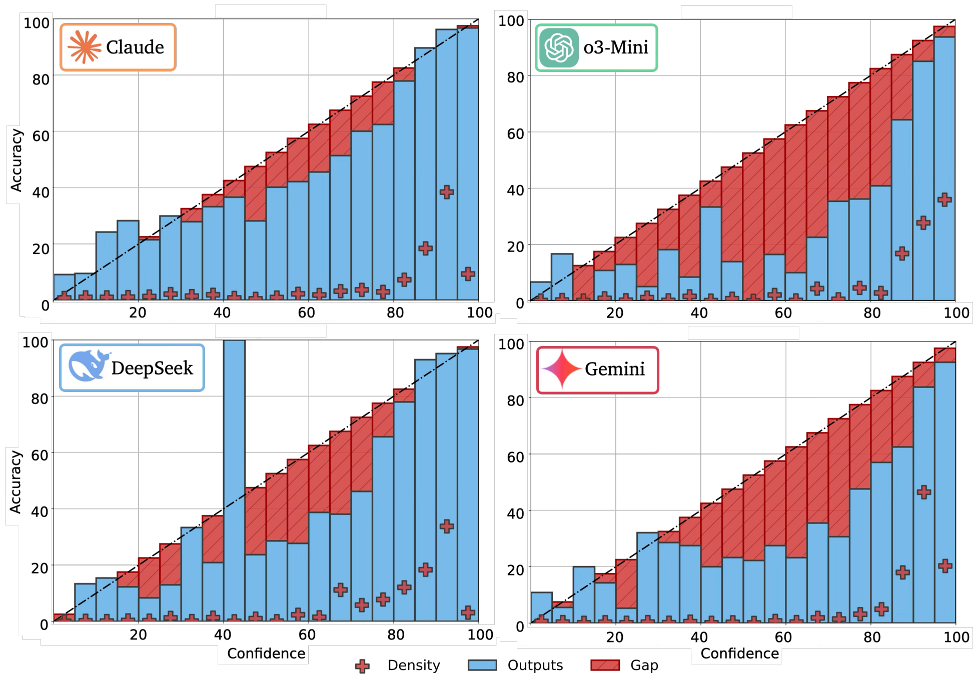


Figure 8: **Are reasoning models consistently underconfident or overconfident?** The reliability diagrams are aggregated over all prompt strategies and all datasets. Reasoning models are overconfident, but it's less obvious in datasets with very high model accuracies.

show stronger correlations compared to Claude and DeepSeek.

Appendix H Reasoning Depth vs. Calibration

We provide the reliability diagrams associated with the reasoning depths aggregated over all datasets in Figure 18, showing that the proportion of highly confident responses increases with reasoning depth, oftentimes without a corresponding increase in accuracy, degrading the calibration of the model. For example, in Figure 18, we see that reasoning models are generally overconfident with most of the model's confidence estimates residing between 85% and 100%. As we increase the reasoning depth, the density of samples in this interval increases while the accuracy of these samples decreases (e.g., samples in the 90%-95% interval), underscoring that *reasoning models become even more overconfident with deeper reasoning.*

Further, MCE decreases with greater reasoning depth (Figure 5). We note that the worst-case calibration error is generally associated with confidence bins containing a single or only a few samples. Deeper reasoning generally increases either the accuracy of these samples or the confidence associated with these samples, since the affected samples typically move to a bin with more samples, ultimately reducing their influence on the MCE. Lastly, from Figure 18, the completion rate of reasoning models decreases with greater reasoning depth. Reasoning models tend to struggle more with following format instructions specified in the system prompt when provided with a greater thinking budget. This observation is more prominent in scientific domains, e.g., GPQA, where reasoning involves significant scientific notation and equations.

Next, for a more detailed analysis, we examine the calibration errors as a function of reasoning depth on *wrongly-answered questions*, (i.e., on questions with incorrect responses at all reasoning depth levels). Figure 19 reveals that model calibration degrades with reasoning depth in this domain with an increase in the average ECE and MCE. This result suggests that deeper reasoning makes reasoning models more confident in their incorrect responses by reinforcing their misguided thought process in these problems. This result bears some similarity to human behavior, since prior work has shown that humans tend to provide higher estimates of their confidence in their answers (Shynkaruk and Thompson, 2006) when given more time to think, even when their answers do not change. Moreover, we find that the relative effect of reasoning depth on calibration varies significantly with the model.

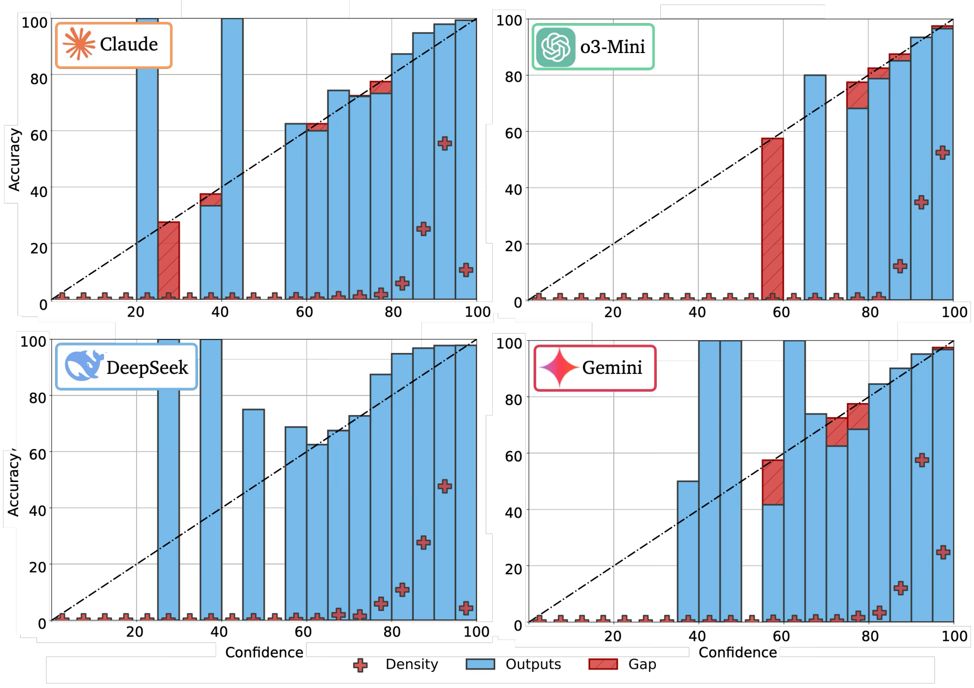


Figure 9: **Are reasoning models consistently underconfident or overconfident?** We show the reliability diagrams on the easier datasets . Overconfidence is difficult to detect in these datasets due to the high accuracy of the models.

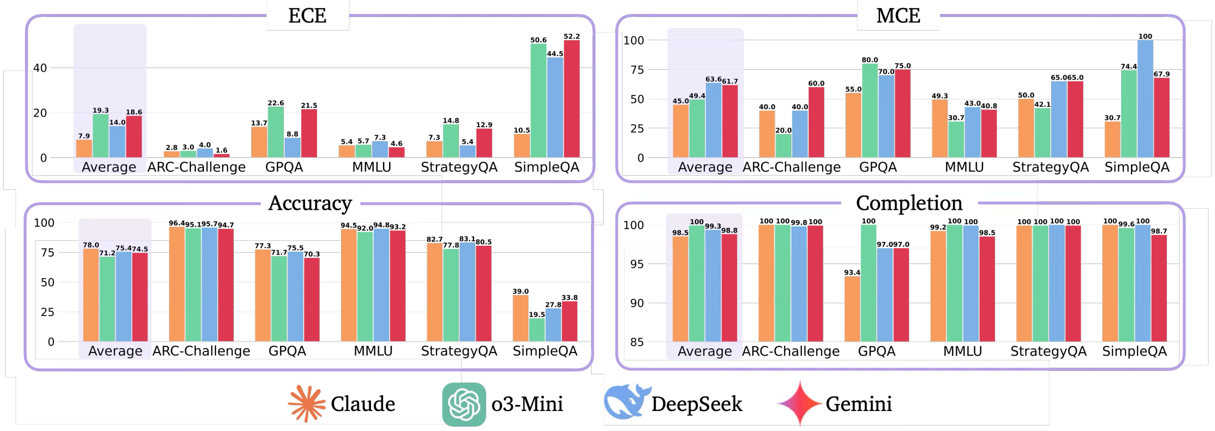


Figure 10: **Basic Prompt Strategy.** Calibration and Accuracy of Reasoning Models.

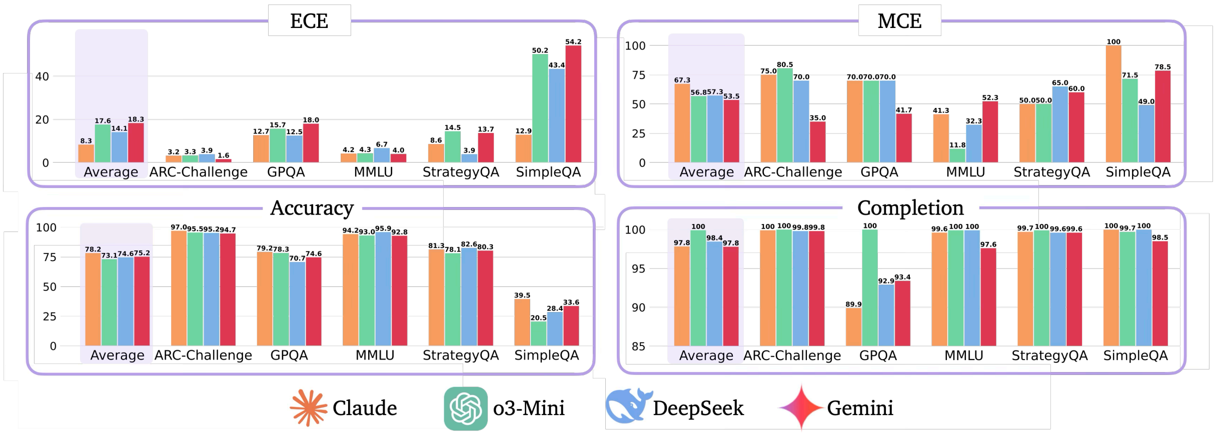


Figure 11: **Chain-of-Thought Prompt Strategy.** Calibration and Accuracy of Reasoning Models.

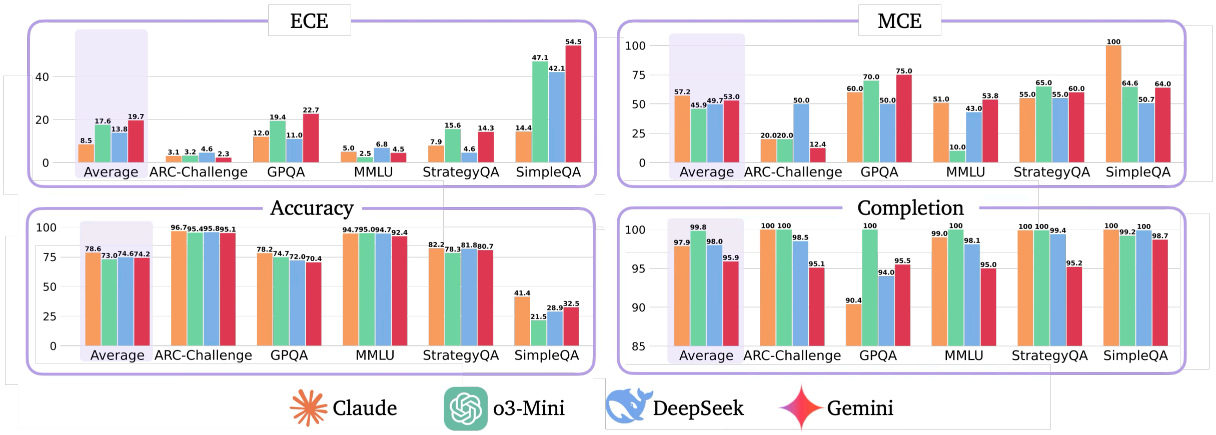


Figure 12: **Multi-Step Prompt Strategy.** Calibration and Accuracy of Reasoning Models.

