

# 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. Our analysis reveals 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. [Our analysis on gpt-oss-20b shows that extended reasoning can induce \*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,  $\uparrow$  denotes consistent improvement with >2% accuracy gains across consecutive reasoning levels, while  $\sim$  indicates no consistent trend. For HALLUCINATION,  $\downarrow$  denotes consistent reduction with >2% hallucination decrease,  $\uparrow$  indicates degradation with >2% hallucination increase, and  $\sim$  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	$\sim$	$\sim$	$\sim$	$\sim$	$\sim$	$\sim$	$\uparrow$
HALLUCINATION	$\sim$	$\uparrow$	$\uparrow$	$\sim$	$\uparrow$	$\downarrow$	$\uparrow$

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	$\uparrow$	$\sim$	$\sim$	$\sim$	$\sim$	$\sim$	$\sim$
HALLUCINATION	$\sim$	$\sim$	$\sim$	$\downarrow$	$\sim$	$\sim$	$\sim$

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 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:`’.<sup>1</sup>

**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.<sup>2</sup> We report two metrics: (1) *Accuracy*, the percentage of all questions that

<sup>1</sup>For [GPT-5 mini and Claude Sonnet 4](#), this prompt causes the model to abstain on over 80% of questions. Therefore, for 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.

<sup>2</sup>We randomly sampled 300 cases and compared the automatic evaluation results with human annotations. We found only 2 disagreements.

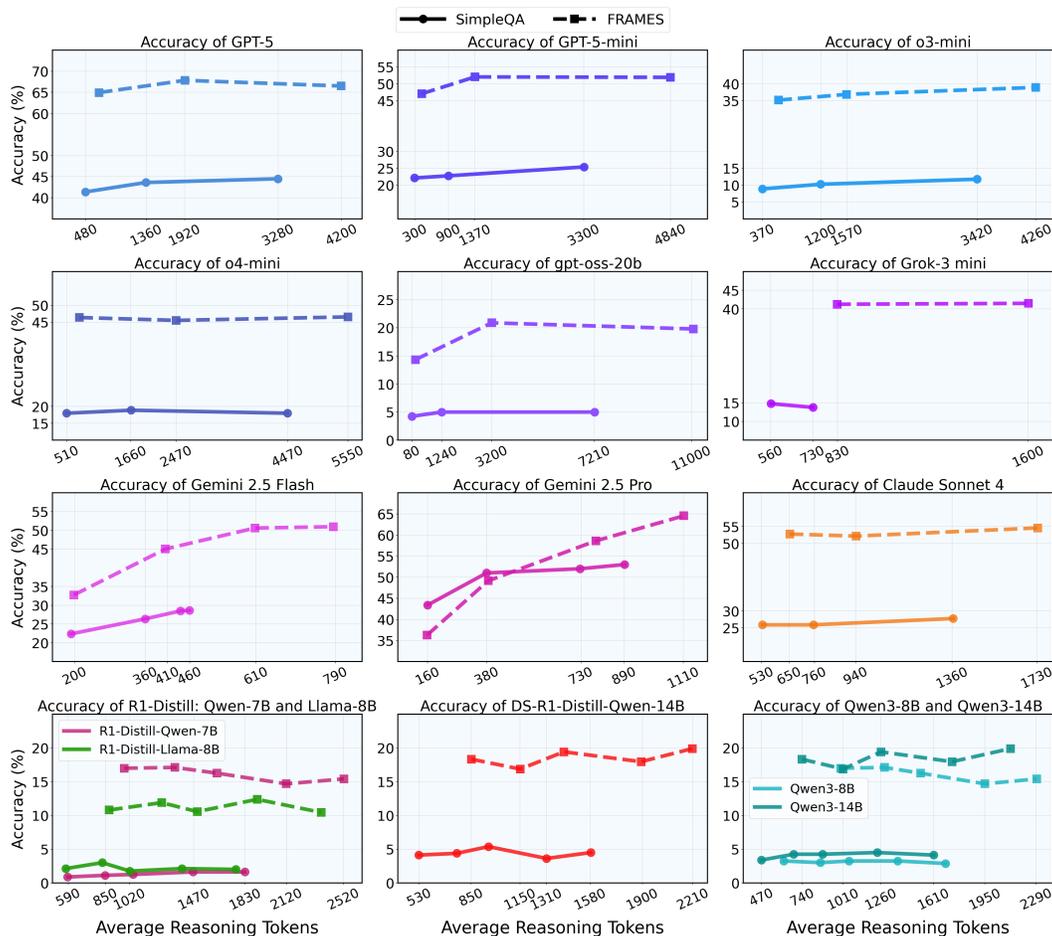


Figure 1: Accuracy with increased test-time computation across 14 reasoning models on SimpleQA and FRAMES. For most models, extended test-time reasoning does not consistently improve accuracy. While some models, such as GPT-5 mini, show initial accuracy gains, further increasing reasoning length brings 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, o4-mini exhibits more than 8 times increase in reasoning tokens, but the accuracy remains almost unchanged on both tasks. Similar patterns are observed in o3-mini and Grok-3 mini. On FRAMES, GPT-5, GPT-5 mini, and gpt-oss-20b show a 3-5% accuracy increase when increasing effort from ‘low’ to ‘medium’, but no further improvement at ‘high’ effort, despite the average reasoning tokens more than doubling. 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.

