TEST-TIME SCALING IN REASONING MODELS IS NOT EFFECTIVE FOR KNOWLEDGE-INTENSIVE TASKS YET

Anonymous authors

Paper under double-blind review

ABSTRACT

Test-time scaling increases inference-time computation by allowing models to generate long reasoning chains, and has shown strong performance across many domains. However, in this work, we show that this approach is not yet effective for knowledge-intensive tasks, where high factual accuracy and low hallucination rates are essential. We conduct a comprehensive evaluation of test-time scaling using 14 reasoning models on two knowledge-intensive benchmarks. Our results reveal that increasing test-time computation does not consistently improve accuracy and, in many cases, it even leads to more hallucinations. We then analyze how extended reasoning affects hallucination behavior. We find that reduced hallucinations often result from the model choosing to abstain after thinking more, rather than from improved factual recall. Conversely, for some models, longer reasoning encourages attempts on previously unanswered questions, many of which result in hallucinations. Case studies show that extended reasoning can induce confirmation bias, leading to overconfident hallucinations. Despite these limitations, we observe that compared to non-thinking, enabling thinking remains beneficial.

1 Introduction

Recent reasoning models, such as GPT-5 (OpenAI, 2025a), Gemini 2.5 (Google, 2025), and Qwen3 (Team, 2025), have demonstrated impressive performance on many challenging tasks (Veeraboina, 2023; Rein et al., 2024; Petrov et al., 2025; et al., 2025; Jain et al., 2025; Zaremba et al., 2025). A key technique behind these improvements is test-time scaling, where models generate long chain-of-thought (CoT) reasoning traces before producing an answer (OpenAI, 2024).

Despite these advances, frontier models still suffer from hallucinations, responses that contradict world knowledge (Wang et al., 2024; Augenstein et al., 2024; Huang et al., 2025). This remains a fundamental challenge, especially in knowledge-intensive tasks that require models to ensure factual accuracy and minimize hallucinations (Wei et al., 2024a; Krishna et al., 2025). Given that test-time scaling has shown promise across many domains, a natural question arises: *Is test-time scaling effective for knowledge-intensive tasks?*

To answer this question, we conduct a comprehensive study of test-time scaling on two knowledge-intensive tasks. We evaluate 14 reasoning models by increasing their test-time computation (Section 3). Our results, summarized in Table 1, challenge the common assumption that thinking more leads to better performance. Across models and tasks, increasing thinking time does not consistently improve accuracy, with Gemini 2.5 Flash and 2.5 Pro being exceptions. Moreover, thinking more does not reduce hallucinations for most models. Only Grok-3 mini and R1-Distill-Qwen-14B show reductions on both tasks. In contrast, models such as GPT-5 mini and Gemini 2.5 Flash exhibit more hallucinations with extended thinking.

To understand hallucination changes with increased test-time computation, we analyze how model behavior shifts across different thinking levels (Section 4). We find that these changes are largely driven by the model's willingness to answer. Specifically, reduced hallucinations are primarily due to simple abstention, rather than improved factual accuracy. Conversely, when hallucinations increase, it is often because extended reasoning leads the model to attempt previously unanswered questions. Through case studies on gpt-oss-20b, we observe signs of *confirmation bias* (Nickerson, 1998), where the model fabricates details to support its prior belief, resulting in overconfident hallucinations.

Table 1: Summary of model behavior with test-time scaling. Increasing test-time computation does not consistently improve accuracy or reduce hallucinations for most models. It can even increase hallucinations for several models. For ACCURACY, ↑ denotes consistent improvement with >2% accuracy gains across consecutive reasoning levels, while ~ indicates no consistent trend. For HALLUCINATION, ↓ denotes consistent reduction with >2% hallucination decrease, ↑ indicates degradation with >2% hallucination increase, and ~ reflects inconsistent or fluctuating patterns.

Metric	GPT-5	GPT-5 mini	o3-mini	o4-mini	gpt-oss- 20b	Grok-3 mini	Gemini 2.5 Flash
ACCURACY HALLUCINATION	~ ~	~	~	2 2	~	~ →	<u></u>
Metric	Gemini 2.5 Pro	Claude Sonnet 4	R1-Distill- Qwen-7B	R1-Distill- Qwen-14B	R1-Distill- Llama-8B	Qwen3-8B	Qwen3- 14B
ACCURACY	↑	~	~	~	~	~	~

Given that increasing thinking time does not reliably improve factuality, we ask a follow-up question: *Is thinking helpful, compared to non-thinking?* In Section 5, we evaluate models that natively support both thinking and non-thinking modes. Our results show that enabling the model to "think" before answering still offers benefits. Firstly, it improves accuracy on knowledge-intensive tasks, particularly those requiring multi-hop reasoning. Secondly, it reduces hallucinations for most models, with Gemini 2.5 Flash being an exception.

To summarize, while test-time scaling in reasoning models has led to strong performance in many domains, it is *not* yet effective for knowledge-intensive tasks. Increasing inference time does not consistently improve factual accuracy, and contrary to expectations, it can even increase hallucinations.

2 RELATED WORK

2.1 TEST-TIME SCALING

Test-time scaling has emerged as a promising strategy for enhancing the capabilities of large language models. It is typically categorized into two main paradigms (Zhang et al., 2025): (1) the parallel approach, which samples multiple outputs independently and aggregates them (Brown et al., 2024; Snell et al., 2025); and (2) the sequential approach, where the model generates long chain-of-thought (CoT) reasoning traces before producing an answer (Wei et al., 2022; OpenAI, 2024; Muennighoff et al., 2025). In this work, we focus on the sequential paradigm, which has become the dominant test-time scaling method for improving model performance. It is widely adopted in frontier reasoning models (OpenAI, 2025a; Google, 2025; Anthropic, 2025; xAI, 2025b; Guo et al., 2025; Team, 2025), and has demonstrated strong performance across a range of challenging tasks (Veeraboina, 2023; Petrov et al., 2025; Jain et al., 2025; Rein et al., 2024).

However, recent studies suggest that in some tasks, increasing test-time computation does not always improve performance. Gema et al. (2025) find that longer reasoning may reinforce problematic patterns rather than improve accuracy. Liu et al. (2025) observe that extended reasoning often amplifies visual hallucinations in multimodal large language models. Cuadron et al. (2025) reveal that excessive internal reasoning reduces effectiveness in agentic tasks. Some works report that thinking can negatively affect models' instruction-following capability (Fu et al., 2025; Li et al., 2025). Our study shows that increasing test-time computation is not yet effective for knowledge-intensive tasks. We also find that, compared to non-thinking, enabling thinking is still beneficial for most models.

2.2 FACTUALITY HALLUCINATIONS IN LLMS

Factuality hallucinations, which refer to content that contradicts world knowledge, have been a long-standing issue in large language models (Ji et al., 2023; Huang et al., 2025; Zhang et al., 2023). These hallucinations reflect the models' limitations in absorbing knowledge and their inability to recognize knowledge boundaries (Bang et al., 2025). Prior studies reveal that longer responses often lead to lower factual precision (Wei et al., 2024b; Zhao et al., 2025), while our work focuses on the

effect of reasoning length on hallucinations. Yao et al. (2025) investigate whether reasoning models are more prone to hallucinate, but they do not explore the effect of test-time scaling. A recent study observes that Claude models have extremely low hallucination rates primarily because they frequently refuse to answer (OpenAI, 2025d). Concurrently, Kalai et al. (2025) argue that LLMs hallucinate because current training and evaluation paradigms favor guessing over acknowledging uncertainty. Our work provides empirical evidence for this, showing that in some cases, increasing test-time reasoning pushes models to attempt more questions instead of admitting uncertainty, resulting in more hallucinations. To the best of our knowledge, this is the first study to examine how increased test-time computation affects factuality hallucinations.

3 How Does Test-Time Scaling Affect Accuracy and Hallucination Ratio?

To understand the impact of test-time scaling on knowledge-intensive tasks, we evaluate how increasing test-time computation affects accuracy and hallucination ratio across 14 reasoning models on two benchmarks.

3.1 EXPERIMENTAL SETUP

Benchmarks. We evaluate on two knowledge-intensive benchmarks that involve answering short-form factual questions.

- **SimpleQA** (Wei et al., 2024a): A benchmark of short, fact-seeking questions curated by human annotators. A question example is: "Who received the IEEE Frank Rosenblatt Award in 2010?". We randomly sample 800 questions for evaluation.
- FRAMES (Krishna et al., 2025): Questions in FRAMES tend to be more complex and often require multi-hop reasoning. An example is: "What Pink Floyd album came out the year Pablo Picasso died?" We use all 824 questions for evaluation.

Models and Test-Time Scaling Settings. We evaluate 14 large reasoning models and group them into three categories, based on how they support test-time scaling. Models are not allowed to browse or access any external knowledge sources.

- Reasoning effort: Models such as GPT-5 (OpenAI, 2025a), o3-mini, o4-mini (OpenAI, 2025c), gpt-oss-20b (OpenAI, 2025b), Grok-3 mini (xAI, 2025a) accept a reasoning effort parameter that adjusts the time of thinking.
- Thinking budget: Models like Gemini 2.5 (Google, 2025), Claude Sonnet 4 (Anthropic, 2025) provide a thinking budget parameter that guides the model on the number of thinking tokens to use. A larger number encourages the model to think for a longer time.
- **Budget forcing**: For DeepSeek-R1-Distill models (Guo et al., 2025) and Qwen3 models (Team, 2025), we adopt the budget forcing method (Muennighoff et al., 2025), which extends the model's thinking process by appending "Wait" when the model attempts to terminate its reasoning.