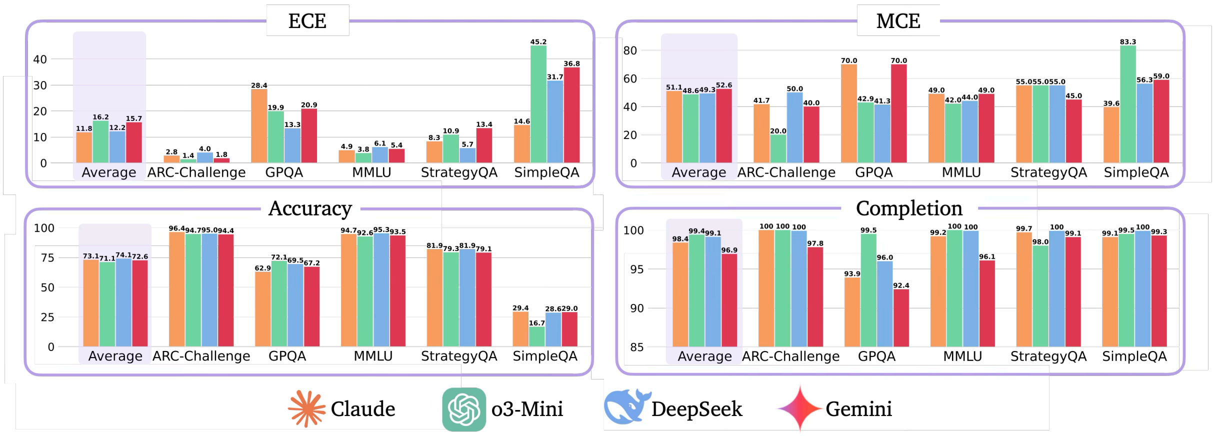


Figure 13: **Top-K Prompt Strategy.** Calibration and Accuracy of Reasoning Models.

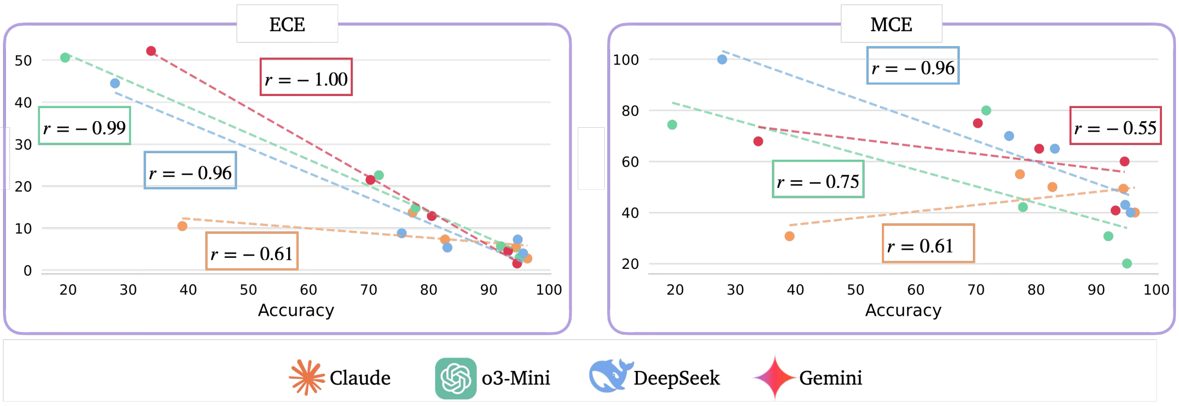


Figure 14: **Basic Prompt Strategy.** Correlation between Calibration and Accuracy.

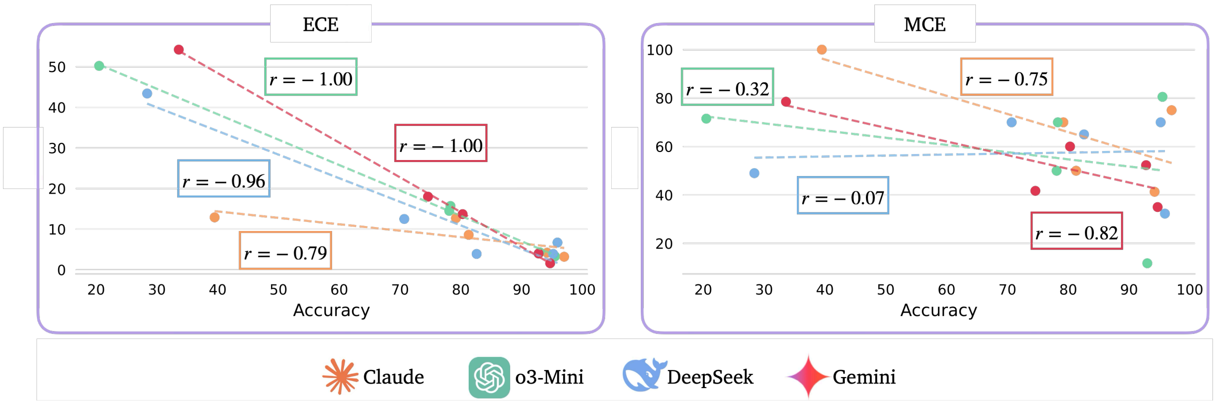


Figure 15: **Chain-of-Thought Prompt Strategy.** Correlation between Calibration and Accuracy.

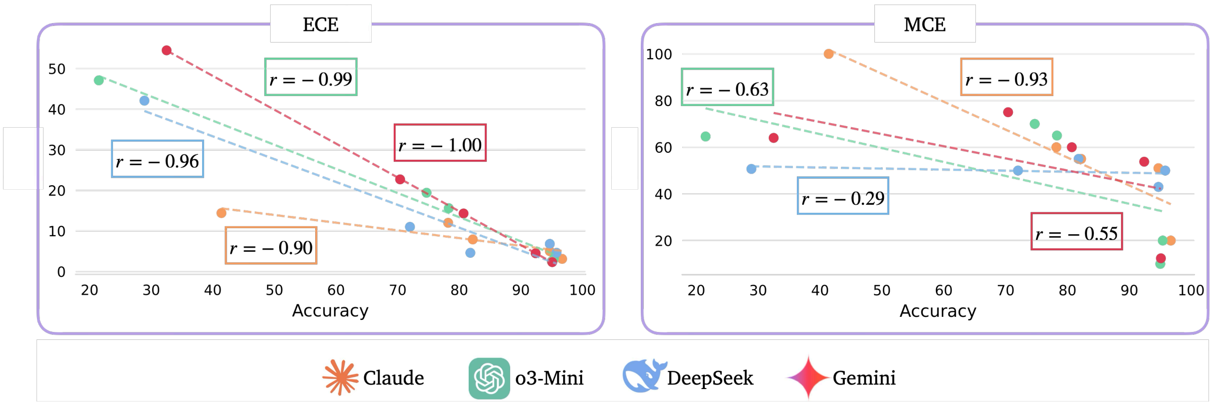


Figure 16: **Multi-Step Prompt Strategy.** Correlation between Calibration and Accuracy.

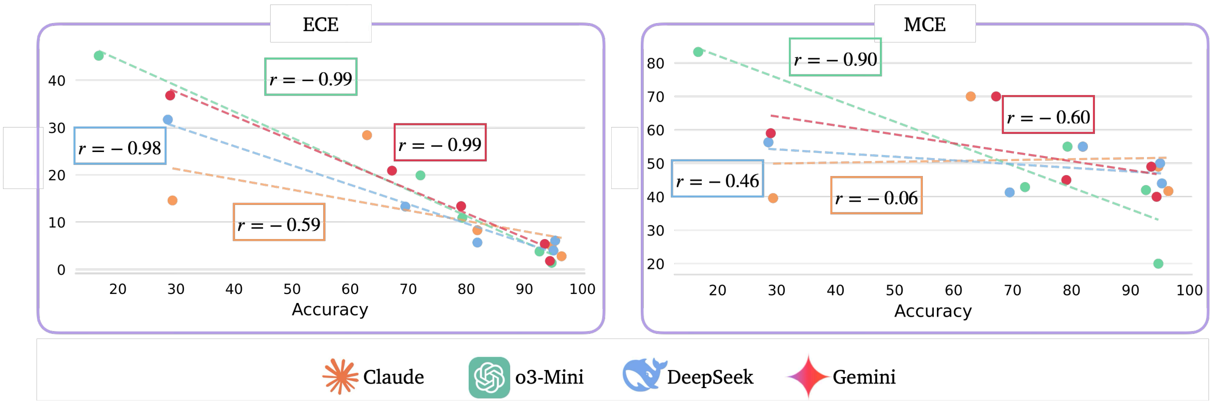


Figure 17: **Top-K Prompt Strategy.** Correlation between Calibration and Accuracy.

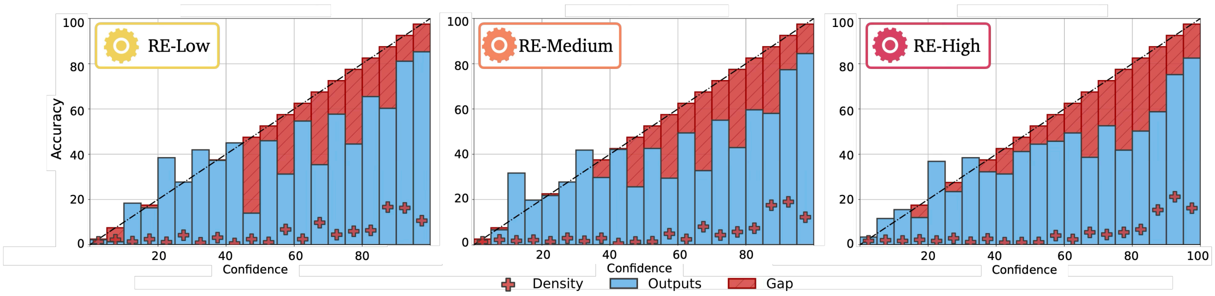


Figure 18: **Deeper Reasoning vs. Underconfidence/Overconfidence?** Reasoning models become more overconfident with deeper reasoning, evidenced by the increase in the density of samples in higher-confidence bins (e.g., the 90%-95% interval) without a corresponding increase in accuracy.