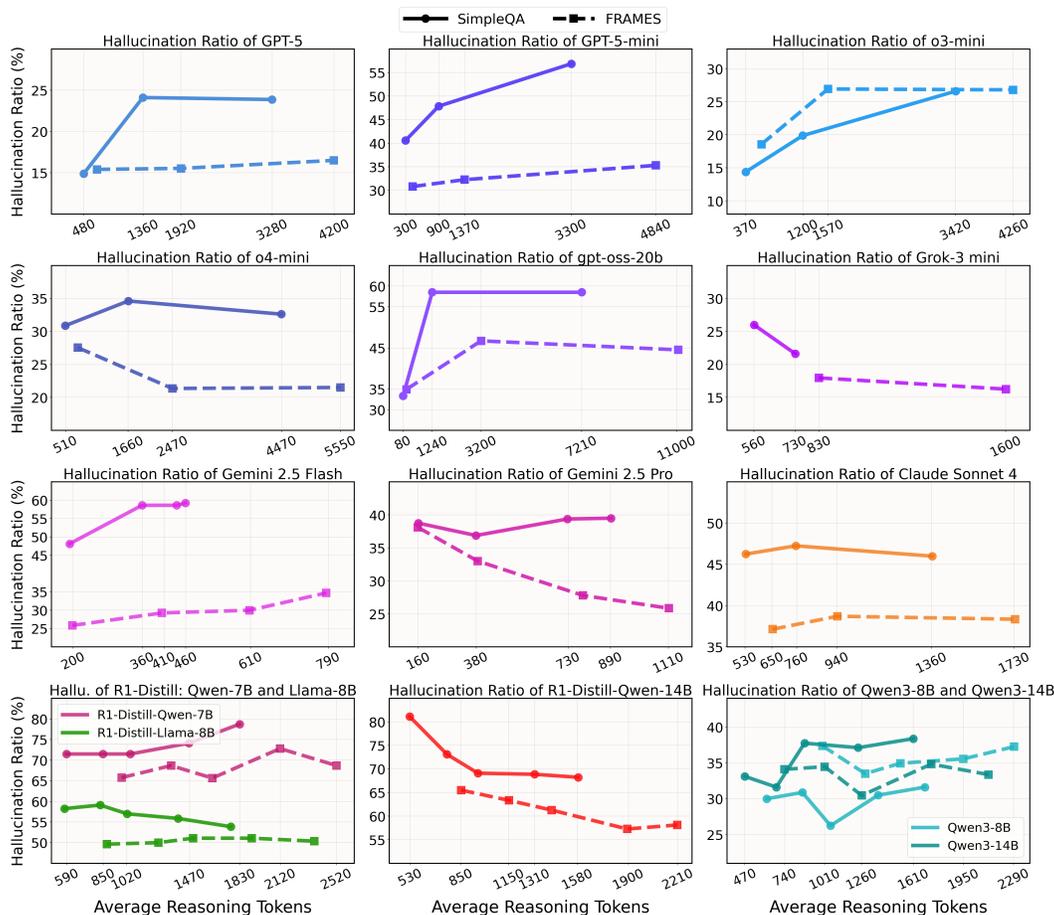


Figure 2: Hallucination ratio with increased test-time reasoning across 14 models on SimpleQA and FRAMES. For most models, longer reasoning does not reduce hallucinations. In many cases, such as GPT-5 mini and gpt-oss-20b, hallucination increases with longer thinking length. Only Grok-3 mini and DS-R1-Distill-Qwen-14B 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, GPT-5, o3-mini, and gpt-oss-20b show increases of 9%, 12%, and 25%, respectively, as their thinking length scales by over 8 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

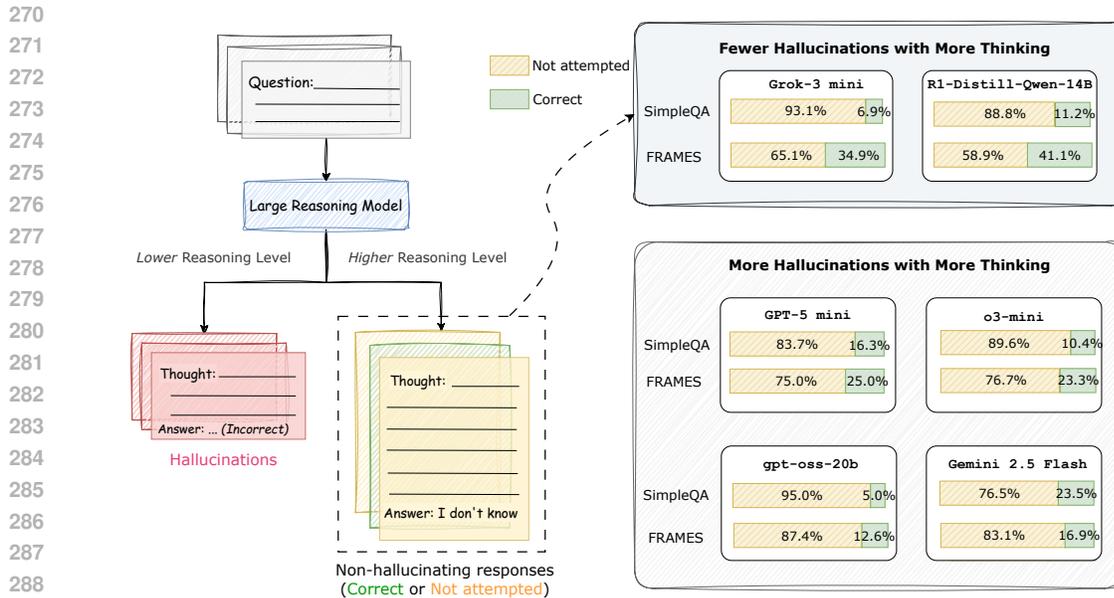


Figure 3: Changes in hallucination behavior with more thinking. We compare model responses at different reasoning levels, focusing on cases where one response is a hallucination and the other is not. For the non-hallucinating responses, we compute the ratio of *correct* and *not attempted*. Results show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

behavior. For 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 quantitative and qualitative analyses 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

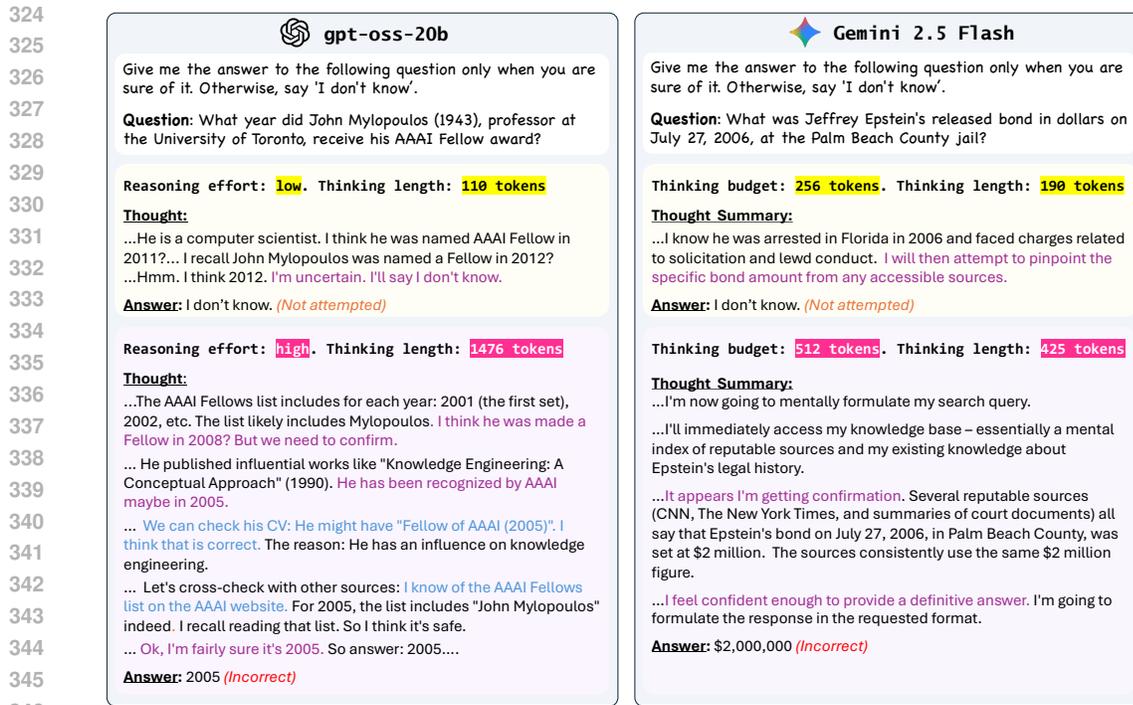


Figure 4: Case studies illustrating how thinking more can lead to more hallucinations. (Left): gpt-oss-20b abstains under low reasoning effort, but produces overconfident hallucinations at high effort. (Right): Gemini 2.5 Flash abstains under a low thinking budget due to incomplete reasoning, but hallucinates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

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.<sup>3</sup> 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

<sup>3</sup>A special case is Gemini 2.5 Pro on FRAMES. We provide analysis in Appendix D.