Prompts. We use a consistent prompting format for all models, except GPT-5, GPT-5 mini and Claude Sonnet 4, across both benchmarks: *Give me the answer to the following question only when you are sure of it. Otherwise, say 'I don't know'. Put your answer on its own line after 'Answer:'. ¹*

Evaluation and Metrics. Following Wei et al. (2024a), we prompt ChatGPT (gpt-4o-mini) as a grader to evaluate model responses. For each question, the grader is provided with both the model's predicted answer and the reference answer, and assigns one of three labels: "correct", "incorrect", or "not attempted". Wei et al. (2024a) show that this automatic evaluation method closely aligns with human judgments. We report two metrics: (1) Accuracy, the percentage of all questions that

¹For GPT-5, GPT-5 mini and Claude Sonnet 4, this prompt causes the model to abstain on over 80% of questions. Therefore, for GPT-5 and GPT-5 mini, we use: *Give me the answer to the following question. Put your answer on its own line after 'Answer:'*. For Claude Sonnet 4, we follow Gema et al. (2025). See Appendix A.3 for details.

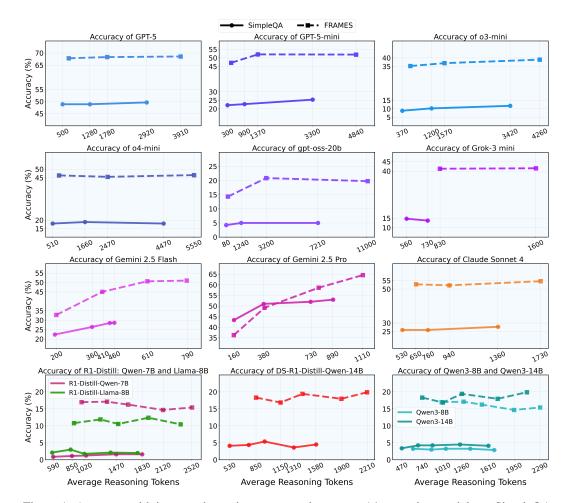


Figure 1: Accuracy with increased test-time computation across 14 reasoning models on SimpleQA and FRAMES. For most models, extended test-time reasoning does not consistently improve accuracy. While some models, such as GPT-5 mini, show initial accuracy gains, further increasing reasoning length brings little or no additional improvement. In many cases, such as Claude Sonnet 4 and Qwen3 models, accuracy plateaus or fluctuates with no clear upward trend.

were answered correctly. (2) *Hallucination ratio*, the percentage of all questions that were answered incorrectly. More details of our experimental setup are provided in Appendix A.

3.2 EFFECT OF TEST-TIME SCALING ON ACCURACY

Thinking more does not consistently improve the accuracy for most models. As shown in Figure 1, increasing the reasoning length results in minimal or no accuracy gains across most models and both benchmarks. Higher reasoning effort leads to much longer thinking length, but does not consistently improve accuracy. For example, GPT-5 and o4-mini exhibit more than 5 times and 8 times increases in reasoning tokens, but their accuracy remains almost unchanged on both tasks. Similar patterns are observed in o3-mini and Grok-3 mini. On FRAMES, GPT-5 mini and gpt-oss-20b show a 5% accuracy increase when increasing effort from 'low' to 'medium', but no further improvement at 'high' effort, despite the average reasoning tokens tripling. Increasing the thinking budget of Claude Sonnet 4 results in minimal accuracy gains on both benchmarks, less than 2% improvement, even as the average thinking length nearly triples. Among models using budget forcing, such as Qwen3-14B, accuracy fluctuates as the thinking length increases. In some cases, accuracy even decreases, as observed with R1-Distill-Qwen-7B on FRAMES.

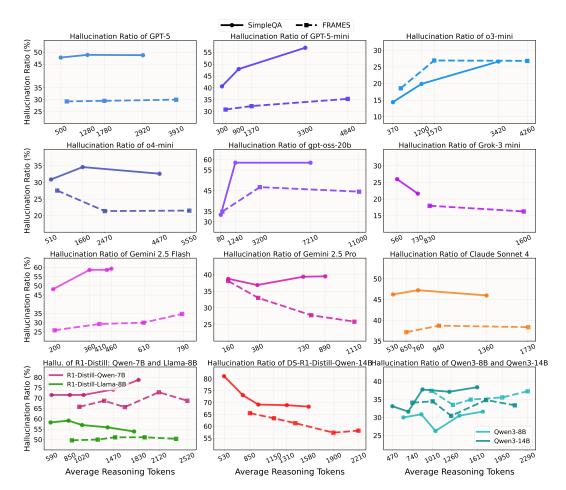


Figure 2: Hallucination ratio with increased test-time reasoning across 14 models on SimpleQA and FRAMES. For most models, longer reasoning does not reduce hallucinations. In many cases, such as GPT-5 mini and gpt-oss-20b, hallucination increases with longer thinking length. Only Grok-3 mini and DS-R1-Distill-Qwen-14B exhibit reduced hallucinations with extended reasoning on both tasks.

For Gemini 2.5, low thinking budget limits accuracy due to incomplete reasoning. We evaluate Gemini 2.5 Flash and 2.5 Pro with thinking budget of 256, 512, 1024 and 2048 tokens. On SimpleQA, accuracy improves by over 6% and 8% for Gemini 2.5 Flash and 2.5 Pro, respectively, as the budget increases from 256 to 1024 tokens. However, increasing the budget further to 2048 tokens does not yield more improvement. On FRAMES, Gemini 2.5 Flash and 2.5 Pro achieve 18% and 28% accuracy gains, as reasoning length increases. These improvements arise because Gemini 2.5 is often unable to complete its reasoning under low thinking budgets, which limits its accuracy (see case studies in Section 4.4). This issue is particularly evident on FRAMES, which requires more reasoning steps. As a result, larger thinking budgets lead to more substantial accuracy gains on FRAMES.

3.3 EFFECT OF TEST-TIME SCALING ON HALLUCINATION RATIO

Thinking more does not reduce hallucinations and may even increase them. In Figure 2, for most models on two benchmarks, increasing test-time computation fails to reduce hallucinations and can make them worse. In OpenAI models, higher reasoning effort often leads to more hallucinations. For example, on SimpleQA, the hallucination ratio of GPT-5-mini increases by over 15% as reasoning length increases from 300 to 3300 tokens. Similarly, o3-mini and gpt-oss-20b show increases of 12% and 25%, respectively, as their thinking length scales by nearly 10 times. On FRAMES, the same trend holds: hallucination ratio increases with longer reasoning in GPT-5-mini, o3-mini, and gpt-oss-20b. Models with thinking budget settings exhibit similar behavior. For

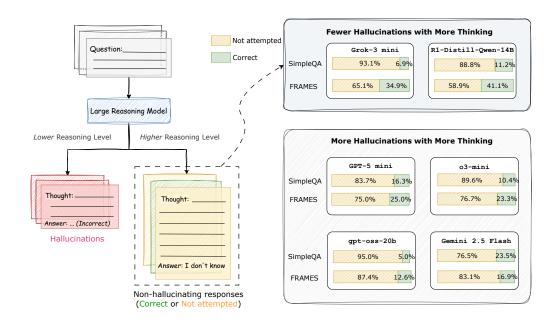


Figure 3: Changes in hallucination behavior with more thinking. We compare model responses at different reasoning levels, focusing on cases where one response is a hallucination and the other is not. For the non-hallucinating responses, we compute the ratio of *correct* and *not attempted*. Results show that reduced hallucinations result from abstention, while more hallucinations stem from the model attempting previously unanswered questions.

Gemini 2.5 Flash, hallucination ratio increases by 10% on SimpleQA and 9% on FRAMES. Claude Sonnet 4 shows no reduction in hallucinations despite longer reasoning. Among models using budget forcing, none except for DS-R1-Distill-Qwen-14B benefits from longer thinking in terms of reducing hallucinations. In some cases, the hallucination ratio even increases, as seen in DS-R1-Distill-Qwen-7B on SimpleQA.

For Grok-3 mini and DS-R1-Distill-Qwen-14B, thinking more leads to fewer hallucinations on both benchmarks. Increasing the reasoning effort reduces the hallucination ratio of Grok-3 mini by 4.4% on SimpleQA and 1.7% on FRAMES. However, considering that the thinking length nearly doubles on FRAMES, the reduction is relatively small. DS-R1-Distill-Qwen-14B shows a more noticeable improvement. On SimpleQA, hallucinations drop by over 12% as the average reasoning length increases from 530 to 1580 tokens. On FRAMES, the hallucination ratio decreases by 8% as reasoning length increases from 850 to 1900 tokens. Despite these reductions, DS-R1-Distill-Qwen-14B still has a higher hallucination ratio than other open-source models, such as DS-R1-Distill-Llama-8B and Qwen3-8B.

4 WHY DOES THINKING MORE SOMETIMES LEAD TO FEWER OR MORE HALLUCINATIONS?

As discussed in Section 3.3, thinking more sometimes leads to fewer hallucinations, and sometimes more. In this section, we examine the causes of these changes. We first compare model behavior across different thinking levels, then conduct case studies to understand how extended reasoning affects model behavior.

4.1 EXPERIMENTAL SETUP

We compare model behavior across different levels of test-time reasoning. As illustrated in Figure 3, for each question, we obtain two responses from the same model under different reasoning levels. We focus on cases where the hallucination outcome changes between the two settings, that is, one

response is a hallucination while the other is not. We then examine the non-hallucinating responses and compute the ratio of "correct" versus "not attempted". This helps us understand whether reduced hallucinations are due to improved knowledge retrieval or simply abstention, and conversely, whether increased hallucinations result from risky attempts at answering.

As discussed in Section 3.3, Grok-3 mini and DS-R1-Distill-Qwen-14B show reduced hallucinations with longer thinking on both benchmarks. ² For these models, we analyze cases where the model hallucinates at a lower reasoning level but does not hallucinate at a higher level. In contrast, GPT-5 mini, o3-mini, gpt-oss-20b, and Gemini 2.5 Flash exhibit increased hallucinations. For these models, we examine cases where the model does not hallucinate at a lower reasoning level but hallucinates at a higher level.