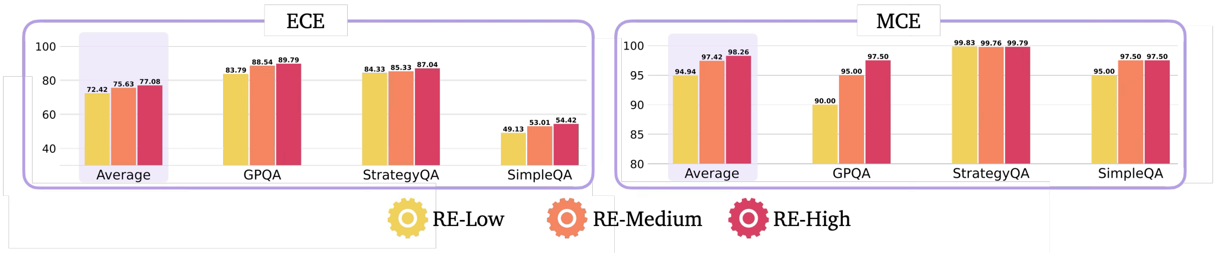


Figure 19: **Does increasing reasoning depth make the models more confidently wrong on average?** On wrongly-answered questions, reasoning models become more overconfident, even though the correctness of their answers does not change.

From Figure 20, o3-Mini is more susceptible to reasoning-induced overconfidence, with a 7% absolute increase in the ECE from the low-reasoning to the high-reasoning setting, compared to a 2% increase for Claude. In fact, not only is Claude more robust, it is also better calibrated across all datasets. Moreover, the MCE increases for all models as all models become more confident without a corresponding increase in their accuracy.

Appendix I Reasoning about Confidence

We visualize the reliability diagrams associated with the first and second stages in Figure 21, showing that less critical introspection, e.g., IUQ-Low, can increase model overconfidence, i.e., a rightward shift in the confidence estimates from low-confidence bins to high-confidence bins. For example, we observe an increase in the density of samples in the 0.85-1.0 confidence bins, in IUQ-Low. However, more critical introspection, e.g., IUQ-Medium and IUQ-High, leads to better calibration, reducing overconfidence.

We further analyze the calibration of introspective UQ estimates for each model, summarized in Figure 22. We find that calibration of Claude significantly degrades when the model reasons about its uncertainty, particularly in StrategyQA and SimpleQA. Specifically, Claude becomes more overconfident with introspective UQ, as highlighted in Figure 23, showing that the proportion of confidence estimates in the 0.9-0.95 confidence bin nearly doubles, with a corresponding decrease in the density of almost all the bins below 0.8. This observation may be due to the implementation details of Claude, which is closed-source.

In contrast, introspection improves the calibration of DeepSeek and o3-Mini, with bigger improvements with IUQ-Medium and IUQ-High in SimpleQA. During the second stage, critical introspection enables these models to identify flaws and hallucinations in their reasoning traces, leading to effective reassessment of their confidence. Figure 23 highlights these findings, showing an increase in the density of the lower-confidence bins with a corresponding decrease in the density of the higher-confidence bins, ultimately reducing overconfidence.

In summary, these results suggest that introspective UQ improves the calibration of reasoning models, particularly in challenging problems where introspection enables these models to identify flaws in their reasoning.

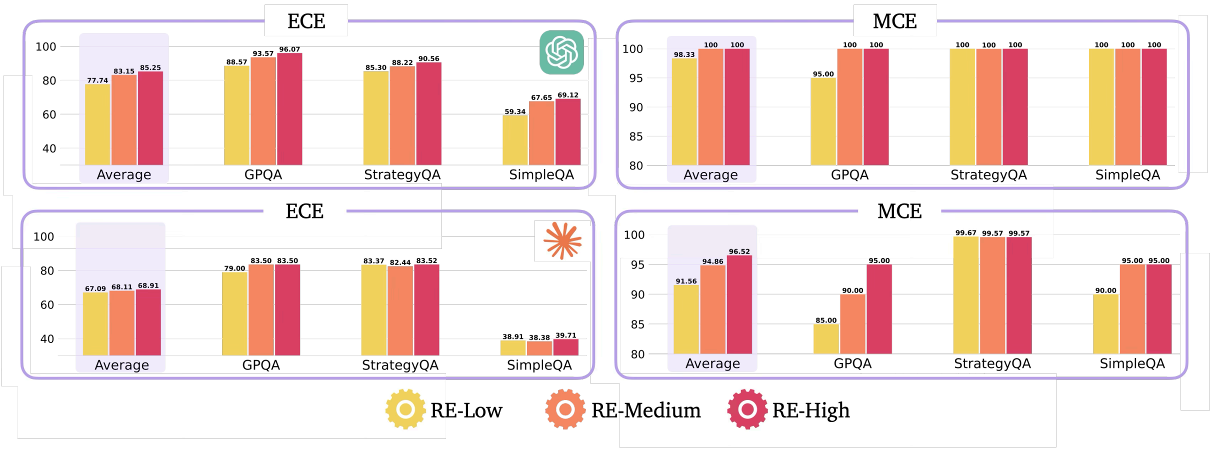


Figure 20: **Does increasing reasoning depth make the models more confidently wrong?** With more reasoning effort, o3-Mini's overconfidence worsens, while Claude's overconfidence remains relatively constant.

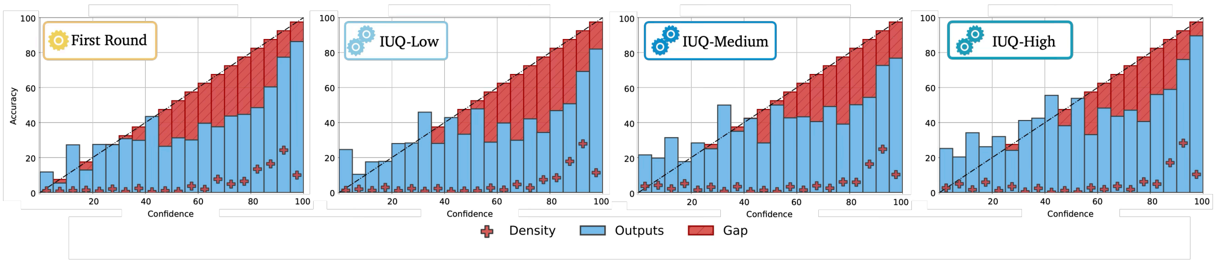


Figure 21: **Introspective UQ vs. Underconfidence/Overconfidence.** More critical introspection, e.g., IUQ-Medium and IUQ-High, improves the calibration of reasoning models, mitigating model overconfidence, unlike IUQ-Low.

Appendix J Effects of LLM Prompt Strategies on Calibration and Accuracy of Reasoning Models

Prior work has shown that prompt engineering can improve the performance of LLMs significantly. For example, existing work has demonstrated that chain-of-thought (CoT) prompting leads to higher accuracy in non-reasoning LLMs, e.g., in zero-shot tasks. Here, we explore the question: do prompt strategies such as chain-of-prompting improve the accuracy and calibration of reasoning models? To do so, we compute the accuracy and calibration over the four prompt strategies: basic, CoT, multi-step, and top-K prompts, averaging the models' performance within each dataset.

From Figure 24, we observe that the performance of reasoning models is relatively independent of the selected prompt strategy. In particular, specialized prompt strategies such as CoT prompting and multi-step prompting which have been shown to be effective in non-reasoning models do not significantly outperform basic prompting techniques that just ask the model directly for its answer and confidence in terms of calibration, accuracy, and completion rates. In non-reasoning models, CoT prompting and multi-step prompting are generally effective due to their ability to elicit reasoning in these models. However, reasoning models are trained to generate reasoning traces by default. Consequently, specialized prompt strategies are not necessary to induce reasoning, explaining the independence between the performance of reasoning models and prompt strategies.

Here, we examine the variability in calibration and accuracy for each model for each prompt strategy. Unlike the other models, Claude suffers a notable dip in its accuracy and calibration in the Top-K strategy, where the model is asked to provide multiple guesses along with its confidence in the correctness of each guess. This finding suggests that Claude might be more likely to change its answers to potentially incorrect ones when asked to reason about false statements/claims. However, Claude still outperforms the other models in the Top-K prompt strategy, albeit by much smaller margins. In fact, we observe that GPT, DeepSeek, and Gemini achieve lower calibration errors with Top-K prompts compared to the other prompts, sug-

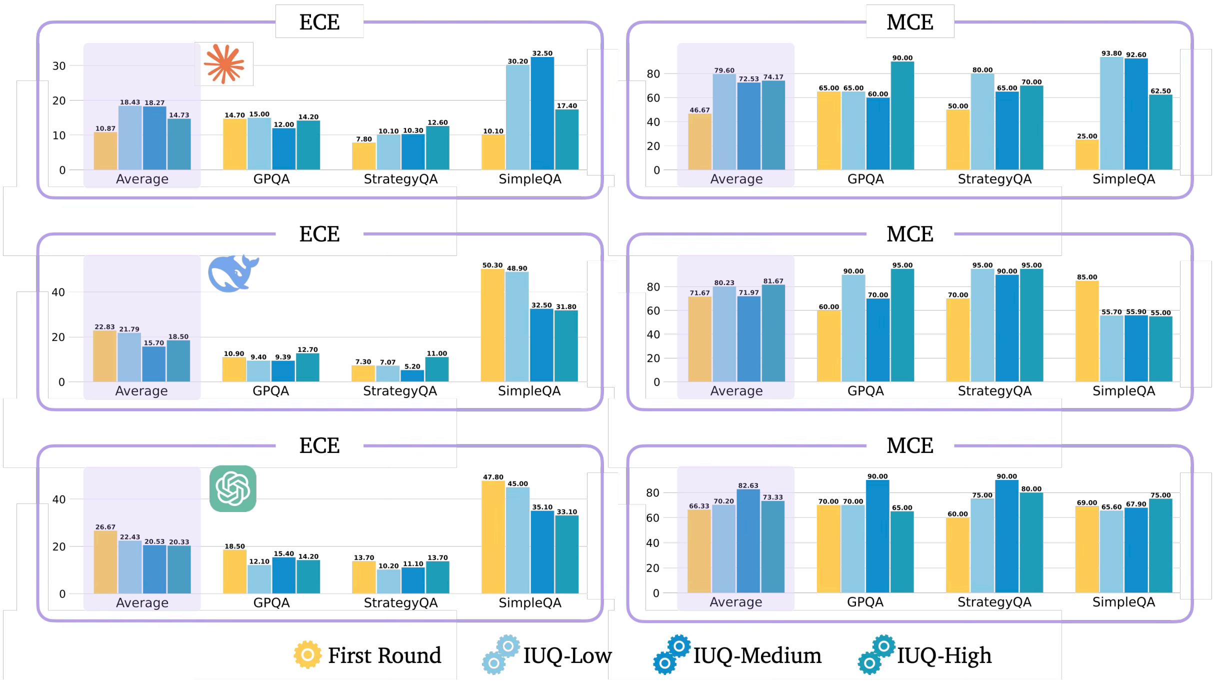


Figure 22: **Introspective UQ per Model.** Introspective UQ improves the calibration of DeepSeek and o3-Mini, especially in the challenging dataset SimpleQA, but degrades the calibration of Claude.

gesting that explicitly reasoning about alternative answers improves the calibration of these models even though the accuracy remains relatively constant. Lastly, Figure 26 shows that chain-of-thought and multi-step prompting provide a slight boost to the accuracy of o3-Mini. Overall, we reiterate that the specialized prompt strategies seem to have minimal effect on the calibration and accuracy of reasoning models.