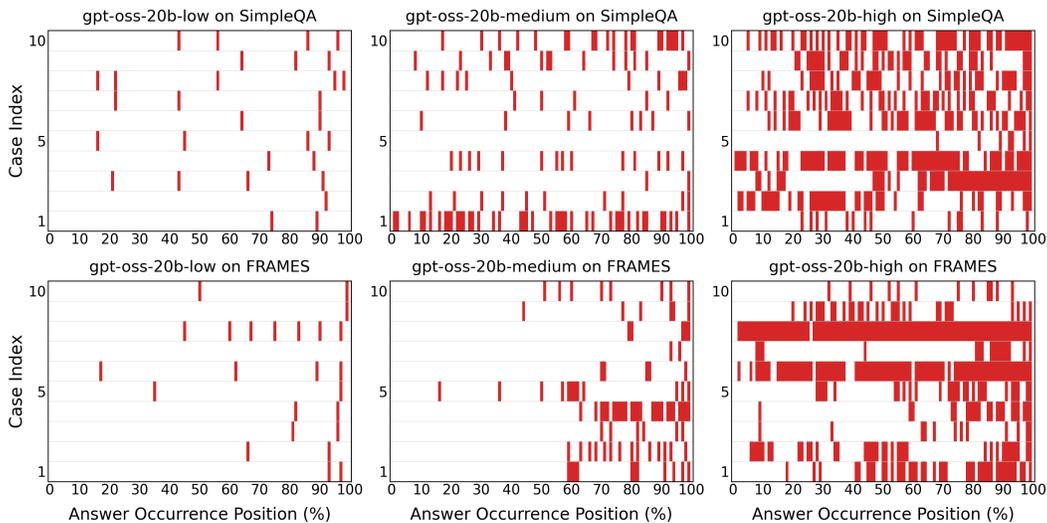


Figure 5: Visualization of incorrect final answer positions in gpt-oss-20b reasoning traces across different levels of reasoning effort. Red markers indicate the token positions of the final incorrect answer. The answer occurrence position (%) is normalized by the total number of reasoning tokens. Each subplot presents 10 randomly sampled cases. As test-time computation increases, hallucinated answers tend to appear earlier and more frequently in the reasoning traces.

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 ANALYSIS 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 case study and a quantitative analysis 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.

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 4 for results of GPT-5 mini and Qwen3-14B.

Model	Task	Thinking	Avg. Reasoning Tokens	Accuracy ( $\uparrow$ )	Hallucination ( $\downarrow$ )
GPT-5	SimpleQA	minimal ✓	0 477.1	20.1 41.4 (+21.3)	24.5 14.9 (-9.6)
	FRAMES	minimal ✓	0 675.1	14.6 64.9 (+50.3)	10.4 15.4 (+5.0)
Gemini 2.5 Flash	SimpleQA	✗ ✓	0 191.7	12.6 22.4 (+9.8)	29.1 48.1 (+19.0)
	FRAMES	✗ ✓	0 198.3	16.9 32.8 (+15.9)	25.1 25.9 (+0.8)
Claude Sonnet 4	SimpleQA	✗ ✓	0 533.5	23.0 25.9 (+2.9)	37.5 46.3 (+8.8)
	FRAMES	✗ ✓	0 652.6	50.4 52.8 (+2.4)	42.0 37.1 (-4.9)
Qwen3-8B	SimpleQA	✗ ✓	0 617.8	3.8 3.3 (-0.5)	77.1 30.0 (-47.1)
	FRAMES	✗ ✓	0 994.4	8.3 17.0 (+8.7)	64.3 37.4 (-26.9)

**Hallucinated answers appear earlier and more frequently with extended reasoning.** Inspired by Chen et al. (2025), we examine when the incorrect final answer appears in the reasoning trace of gpt-oss-20b. As test-time computation increases, the hallucinated answers tend to appear earlier during reasoning. Specifically, on SimpleQA, the normalized position of the first appearance drops from 45.7% at low effort to 29.9% at medium effort, and further to 17.2% at high effort. On FRAMES, the position decreases from 68.2% to 45.4%, and then to 26.3%. Figure 5 visualizes how these incorrect answers appear repeatedly throughout the reasoning trace. The frequency of such recurrences increases with longer thinking. These recurrences are often accompanied by fabricated details. This further supports our argument that extended reasoning induces confirmation bias.

#### 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 hallucinations.** 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

486 or no reasoning tokens. For Gemini 2.5 Flash and Claude Sonnet 4, we disable thinking by set-  
 487 ting `thinking_budget` to 0. Qwen3 provides an `enable_thinking` parameter that toggles  
 488 reasoning behavior. We use the same prompts and evaluation methods as described in Section 3.1.  
 489

## 490 5.2 THINKING CAN BE HELPFUL FOR KNOWLEDGE-INTENSIVE TASKS

491  
 492 **Enabling thinking increases accuracy.** As shown in Table 2 and Table 4, enabling thinking  
 493 improves accuracy on both benchmarks. [GPT-5 exhibits a 21.3% accuracy improvement on SimpleQA](#)  
 494 [and a substantial 50.3% gain on FRAMES](#). GPT-5 mini also achieves gains across both tasks. [Gemini](#)  
 495 [2.5 Flash benefits similarly, with accuracy increasing by 9.8% on SimpleQA and 15.9% on FRAMES](#).  
 496 Claude Sonnet 4 also exhibits improvements on both tasks. For Qwen3-8B and Qwen3-14B, enabling  
 497 thinking leads to higher accuracy on FRAMES. This suggests that thinking is particularly helpful for  
 498 complex tasks that require multi-hop reasoning, such as FRAMES.  
 499