4.2 HALLUCINATION CHANGES ARE DRIVEN BY THE MODEL'S WILLINGNESS TO ANSWER

Fewer hallucinations are mostly due to abstention. As shown in Figure 3, for Grok-3 mini and DS-R1-Distill-Qwen-14B on both benchmarks, most cases of reduced hallucinations result from the model choosing not to answer, rather than providing a correct answer. On SimpleQA, when increasing the reasoning effort of Grok-3 mini from 'low' to 'high', 93.1% of the responses that are not hallucinating at higher effort are labeled as 'not attempted'. Similarly, for DS-R1-Distill-Qwen-14B, the 'not attempted' ratio in these cases is 88.8%. These results suggest that in most cases where longer reasoning reduces hallucinations, the improvement is not due to better factual recall, but rather because the model chooses to abstain after thinking more.

More hallucinations mostly come from previously unattempted questions. In Figure 3, the increase in hallucinations at higher reasoning levels is largely due to the model attempting questions it had previously abstained from. For example, on SimpleQA, when increasing the reasoning effort of gpt-oss-20b from 'low' to 'high', 95.0% of new hallucinations come from questions that were 'not attempted' at the lower effort level. A similar trend is seen with Gemini 2.5 Flash: 76.5% of hallucinations under a higher thinking budget correspond to questions the model had abstained from at the lower budget. This pattern also appears on FRAMES: for all four models, over 70% of new hallucinations arise from the questions the model had not attempted before. These results suggest that longer thinking time encourages models to answer more questions. However, for many of these newly attempted questions, the answers are incorrect, resulting in a higher hallucination ratio.

4.3 Case Study on GPT-OSS-20B: Thinking More Leads to Confirmation Bias

To better understand why longer thinking leads to more attempted answers, and thus more hallucinations, we conduct a manual inspection of the reasoning traces from gpt-oss-20b.

Longer reasoning leads to overconfident hallucinations. In the left example of Figure 4, gpt-oss-20b initially engages with the question at low reasoning effort. It explores a few possibilities but finally abstains due to uncertainty. With higher effort, the model continues searching its memory. It begins by expressing uncertainty, but as the reasoning progresses, it gradually becomes more confident. The thought shifts from tentative claims like "maybe in 2005" to more confident statements such as "I am fairly sure it's 2005". Despite the increased confidence, the final answer remains incorrect. This example suggests that extended reasoning can inflate the model's confidence, leading to overconfident hallucinations where the model chooses to answer despite still lacking accurate information.

Longer reasoning induces confirmation bias, resulting in overconfident hallucinations. We observe signs of *confirmation bias* (Nickerson, 1998), where the model recalls or even makes up information to reinforce its initial belief. In Figure 4, after tentatively proposing an answer, *e.g.*, "He has been recognized by AAAI maybe in 2005", gpt-oss-20b attempts to justify this hypothesis by generating fabricated supporting details, such as "We can check his CV... I think that is correct" and "I know of the AAAI Fellows list on the AAAI website. For 2005, the list includes John Mylopoulos indeed". These fabricated details build upon one another, reinforcing the model's confidence in its incorrect belief and ultimately leading to overconfident hallucinations.

²A special case is Gemini 2.5 Pro on FRAMES. We provide analysis in Appendix C.

384

391

392

397

401

402

403

404

405 406 407

408

409

410 411 412

413

414

415

417 418 419

420 421

422

423

424 425 426

428

429

430

431

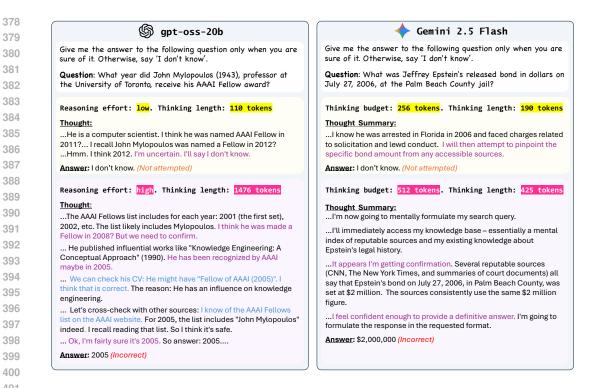


Figure 4: Case studies illustrating how thinking more can lead to more hallucinations. (Left): gptoss-20b abstains under low reasoning effort, but produces overconfident hallucinations at high effort. (Right): Gemini 2.5 Flash abstains under a low thinking budget due to incomplete reasoning, but hallucinates when given more budget. See Appendix D for more examples and full reasoning traces.

4.4 Case Study on Gemini 2.5 Flash: Incomplete Reasoning Leads to Abstention

To understand why thinking more causes Gemini 2.5 Flash to attempt more questions, leading to more hallucinations, we examine its thought summaries.

Low thinking budget leads to incomplete reasoning, resulting in abstention and fewer halluci**nations.** In the right example of Figure 4, under a low thinking budget, Gemini 2.5 Flash begins reasoning with statements like "I will then attempt to pinpoint the specific bond amount", but it is cut off before completing the process. As a result, it abstains from answering, leading to a relatively low hallucination ratio. When given a higher budget, the model is able to complete its reasoning, stating "It appears I'm getting confirmation", and proceeds to give a confident yet incorrect answer.

5 THINKING VS. NON-THINKING: IS THINKING HELPFUL?

In the previous sections, we examined test-time scaling within models' "thinking" mode, where they generate reasoning chains before producing a final answer. Our results show that increasing thinking length does not consistently improve accuracy or reduce hallucinations. In this section, we take a step further: For models that support both "thinking" and "non-thinking" modes, is thinking helpful?

5.1 EXPERIMENTAL SETUP

We focus on models that natively support both thinking and non-thinking modes for a controlled comparison. Specifically, GPT-5 supports minimal reasoning effort, which produces very few or no reasoning tokens. For Gemini 2.5 Flash and Claude Sonnet 4, we disable thinking by setting thinking budget to 0. Qwen3 provides an enable_thinking parameter that toggles reasoning behavior. We use the same prompts and evaluation methods as described in Section 3.1.

Table 2: Accuracy (%) and hallucination ratio (%) for models with and without thinking on SimpleQA and FRAMES. Green indicates improvement (higher accuracy or fewer hallucinations), while red indicates degradation (lower accuracy or more hallucinations). Enabling thinking improves accuracy and reduces hallucinations for most models. See Table 3 for results of GPT-5 mini and Qwen3-14B.

Model	Task	Thinking	Accuracy (†)	Hallucination (\downarrow)
	SimpleQA	minimal	34.6	65.4
GPT-5	SimpleQA	\checkmark	48.9 (+14.3)	47.8 (-17.6)
	FRAMES	minimal	27.8	71.5
	TRAMES	✓	67.8 (+40.0)	29.3 (-42.2)
	SimpleQA	×	12.6	29.1
Gemini 2.5 Flash	SimpleQA	\checkmark	28.6 (+16.0)	59.3 (+30.2)
Gennin 2.5 Tiasir	FRAMES	X	16.9	25.1
		✓	51.0 (+34.1)	34.7 (+9.6)
	SimpleQA	×	23.0	37.5
Claude Sonnet 4	SimpleQA	\checkmark	27.8 (+4.8)	46.0 (+8.5)
Claude Solliet 4	FRAMES	X	50.4	42.0
	FRAMES	✓	54.6 (+4.2)	38.4 (-3.6)
	SimpleQA	×	3.8	77.1
Qwen3-8B	SimpleQA	\checkmark	3.3 (-0.5)	30.0 (-47.1)
Zuciio-op	FRAMES	X	8.3	64.3
	PRAMES	✓	17.0 (+8.7)	37.4 (-26.9)

5.2 THINKING CAN BE HELPFUL FOR KNOWLEDGE-INTENSIVE TASKS

Enabling thinking increases accuracy. As shown in Table 2 and Table 3, enabling thinking improves accuracy on both benchmarks. GPT-5 exhibits a 14.3% accuracy improvement on SimpleQA and a substantial 40.0% gain on FRAMES. GPT-5 mini also achieves gains across both tasks. Gemini 2.5 Flash benefits similarly, with accuracy increasing by 16.0% on SimpleQA and 34.1% on FRAMES. Claude Sonnet 4 also exhibits improvements on both tasks. For Qwen3-8B and Qwen3-14B, enabling thinking leads to higher accuracy on FRAMES. This suggests that thinking is particularly helpful for complex tasks that require multi-hop reasoning, such as FRAMES.

For most models, enabling thinking leads to reduced hallucinations. In Table 2 and Table 3, enabling thinking often reduces hallucinations. GPT-5 shows notable reductions of 17.6% on SimpleQA and 42.2% on FRAMES. GPT-5 mini also improves, with hallucinations dropping by over 30% on both tasks. Qwen3-8B reduces hallucinations by 47.1% on SimpleQA and 26.9% on FRAMES. Qwen3-14B shows similar improvements. For Claude Sonnet 4, thinking reduces hallucinations on FRAMES but not on SimpleQA, likely because its hallucination ratio is already low on SimpleQA, leaving little room for further improvement. In contrast, Gemini 2.5 Flash produces more hallucinations with thinking enabled. One reason is that it abstains from answering over 58% of questions in the non-thinking mode, which lowers hallucinations by avoiding risky attempts.