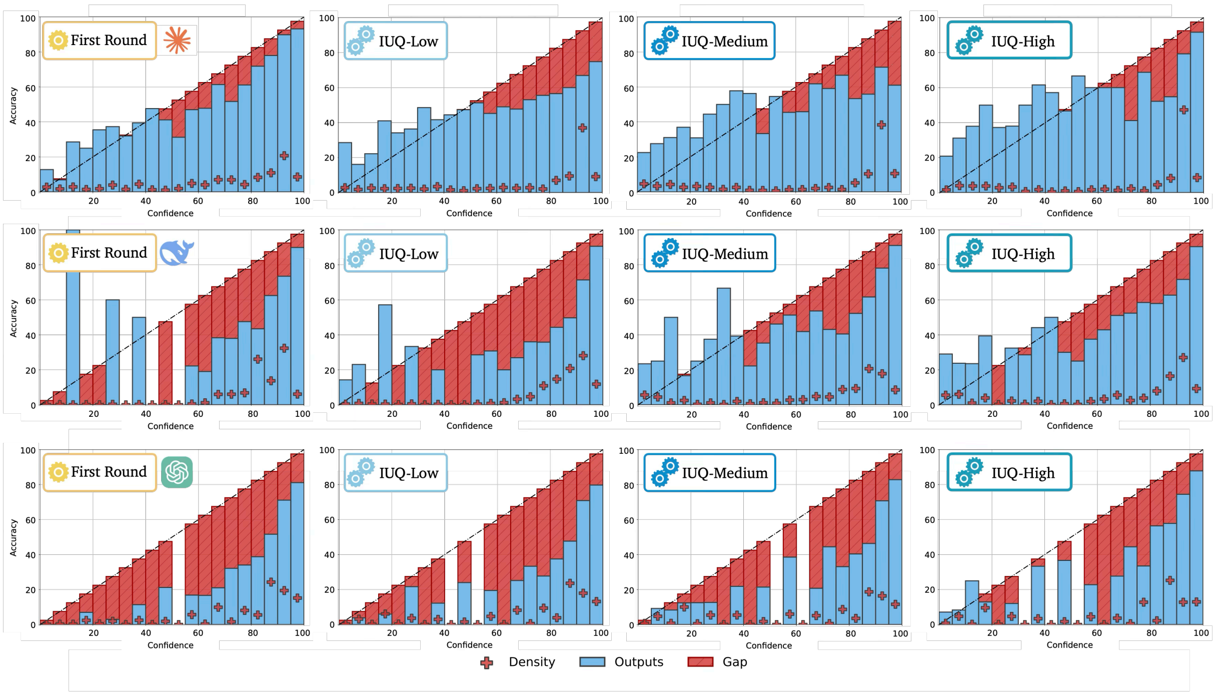


Figure 23: **Introspective UQ per Model vs. Underconfidence/Overconfidence.** Through introspection, DeepSeek and o3-Mini become less overconfident, especially with IUQ-Medium and IUQ-High, unlike Claude, which becomes more overconfident.

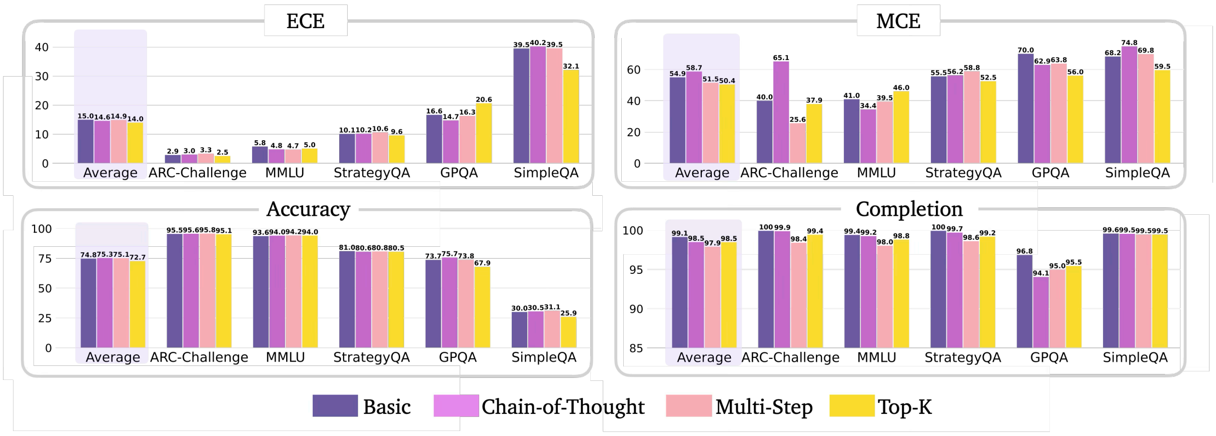


Figure 24: **Do specialized prompt strategies (e.g., chain-of-thought prompting) improve calibration and accuracy?** Chain-of-thought prompting and multi-step prompting do not outperform the basic prompt strategy. On average, calibration and accuracy of reasoning models do not depend on the prompt strategy.

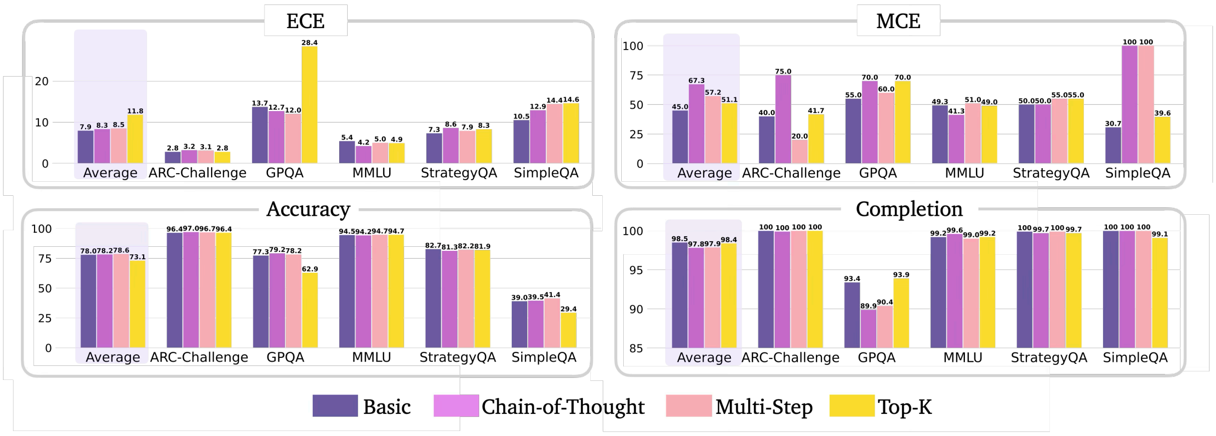


Figure 25: **Claude 3.7 Sonnet.** Calibration and accuracy for each prompt strategy.

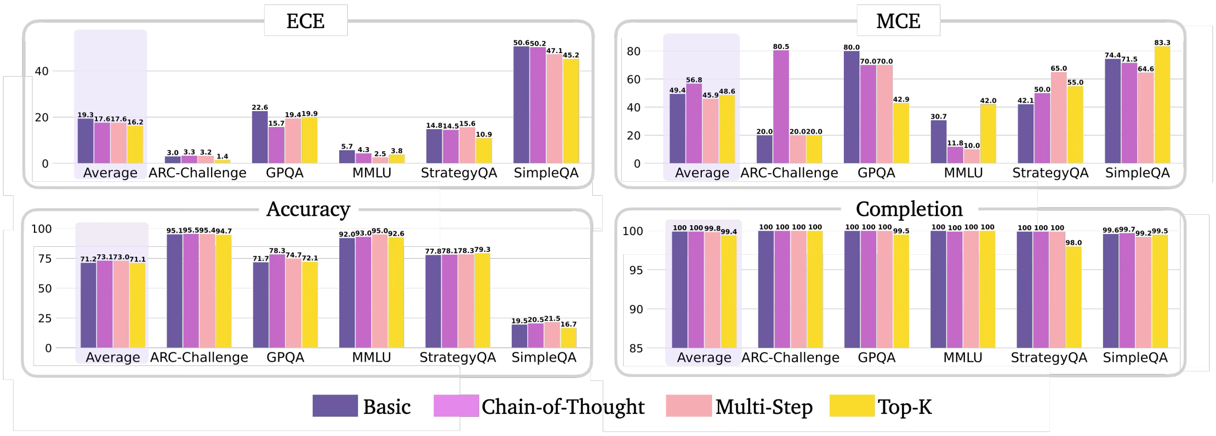


Figure 26: **o3-Mini.** Calibration and accuracy for each prompt strategy.

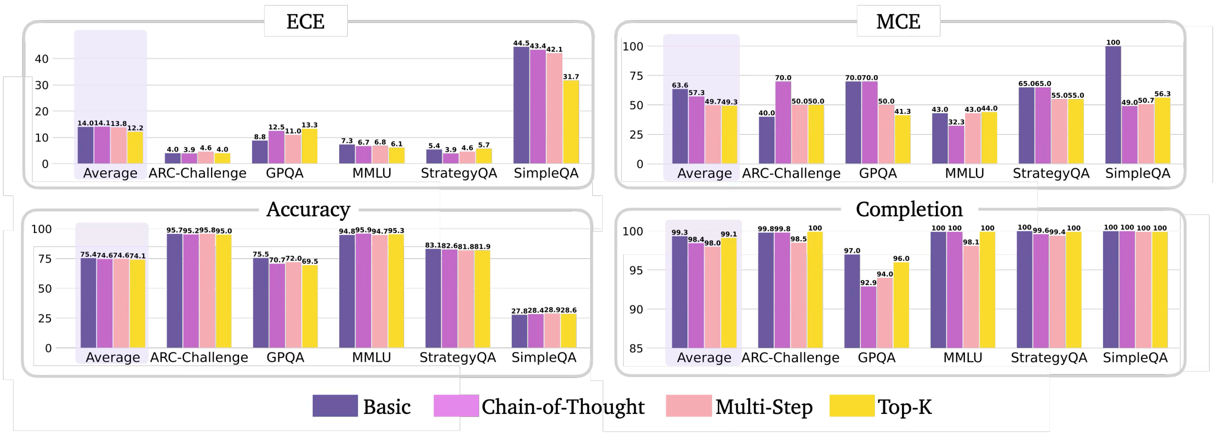


Figure 27: **DeepSeek R1.** Calibration and accuracy for each prompt strategy.

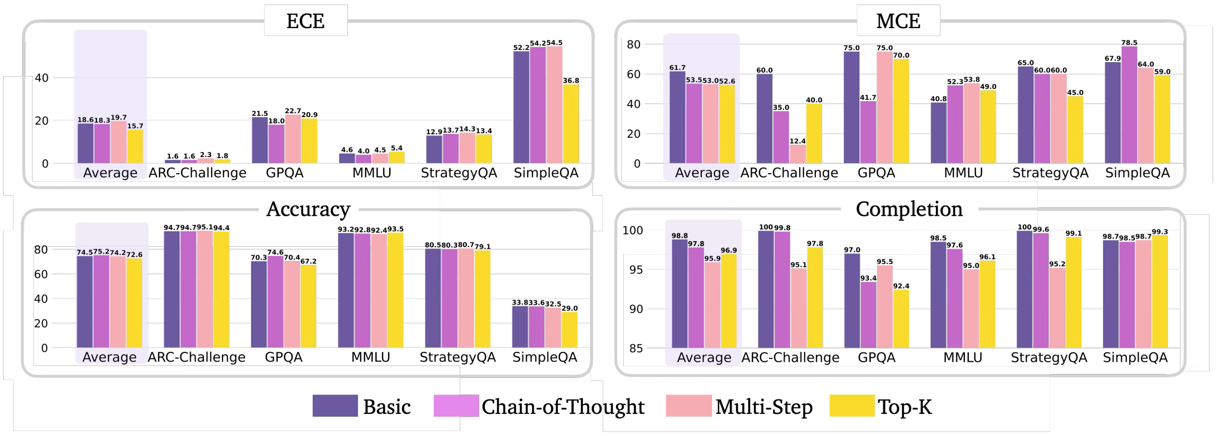


Figure 28: **Gemini 2.0 Flash-Thinking-Exp.** Calibration and accuracy for each prompt strategy.