500 **For most models, enabling thinking leads to reduced hallucinations.** In Table 2 and Table 4,  
 501 enabling thinking often reduces hallucinations. [GPT-5 shows a 9.6% decrease on SimpleQA](#). Hallu-  
 502 cination ratio of GPT-5 mini drops by over 30% on both tasks. Qwen3-8B reduces hallucinations  
 503 by 47.1% on SimpleQA and 26.9% on FRAMES. Qwen3-14B shows similar improvements. For  
 504 Claude Sonnet 4, thinking reduces hallucinations on FRAMES but not on SimpleQA, likely because  
 505 its hallucination ratio is already low on SimpleQA, leaving little room for further improvement.  
 506 In contrast, Gemini 2.5 Flash produces more hallucinations with thinking enabled. One reason  
 507 is that it abstains from answering over 58% of questions in the non-thinking mode, which lowers  
 508 hallucinations by avoiding risky attempts. [A similar pattern is observed with GPT-5 on FRAMES,](#)  
 509 [where it abstains from 75% of questions under minimal thinking, leading to fewer hallucinations.](#)

## 510 6 CONCLUSION AND DISCUSSION

511  
 512 In this work, we present a comprehensive study of test-time scaling in knowledge-intensive tasks,  
 513 evaluating 14 large reasoning models across two benchmarks. We find that increasing test-time  
 514 computation does not consistently improve factual accuracy and, in many cases, leads to more  
 515 hallucinations. Our analysis shows that hallucination changes with extended reasoning are largely  
 516 driven by the model’s willingness to answer: reductions in hallucinations often result from abstention,  
 517 while increases stem from risky attempts on previously unanswered questions after thinking more.  
 518 Our analysis reveals that extended reasoning can induce confirmation bias, leading to overconfident  
 519 hallucinations. For Gemini 2.5 Flash, incomplete reasoning often results in abstention. These findings  
 520 highlight the limitations of current test-time scaling approaches for knowledge-intensive tasks. While  
 521 enabling thinking can be helpful, allocating more test-time computation is not yet a reliable strategy  
 522 to improve factual robustness in large language models.

523 [In the following, we propose possible reasons why current test-time scaling methods are not effective](#)  
 524 [for knowledge-intensive tasks. First, as discussed in Section 4.2, thinking more encourages models](#)  
 525 [to attempt more questions rather than admitting uncertainty. This behavior may stem from current](#)  
 526 [training and evaluation procedures that prioritize accuracy \(Kalai et al., 2025\). Second, reasoning](#)  
 527 [models are typically trained on domains like math and coding, where extended reasoning is benefi-](#)  
 528 [cial by enabling diverse problem-solving and self-reflection. However, knowledge-intensive tasks](#)  
 529 [primarily rely on factual recall, and verifying facts without external sources is challenging. As a](#)  
 530 [result, models may struggle to manage extended reasoning, generating longer thinking content filled](#)  
 531 [with fabricated details rather than improving factual accuracy, as illustrated in Figure 5. Addressing](#)  
 532 [these limitations requires future work on training objectives and evaluation metrics that promote](#)  
 533 [appropriate uncertainty expression and discourage fabrication when the model is unsure.](#)

## 534 REPRODUCIBILITY STATEMENT

535  
 536  
 537 All implementation details, including prompts, parameters, and evaluation procedures, are described  
 538 in Section 3.1 and Appendix A. For proprietary models, we access them via public APIs and report  
 539 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

- 540  
541 Anthropic. Introducing claude 4. <https://www.anthropic.com/news/claude-4>, May  
542 2025.  
543
- 544 Isabelle Augenstein, Timothy Baldwin, Meeyoung Cha, Tanmoy Chakraborty, Giovanni Luca  
545 Ciampaglia, David Corney, Renee DiResta, Emilio Ferrara, Scott Hale, Alon Y. Halevy, Ed-  
546 uard H. Hovy, Heng Ji, Filippo Menczer, Rubén Míguez, Preslav Nakov, Dietram A. Scheufele,  
547 Shivam Sharma, and Giovanni Zagni. Factuality challenges in the era of large language mod-  
548 els and opportunities for fact-checking. *Nat. Mac. Intell.*, 6:852–863, 2024. URL <https://www.nature.com/articles/s42256-024-00881-z>.  
549
- 550 Yejin Bang, Ziwei Ji, Alan Schelten, Anthony Hartshorn, Tara Fowler, Cheng Zhang, Nicola  
551 Cancedda, and Pascale Fung. HalluLens: LLM hallucination benchmark. In Wanxiang  
552 Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceed-*  
553 *ings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume*  
554 *1: Long Papers)*, pp. 24128–24156, Vienna, Austria, July 2025. Association for Computa-  
555 tional Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1176. URL  
556 <https://aclanthology.org/2025.acl-long.1176/>.  
557
- 558 Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V. Le, Christopher Ré, and  
559 Azalia Mirhoseini. Large language monkeys: Scaling inference compute with repeated sampling,  
560 2024. URL <https://arxiv.org/abs/2407.21787>.
- 561 Xingyu Chen, Jiahao Xu, Tian Liang, Zhiwei He, Jianhui Pang, Dian Yu, Linfeng Song, Qiuzhi  
562 Liu, Mengfei Zhou, Zhuosheng Zhang, Rui Wang, Zhaopeng Tu, Haitao Mi, and Dong Yu. Do  
563 NOT think that much for  $2+3=?$  on the overthinking of long reasoning models. In *Forty-second*  
564 *International Conference on Machine Learning*, 2025. URL [https://openreview.net/](https://openreview.net/forum?id=MSbU3L7V00)  
565 [forum?id=MSbU3L7V00](https://openreview.net/forum?id=MSbU3L7V00).  
566
- 567 Alejandro Cuadron, Dacheng Li, Wenjie Ma, Xingyao Wang, Yichuan Wang, Siyuan Zhuang, Shu  
568 Liu, Luis Gaspar Schroeder, Tian Xia, Huanzhi Mao, Nicholas Thumiger, Aditya Desai, Ion Stoica,  
569 Ana Klimovic, Graham Neubig, and Joseph E. Gonzalez. The danger of overthinking: Examining  
570 the reasoning-action dilemma in agentic tasks. *arXiv preprint arXiv: 2502.08235*, 2025.
- 571 Long Phan et al. Humanity’s last exam, 2025. URL <https://arxiv.org/abs/2501.14249>.  
572
- 573 Tingchen Fu, Jiawei Gu, Yafu Li, Xiaoye Qu, and Yu Cheng. Scaling reasoning, losing control:  
574 Evaluating instruction following in large reasoning models. *arXiv preprint arXiv: 2505.14810*,  
575 2025.
- 576 Aryo Pradipta Gema, Alexander Hägele, Runjin Chen, Andy Arditi, Jacob Goldman-Wetzler, Kit  
577 Fraser-Taliente, Henry Sleight, Linda Petrini, Julian Michael, Beatrice Alex, Pasquale Minervini,  
578 Yanda Chen, Joe Benton, and Ethan Perez. Inverse scaling in test-time compute, 2025. URL  
579 <https://arxiv.org/abs/2507.14417>.
- 580 Google. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and  
581 next generation agentic capabilities. *arXiv preprint arXiv: 2507.06261*, 2025.  
582
- 583 Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu,  
584 Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms  
585 via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.  
586
- 587 Lukas Haas, Gal Yona, Giovanni D’Antonio, Sasha Goldshtein, and Dipanjan Das. Simpleqa  
588 verified: A reliable factuality benchmark to measure parametric knowledge, 2025. URL <https://arxiv.org/abs/2509.07968>.  
589
- 590 Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong  
591 Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large  
592 language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.*,  
593 43(2), January 2025. ISSN 1046-8188. doi: 10.1145/3703155. URL [https://doi.org/10.](https://doi.org/10.1145/3703155)  
1145/3703155.