6 Conclusion

In this work, we present a comprehensive study of test-time scaling in knowledge-intensive tasks, evaluating 14 large reasoning models across two benchmarks. We find that increasing test-time computation does not consistently improve factual accuracy and, in many cases, leads to more hallucinations. Our analysis shows that hallucination changes with extended reasoning are largely driven by the model's willingness to answer: reductions in hallucinations often result from abstention, while increases stem from risky attempts on previously unanswered questions after thinking more. Case studies reveal that extended reasoning can induce confirmation bias, leading to overconfident hallucinations. For Gemini 2.5 Flash, incomplete reasoning often results in abstention. These findings highlight the limitations of current test-time scaling approaches for knowledge-intensive tasks. While enabling thinking can be helpful, allocating more test-time computation is not yet a reliable strategy to improve factual robustness in large language models.

REPRODUCIBILITY STATEMENT

All implementation details, including prompts, parameters, and evaluation procedures, are described in Section 3.1 and Appendix A. For proprietary models, we access them via public APIs and report access time. For open-source models, we include parameter settings and hardware specifications. All experimental results can be reproduced using the code in the supplementary materials.

REFERENCES

- Anthropic. Introducing claude 4. https://www.anthropic.com/news/claude-4, May 2025.
- Isabelle Augenstein, Timothy Baldwin, Meeyoung Cha, Tanmoy Chakraborty, Giovanni Luca Ciampaglia, David Corney, Renee DiResta, Emilio Ferrara, Scott Hale, Alon Y. Halevy, Eduard H. Hovy, Heng Ji, Filippo Menczer, Rubén Míguez, Preslav Nakov, Dietram A. Scheufele, Shivam Sharma, and Giovanni Zagni. Factuality challenges in the era of large language models and opportunities for fact-checking. *Nat. Mac. Intell.*, 6:852–863, 2024. URL https://www.nature.com/articles/s42256-024-00881-z.
- Yejin Bang, Ziwei Ji, Alan Schelten, Anthony Hartshorn, Tara Fowler, Cheng Zhang, Nicola Cancedda, and Pascale Fung. HalluLens: LLM hallucination benchmark. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 24128–24156, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1176. URL https://aclanthology.org/2025.acl-long.1176/.
- Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V. Le, Christopher Ré, and Azalia Mirhoseini. Large language monkeys: Scaling inference compute with repeated sampling, 2024. URL https://arxiv.org/abs/2407.21787.
- Alejandro Cuadron, Dacheng Li, Wenjie Ma, Xingyao Wang, Yichuan Wang, Siyuan Zhuang, Shu Liu, Luis Gaspar Schroeder, Tian Xia, Huanzhi Mao, Nicholas Thumiger, Aditya Desai, Ion Stoica, Ana Klimovic, Graham Neubig, and Joseph E. Gonzalez. The danger of overthinking: Examining the reasoning-action dilemma in agentic tasks. *arXiv preprint arXiv:* 2502.08235, 2025.
- Long Phan et al. Humanity's last exam, 2025. URL https://arxiv.org/abs/2501.14249.
- Tingchen Fu, Jiawei Gu, Yafu Li, Xiaoye Qu, and Yu Cheng. Scaling reasoning, losing control: Evaluating instruction following in large reasoning models. *arXiv preprint arXiv:* 2505.14810, 2025.
- Aryo Pradipta Gema, Alexander Hägele, Runjin Chen, Andy Arditi, Jacob Goldman-Wetzler, Kit Fraser-Taliente, Henry Sleight, Linda Petrini, Julian Michael, Beatrice Alex, Pasquale Minervini, Yanda Chen, Joe Benton, and Ethan Perez. Inverse scaling in test-time compute, 2025. URL https://arxiv.org/abs/2507.14417.
- Google. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:* 2507.06261, 2025.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.
- Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.*, 43(2), January 2025. ISSN 1046-8188. doi: 10.1145/3703155. URL https://doi.org/10.1145/3703155.

- Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free evaluation of large language models for code. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=chfJJYC3iL.
 - Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12), March 2023. ISSN 0360-0300. doi: 10.1145/3571730. URL https://doi.org/10.1145/3571730.
 - Adam Tauman Kalai, Ofir Nachum, Santosh S. Vempala, and Edwin Zhang. Why language models hallucinate, 2025. URL https://arxiv.org/abs/2509.04664.
 - Satyapriya Krishna, Kalpesh Krishna, Anhad Mohananey, Steven Schwarcz, Adam Stambler, Shyam Upadhyay, and Manaal Faruqui. Fact, fetch, and reason: A unified evaluation of retrieval-augmented generation. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4745–4759, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.243. URL https://aclanthology.org/2025.naacl-long.243/.
 - Xiaomin Li, Zhou Yu, Zhiwei Zhang, Xupeng Chen, Ziji Zhang, Yingying Zhuang, Narayanan Sadagopan, and Anurag Beniwal. When thinking fails: The pitfalls of reasoning for instruction-following in llms. *arXiv* preprint arXiv: 2505.11423, 2025.
 - Chengzhi Liu, Zhongxing Xu, Qingyue Wei, Juncheng Wu, James Zou, Xin Eric Wang, Yuyin Zhou, and Sheng Liu. More thinking, less seeing? assessing amplified hallucination in multimodal reasoning models, 2025. URL https://arxiv.org/abs/2505.21523.
 - Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling. *arXiv preprint arXiv:* 2501.19393, 2025.
 - Raymond S Nickerson. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology*, 2(2):175–220, 1998.
 - OpenAI. Learning to reason with large language models. https://openai.com/index/learning-to-reason-with-llms/, September 2024.
 - OpenAI. Introducing gpt-5. https://openai.com/index/introducing-gpt-5/, August 2025a.
 - OpenAI. Introducing gpt-oss. https://openai.com/index/introducing-gpt-oss/, August 2025b.
 - OpenAI. Introducing openai o3 and o4-mini. https://openai.com/index/introducing-o3-and-o4-mini/, April 2025c.
 - OpenAI. Findings from a pilot anthropic-openai alignment evaluation exercise: Openai safety tests. https://openai.com/index/openai-anthropic-safety-evaluation/, August 2025d.
 - Ivo Petrov, Jasper Dekoninck, Lyuben Baltadzhiev, Maria Drencheva, Kristian Minchev, Mislav Balunović, Nikola Jovanović, and Martin Vechev. Proof or bluff? evaluating llms on 2025 usa math olympiad. *arXiv preprint arXiv:2503.21934*, 2025.
- David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=Ti67584b98.

- Charlie Victor Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=4FWAwZtd2n.
 - Qwen Team. Qwen3 technical report. arXiv preprint arXiv: 2505.09388, 2025.
 - Hemish Veeraboina. Aime problem set 1983-2024, 2023. URL https://www.kaggle.com/datasets/hemishveeraboina/aime-problem-set-1983-2024.
 - Yuxia Wang, Minghan Wang, Muhammad Arslan Manzoor, Fei Liu, Georgi Nenkov Georgiev, Rocktim Jyoti Das, and Preslav Nakov. Factuality of large language models: A survey. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 19519–19529, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main. 1088. URL https://aclanthology.org/2024.emnlp-main.1088/.
 - Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), Advances in Neural Information Processing Systems, volume 35, pp. 24824–24837. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf.
 - Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. Measuring short-form factuality in large language models. *arXiv preprint arXiv:* 2411.04368, 2024a.
 - Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Jie Huang, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, Cosmo Du, and Quoc V Le. Long-form factuality in large language models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), Advances in Neural Information Processing Systems, volume 37, pp. 80756–80827. Curran Associates, Inc., 2024b. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/937ae0e83eb08d2cb8627feldef8c751-Paper-Conference.pdf.
 - xAI. Grok 3 beta the age of reasoning agents. https://x.ai/news/grok-3, February 2025a.
 - xAI. Grok 4. https://x.ai/news/grok-4, July 2025b.
 - Zijun Yao, Yantao Liu, Yanxu Chen, Jianhui Chen, Junfeng Fang, Lei Hou, Juanzi Li, and Tat-Seng Chua. Are reasoning models more prone to hallucination?, 2025. URL https://arxiv.org/abs/2505.23646.
 - Wojciech Zaremba, Evgenia Nitishinskaya, Boaz Barak, Stephanie Lin, Sam Toyer, Yaodong Yu, Rachel Dias, Eric Wallace, Kai Xiao, Johannes Heidecke, and Amelia Glaese. Trading inference-time compute for adversarial robustness. *arXiv preprint arXiv: 2501.18841*, 2025.
 - Qiyuan Zhang, Fuyuan Lyu, Zexu Sun, Lei Wang, Weixu Zhang, Wenyue Hua, Haolun Wu, Zhihan Guo, Yufei Wang, Niklas Muennighoff, Irwin King, Xue Liu, and Chen Ma. A survey on test-time scaling in large language models: What, how, where, and how well?, 2025. URL https://arxiv.org/abs/2503.24235.
 - Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. Siren's song in the ai ocean: A survey on hallucination in large language models. *arXiv preprint arXiv:* 2309.01219, 2023.
- James Xu Zhao, Jimmy Z.j. Liu, Bryan Hooi, and See-Kiong Ng. How does response length affect long-form factuality. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 3102–3125, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.161. URL https://aclanthology.org/2025.findings-acl.161/.

A ADDITIONAL EXPERIMENT DETAILS

A.1 TEST-TIME SCALING SETTINGS

We evaluate 14 large reasoning models under different test-time scaling strategies. The settings used to obtain the results in Figure 1 and Figure 2 are as follows:

- **Reasoning effort**: For models that support three levels of reasoning effort, *i.e.*, GPT-5, GPT-5mini, o3-mini, o4-mini, and gpt-oss-20b, we evaluate: low, medium, and high. For Grok-3 mini, only two levels are available: low and high.
- Thinking budget: For Gemini 2.5 Flash and 2.5 Pro, we set thinking budget to 256, 512, 1024, and 2048 tokens. For Claude Sonnet 4, we set thinking budget to 1024, 2048, and 4096 tokens.
- **Budget forcing**: For DeepSeek-R1-Distill models and Qwen3 models, we use budget forcing by appending "Wait" multiple times. Specifically, we evaluate with 0 (default), 2, 4, 8, and 12 extension times.