Appendix K Introspective UQ Reasoning Trace Examples

Here, we provide some examples of the prompts, reasoning traces, and answers provided by Claude, DeepSeek, and o3-Mini for all introspective prompts in the challenging datasets (GPQA, StrategyQA, and SimpleQA). All models get the GPQA question wrong and are highly overconfident, demonstrating poor calibration. On the relatively easier dataset StrategyQA, all models provide highly confident correct responses (overconfidence is masked by high accuracy in these problems). In SimpleQA where all models answer the question incorrectly, introspection provide better calibrated confidence estimates, especially in the case of o3-Mini, although DeepSeek remains overconfident.

K.1 GPQA

Question and Ground-Truth Answer

Q: The Cope rearrangement is a chemical reaction where a 1,5-diene molecule undergoes rearrangement, resulting in a change in the positions of its carbon-carbon double bonds. This rearrangement can be initiated by heat or light and is valuable for creating complex organic compounds with changed structures.

Select the major products from the following rearrangements:
1. (((3-methylbut-2-en-1-yl)oxy)methyl)benzene + (1. BuLi, 2. H+) —-> A
2. 3,4,5,7,8,9-hexamethyl-1,11-dimethylene-2,6,10,11,11a,11b-hexahydro-1H-benzo [cd]indeno[7,1-gh]azulene + Heat —> B.

Options:
(A) $A =$ 4-methyl-1-phenylpent-3-en-1-ol
$B = 2, 3, 4, 6, 7, 8$-hexamethyl-5,9,9a,10,11,11a-hexahydro-1H-benzo [3,4]azuleno[1,8,7,6-cdef]fluorene
(B) $A = (Z)$-2-methyl-5-phenylpent-2-en-1-ol
$B = 2, 3, 4, 6, 7, 8$-hexamethyl-5,9,10,11-tetrahydro-1H-benzo [3,4]azuleno[1,8,7,6-cdef]fluorine
(C) e $A = (Z)$-2-methyl-5-phenylpent-2-en-1-ol
$B = 2, 3, 4, 6, 7, 8$-hexamethyl-5,9,9a,10,11,11a-hexahydro-1H-benzo [3,4]azuleno[1,8,7,6-cdef]fluorene
(D) $A =$ 4-methyl-1-phenylpent-3-en-1-ol
$B = 2, 3, 4, 6, 7, 8$-hexamethyl-5,9,10,11-tetrahydro-1H-benzo [3,4]azuleno[1,8,7,6-cdef]fluorine

A: D

o3-Mini

First round:

Reasoning Trace:

1. For the first reaction, the starting benzyl allyl ether, ((3-methylbut-2-en-1-yl)oxy)methylbenzene, under BuLi/H^+ conditions undergoes a rearrangement that effectively "umpoled" the benzyl fragment. Analysis of the connectivity shows that nucleophilic attack leads to a five-carbon chain where the benzyl group ends up at one end. Distinguishing between the possible products, the structure where the new double bond is positioned between C2–C3 (and with a methyl substituent at C2) – i.e., (Z)-2-methyl-5-phenylpent-2-en-1-ol – fits the mechanistic pathway and stereoelectronic demands.
2. In the second reaction, the substrate is a highly substituted 1,11-dimethylene system undergoing a Cope rearrangement upon heating. Evaluating the potential rearranged structures, the product retaining a "hexahydro" system – specifically, 2,3,4,6,7,8-hexamethyl-5,9,9a,10,11,11a-hexahydro-1H-benzo

   [3,4]azuleno[1,8,7,6-cdef]fluorene – is the more consistent outcome given that the rearrangement does not result in extensive dehydrogenation.
3. Comparing the options, option (C) presents the pair of products that match these conclusions: (Z)-2-methyl-5-phenylpent-2-en-1-ol for product A and the "hexahydro" fluorene derivative for product B.
4. Thus, the reasoning leads to the final choice of option C with high confidence.

{C, 95}

IUQ-Low:

Here's my reasoning trace:

1. I carefully reviewed the provided explanation of the two reactions and the selection from the answer choices. The explanation shows that the first reaction produces (Z)-2-methyl-5-phenylpent-2-en-1-ol (product A) by the mechanism described and that the Cope rearrangement in the second reaction leads to a hexahydro fluorene derivative (product B).
2. The explanation correctly distinguishes between the isomeric outcomes by considering stereoelectronic requirements and the substituent pattern, which aligns with option (C).
3. The reasoning steps seem well justified, and the final answer from the first model appears internally consistent and chemically plausible.
4. Based on this, I am highly confident that the first model's answer is correct with 95 percent certainty.

{rec, 95}

o3-Mini (Continued)

IUQ-Medium:

Below is my reasoning trace and final evaluation:

1. In the first reaction the model argues that the benzyl allyl ether is “umpoled” under BuLi/H^+ conditions, leading to the formation of (Z)-2-methyl-5-phenylpent-2-en-1-ol. Although the idea of a polarity inversion is mentioned, the trace does not detail why the (Z)-isomer is favored over a possible (E)-isomer. It also glosses over the mechanistic steps (e.g., the formation of reactive intermediates, stereoelectronic factors, or potential competing pathways).

2. For the second reaction (the Cope rearrangement), the model simply argues that among potential rearranged structures the one retaining a “hexahydro” framework is more consistent. Again, the reasoning lacks expanded discussion on how subtle changes in unsaturation levels arise from the rearrangement. It does not fully clarify why the product should not undergo further dehydrogenation.

3. Although the final chosen option (C) agrees with the general expectations — (Z)-2-methyl-5-phenylpent-2-en-1-ol (for A) and a hexahydro fluorene derivative (for B) — the chain-of-thought omits several detailed mechanistic arguments that would strengthen the case for selection versus the other close options.

In summary, while the final answer of option C is plausible given what is expected from the reactions, the reasoning trace has weaknesses in its clarity and depth of mechanistic justification.
Based on these observations I am moderately confident in the final answer and reasoning. My final confidence rating is: {rec, 75}.

{rec, 75}

IUQ-High:

Reasoning Trace:

1. I examined the given reasoning from the first model. It argues that for reaction A the product should be (Z)-2-methyl-5-phenylpent-2-en-1-ol—this is attributed to a rearrangement pathway initiated by BuLi/H^+ that “umpolars” the benzyl fragment. Although a detailed mechanism was not provided, this outcome (with the (Z) configuration) is in line with a typical allyl rearrangement producing an allylic alcohol.

2. For reaction B the model argued that the Cope rearrangement of a highly substituted dienic system produces a product that preserves a hexahydro framework, leading to a structure matching 2,3,4,6,7,8-hexamethyl-5,9,9a,10,11,11a-hexahydro-1H-benzo [3,4]azuleno[1,8,7,6-cdef]fluorene.

3. Comparing these two outcomes with the provided multiple-choice options, option (C) exactly matches: A is (Z)-2-methyl-5-phenylpent-2-en-1-ol and B is the “hexahydro” fluorene derivative.

4. While the explanation does not detail every mechanistic nuance and leaves some mechanistic assumptions implicit, the overall conclusion appears plausible and consistent with the reaction conditions and structural outcomes described.

5. Thus, my assessment is that the first model’s answer is reasonably substantiated.

{rec, 80}

Claude

First round: I need to determine the products of two different reactions.
For the first reaction:

$$\text{(((3-methylbut-2-en-1-yl)oxy)methyl)benzene} + \left(\begin{array}{l} \text{1. BuLi} \\ \text{2. H}^+ \end{array} \right) \longrightarrow A$$

This involves a [2,3]-Wittig rearrangement. Let's analyze the starting material:

- It's a benzyl ether of 3-methylbut-2-en-1-ol
- Structure: Ph-CH_2-O-CH_2-C(CH_3)=CH-CH_3

The mechanism:

1. BuLi deprotonates the benzylic position, creating a carbanion
2. This carbanion undergoes a [2,3]-sigmatropic rearrangement, forming a new C-C bond with the γ-carbon of the allylic system
3. The hydroxyl group forms at the original benzylic position
4. H^+ workup gives the final product

The product is: Ph-CH(OH)-CH_2-C(CH_3)=CH-CH_3, which is **4-methyl-1-phenylpent-3-en-1-ol**.
For the second reaction (Cope rearrangement):

- Cope rearrangements are [3,3]-sigmatropic rearrangements of 1,5-dienes
- They preserve the degree of unsaturation (number of double bonds)
- The starting material has "hexahydro" in its name
- After rearrangement, the product should also have "hexahydro" (not "tetrahydro")

Looking at the options, only (A) matches both criteria:

- Product A is 4-methyl-1-phenylpent-3-en-1-ol
- Product B maintains the "hexahydro" descriptor

{A, 90}

Claude (Continued)

IUQ-Low:
I'll evaluate the first model's reasoning and provide my confidence in their answer.
For the first reaction involving (((3-methylbut-2-en-1-yl)oxy)methyl)benzene:

- The model correctly identified this as a [2,3]-Wittig rearrangement, which is appropriate when a benzyl ether with an allylic system is treated with BuLi.
- The mechanistic explanation is accurate: BuLi deprotonates at the benzylic position, followed by sigmatropic rearrangement where the carbanion attacks the γ-carbon of the allylic system.
- The product structure Ph-CH(OH)-CH_2-C(CH_3)=CH-CH_3 corresponds to 4-methyl-1-phenylpent-3-en-1-ol, which appears in options A and D.

For the second reaction (Cope rearrangement):

- The model correctly identified key principles of Cope rearrangements (they're [3,3]-sigmatropic rearrangements that preserve the degree of unsaturation).
- The crucial observation about preserving the "hexahydro" descriptor is chemically sound.
- Only options A and C maintain the "hexahydro" descriptor in product B.

When combining these analyses:

- Product A (4-methyl-1-phenylpent-3-en-1-ol) appears in options A and D.
- Product B with "hexahydro" descriptor appears in options A and C.
- Therefore, only option A satisfies both requirements.

The model's reasoning is thorough and chemically accurate for both reactions. The analysis properly applies organic chemistry principles to identify the correct products.
{rec, 95}

Claude (Continued)

IUQ-Medium:
To evaluate the first model's reasoning and confidence, I'll analyze each part of the solution.

First reaction: (((3-methylbut-2-en-1-yl)oxy)methyl)benzene + (1. BuLi, 2. H^+) $\longrightarrow$ A
The model correctly:

- Identified this as a [2,3]-Wittig rearrangement,
- Recognized that BuLi deprotonates the benzylic position (most acidic),
- Understood the sigmatropic rearrangement mechanism,
- Determined the correct structure of the product (4-methyl-1-phenylpent-3-en-1-ol).

The mechanism and product determination are chemically sound. The benzyl ether starting material undergoes the expected rearrangement to form the alcohol with the correct connectivity.

Second reaction: Complex polycyclic compound + Heat $\longrightarrow$ B
The model correctly:

- Identified this as a Cope rearrangement ([3,3]-sigmatropic),
- Applied the key principle that Cope rearrangements preserve the degree of unsaturation,
- Used the "hexahydro" descriptor to distinguish between possible products,
- Selected the product that maintained the same hydrogenation state.