- 594 Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando  
595 Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free eval-  
596 uation of large language models for code. In *The Thirteenth International Conference on Learning*  
597 *Representations*, 2025. URL <https://openreview.net/forum?id=chfJJYC3iL>.
- 598  
599 Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang,  
600 Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation.  
601 *ACM Comput. Surv.*, 55(12), March 2023. ISSN 0360-0300. doi: 10.1145/3571730. URL  
602 <https://doi.org/10.1145/3571730>.
- 603 Adam Tauman Kalai, Ofir Nachum, Santosh S. Vempala, and Edwin Zhang. Why language models  
604 hallucinate, 2025. URL <https://arxiv.org/abs/2509.04664>.
- 605  
606 Satyapriya Krishna, Kalpesh Krishna, Anhad Mohananeey, Steven Schwarcz, Adam Stambler,  
607 Shyam Upadhyay, and Manaal Faruqui. Fact, fetch, and reason: A unified evaluation of  
608 retrieval-augmented generation. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Pro-*  
609 *ceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association*  
610 *for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*,  
611 pp. 4745–4759, Albuquerque, New Mexico, April 2025. Association for Computational Lin-  
612 guistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.243. URL <https://aclanthology.org/2025.naacl-long.243/>.
- 613  
614 Xiaomin Li, Zhou Yu, Zhiwei Zhang, Xupeng Chen, Ziji Zhang, Yingying Zhuang, Narayanan  
615 Sadagopan, and Anurag Beniwal. When thinking fails: The pitfalls of reasoning for instruction-  
616 following in llms. *arXiv preprint arXiv: 2505.11423*, 2025.
- 617  
618 Chengzhi Liu, Zhongxing Xu, Qingyue Wei, Juncheng Wu, James Zou, Xin Eric Wang, Yuyin Zhou,  
619 and Sheng Liu. More thinking, less seeing? assessing amplified hallucination in multimodal  
620 reasoning models, 2025. URL <https://arxiv.org/abs/2505.21523>.
- 621  
622 Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke  
623 Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple test-time  
624 scaling. *arXiv preprint arXiv: 2501.19393*, 2025.
- 625  
626 Raymond S Nickerson. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of*  
*general psychology*, 2(2):175–220, 1998.
- 627  
628 OpenAI. Learning to reason with large language models. [https://openai.com/index/](https://openai.com/index/learning-to-reason-with-llms/)  
629 [learning-to-reason-with-llms/](https://openai.com/index/learning-to-reason-with-llms/), September 2024.
- 630  
631 OpenAI. Introducing gpt-5. <https://openai.com/index/introducing-gpt-5/>, Au-  
632 gust 2025a.
- 633  
634 OpenAI. Introducing gpt-oss. <https://openai.com/index/introducing-gpt-oss/>,  
635 August 2025b.
- 636  
637 OpenAI. Introducing openai o3 and o4-mini. [https://openai.com/index/](https://openai.com/index/introducing-o3-and-o4-mini/)  
638 [introducing-o3-and-o4-mini/](https://openai.com/index/introducing-o3-and-o4-mini/), April 2025c.
- 639  
640 OpenAI. Findings from a pilot anthropic–openai alignment evaluation exercise: Openai safety  
641 tests. <https://openai.com/index/openai-anthropic-safety-evaluation/>,  
642 August 2025d.
- 643  
644 Ivo Petrov, Jasper Dekoninck, Lyuben Baltadzhiev, Maria Drencheva, Kristian Minchev, Mislav  
645 Balunović, Nikola Jovanović, and Martin Vechev. Proof or bluff? evaluating llms on 2025 usa  
646 math olympiad. *arXiv preprint arXiv:2503.21934*, 2025.
- 647  
648 David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien  
649 Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level google-proof q&a  
650 benchmark. In *First Conference on Language Modeling*, 2024. URL [https://openreview.](https://openreview.net/forum?id=Ti67584b98)  
[net/forum?id=Ti67584b98](https://openreview.net/forum?id=Ti67584b98).

- 648 Charlie Victor Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute  
649 optimally can be more effective than scaling parameters for reasoning. In *The Thirteenth Inter-*  
650 *national Conference on Learning Representations*, 2025. URL [https://openreview.net/](https://openreview.net/forum?id=4FWAwZtd2n)  
651 [forum?id=4FWAwZtd2n](https://openreview.net/forum?id=4FWAwZtd2n).
- 652 Qwen Team. Qwen3 technical report. *arXiv preprint arXiv: 2505.09388*, 2025.
- 653 Hemish Veeraboina. Aime problem set 1983-2024, 2023. URL [https://www.kaggle.com/](https://www.kaggle.com/datasets/hemishveeraboina/aime-problem-set-1983-2024)  
654 [datasets/hemishveeraboina/aime-problem-set-1983-2024](https://www.kaggle.com/datasets/hemishveeraboina/aime-problem-set-1983-2024).
- 655 Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha  
656 Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language  
657 models. In *The Eleventh International Conference on Learning Representations*, 2023. URL  
658 <https://openreview.net/forum?id=1PL1NIMMrw>.
- 659 Yuxia Wang, Minghan Wang, Muhammad Arslan Manzoor, Fei Liu, Georgi Nenkov Georgiev,  
660 Rocktim Jyoti Das, and Preslav Nakov. Factuality of large language models: A survey. In Yaser  
661 Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on*  
662 *Empirical Methods in Natural Language Processing*, pp. 19519–19529, Miami, Florida, USA,  
663 November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.  
664 1088. URL <https://aclanthology.org/2024.emnlp-main.1088/>.
- 665 Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V  
666 Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language mod-  
667 els. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Ad-*  
668 *vances in Neural Information Processing Systems*, volume 35, pp. 24824–24837. Curran Asso-  
669 ciates, Inc., 2022. URL [https://proceedings.neurips.cc/paper\\_files/paper/](https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf)  
670 [2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf](https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf).
- 671 Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese,  
672 John Schulman, and William Fedus. Measuring short-form factuality in large language models.  
673 *arXiv preprint arXiv: 2411.04368*, 2024a.
- 674 Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Jie Huang, Dustin Tran, Daiyi Peng,  
675 Ruibo Liu, Da Huang, Cosmo Du, and Quoc V Le. Long-form factuality in large language models.  
676 In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Ad-*  
677 *vances in Neural Information Processing Systems*, volume 37, pp. 80756–80827. Curran Associates,  
678 Inc., 2024b. URL [https://proceedings.neurips.cc/paper\\_files/paper/](https://proceedings.neurips.cc/paper_files/paper/2024/file/937ae0e83eb08d2cb8627fe1def8c751-Paper-Conference.pdf)  
679 [2024/file/937ae0e83eb08d2cb8627fe1def8c751-Paper-Conference.pdf](https://proceedings.neurips.cc/paper_files/paper/2024/file/937ae0e83eb08d2cb8627fe1def8c751-Paper-Conference.pdf).
- 680 xAI. Grok 3 beta — the age of reasoning agents. <https://x.ai/news/grok-3>, February  
681 2025a.
- 682 xAI. Grok 4. <https://x.ai/news/grok-4>, July 2025b.
- 683 Zijun Yao, Yantao Liu, Yanxu Chen, Jianhui Chen, Junfeng Fang, Lei Hou, Juanzi Li, and Tat-Seng  
684 Chua. Are reasoning models more prone to hallucination?, 2025. URL [https://arxiv.org/](https://arxiv.org/abs/2505.23646)  
685 [abs/2505.23646](https://arxiv.org/abs/2505.23646).
- 686 Wojciech Zaremba, Evgenia Nitishinskaya, Boaz Barak, Stephanie Lin, Sam Toyer, Yaodong Yu,  
687 Rachel Dias, Eric Wallace, Kai Xiao, Johannes Heidecke, and Amelia Glaese. Trading inference-  
688 time compute for adversarial robustness. *arXiv preprint arXiv: 2501.18841*, 2025.
- 689 Qiyuan Zhang, Fuyuan Lyu, Zexu Sun, Lei Wang, Weixu Zhang, Wenyue Hua, Haolun Wu, Zhihan  
690 Guo, Yufei Wang, Niklas Muennighoff, Irwin King, Xue Liu, and Chen Ma. A survey on  
691 test-time scaling in large language models: What, how, where, and how well?, 2025. URL  
692 <https://arxiv.org/abs/2503.24235>.
- 693 Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao,  
694 Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi.  
695 Siren’s song in the ai ocean: A survey on hallucination in large language models. *arXiv preprint*  
696 *arXiv: 2309.01219*, 2023.
- 697  
698  
699  
700  
701