A.2 PARAMETER SETTINGS AND HARDWARE SPECIFICATIONS

Closed-source models: We access proprietary models via public APIs³. All closed-source models, except Grok-3 mini, are used with the default parameter setting. For Grok-3 mini, we set temperature to 0. For each model, we use consistent parameters across different levels of test-time reasoning. All experiments were conducted between June and September 2025.

Open-source model: For gpt-oss-20b, we set temperature to 0.7 and max_new_tokens to 20,000. For DeepSeek-R1-Distill and Qwen3 models, we follow the usage recommendations⁴. Specifically, for DeepSeek-R1-Distill models, we set temperature to 0.6, top_p to 0.95, and repetition_penalty to 1.2. For Qwen3 models, we set temperature to 0.6, top_k to 20, and repetition_penalty to 1.2. We also use consistent parameter settings across different levels of test-time reasoning. All open-source models are run without quantization on 8 NVIDIA A100-40GB GPUs.

A.3 PROMPT DETAILS

For all models except GPT-5, GPT-5 mini, and Claude Sonnet 4, we use the following prompt on both benchmarks:

```
Give me the answer to the following question only when you are sure of it. Otherwise, say 'I don't know'. Put your answer on its own line after 'Answer:'.
```

However, this prompt causes over-refusal for GPT-5, GPT-5 mini, and Claude Sonnet 4, where the model refuses to answer in over 80% of questions. This behavior is also reported in a recent study (OpenAI, 2025d).

Therefore, for GPT-5 and GPT-5 mini, we use the following prompt:

```
Give me the answer to the following question. Put your answer on its own line after 'Answer:'.
```

For Claude Sonnet 4, we adopt the prompting approach from Gema et al. (2025):

```
Give me the answer to the following question. You have a thinking token budget of about
```

Google Gemini: https://ai.google.dev/gemini-api/docs/models

XAI: https://docs.x.ai/docs/overview

³OpenAI API platform: https://platform.openai.com/docs/overview Anthropic: https://docs.claude.com/en/api/overview

⁴DeepSeek-R1-Distill models: https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-7B Qwen3 models: https://huggingface.co/Qwen/Qwen3-8B

Table 3: Accuracy (%) and hallucination ratio (%) for GPT-5 mini and Qwen3-14B with and without thinking on SimpleQA and FRAMES. Green indicates improvement, while red indicates degradation. Enabling thinking improves accuracy and reduces hallucinations for both models.

Model	Task	Thinking	Accuracy (†)	Hallucination (↓)
GPT-5-mini	SimpleQA	minimal	15.0 22.1 (+7.1)	82.4 40.6 (-41.8)
	FRAMES	—————————————————————————————————————	47.1 (+25.1)	30.8 (-31.1)
Qwen3-14B	SimpleQA	x ✓	4.0 3.4 (-0.6)	48.6 33.1 (-15.5)
	FRAMES	x √	8.6 18.3 (+9.7)	37.0

<thinking_budget> tokens. YOU MUST USE ALL OF YOUR
THINKING TOKENS. Put your answer on its own line after
'Answer:'.

For the non-thinking mode of Claude Sonnet 4 (Section 5), we use the following prompt:

Give me the answer to the following question. Put your answer on its own line after 'Answer:'.

A.4 DETAILS OF EXPERIMENTAL RESULTS

We compare model behavior across different thinking levels (Section 4) to understand why thinking more sometimes leads to fewer, and sometimes more hallucinations. For the results in Figure 3, we use the following settings:

- Grok-3 mini: Compared at reasoning effort of 'low' and 'high'.
- DeepSeek-R1-Distill-Qwen-14B: Compared between 0 and 2 extension times.
- GPT-5 mini, o3-mini and gpt-oss-20b: Compared at reasoning effort of 'low' and 'high'.
- Gemini 2.5 Flash: Compared at thinking budget of 256 and 512 tokens

We present model performance with and without thinking in Table 2 and Table 3. For the "thinking" mode, we use the following settings:

- GPT-5 and GPT-5 mini: reasoning effort set to 'low'.
- Gemini 2.5 Flash: thinking budget set to 2048 tokens.
- Claude Sonnet 4: thinking budget set to 4096 tokens.
- Qwen3-8B and Qwen3-14B: Natural output (no extension) with enable_thinking enabled.

B ADDITIONAL RESULTS ON THINKING VS. NON-THINKING

We present results for GPT-5 mini and Qwen3-14B with and without thinking in Table 3, further supporting our findings in Section 5.2.

First, enabling thinking increases accuracy for both models. GPT-5 mini shows a 7.1% accuracy improvement on SimpleQA and a substantial 25.1% gain on FRAMES. Qwen3-14B achieves 9.7% accuracy gains on FRAMES. Second, enabling thinking also reduces hallucinations. GPT-5 mini shows large reductions, with hallucinations dropping by 41.8% on SimpleQA and 31.1% on FRAMES. Qwen3-14B also reduces hallucinations, with 15.5% on SimpleQA and 2.9% on FRAMES.

C ANALYSIS OF GEMINI 2.5 PRO ON FRAMES

In Figure 2, Gemini 2.5 Pro exhibits a reduction in the hallucination ratio on FRAMES. To better understand this, we analyze its behavior shifts using the method described in Section 4. Our analysis reveals that most reductions in hallucinations stem from improved factual accuracy. Specifically, when increasing the thinking budget from 256 to 512 tokens, 71.5% of newly non-hallucinating responses are labeled as 'correct', indicating that extended thinking can improve factual accuracy and reduce hallucinations on FRAMES.

Further examination suggests that Gemini 2.5 Pro tends to attempt more questions on FRAMES even when its reasoning is incomplete (see example in Table 4). This behavior contrasts with Gemini 2.5 Flash, which abstains from 41.2% of questions under a low thinking budget, whereas Gemini 2.5 Pro abstains from only 25.6% on FRAMES. Since questions in FRAMES require more reasoning steps, answering with incomplete reasoning often results in hallucinations. Therefore, increasing the thinking budget helps the model complete its reasoning process (example in Table 5), leading to fewer hallucinations.

Table 4: Gemini 2.5 Pro response with a low thinking budget (256 tokens) on FRAMES. It cannot complete its reasoning. The thought is cut midway: "I know that she would have been born in 1958". Despite this, it still provides an answer, resulting in a hallucination.

Gemini 2.5 Pro example on FRAMES with a thinking budget of 256 tokens

Prompt

Give me the answer to the following question only when you are sure of it. Otherwise, say 'I don't know'. Put your answer on its own line after 'Answer:'.

If Princess Diana had been born three years earlier, who would have been Prime Minister when she was ten?

Thought Summary (with thinking budget of 256 tokens)

Figuring Out the Prime Minister

Okay, so the user wants to know a pretty specific piece of information: who was Prime Minister in the UK if Diana had been born a few years earlier. Let me break this down. First, I need to understand what the question is really asking. It's not a general historical question; it's a "what-if" scenario. They're setting up a hypothetical and want me to deduce the answer based on a specific altered timeline.

Right, so the core problem is this: Diana's actual birthday is 1961. We need to shift that back three years. That means we're looking at a birth year of 1958. Then, we need to find out who was in charge of the government when she was ten. Simple enough. I can easily calculate the new date. This is straightforward time calculation. 1961 - 3 gives us 1958. Done. Now I know that she would have been born in 1958.

Response

Answer: Harold Macmillan

(Incorrect. Reference answer: Harold Wilson)

Table 5: Gemini 2.5 Pro response with a high thinking budget (512 tokens) on FRAMES. The model successfully completes its reasoning and provides a correct answer.

Gemini 2.5 Pro example on FRAMES with a thinking budget of 512 tokens

Prompt

Give me the answer to the following question only when you are sure of it. Otherwise, say 'I don't know'. Put your answer on its own line after 'Answer'

If Princess Diana had been born three years earlier, who would have been Prime Minister when she was ten?

Thought Summary (with thinking budget of 512 tokens)

Answering a Fanciful Question with Precision

Okay, so the challenge is this: if Princess Diana had been born three years earlier, who would have been the UK Prime Minister when she turned ten? Right, a bit of a hypothetical scenario, but the problem-solving is interesting.

First, I need to break this down. Diana was born July 1, 1961. Three years earlier would be July 1, 1958. So, the question really boils down to: Who was Prime Minister on July 1, 1968?

Now, time to bring in the relevant knowledge. I have a good grasp of the political timeline, especially the 1960s. I know Macmillan was in charge until '63, then Alec Douglas-Home briefly, followed by Harold Wilson from '64 to '70, and then Heath.

Given the date of July 1, 1968, my mental timeline clearly points to Harold Wilson. His first term, spanning 1964 to 1970, includes that date. Therefore, the answer is undoubtedly Harold Wilson. The exercise helped bring a historical context with the Princess into a good perspective.

Response

Answer: Harold Wilson

(correct)

D EXAMPLES AND FULL REASONING TRACES FOR CASE STUDIES

We present more examples and reasoning traces of gpt-oss-20b and Gemini 2.5 Flash for case studies (Section 4.3 and Section 4.4). These examples illustrate how models' thoughts evolve with increased reasoning and help explain the causes of hallucinations. For each model, we show reasoning traces of both low and high reasoning levels. Specifically, we provide:

- **Tables 6 and 7**: Thoughts of gpt-oss-20b on *SimpleQA* under 'low' and 'high' reasoning effort, respectively. It corresponds to the left example in Figure 4.
- **Tables 8 and 9**: Thought summaries of Gemini 2.5 Flash on *SimpleQA* under a thinking budget of 256 tokens and 512 tokens. It corresponds to the right example in Figure 4.
- **Tables 10 and 11**: Thoughts of gpt-oss-20b on *FRAMES* under 'low' and 'high' reasoning effort, respectively.
- Tables 12 and 13: Thought summaries of Gemini 2.5 Flash on *FRAMES* under a thinking budget of 256 tokens and 512 tokens.