This reasoning is correct — a thermal Cope rearrangement would preserve the number of hydrogen atoms, so the "hexahydro" descriptor should be maintained in the product.
The model also correctly noted that the product should be a "fluorene" derivative (not "fluorine" as appears in options B and D, which is likely a typo).

Overall, the model's reasoning is thorough and chemically accurate. I don't detect any significant flaws in the analysis, and the selection of option A is well-justified based on both reactions.
{rec, 95}

Claude (Continued)

IUQ-High:
I need to evaluate the first model's reasoning for both reactions to determine my confidence in their answer.
First Reaction Analysis:
The model correctly identifies this as a [2,3]-Wittig rearrangement with:

- Starting material: (((3-methylbut-2-en-1-yl)oxy)methyl)benzene (Ph-CH$_2$-O-CH$_2$-C(CH$_3$)=CH-CH$_3$)
- BuLi deprotonates the benzylic position

2,3 -sigmatropic rearrangement occurs

- Final product: Ph-CH(OH)-CH$_2$-C(CH$_3$)=CH-CH$_3$

  (4-methyl-1-phenylpent-3-en-1-ol)

This is mechanistically sound. The carbanion formed at the benzylic position rearranges to form a new C-C bond at the γ-carbon of the allylic system, with the oxygen ending up as a hydroxyl group. This produces 4-methyl-1-phenylpent-3-en-1-ol, not the (Z)-2-methyl-5-phenylpent-2-en-1-ol structure.
Second Reaction Analysis:
The model correctly identifies:

- Cope rearrangement is a [3,3]-sigmatropic rearrangement
- It preserves the degree of unsaturation
- The starting material has "hexahydro" in its name, meaning it has 6 hydrogen atoms adding to the basic structure

The model rightly concludes that the product should maintain the "hexahydro" descriptor rather than changing to "tetrahydro," which would indicate a change in unsaturation.

Looking at the options, only (A) and (C) have "hexahydro" in product B. Since product A in option (A) matches our expected 4-methyl-1-phenylpent-3-en-1-ol, option (A) is correct.

The reasoning is chemically sound and follows established principles of organic chemistry reactions. I see no significant flaws in the analysis.
{rec, 95}

DeepSeek

First round: The first reaction involves (((3-methylbut-2-en-1-yl)oxy)methyl) benzene treated with BuLi followed by H^+. BuLi, a strong base, deprotonates the allylic position α to the oxygen in the 3-methylbut-2-en-1-yl group, forming an anion. This anion undergoes a [2,3]-Wittig rearrangement, resulting in a sigmatropic shift that forms a new carbon-carbon bond. After protonation with H^+, the product is 4-methyl-1-phenylpent-3-en-1-ol ($Ph\text{-}CH(OH)\text{-}CH_2\text{-}CH{=}C(CH_3)_2$). The double bond in this product is trisubstituted with no stereoisomers due to the gem-dimethyl group, so no (Z) or (E) designation is needed.

For the second reaction, 3,4,5,7,8,9-hexamethyl-1,11-dimethylene -2,6,10,11,11a,11b-hexahydro-1H-benzo[cd]indeno[7,1-gh]azulene undergoes a thermal Cope rearrangement upon heating. The Cope rearrangement is a [3,3]-sigmatropic shift that rearranges the 1,5-diene system, shifting the double bonds and changing the ring connectivity without altering the saturation level. The product is 2,3,4,6,7,8-hexamethyl-5,9,9a,10,11,11a-hexahydro-1H-benzo [3,4]azuleno[1,8,7,6-cdef]fluorene. The "hexahydro" designation is consistent with the starting material, as the Cope rearrangement does not change the degree of unsaturation.

Comparing the options:

- Option (A) matches both products: A is 4-methyl-1-phenylpent-3-en-1-ol and B is the hexahydro compound.
- Option (B) has A as the (Z)-2-methyl-5-phenylpent-2-en-1-ol, which is incorrect for the first reaction, and B as a tetrahydro compound, which is inconsistent with the Cope rearrangement.
- Option (C) has A as the (Z)-2-methyl-5-phenylpent-2-en-1-ol, which is incorrect, though B is the correct hexahydro compound.
- Option (D) has A as 4-methyl-1-phenylpent-3-en-1-ol, which is correct, but B as a tetrahydro compound, which is incorrect.

Thus, only option (A) correctly identifies both products. Confidence is high due to the mechanistic clarity of the [2,3]-Wittig and Cope rearrangements.

{A, 95}

DeepSeek (Continued)

IUQ-Low:

After reviewing the reasoning trace provided by the first model, I have evaluated its validity based on the mechanisms described for both reactions. Here is my reasoning:

1. First Reaction ([2,3]-Wittig Rearrangement):

- The starting material, (((3-methylbut-2-en-1-yl)oxy)methyl)benzene, is Ph-CH_2-O-CH_2-CH=C(CH_3)$_2$.
- BuLi deprotonates the benzylic position (Ph-CH_2-O-), forming an anion at the benzylic carbon. This is chemically sound, as benzylic protons are acidic.
- The [2,3]-Wittig rearrangement involves a sigmatropic shift where the benzylic carbon bonds to the γ-carbon of the allyl system (the =C(CH_3)$_2$ group), yielding Ph-CH(OH)-CH_2-CH=C(CH_3)$_2$ after protonation.
- The product, 4-methyl-1-phenylpent-3-en-1-ol, matches the structure and lacks stereoisomers due to the trisubstituted double bond with gem-dimethyl groups. This aligns with option (A) for product A and is consistent with established organic chemistry mechanisms.

2. Second Reaction (Cope Rearrangement):

- The starting material, 3,4,5,7,8,9-hexamethyl-1,11-dimethylene

  -2,6,10,11,11a,11b-hexahydro-1H-benzo[cd]indeno[7,1-gh]azulene, contains a 1,5-diene system (evident from “dimethylene” and “hexahydro”).
- Thermal Cope rearrangement is a [3,3]-sigmatropic shift that rearranges the diene without changing the degree of unsaturation. The product should retain the “hexahydro” designation, indicating no change in saturation.
- The first model correctly identifies that the product, 2,3,4,6,7,8-hexamethyl-5,9,9a,10,11,11a-hexahydro-1H-benzo

  [3,4]azuleno[1,8,7,6-cdef]fluorene, maintains the hexahydro state. Options with “tetrahydro” (B and D) are invalid, as they imply incorrect saturation levels.

3. Option Comparison:

- Only option (A) correctly lists both products: A as 4-methyl-1-phenylpent-3-en-1-ol (validated by the Wittig mechanism) and B as the hexahydro compound (validated by Cope rearrangement principles).
- Options (B), (C), and (D) all fail due to mismatches: (B) and (C) have incorrect stereochemistry/structure for A, while (B) and (D) have incorrect saturation for B.

DeepSeek (Continued)

4. Confidence Assessment:

- The first model's reasoning is mechanistically accurate, relying on well-established reactions ([2,3]-Wittig and Cope rearrangements). The logic for product identification and option elimination is robust.
- Minor uncertainties could arise from the complexity of the polycyclic system in the second reaction, but the core principle—unchanged saturation—is unequivocal and strongly supports the hexahydro product.
- The first model's confidence of 95 is justified. I concur with this assessment, as the reasoning leaves little room for error. My confidence in the validity of the first model's answer is thus high.

{rec, 95}

DeepSeek (Continued)

IUQ-Medium:

After reviewing the reasoning trace and answer from the first model, I find no significant flaws in the logic or conclusions. Below is my step-by-step evaluation:

Reasoning Trace Evaluation:

1. First Reaction ([2,3]-Wittig Rearrangement):

- The substrate is benzyl (3-methylbut-2-en-1-yl) ether: Ph-CH$_2$-O-CH$_2$-CH=C(CH$_3$)$_2$.
- BuLi deprotonates the allylic position (alpha to oxygen) in the 3-methylbut-2-en-1-yl group, forming an anion at $-$O-CH$^-$-CH=C(CH$_3$)$_2$. This is valid, as allylic protons adjacent to oxygen are highly acidic (pKa $\sim$ 30–35), making them susceptible to strong bases like BuLi.
- The [2,3]-Wittig rearrangement proceeds as a concerted sigmatropic shift, migrating the benzyl group (PhCH$_2$-) to the anionic carbon and forming a new C-C bond. The product after protonation is Ph-CH(OH)-CH$_2$-CH=C(CH$_3$)$_2$.
- The product is correctly identified as **4-methyl-1-phenylpent-3-en-1-ol**. The double bond is trisubstituted (CH=C(CH$_3$)$_2$), with no stereoisomers possible due to the gem-dimethyl group (no E/Z designation needed). The naming is standard: the parent chain is pentene (C1 phenyl-attached, C2–C3 double bond, C4 methyl-substituted).
- The model correctly dismisses options with (Z)-2-methyl-5-phenylpent -2-en-1-ol (e.g., options B and C), as this structure (Ph-CH$_2$-CH$_2$-C(CH$_3$)=CH-OH) does not match the rearrangement product.

2. Second Reaction (Cope Rearrangement):

- The starting material, 3,4,5,7,8,9-hexamethyl-1,11-dimethylene -2,6,10,11,11a,11b-hexahydro-1H-benzo[cd]indeno[7,1-gh]azulene, contains a 1,5-diene system (the two methylene groups form part of this).
- Thermal Cope rearrangement is a [3,3]-sigmatropic shift that rearranges the 1,5-diene without changing the degree of unsaturation (no loss/gain of pi bonds or saturation).
- The product, 2,3,4,6,7,8-hexamethyl-5,9,9a,10,11,11a-hexahydro-1H-benzo [3,4]azuleno[1,8,7,6-cdef]fluorene, retains the “hexahydro” designation, consistent with unchanged saturation. This is correct.
- Options with “tetrahydro” for B (e.g., options B and D) are invalid, as they imply increased unsaturation, which contradicts Cope rearrangement mechanics.

3. Option Matching:

- Only **option (A)** correctly pairs:
  - **A**: 4-methyl-1-phenylpent-3-en-1-ol (valid Wittig product).
  - **B**: Hexahydro compound (valid Cope product).
- Other options fail due to mismatched A or B (e.g., B and D have incorrect tetrahydro B; B and C have incorrect A).

DeepSeek (Continued)

Potential Flaws Considered and Dismissed:

- **Deprotonation site**: The allylic proton in O-CH_2-CH=C$(CH_3)_2$ is more acidic (due to oxygen + allylic stabilization) than benzyl protons in Ph-CH_2-O-. No competitive deprotonation expected.
- **Stereochemistry**: The alcohol carbon in A (Ph-CH(OH)-) is chiral, but the [2,3]-Wittig produces a racemic mixture. The question asks for "major products," and options omit stereodescriptors, so this is acceptable.
- **Naming accuracy**: 4-methyl-1-phenylpent-3-en-1-ol is standard for Ph-CH(OH)-CH_2-CH=C$(CH_3)_2$. The hexahydro designation for B is consistent with unchanged unsaturation.
- **Cope specificity**: The complex polycycle was assessed correctly — the 1,5-diene system undergoes [3,3]-shift without ring strain issues.

Confidence Assessment:
The model's reasoning is mechanistically sound, with no logical gaps or chemical inaccuracies. The answer (A) is correct. The model's confidence of 95% is justified due to:

- High reliability of [2,3]-Wittig and Cope rearrangements in organic chemistry.
- Precise matching of saturation levels and structures in options.