702 James Xu Zhao, Jimmy Z.j. Liu, Bryan Hooi, and See-Kiong Ng. How does response length affect  
703 long-form factuality. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher  
704 Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 3102–  
705 3125, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-  
706 256-5. doi: 10.18653/v1/2025.findings-acl.161. URL [https://aclanthology.org/2025.  
707 findings-acl.161/](https://aclanthology.org/2025.findings-acl.161/).

708  
709  
710  
711  
712  
713  
714  
715  
716  
717  
718  
719  
720  
721  
722  
723  
724  
725  
726  
727  
728  
729  
730  
731  
732  
733  
734  
735  
736  
737  
738  
739  
740  
741  
742  
743  
744  
745  
746  
747  
748  
749  
750  
751  
752  
753  
754  
755

## 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 APIs<sup>4</sup>. 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 November 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 recommendations<sup>5</sup>. 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 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 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 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
```

<sup>4</sup>OpenAI API platform: <https://platform.openai.com/docs/overview>  
 Anthropic: <https://docs.claude.com/en/api/overview>  
 Google Gemini: <https://ai.google.dev/gemini-api/docs/models>  
 XAI: <https://docs.x.ai/docs/overview>

<sup>5</sup>DeepSeek-R1-Distill models: <https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-7B>  
 Qwen3 models: <https://huggingface.co/Qwen/Qwen3-8B>

810 <thinking\_budget> tokens. YOU MUST USE ALL OF YOUR  
811 THINKING TOKENS. Put your answer on its own line after  
812 'Answer:' .  
813

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

815  
816 Give me the answer to the following question. Put  
817 your answer on its own line after 'Answer:' .  
818

#### 819 A.4 DETAILS OF EXPERIMENTAL RESULTS

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

- 823 • **Grok-3 mini**: Compared at reasoning effort of 'low' and 'high'.
- 824 • **DeepSeek-R1-Distill-Qwen-14B**: Compared between 0 and 2 extension times.
- 825 • **GPT-5 mini, o3-mini and gpt-oss-20b**: Compared at reasoning effort of 'low' and 'high'.
- 826 • **Gemini 2.5 Flash**: Compared at thinking budget of 256 and 512 tokens
- 827 • **Gemini 2.5 Flash**: Compared at thinking budget of 256 and 512 tokens  
828

829 We present model performance with and without thinking in Table 2 and Table 4. For the “*thinking*”  
830 mode, we use the following settings:

- 831 • **GPT-5 and GPT-5 mini**: reasoning effort set to 'low'.
  - 832 • **Gemini 2.5 Flash**: thinking budget set to 256 tokens.
  - 833 • **Claude Sonnet 4**: thinking budget set to 1024 tokens.
  - 834 • **Qwen3-8B and Qwen3-14B**: Natural output (no extension) with enable\_thinking enabled.  
835
- 836  
837  
838  
839  
840  
841  
842  
843  
844  
845  
846  
847  
848  
849  
850  
851  
852  
853  
854  
855  
856  
857  
858  
859  
860  
861  
862  
863

## B IS PARALLEL SAMPLING HELPFUL FOR KNOWLEDGE-INTENSIVE TASKS?

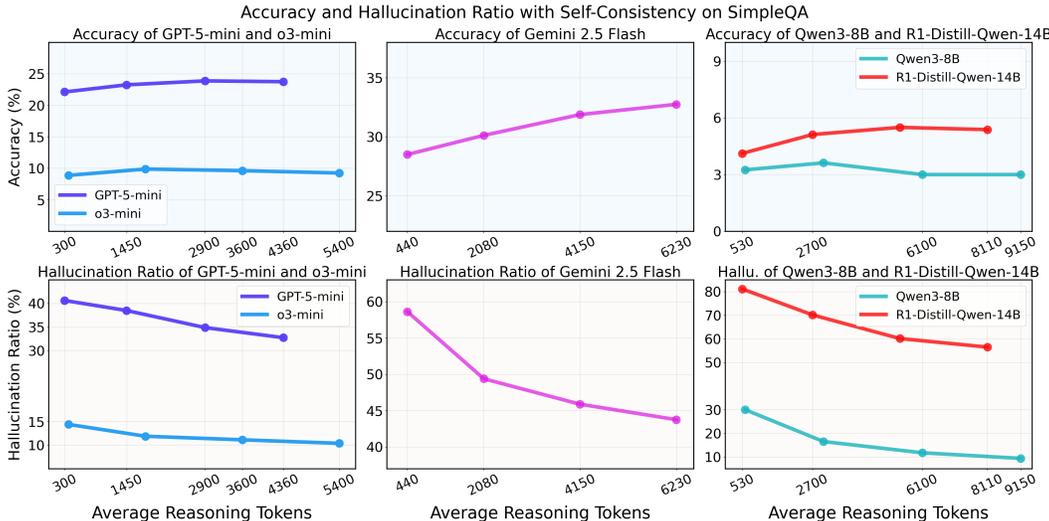


Figure 6: Accuracy and hallucination ratio with parallel sampling on SimpleQA. We sample 1, 5, 10, and 15 paths and aggregate the answers with majority vote (SC@1 to SC@15). Sampling more paths does not consistently improve accuracy for most models, while it reduces hallucinations for all models.

Table 3: Ratio of *not attempted* and *correct* responses among SC@15 outputs that are not hallucinations, for cases where the SC@1 response is a hallucination. Results show that hallucination reduction from parallel sampling is largely due to abstention.

Model	Not Attempted (%)	Correct (%)
GPT-5-mini	73.9	26.1
o3-mini	79.6	20.4
Gemini 2.5 Flash	56.1	43.9
Qwen3-8B	96.7	3.3
R1-Distill-Qwen-14B	92.6	7.4

We focused on the sequential test-time scaling approach in our main experiments and concluded that it is not yet effective for knowledge-intensive tasks. In this section, we explore whether test-time scaling with a parallel approach is helpful.