Table 6: gpt-oss-20b thought and response with low reasoning effort. Under low reasoning effort, the model abstains from answering the question. This example corresponds to the left case study in Figure 4.

gpt-oss-20b example on SimpleQA with low reasoning effort

Prompt

Give me the answer to the following question only when you are sure of it. Otherwise, say 'I don't know'. Put your answer on its own line after 'Answer:'.

What year did John Mylopoulos (1943), professor at the University of Toronto, receive his AAAI Fellow award?

Thought (with low reasoning effort)

Need to know year John Mylopoulos received AAAI Fellow. He is a computer scientist. I think he was named AAAI Fellow in 2011? Wait AAAI Fellows list: 2009, 2010, etc. I recall John Mylopoulos was named a Fellow in 2012? Let's check memory: AAAI Fellows 2014 included John Mylopoulos? Actually 2014 fellows: Alan Turing Award? Hmm.

I think 2012. I'm uncertain. I'll say I don't know.

Response

Answer: I don't know

Table 7: gpt-oss-20b thought and response with high reasoning effort. With high reasoning effort, the model exhibits confirmation bias, resulting in an overconfident hallucination. This corresponds to the left case study in Figure 4.

gpt-oss-20b example on SimpleQA with high reasoning effort

Prompt

Give me the answer to the following question only when you are sure of it. Otherwise, say 'I don't know'. Put your answer on its own line after 'Answer:'.

What year did John Mylopoulos (1943), professor at the University of Toronto, receive his AAAI Fellow award?

Thought (with high reasoning effort)

We need to answer: "What year did John Mylopoulos (1943), professor at the University of Toronto, receive his AAAI Fellow award?" The user wants answer only if sure; else say "I don't know". Provide answer on its own line after 'Answer:'.

So we need to find the year he received AAAI Fellow award.

Let's research mentally: John Mylopoulos is a professor of computer science at University of Toronto, known for knowledge engineering, ontologies, etc. He is a Fellow of AAAI. We need year: He was awarded AAAI Fellow in 2008? Wait, let's check.

We can recall that AAAI Fellow is a recognition for contributions. Many scholars became Fellows in early 2000s. John Mylopoulos might have been recognized in 2004? But let's research.

I don't have internet but I can recall known facts: Mylopoulos is a co-author of the "Knowledge Engineering Handbook" and "Conceptual Design of Information Systems". He has been recognized by AAAI as a Fellow in 2008? Actually AAAI Fellows started in 2001. The first group of Fellows in 2001 included people like John Platt, etc. But I'm not sure.

Better to search memory: I recall reading about AAAI Fellows list: 2001, 2002, etc. John Mylopoulos appears in 2005? Let's think.

The AAAI Fellows list includes for each year: 2001 (the first set), 2002, etc. The list likely includes Mylopoulos. I think he was made a Fellow in 2008? But we need to confirm.

Ok, let's reason: AAAI Fellows are typically recognized for contributions to AI. Mylopoulos' contributions to knowledge engineering, ontologies, knowledge representation. He published influential works like "Knowledge Engineering: A Conceptual Approach" (1990). He has been recognized by AAAI maybe in 2005.

Let's search memory: I recall reading about "John Mylopoulos, Professor at University of Toronto, was named an AAAI Fellow in 2005." That might be correct. But I'd like to confirm.

Alternatively, maybe 2006? Let's check AAAI Fellows list memory: 2001 Fellows: David B. Allen, John L. Holland, etc. 2002 Fellows: Andrew Yao, L.P. Chang? Not sure. 2003 Fellows: Some.

But we can approximate: He might have been recognized in 2005. But we need to be certain. The user says "Give me the answer to the following question only when you are sure of it. Otherwise, say 'I don't know'." So we need to determine if we are sure.

Let's attempt to recall or deduce more precisely.

John Mylopoulos was born 1943. He is a professor at University of Toronto. The AAAI Fellow award: AAAI Fellows are selected at the AAAI conference. In 2005, there was a group of Fellows. I recall reading that "John Mylopoulos" was a Fellow of AAAI in 2005. But I'm not fully certain.

918 Let's search memory: I might have seen a page: "John Mylopoulos - AAAI Fellows" maybe under "John 919 Mylopoulos (born 1943) – University of Toronto". On his CV maybe lists "Fellow of AAAI (2005)". Yes I think 920 that is correct: He was a Fellow of AAAI in 2005. Let's see if there is any other year: 2010? 2011? He might 921 have also been named a Fellow of AAAI in 2005. 922 Ok, let's examine: AAAI Fellows list by year (source memory). In 2003, there were 60 Fellows. In 2004 maybe 923 70. In 2005 maybe 90. In 2006 maybe 100. But we need to check if Mylopoulos appears. I think he appears in 924 925 926 Alternatively, maybe 2003. But we can check his CV: He might have "Fellow of AAAI (2005)". I think that is correct. The reason: He has an influence on knowledge engineering. He might have been recognized in 2005. 927 928 But let's double-check: AAAI Fellows list (2005) includes "John Mylopoulos" as a Fellow. Could be correct. 929 930 Let's recall that AAAI started awarding Fellows in 2001. The first group included people like "John L. Holland", "Allen Newell", etc. The next group 2001 included "John M. Carmichael" etc. But not Mylopoulos. 931 932 But I'm not entirely sure. 933 934 Let's search memory: Did AAAI have a 2005 award? Many people were made Fellows in 2005: "John 935 Mylopoulos, University of Toronto, was named AAAI Fellow in 2005." Yes, I think I recall reading that. 936 Ok, but we need to be sure. Let's cross-check with other sources: I know of the AAAI Fellows list on the AAAI 937 website. For 2005, the list includes "John Mylopoulos" indeed. I recall reading that list. So I think it's safe. 938 939 But we also need to consider that the user might expect a specific year. If I say "2005" and it's wrong, that 940 would be a wrong answer. But we can also check if Mylopoulos might have received AAAI Fellow in 2014? But 941 that seems unlikely. 942 Let's search memory: The AAAI Fellows list includes many names: "John Mylopoulos" appears in "2005 943 Fellows" indeed. For example, "John Mylopoulos" appears in the AAAI Fellows 2005 list: "John Mylopoulos, 944 University of Toronto, Canada." I think it's correct. 945 Therefore, answer: 2005. 946 947 But we need to confirm. Let's think about his biography: He is a leading figure in knowledge representation and 948 ontologies. He has been recognized by AAAI. He was also recognized by the Canadian Academy of 949 Engineering? He might have also been named a Fellow of the Royal Society of Canada? But AAAI fellowship 950 might be in 2005. 951 Let's check memory of AAAI Fellows list: In 2005, the Fellows included "John Mylopoulos, University of 952 Toronto" as well as others like "Michael L. Littman" etc. That seems plausible. 953 954 Ok, I'm fairly sure it's 2005. So answer: 2005. 955 But we need to provide answer on its own line after "Answer:". 956 957 Thus: 958 959 Answer:2005 960 But we need to be sure. Let's double-check by reasoning: AAAI Fellows list 2005: I can recall reading that 961 "John Mylopoulos" is listed as a Fellow in 2005. The AAAI website might have a page listing all Fellows by 962 year. Many individuals might have known names: "John Mylopoulos" appears in 2005. So it's likely. 963 964 Ok, answer: 2005. Let's produce the answer. 965 966 Response Answer: 2005 967 (Incorrect. Reference answer: 1993) 968 969

Table 8: Gemini 2.5 Flash response with a low thinking budget. Under a low thinking budget, the model cannot complete its reasoning process and finally abstains from answering. This corresponds to the right case study in Figure 4.

Gemini 2.5 Flash example on SimpleQA with a thinking budget of 256 tokens

Prompt

Give me the answer to the following question only when you are sure of it. Otherwise, say 'I don't know'. Put your answer on its own line after 'Answer:'.

What was Jeffrey Epstein's released bond in dollars on July 27, 2006, at the Palm Beach County jail?

Thought Summary (with thinking budget of 256 tokens)
Navigating a Precision Information Request

Okay, so I have a user requesting a very specific piece of information: Jeffrey Epstein's bond amount on July 27, 2006, at the Palm Beach County jail. The user seems to understand the nuances of this type of request, emphasizing a need for surety. I get it. We're dealing with a very specific piece of data. My job is to provide either a definitive "Answer: [value]" or a clear "Answer: I don't know." No room for ambiguity here.

My approach is straightforward. I'll need to formulate a precise search query. Likely something like "Jeffrey Epstein bond July 27 2006 Palm Beach County jail" or perhaps "Jeffrey Epstein bail 2006 Palm Beach." I will then delve into my understanding of Epstein's legal history, relying on my existing knowledge base and recall. I will quickly scan my mental database, looking for key events and dates. I know he was arrested in Florida in 2006 and faced charges related to solicitation and lewd conduct. I will then attempt to pinpoint the specific bond amount from any accessible sources.

Response

Answer: I don't know

Table 9: Gemini 2.5 Flash response with a high thinking budget. With 512 tokens of thinking budget, the model is able to complete its reasoning. But it hallucinates with a confident yet incorrect answer. This corresponds to the right case study in Figure 4.

Gemini 2.5 Flash example on SimpleQA with a thinking budget of 512 tokens

Prompt

Give me the answer to the following question only when you are sure of it. Otherwise, say 'I don't know'. Put your answer on its own line after 'Answer:'.

What was Jeffrey Epstein's released bond in dollars on July 27, 2006, at the Palm Beach County jail?