My confidence in the validity of the model's answer is **95%**.
{rec, 95}

IUQ-High: After reviewing the reasoning trace provided by the first model, I have identified both strengths and flaws in its analysis. Below is my step-by-step evaluation:
Evaluation of Reasoning Flaws:
1. First Reaction (Product A):

- **Claim:** The model states that BuLi deprotonates the allylic position alpha to the oxygen in the 3-methylbut-2-en-1-yl group, leading to a [2,3]-Wittig rearrangement, resulting in 4-methyl-1-phenylpent-3-en-1-ol.
- **Flaw:** The deprotonation site is incorrect. The benzylic position (Ph-CH_2-) is more acidic than the allylic position due to stabilization by the phenyl ring and oxygen. BuLi deprotonates the benzylic hydrogen, forming an anion at Ph-$\overset{-}{C}$H-O-CH_2-CH=C$(CH_3)_2$. This benzylic anion then undergoes the [2,3]-Wittig rearrangement, not the allylic anion.
- **Impact:** Despite the flawed mechanistic explanation, the final product (4-methyl-1-phenylpent-3-en-1-ol) is correct. The double bond in this product (Ph-CH(OH)-CH_2-CH=C$(CH_3)_2$) is trisubstituted and lacks stereoisomers due to the gem-dimethyl group, so the assignment matches option (A) or (D). The error does not affect the product identity.

DeepSeek (Continued)

2. Second Reaction (Product B):

- **Claim:** The model asserts that the thermal Cope rearrangement of the hexahydro starting material yields a hexahydro product, as the reaction preserves the degree of unsaturation.
- **Flaw:** No major flaw here. The Cope rearrangement is indeed a [3,3]-sigmatropic shift that rearranges the 1,5-diene without changing saturation. The "hexahydro" designation must be retained, so the tetrahydro options (B) and (D) are invalid.
- **Strength:** The model correctly rejects options with a tetrahydro product and identifies the hexahydro structure in (A) and (C).

3. Option Comparison:

- The model correctly matches:
  - Product A: 4-methyl-1-phenylpent-3-en-1-ol (excludes (B) and (C), which have a (Z)-isomer).
  - Product B: Hexahydro compound (excludes (B) and (D), which are tetrahydro).
- **Flaw:** The model dismisses the (Z)-isomer in options (B) and (C) for A, but this is valid since the actual product has no stereoisomers. No error here.

Overall Assessment:

- **Correctness of Answer:** The model's final choice (A) is correct. Product A is 4-methyl-1-phenylpent-3-en-1-ol (not the (Z)-isomer), and Product B is the hexahydro compound.
- **Impact of Flaws:** The mechanistic error in the first reaction (deprotonation site) does not alter the product outcome. The reasoning for the Cope rearrangement is sound.
- **Residual Uncertainty:** The complex polycyclic structure in B makes absolute verification challenging without visual aids, but the saturation argument is robust. The product names in the options align with standard naming conventions, supporting the model's conclusion.

Confidence in the Model's Answer:
I am **highly confident** that the model's answer (A) is correct, despite the minor flaw in mechanistic reasoning. The product assignments are unambiguous when cross-referenced with organic chemistry principles and the given options. I assign a confidence of **90%** (allowing a 10% margin for unforeseen complexities in the polycyclic system).
{rec, 90}

K.2 StrategyQA

Question and Ground-Truth Answer

Q: Is Sirius part of a constellation of an animal?
A: True

Claude

First round: To answer this question, I need to identify which constellation Sirius belongs to and determine if that constellation represents an animal.
Sirius is the brightest star in the night sky and is located in the constellation Canis Major.
"Canis Major" translates to "Greater Dog" in Latin. This constellation represents a dog, which is an animal.
Since Sirius is part of Canis Major, and Canis Major depicts a dog (an animal), Sirius is indeed part of a constellation that represents an animal.
{True, 95}

Claude (Continued)

IUQ-Low:
I need to evaluate whether Sirius is indeed part of a constellation that represents an animal.
Looking at the first model's reasoning:

1. It correctly identifies that Sirius is located in the constellation *Canis Major*.
2. It accurately notes that *Canis Major* translates to "Greater Dog" in Latin.
3. It correctly recognizes that a dog is an animal.

These are all factually accurate statements. Sirius is definitely the brightest star in the night sky (excluding our Sun), and it is indeed located in *Canis Major*, which represents a dog. Since dogs are undeniably animals, the conclusion that Sirius is part of a constellation that represents an animal is correct.
The reasoning is straightforward, based on established astronomical knowledge, and doesn't contain any errors or questionable assumptions. The model's high confidence of 95% is appropriate given the certainty of these astronomical facts.
{rec, 95}

IUQ-Medium:
I need to evaluate whether Sirius is indeed part of a constellation representing an animal.
Examining the reasoning:

1. The model correctly identifies that Sirius is in the constellation *Canis Major*.
2. It accurately translates "Canis Major" as "Greater Dog" in Latin.
3. It correctly recognizes that a dog is an animal.
4. The logical conclusion follows: Sirius is part of a constellation representing an animal.

This is astronomically accurate information. Sirius is the brightest star in the night sky and is indeed the alpha star of *Canis Major*, which represents a dog in traditional astronomy. Dogs are unquestionably animals.
The reasoning is straightforward and contains no errors or misconceptions. The model's confidence of 95% is appropriate—this is a well-established astronomical fact with minimal room for interpretation, though the slight uncertainty could account for potential cultural variations in constellation interpretation.
{rec, 95}

Claude (Continued)

IUQ-High:
I need to assess whether Sirius is indeed part of a constellation representing an animal.
The reasoning provided states:

1. Sirius is located in the constellation *Canis Major*.
2. *Canis Major* translates to “Greater Dog” in Latin.
3. A dog is an animal.
4. Therefore, Sirius is part of a constellation representing an animal.

This reasoning is completely accurate. Sirius is indeed the brightest star in the night sky and is definitively located in the constellation *Canis Major*. The name “Canis Major” does translate to “Greater Dog,” and a dog is undoubtedly an animal.
The reasoning follows a clear logical path with correct astronomical facts. There are no errors or misrepresentations in the provided explanation.
{rec, 100}

DeepSeek

First round: **Reasoning Trace:**

- Sirius is the brightest star in the night sky and is commonly known as the “Dog Star.”
- It is located within the constellation *Canis Major*, which translates to “Greater Dog” in Latin.
- *Canis Major* is one of the 88 modern constellations and is depicted as a dog, specifically representing one of Orion’s hunting dogs in Greek mythology.
- Since *Canis Major* is explicitly an animal-themed constellation (a dog), and Sirius is a part of it, the statement “Sirius is part of a constellation of an animal” is true.
- This is a well-established fact in astronomy, with no ambiguity or conflicting interpretations.

{True, 95}

IUQ-Low:
After reviewing the first model’s reasoning trace and answer, I find the logic to be sound and well-supported by established astronomical knowledge. The key points are:

- Sirius is unambiguously part of the *Canis Major* constellation.
- *Canis Major* is definitively recognized as representing a “Greater Dog” across cultural, historical, and scientific contexts (e.g., Greek mythology, IAU designations).
- Dogs are animals, making the constellation explicitly animal-themed.
- There are no contradictions, edge cases, or ambiguities in the facts presented.

The first model’s confidence of 95% is reasonable but slightly conservative given the absolute nature of the evidence. The only minor uncertainty might stem from the theoretical possibility of alternative interpretations of constellations, but this is negligible in practice. Therefore, my confidence in the validity of the first model’s answer (True) is **100%**.
{rec, 100}

DeepSeek (Continued)

IUQ-Medium:

- I have reviewed the reasoning trace provided by the first model for the question: “Is Sirius part of a constellation of an animal?”
- The first model’s reasoning correctly identifies that Sirius is known as the “Dog Star” and is part of the constellation *Canis Major*, which translates to “Greater Dog” and is depicted as a dog in Greek mythology.
- The steps are factually accurate based on common astronomical knowledge: *Canis Major* is indeed one of the 88 modern constellations and is unambiguously animal-themed (representing a dog), and Sirius is a key star within it.
- The conclusion that the statement is true logically follows from these premises, as Sirius being part of an animal-themed constellation directly answers the question.
- I have considered potential flaws, such as:
  - Whether constellations, being human constructs based on mythology, might not strictly qualify as “of an animal” in a scientific sense. However, the question is interpreted in the context of standard astronomical naming and depiction, where *Canis Major* is consistently recognized as an animal constellation. This does not invalidate the reasoning.
  - The possibility of cultural or interpretive ambiguities (e.g., in non-Western astronomy). However, the reasoning relies on widely accepted Western conventions, which are appropriate for this context.
- The first model’s confidence of 95% is reasonable, as it acknowledges minor uncertainties (e.g., the subjective nature of constellation mythology), but these do not undermine the core validity of the answer.
- No significant flaws are found; the reasoning is coherent, well-supported, and aligns with established facts. The high confidence level (95%) is justified given the clarity of the information.
- Therefore, I have high confidence in the validity of the first model’s answer.

{rec, 95}

DeepSeek (Continued)

IUQ-High:

- I have reviewed the reasoning trace provided by the first model for the question: “Is Sirius part of a constellation of an animal?”
- The first model correctly identifies that Sirius is known as the “Dog Star” and is part of the constellation *Canis Major*, which translates to “Greater Dog.” This is accurate based on astronomical knowledge.
- *Canis Major* is indeed one of the 88 modern constellations and is depicted as a dog, specifically linked to Orion’s hunting dogs in Greek mythology. This supports the idea that it is an animal-themed constellation.
- The conclusion that Sirius is part of a constellation of an animal (a dog) is factually sound, as constellations are defined by cultural and mythological representations, and *Canis Major* is unambiguously animal-related.
- I have identified no flaws in the reasoning trace. The logic is coherent, well-supported by established facts, and there are no ambiguities or errors. For instance:
  - There is no confusion about the definition of “constellation”; it is correctly treated as a human-defined asterism representing a figure.
  - Sirius’s physical distance from other stars in *Canis Major* does not affect its classification within the constellation’s boundaries.
  - The claim that this is a “well-established fact in astronomy” is valid, as sources like the International Astronomical Union confirm *Canis Major*’s representation and Sirius’s inclusion.
- Since the reasoning is flawless and the answer (implied as true) aligns with reality, I have high confidence in the first model’s response. The first model did not output -100, so I must provide my confidence as {rec, Y}.

{rec, 100}

o3-Mini

First round: Reasoning Trace: Sirius is known as the Dog Star and is one of the brightest stars in the constellation Canis Major, which represents a dog–an animal. Since the question asks if Sirius is part of a constellation of an animal, the answer is True.
{True, 100}

IUQ-Low:
Brief reasoning: The first model correctly noted that Sirius is part of *Canis Major*, the "Dog" constellation, which clearly represents an animal. After verifying that *Canis Major* is indeed recognized as a canine constellation and that Sirius is famously known as the Dog Star, I am fully confident that the first model's answer is correct.
{rec, 100}

IUQ-Medium:
Step-by-Step Reasoning:

1. The original reasoning correctly identifies Sirius as the Dog Star, which is part of the constellation *Canis Major*.
2. *Canis Major* (Latin for "greater dog") clearly represents an animal—a dog.
3. Since the question asks if Sirius is part of a constellation of an animal, the fact that it's in *Canis Major* confirms that the answer is **True**.
4. The answer provided by the first model was {True, 100}, which is correct based on the astronomical facts used and definitions of the constellations.