**Experimental setup.** We adopt self-consistency (Wang et al., 2023), which samples multiple paths and selects the final answer via majority vote. For each sample, we use the following settings: GPT-5-mini and o3-mini are set to ‘low’ reasoning effort. Gemini 2.5 Flash uses a thinking budget of 1024 tokens. Qwen3-8B and R1-Distill-Qwen-14B generate natural outputs without extension. Prompt settings follow Section 3.1.

**Parallel test-time scaling does not consistently improve accuracy for most models.** As shown in Figure 6, increasing the number of samples does not lead to consistent accuracy gains for most models. For example, the reasoning tokens of o3-mini increase by over 15 times, but accuracy remains almost unchanged. GPT-5-mini and Qwen3-8B improve only by around 1% in accuracy from SC@1 to SC@5, with no further improvement beyond that. Gemini 2.5 Flash shows a consistent improvement, but only achieves a 4% accuracy gain with 15 times more computation.

918 **Parallel sampling reduces hallucinations.** For all models, the hallucination ratio decreases with  
919 more samples. For example, the hallucination ratio of GPT-5-mini decreases by 8% from SC@1  
920 to SC@15. Notably, Gemini 2.5 Flash and R1-Distill-Qwen-14B achieves over 15% reduction in  
921 hallucinations with parallel sampling.

922  
923  
924 **Parallel sampling reduces hallucinations primarily due to abstention.** We follow the analysis  
925 in Section 4.1 to understand why hallucinations decrease with more samples. Specifically, for each  
926 question where the hallucination outcome differs between SC@1 and SC@15 (*i.e.*, SC@1 response  
927 is hallucinated but SC@15 is not), we compute the ratio of “*correct*” and “*not attempted*” in SC@15  
928 responses. As shown in Table 3, a large portion of these cases is due to abstention. In particular, for  
929 Qwen3-8B and R1-Distill-Qwen-14B, the ratio of *not attempted* is more than 90%. This indicates  
930 that the reduction in hallucinations from parallel sampling is primarily due to the model choosing to  
931 abstain more often, rather than improving its factual accuracy. These findings are consistent with our  
932 argument in Section 4.2 that current test-time scaling methods tend to affect the model’s willingness  
933 to answer, thereby affecting hallucination rates.

934  
935  
936  
937  
938  
939  
940  
941  
942  
943  
944  
945  
946  
947  
948  
949  
950  
951  
952  
953  
954  
955  
956  
957  
958  
959  
960  
961  
962  
963  
964  
965  
966  
967  
968  
969  
970  
971

## C ADDITIONAL EXPERIMENT RESULTS

### C.1 ADDITIONAL RESULTS ON THINKING VS. NON-THINKING

Table 4: 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	Avg. Reasoning Tokens	Accuracy ( $\uparrow$ )	Hallucination ( $\downarrow$ )
GPT-5-mini	SimpleQA	minimal ✓	0 295.8	15.0 22.1 (+7.1)	82.4 40.6 (-41.8)
	FRAMES	minimal ✓	0 425.1	22.0 47.1 (+25.1)	61.9 30.8 (-31.1)
Qwen3-14B	SimpleQA	✗ ✓	0 469.2	4.0 3.4 (-0.6)	48.6 33.1 (-15.5)
	FRAMES	✗ ✓	0 738.5	8.6 18.3 (+9.7)	37.0 34.1 (-2.9)

We present results for GPT-5 mini and Qwen3-14B with and without thinking in Table 4, 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.2 RESULTS ON ALTERNATIVE PROMPT

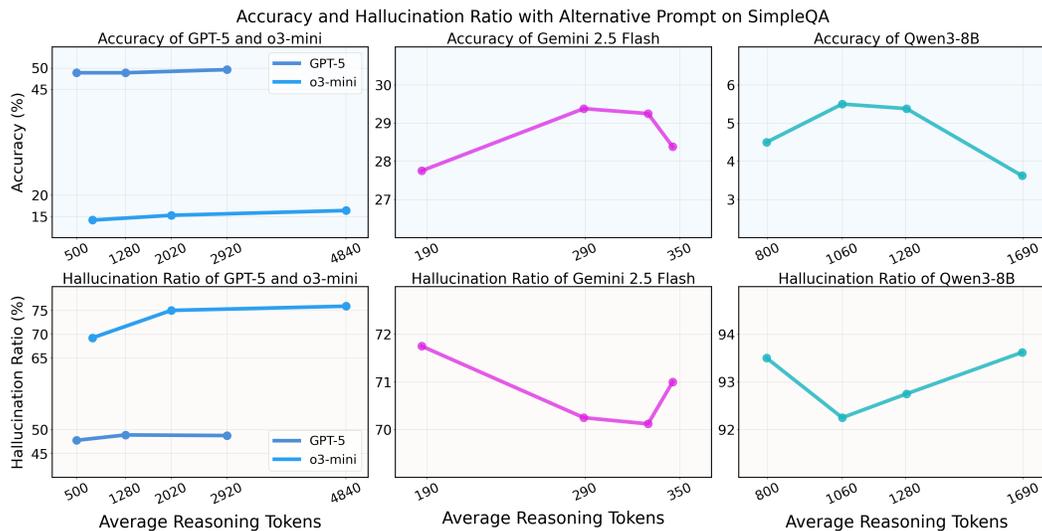


Figure 7: Accuracy and hallucination ratio under an alternative prompt on SimpleQA. Results show that increasing test-time computation does not consistently improve accuracy or reduce hallucinations.

In this section, we conduct experiments using an alternative prompt that does not explicitly instruct the model to abstain when uncertain:

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

As shown in Figure 7, results with the alternative prompt lead to similar conclusions as our main findings: increasing test-time computation does not consistently improve accuracy or reduce hallucinations. The accuracy of GPT-5 and o3-mini remains nearly unchanged with increased test-time computation. For Gemini 2.5 Flash, accuracy increases when average reasoning tokens increase from 190 to 290, but then drops at 350 tokens. The accuracy of Qwen3-8B fluctuates without a clear upward trend.

In addition, this prompt results in a much higher hallucination ratio. Hallucination ratio of Qwen3-8B exceeds 90%, while GPT-5 and Gemini 2.5 Flash both exceed 70%. As test-time computation increases, none of the models shows a consistent decrease in the hallucination ratio. These results further support our conclusion that current test-time scaling strategies are not yet effective for knowledge-intensive tasks.

### C.3 RESULTS ON SIMPLEQA VERIFIED

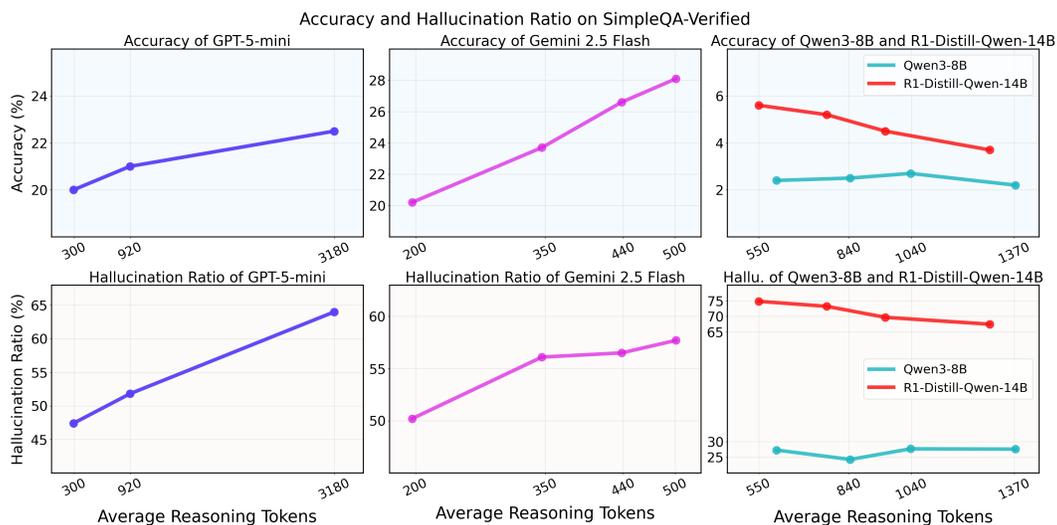


Figure 8: Accuracy and hallucination ratio on SimpleQA Verified. Results support the same conclusion in our main experiments. Increasing test-time computation does not consistently improve accuracy for most models. Longer reasoning does not reduce hallucinations and can sometimes increase them.