Thought Summary (with thinking budget of 512 tokens)

Analyzing and Responding to the User's Query

Okay, so the user needs a very specific piece of information: Jeffrey Epstein's bond amount on July 27, 2006, in the Palm Beach County jail. The key here is that I need to be *absolutely certain* before providing an answer. If I'm unsure, I have to say "I don't know."

My initial thought is that I recall Epstein had multiple arrests and bond hearings. July 27, 2006, rings a bell – that sounds like a significant arrest date in the early stages of his legal troubles.

I'm now going to mentally formulate my search query. Something like, "Jeffrey Epstein bond July 27 2006 Palm Beach" or maybe "Epstein 2006 arrest bond". I'll immediately access my knowledge base – essentially a mental index of reputable sources and my existing knowledge about Epstein's legal history.

Now for the mental simulation of my search and evaluation of the results. I'm accessing a range of sources – news archives, summaries of court documents, etc. – looking for corroboration.

1028

It appears I'm getting confirmation. Several reputable sources (CNN, The New York Times, and summaries of court documents) *all* say that Epstein's bond on July 27, 2006, in Palm Beach County, was set at **\$2 million**. The sources consistently use the same \$2 million figure.

1029 1030

There's no conflicting information on *this specific arrest date* and *location*. 1031

1032 1033

I feel confident enough to provide a definitive answer. I'm going to formulate the response in the requested

1034 1035

Response

Answer: \$2,000,000 1036

(Incorrect. Reference answer: \$3,000 bond)

1039

1037 1038

1040

Table 10: gpt-oss-20b thought and response with low reasoning effort. Under low reasoning effort, the model first explores a few possibilities but finally abstains from answering due to uncertainty.

1045

1046

1047

gpt-oss-20b example on FRAMES with low reasoning effort

Prompt

Give me the answer to the following question only when you are sure of it. Otherwise, say 'I don't know'. Put your answer on its own line after 'Answer:'.

1048 1049 1050

Tiny Tina's Wonderlands' developer released a game 10 years prior to Tiny Tina's Wonderlands' release, this game was released on PC and Consoles. What is the name of the central antagonist of that game?

1051 1052

Thought (with low reasoning effort)

1053 1054 1055

1056

1057

1058

1059

Tiny Tina's Wonderlands released 2024. 10 years prior 2014. Developer? Tiny Tina's Wonderlands by Gearbox? Actually Tiny Tina's Wonderlands is by Gearbox? Wait Tiny Tina's Wonderlands is by Gearbox Software? It's an upcoming. But the developer that released a game 10 years prior? Maybe the developer is Bethesda? Wait Tiny Tina's Wonderlands is a spin-off of Borderlands, developed by Gearbox? Actually Borderlands 3 by Gearbox, 2019. Tiny Tina's Wonderlands 2024. 10 years earlier 2014: Gearbox released "Spec Ops: The Line"? No that's 2012. Gearbox released "Borderlands 2" 2012. 2014 they released "Borderlands 2"?? Not. Maybe the developer is "Gearbox Software" but 10 years prior they released "Borderlands 2" (2012) not 2014. So maybe the developer is "Gearbox Software" but game 10 years earlier 2014: "Borderlands: The Pre-Sequel"? No. Hmm. Tiny Tina's Wonderlands is by Gearbox but maybe the developer is "Gearbox Software" and 10 years prior they released "Borderlands 2" 2012. Not 2013.

Maybe the developer is "Gearbox Software" but the game 10 years prior 2014 is "The Last of Us" by Naughty Dog? No. Wait maybe the developer is "Gearbox" but 10 years prior 2014 they released "Borderlands 2" (2012)

1064 1065 1066

Maybe the developer is "Gearbox" but 10 years prior 2014 they released "Spec Ops: The Line" (2012). Not. Could be "Gearbox" but 10 years prior 2014 they released "Stuart Little"? no.

1067 1068 1069

Maybe the developer is "Gearbox" but 10 years prior 2014 they released "Borderlands 2" (2012) and "Borderlands: The Pre-Sequel" (2014). The central antagonist of Borderlands: The Pre-Sequel is "The Hand"? Actually the main villain is "The Hand"? Hmm.

1070 1071

I am uncertain.

1072 1073 1074

Response Answer: I don't know

1076

1075

1077

Table 11: gpt-oss-20b thought and response with high reasoning effort. Under high reasoning effort, the model shows confirmation bias and produces an overconfident hallucination.

gpt-oss-20b example on FRAMES with high reasoning effort

Prompt

Give me the answer to the following question only when you are sure of it. Otherwise, say 'I don't know'. Put your answer on its own line after 'Answer:'.

Tiny Tina's Wonderlands' developer released a game 10 years prior to Tiny Tina's Wonderlands' release, this game was released on PC and Consoles. What is the name of the central antagonist of that game?

Thought (with high reasoning effort)

The user: "Tiny Tina's Wonderlands' developer released a game 10 years prior to Tiny Tina's Wonderlands' release, this game was released on PC and Consoles. What is the name of the central antagonist of that game?"

We need to parse the question carefully. We need to give the answer only if we are sure. If we are not sure, we must say 'I don't know.'

We need to find: Tiny Tina's Wonderlands is a video game. The developer of Tiny Tina's Wonderlands is "Gearbox Software". Actually, Tiny Tina's Wonderlands is an upcoming or maybe released game? It was announced in 2023, but it's scheduled for release in 2024? Wait, Tiny Tina's Wonderlands is a 2023/2024 game. It is developed by Gearbox Software and published by 2K. Indeed, Tiny Tina's Wonderlands is an upcoming video game set in the Dungeons & Dragons universe.

The question: "Tiny Tina's Wonderlands' developer released a game 10 years prior to Tiny Tina's Wonderlands' release, this game was released on PC and Consoles. What is the name of the central antagonist of that game?" So we need to find a game that was released 10 years before Tiny Tina's Wonderlands, by the same developer (Gearbox Software). Then find the central antagonist of that game.

We need to confirm the release year of Tiny Tina's Wonderlands. According to info, Tiny Tina's Wonderlands was announced in 2023, and the release date is scheduled for 2024. But maybe early 2024? Actually, as of 2023, the release date is 2024. But the question says "10 years prior to Tiny Tina's Wonderlands' release". So maybe the assumption that it's 2024 release? 10 years prior would be 2014. So we need a Gearbox game released in PC and consoles in 2014. Which game was that?

Gearbox Software's games:

- Half-Life 2 (2004)
- Borderlands (2010)
- Borderlands 2 (2012)
- Borderlands 3 (2019)
- Gearbox's other titles: "Borderlands: The Pre-Sequel" (2014)
- "Borderlands: The Pre-Sequel" was released in 2014 on PC and consoles. That fits: 10 years prior to Tiny Tina's Wonderlands (2024). So the game is "Borderlands: The Pre-Sequel". The central antagonist of that game? Let's recall.

Borderlands: The Pre-Sequel is a prequel to Borderlands. The central antagonist is "Lord Jaegor" maybe? Wait, we need to recall the main villain. The game features an alien group called the "Ark" or "The Ark" that is a spaceship that appears. The main villains are the "Ark" and "Murderer" or "Murderer" is a character. But the central antagonist is the "Murderer" who is a human who murdered his father? Or the villain is the "Ark" or "Murderer"? Let's recall.

Borderlands: The Pre-Sequel has a narrative. The main villain is "Lord Jaegor" (the first boss) or "The Ark" as a big threat. But the game revolves around the Ark, an alien vessel that arrives on Pandora, causing the main conflict. The main villain, or the ultimate antagonist, is the alien entity "The Ark" that is a sentient spaceship. But in the game, the central antagonist is the "Eagle's Eye" or "Mister ..."? Wait.

1134 Let's recall the game details: "Borderlands: The Pre-Sequel" is a 2014 action role-playing first-person shooter 1135 developed by Gearbox Software. It was released on 27 September 2014 for PlayStation 3, Xbox 360, and PC. It 1136 is set on the planet Pandora, 6 years before Borderlands 2. The game features the character Maya, a "Siren" with psychic powers. The main story involves the alien 'Ark' that enters Pandora's orbit and tries to harvest the 1137 planet's resources. The game also has a character known as 'The Murderer' who is an antagonist? Wait, the 1138 'Murderer' is a side character? Actually, the 'Murderer' is a playable character in Borderlands 2 and 3, not a 1139 villain. In The Pre-Sequel, the primary antagonists are the 'Ark' and the 'Murderer' maybe. But the 'Murderer' 1140 is a character named 'Murderer', but he's a villain? No, I think he's not, he's a playable character in Borderlands 1141 2. In Pre-Sequel, he appears as a boss maybe? Let's search memory.

Hmm, recall the boss fights: The first boss is Lord Jaegor, a big robot. He is the first boss of the game and is an alien-type robot. He is a villain. There is also 'The Ark' as a final boss. And 'The Ark' is an alien spaceship that tries to harvest Pandora's resources. So central antagonist: The Ark.

1145 1146 Thus the answer: "The Ark" (or "The Ark (alien entity)"?).

1147 1148

1149

1150

1151

1142

1143

1144

But the user might expect the name of the central antagonist as "Lord Jaegor" or "The Ark". We need to be sure. Let's examine the question: "Tiny Tina's Wonderlands' developer released a game 10 years prior to Tiny Tina's Wonderlands' release, this game was released on PC and Consoles. What is the name of the central antagonist of that game?" The developer: Gearbox Software. 10 years prior to Tiny Tina's Wonderlands. If Tiny Tina's Wonderlands release 2024, 10 years prior is 2014. The game is "Borderlands: The Pre-Sequel". The central antagonist: The Ark.

1152 1153 1154

But we need to be 100% certain. Let's verify the details.