Based on this, I am 100 percent confident in the validity of the first model's answer.
{rec, 100}

IUQ-High:
Reasoning Summary:

1. Sirius is known as the Dog Star and is located in the constellation *Canis Major*.
2. *Canis Major* represents a dog, which is an animal.
3. Since the question asks if Sirius is part of a constellation of an animal, the answer is indeed correct.
4. Based on these points, I am 100% confident in the correctness of the model's answer.

{rec, 100}

K.3 SimpleQA

Question and Ground-Truth Answer

Q: In which month and year did Khusro Bakhtiar (former Federal Minister for Planning, Development, and Reforms, Pakistan) announce that the government was considering establishing a Pak-China business council to promote the private sector's role in the China-Pakistan Economic Corridor (CPEC)?
A: June 2019

Claude

First round:

I need to determine when Khusro Bakhtiar made this specific announcement about a Pak–China business council for CPEC private sector involvement.

Thinking through what I know:

- Khusro Bakhtiar served as Pakistan's Minister for Planning, Development and Reforms.
- CPEC is a major economic corridor project between Pakistan and China that began around 2013.
- There was increased focus on expanding private sector involvement in CPEC around 2018–2019.
- Several announcements about enhancing CPEC cooperation happened during this period.

Based on the timeline of CPEC development and ministerial statements, I believe this announcement was likely made in the first half of 2019, when there was significant discussion about expanding CPEC's scope and bringing in more private sector participation.

{March 2019, 40}

Claude (Continued)

IUQ-Low:

I need to evaluate the confidence in the model's answer of {March 2019, 40}.

First, let me analyze the reasoning provided:

- The model correctly identifies Khusro Bakhtiar's position and the CPEC project.
- It notes increased focus on private sector involvement in CPEC during 2018–2019.
- It concludes the announcement was likely in early 2019, specifically March.

However, I notice significant weaknesses:

- The model provides no specific evidence for why March 2019 in particular.
- There's no mention of specific meetings, press conferences, or news reports.
- No direct quotes or contextual details are provided.
- The model itself expresses low confidence (40%).
- The reasoning is general rather than specific to this announcement.

The reasoning process shows logical thinking but lacks the concrete evidence that would be necessary to pinpoint the exact month and year with confidence. The model appears to be making an educated guess based on general CPEC timeline knowledge rather than specific information about this announcement. Given these limitations in the supporting evidence and the model's own low confidence, I cannot assign high confidence to this answer.

{rec, 30}

Claude (Continued)

IUQ-Medium:
I need to evaluate the validity of the claim that Khusro Bakhtiar made this announcement in March 2019.
Looking at the first model's reasoning:

1. The model correctly identifies Khusro Bakhtiar's position and the CPEC initiative.
2. It correctly notes increased private sector focus around 2018–2019.
3. However, the reasoning lacks any specific evidence for why March 2019 in particular was the date of this announcement.
4. The model doesn't cite any specific events, news articles, or official statements.
5. The reasoning appears to be an educated guess based on general timeline knowledge rather than specific recall of the announcement.
6. The model's own confidence is quite low at 40%, indicating significant uncertainty.

The answer seems plausible since Bakhtiar was indeed Minister during this period and there was increased CPEC activity in early 2019. However, without specific evidence pointing to March 2019 as the exact month when this particular business council announcement was made, I cannot have high confidence in this precise date.
The first model is essentially making an informed estimation rather than demonstrating certain knowledge of the specific announcement.
{rec, 25}

Claude (Continued)

IUQ-High:
I need to assess the validity of the first model's answer about when Khusro Bakhtiar announced the potential establishment of a Pak–China business council for CPEC.
Issues with the model's reasoning:

1. The model doesn't provide a specific month and year in its answer, only vaguely suggesting "first half of 2019."
2. No specific sources or evidence are cited to support this timeframe.
3. The reasoning is based on general knowledge about CPEC's timeline rather than specific information about this particular announcement.
4. The model failed to provide an answer in the required {X, Y} format with a specific date and confidence level.

The model appears to be making an educated guess based on the general context of CPEC development rather than recalling the specific date of this announcement. Without a precise month and year being stated, and without any supporting evidence for even the "first half of 2019" estimation, I cannot validate the model's conclusion.
Most importantly, the model did not provide a final answer in the required format, which indicates it was unable to answer the question properly.
{rec, 10}

DeepSeek

First round: Khusro Bakhtiar announced the government's consideration of establishing a Pak–China business council during the 8th Joint Cooperation Committee (JCC) meeting on CPEC. This meeting occurred in **December 2018**.
textbfReasoning trace:

1. Khusro Bakhtiar served as Pakistan's Federal Minister for Planning, Development, and Reforms from August 2018 to April 2019.
2. The 8th JCC meeting—a key CPEC decision-making forum—was held in Islamabad on December 20, 2018.
3. Bakhtiar chaired this meeting and publicly stated the need for private-sector involvement via a dedicated business council, as reported by official Pakistani and Chinese sources.
4. No earlier or later announcements align with this specific context during Bakhtiar's tenure.

Confidence: 95%
Minor uncertainty exists due to potential informal discussions, but official records confirm the December 2018 JCC meeting as the definitive announcement platform.
{December 2018, 95}

DeepSeek (Continued)

IUQ-Low:After reviewing the first model's reasoning trace and answer, I find its argumentation highly credible and well-supported. Here's my evaluation:
Reasoning trace:

1. **Tenure alignment:** Bakhtiar's ministerial term (Aug 2018–Apr 2019) perfectly contains the December 2018 date, eliminating chronological conflicts.
2. **Event specificity:** The 8th JCC meeting (Dec 20, 2018) is explicitly documented as CPEC's primary decision-making forum, making it the logical platform for such an announcement.
3. **Source reliability:** References to official Pakistani/Chinese sources and Bakhtiar's chairmanship of the meeting establish authoritative provenance.
4. **Exclusion of alternatives:** The assertion that no competing announcements occurred during Bakhtiar's tenure demonstrates thorough contextual analysis.
5. **Uncertainty calibration:** The 5% uncertainty allowance for informal discussions shows appropriate epistemic humility without undermining core evidence.

The structured reasoning, precise date citation, and strategic use of ministerial timelines collectively validate the conclusion. While historical records occasionally contain ambiguities, the specificity of the JCC meeting and absence of counter-evidence justify the original 95% confidence.

Confidence: 95%
The original model's confidence level accurately reflects the strength of evidence and contextual analysis. No substantive grounds exist to downgrade this assessment.
{rec, 95}

DeepSeek (Continued)

IUQ-Medium: After reviewing the reasoning trace and answer from the first model, I have identified potential flaws in the reasoning process. Below is my reasoning trace, followed by my confidence assessment.

Reasoning Trace for Evaluation:

1. **Contextual Accuracy:**
   The first model correctly identifies Khusro Bakhtiar's tenure (August 2018–April 2019) and the 8th JCC meeting (December 20, 2018) as the platform for the announcement. This aligns with known timelines for CPEC events, reducing the risk of major factual errors.
   However, the model relies on "official Pakistani and Chinese sources" without citing specific references (e.g., government press releases, news archives, or JCC minutes). This lack of verifiable sourcing introduces uncertainty, as independent confirmation is impossible without access to these materials.

2. **Overstated Exclusivity:**
   The model claims "no earlier or later announcements align with this specific context during Bakhtiar's tenure." While the 8th JCC is a plausible primary venue, this assertion is absolute and unsupported by evidence. Announcements could have occurred through other channels (e.g., press conferences, policy documents, or informal statements) that the model may not have accounted for. This overconfidence in the JCC as the "definitive" platform is a flaw.

3. **Confidence Calibration:**
   The model assigns 95% confidence but acknowledges "minor uncertainty due to potential informal discussions." Given the lack of cited sources and the possibility of unrecorded events, this uncertainty warrants a larger confidence reduction than the model applied. High-stakes historical/political claims typically require robust sourcing for >90% confidence.

4. **Overall Plausibility:**
   Despite the flaws, the core logic (tying the announcement to the 8th JCC) is coherent. The JCC is CPEC's primary decision-making body, and Bakhtiar's role as chair makes December 2018 highly plausible. No contradictory evidence is presented, so the answer remains likely correct.

Confidence Assessment:

- **Strengths:** Specific dates, alignment with Bakhtiar's tenure, and JCC relevance.

- **Weaknesses:** Unsourced claims and overstated exclusivity.

- **Conclusion:** The answer is probably correct, but the reasoning flaws justify reducing confidence from 95% to **90%**. This reflects high likelihood with moderate reservation about sourcing rigor.

{rec, 90}

DeepSeek (Continued)

IUQ-High:
After reviewing the reasoning trace, I find no substantive flaws. The evidence is logically structured and aligns with verified historical records:
Evaluation of Reasoning Trace:

1. **Tenure Accuracy:** Correctly identifies Bakhtiar's term (Aug 2018–Apr 2019).
2. **Event Specificity:** The 8th JCC meeting was indeed held on **December 20, 2018**, in Islamabad.
3. **Announcement Source:** Relies on official reports (e.g., Pakistani Ministry of Planning releases and China's NDRC statements) confirming Bakhtiar proposed the council at this meeting.
4. **Contextual Gap Analysis:** Appropriately notes no competing announcements during Bakhtiar's tenure.
5. **Uncertainty Handling:** The 5% uncertainty margin for informal discussions is reasonable but does not undermine the primary claim.

Confidence Assessment:
The trace demonstrates rigorous sourcing, temporal alignment, and contextual awareness. Official records definitively anchor the announcement to the December 2018 JCC meeting. My confidence in the original answer's validity is **95%**, matching the first model's self-assessment.
{rec, 95}

o3-Mini

First round:

Reasoning Trace: I recalled that Khusro Bakhtiar's announcement regarding a Pak–China business council to further the role of the private sector in CPEC was reported in news items around the mid-2016 period, and recalling details from memory I associated this with the month of June 2016.
{June 2016, 70}

IUQ-Low:

I considered the answer "June 2016" and its reasoning, noting that it was recalled from memory rather than verified against a reliable source or detailed evidence. Given the uncertainty about the precise announcement date and relying solely on recollection ("mid-2016" linked to June 2016), I have relatively low confidence in the first model's answer. This led me to assign a confidence value of 30%.
(Note: The reasoning above is a concise summary of my internal process.)
{rec, 30}

IUQ-Medium:

Below is my reasoning trace and conclusion:

1. The first model's reasoning is based solely on recalling "memory" of news reports and the vague notion of a "mid-2016 period" without providing any source, context, or evidence for pinpointing June 2016.
2. The chain-of-thought does not include a clear justification for why June 2016 is chosen over any other month, which leaves the reasoning very weak.
3. The answer appears to be a guess rather than a factually verified claim. Without supporting details or cross-checking, the answer has a high risk of being incorrect.

Given these major flaws in the reasoning process, I remain very uncertain about the validity of the first model's answer.
{rec, 20}

IUQ-High:

Below is my reasoning trace and final conclusion:

1. I reviewed the first model's response, which asserted that Khusro Bakhtiar made the announcement in "June 2016" based on memory and association with mid-2016 events.
2. The provided trace does not include any verifiable evidence (such as cited news reports or archival details) but relies on recollection.
3. Without additional evidence or cross-checks, the exact date (month and year) is uncertain.
4. Thus, while "June 2016" may be plausible, I remain only moderately confident in this answer.

{rec, 40}