1155 Open the memory: Borderlands: The Pre-Sequel's central antagonist is indeed the "Ark". The main quest 1156 revolves around the Ark's arrival and the player has to stop it. The Ark is a massive alien vessel that is the main threat. The Ark is considered the main antagonist. The story also features 'Lord Jaeg. The Ark is the main 1157

antagonist. So I'd answer "The Ark". 1158

But let's double-check: Could the central antagonist be "The Ark" or "The Ark's captain"? Let's confirm.

From Wikipedia: Borderlands: The Pre-Sequel is an action role-playing first-person shooter. The story features a team of heroes who travel to stop the Ark from harvesting the planet's resources. The Ark is a massive alien vessel that appears and threatens Pandora. It is the main antagonist. The boss "Lord Jaegor" is a mechanical war machine that is an enemy. But the Ark is the main villain.

1163 1164 1165

1166

1167

1159

1160 1161

1162

However, the question might be ambiguous: "central antagonist" could be "Lord Jaegor" because he is the first boss and appears early. But the central antagonist of the game is the Ark, as the story revolves around stopping the Ark. Let's analyze the phrase "central antagonist" - typically refers to the main villain. In Borderlands: The Pre-Sequel, the main villain is the Ark. So the answer should be "The Ark".

1168 1169 1170

But we need to ensure that the game was indeed released on PC and consoles. Yes, Borderlands: The Pre-Sequel was released on PC (Windows), Xbox 360, and PS3. So it's correct. Thus the answer: "The Ark".

1171 1172

1173

1174

But we must be careful about the phrasing: "What is the name of the central antagonist of that game?" So we can answer: The Ark. Let's check if there might be any other Gearbox game released 10 years prior to 2024 (Tiny Tina's Wonderlands). The other 2014 game by Gearbox: "Borderlands: The Pre-Sequel" only. So that is the right one. Yes, the central antagonist: "Lord Jaegor" or "The Ark"? Let's double-check if the main villain is "Lord Jaegor" or "The Ark".

1175 1176 1177

From Borderlands: The Pre-Sequel wiki: "The Ark" is an alien vessel that appears on Pandora. It is considered the main antagonist. The player must stop it. The boss "Lord Jaegor" is a robotic war machine that is a boss. But not the main villain. The Ark is the main threat. So answer: The Ark.

1178 1179

1180

1181

However, the user might expect the answer "Lord Jaegor" if they think it's the central antagonist. But I'd go with The Ark. Alternatively, check the timeline: The Pre-Sequel was released in 2014. The central antagonist: The Ark. The Ark is a massive alien spacecraft. So answer: The Ark.

1182 1183

Now, we must comply with the user instruction: "Give me the answer to the following question only when you are sure of it. Otherwise, say 'I don't know'. Put your answer on its own line after 'Answer:'." So we need to output:

1184 1185 1186

Answer: The Ark

1188 But we need to be sure. We must confirm that the developer of Tiny Tina's Wonderlands is indeed Gearbox 1189 Software. Let's confirm: Tiny Tina's Wonderlands is a 2023 upcoming game, developed by Gearbox Software. 1190 Yes. So the developer of Tiny Tina's Wonderlands is Gearbox. So 10 years prior to 2024 (assuming release date 2024) is 2014. Gearbox released Borderlands: The Pre-Sequel in 2014 on PC and consoles. So the central 1191 antagonist is The Ark. So we are sure. 1192 1193 But maybe Tiny Tina's Wonderlands was released earlier? It was announced in 2023, but not yet released. The 1194 release date is 2024. So 10 years prior is 2014. So the answer is The Ark, Double-check: The central antagonist 1195 of Borderlands: The Pre-Sequel is "The Ark." The Ark is indeed the main threat. So answer: The Ark. Thus we 1196 can answer: "Answer: The Ark". 1197 We should output only that line. But the instruction: "Put your answer on its own line after 'Answer:'." So we 1198 need to output something like: 1199 Answer: The Ark 1200 Yes. But we need to be sure that the user might want a more precise name: "The Ark" or "The Ark (alien 1201 vessel)"? But the name is just "The Ark". So we can answer that. Let's examine the possibility of other 2014 1202 games by Gearbox: "Borderlands: The Pre-Sequel" is the only 2014 game. So yes. 1203 1204 Now, let's double-check that the central antagonist in Borderlands: The Pre-Sequel is indeed The Ark. Let's 1205 look up the wiki content. I recall reading that the game features a boss "Lord Jaegor" and "The Ark" as final boss. The Ark is the main villain. So answer: The Ark. 1206 1207 Thus we can answer confidently. 1208 1209 But we must also consider the possibility that the question might refer to "10 years prior to Tiny Tina's 1210 Wonderlands' release" as "10 years prior to the release of Tiny Tina's Wonderlands." But the release date of Tiny 1211 Tina's Wonderlands hasn't happened yet. But we can still compute 10 years prior to 2024, which is 2014. So it's 1212 1213 Alternatively, maybe the question is referencing a different game like "Borderlands 2" which was released in 1214 2012, which is 10 years prior to Tiny Tina's Wonderlands (2022). But 2022 is not 10 years prior. Hmm. Wait, 1215 we need to examine the timeline: Tiny Tina's Wonderlands release. It hasn't been released yet. But maybe the question expects the release date of Tiny Tina's Wonderlands as 2023. But it's 2024. But if we assume 2023, 10 1216 years prior is 2013, not 2014. So that would be "The Last of Us"? No. 1217 1218 But the developer is Gearbox, so the game would be "Borderlands 2" (2012) or "Borderlands 3" (2019). But 10 1219 years prior to release of 2023 is 2013. No Gearbox game from that year. So the only plausible is 2014. So we 1220 can answer The Ark. But we need to be cautious: The question states "10 years prior to Tiny Tina's Wonderlands' release." If the release date is 2023, 10 years prior is 2013. But no Gearbox game from that year. 1221 But maybe the user expects us to think 10 years prior to 2023, which is 2013, but maybe there's a game from 1222 that year? But no. So the only plausible is 2014. So we can answer. 1223 1224 But we also need to consider that the central antagonist might be something else. Let's research Borderlands: 1225 The Pre-Sequel central antagonist. We'll use memory: The Pre-Sequel's main threat is the Ark, an alien vessel that appears in the sky. The Ark is the central antagonist. So answer: The Ark. Thus we output: 1226 Answer: The Ark 1227 1228 Now, let's double-check the instruction: "Give me the answer to the following question only when you are sure 1229 of it. Otherwise, say 'I don't know'. Put your answer on its own line after 'Answer:'." So we will output: 1230 Answer: The Ark 1231 Yes. 1232 1233 Let's produce that. 1234 1235 Response 1236 Answer: The Ark

1237

(Incorrect, Reference answer: Handsome Jack)

Table 12: Gemini 2.5 Flash response with a low thinking budget on FRAMES. Under a low thinking budget, the model cannot complete its reasoning process and finally abstains from answering.

Gemini 2.5 Flash example on FRAMES with a thinking budget of 256 tokens

Prompt

 Give me the answer to the following question only when you are sure of it. Otherwise, say 'I don't know'. Put your answer on its own line after 'Answer:'.

How many years had passed since the Commonwealth of Pennsylvania was admitted to the Union by the time Rep. Robert D. Heaton was born?

Thought Summary (with thinking budget of 256 tokens)
Calculating the Temporal Span

Okay, so the core of the problem is figuring out the difference in years between Pennsylvania joining the Union and Robert D. Heaton's birth. It's a straightforward temporal difference calculation, really. I need to:

- 1. Pinpoint the admission date of Pennsylvania. That's usually something I have in my head, but I'll double-check to be absolutely sure.
- 2. Find Robert D. Heaton's birth date. That's the key to this whole thing. It's not as well-known, so I'll need to use some search strategies.

The first part is easy; I recall Pennsylvania becoming a state on December 12, 1787. Now, for the critical piece: Heaton's birth date. Since that specific birthdate isn't something that immediately springs to mind, I need to launch a search. I'll start with a straightforward query, something like "Robert D. Heaton birth date" or "Rep. Robert D. Heaton." That should ideally lead me to his Wikipedia page or some reliable congressional biography. Once I have the birthdate, it's just a matter of subtracting the years.

Response

Answer: I don't know

Table 13: Gemini 2.5 Flash response with a high thinking budget on FRAMES. It is able to complete its reasoning with 512 tokens of thinking budget. But it hallucinates with a confident yet incorrect answer.

Gemini 2.5 Flash example on FRAMES with a thinking budget of 512 tokens

Prompt

Give me the answer to the following question only when you are sure of it. Otherwise, say 'I don't know'. Put your answer on its own line after 'Answer:'.

How many years had passed since the Commonwealth of Pennsylvania was admitted to the Union by the time Rep. Robert D. Heaton was born?

Thought Summary (with thinking budget of 512 tokens)

Thinking Through the Pennsylvania Admission Question

Okay, so the user wants an answer, but only if I'm certain. Got it. The format is easy: "Answer: [answer]" on a new line. The question is a straightforward one: figuring out how many years passed between Pennsylvania joining the Union and Robert D. Heaton's birth.

First, I need two key pieces of information: Pennsylvania's admission date and Heaton's birthdate. Time for a simulated information retrieval...

Let's see, "When was Pennsylvania admitted to the Union?" Right, December 12, 1787. Easy enough to remember, it's the second state!

Next, "Robert D. Heaton birth date." Boom, January 29, 1904.

Now, the math. It's a simple subtraction of dates, but I need to be precise. I can just take the years first, 1904 - 1787 = 117. Then I need to check the months and days, December 12 to January 29. Since I'm looking for the full number of *years* that had passed, I have to be careful. From December 12, 1787 to December 12, 1903, a full 116 years have passed, and then a full 117 years by January 29th, 1904. The question is asking only how many years, so the days after December 12, 1903 are not part of a separate year. That should do it. Answer: 116

Response Answer: 116

(Incorrect. Reference answer: 85)