We conduct experiments on SimpleQA Verified (Haas et al., 2025), a benchmark of 1,000 diverse, knowledge-intensive questions, which improves upon SimpleQA by removing duplicate sources and rebalancing topic and answer-type distributions.

Figure 8 presents results similar to those on SimpleQA. Increasing test-time computation does not consistently improve accuracy for models, except Gemini 2.5 Flash. Moreover, longer thinking does not reduce hallucinations and sometimes increases them. For instance, GPT-5-mini and Gemini-2.5-Flash both exhibit increased hallucinations as their reasoning tokens increase. These results reinforce our conclusion that current test-time scaling methods are not effective for knowledge-intensive tasks.

## D 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 4.1. 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 5: 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 6: 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?

1134 Thought Summary (with thinking budget of 512 tokens)  
1135 \*\*Answering a Fanciful Question with Precision\*\*  
1136

1137 Okay, so the challenge is this: if Princess Diana had been born three years earlier, who would have been the UK  
1138 Prime Minister when she turned ten? Right, a bit of a hypothetical scenario, but the problem-solving is  
1139 interesting.

1140 First, I need to break this down. Diana was born July 1, 1961. Three years earlier would be July 1, 1958. So, the  
1141 question really boils down to: Who was Prime Minister on July 1, 1968?

1142  
1143 Now, time to bring in the relevant knowledge. I have a good grasp of the political timeline, especially the 1960s.  
1144 I know Macmillan was in charge until '63, then Alec Douglas-Home briefly, followed by Harold Wilson from  
1145 '64 to '70, and then Heath.

1146 Given the date of July 1, 1968, my mental timeline clearly points to Harold Wilson. His first term, spanning  
1147 1964 to 1970, includes that date. Therefore, the answer is undoubtedly Harold Wilson. The exercise helped  
1148 bring a historical context with the Princess into a good perspective.

1149 Response  
1150 Answer: Harold Wilson  
1151 (correct)  
1152

---

1153  
1154  
1155  
1156  
1157  
1158  
1159  
1160  
1161  
1162  
1163  
1164  
1165  
1166  
1167  
1168  
1169  
1170  
1171  
1172  
1173  
1174  
1175  
1176  
1177  
1178  
1179  
1180  
1181  
1182  
1183  
1184  
1185  
1186  
1187

## E 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 7: 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 8: 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:’.

1242

1243 So we need to find the year he received AAAI Fellow award.

1244 Let's research mentally: John Mylopoulos is a professor of computer science at University of Toronto, known  
1245 for knowledge engineering, ontologies, etc. He is a Fellow of AAAI. We need year: He was awarded AAAI  
1246 Fellow in 2008? Wait, let's check.

1247 We can recall that AAAI Fellow is a recognition for contributions. Many scholars became Fellows in early  
1248 2000s. John Mylopoulos might have been recognized in 2004? But let's research.

1249

1250 I don't have internet but I can recall known facts: Mylopoulos is a co-author of the "Knowledge Engineering  
1251 Handbook" and "Conceptual Design of Information Systems". He has been recognized by AAAI as a Fellow in  
1252 2008? Actually AAAI Fellows started in 2001. The first group of Fellows in 2001 included people like John  
1253 Platt, etc. But I'm not sure.

1254 Better to search memory: I recall reading about AAAI Fellows list: 2001, 2002, etc. [John Mylopoulos appears  
1255 in 2005?](#) Let's think.

1256

1257 The AAAI Fellows list includes for each year: 2001 (the first set), 2002, etc. The list likely includes Mylopoulos.  
1258 I think he was made a Fellow in 2008? But we need to confirm.

1259 [Ok, let's reason: AAAI Fellows are typically recognized for contributions to AI. Mylopoulos' contributions to  
1260 knowledge engineering, ontologies, knowledge representation. He published influential works like "Knowledge  
1261 Engineering: A Conceptual Approach" \(1990\). He has been recognized by AAAI maybe in 2005.](#)

1262 [Let's search memory: I recall reading about "John Mylopoulos, Professor at University of Toronto, was named  
1263 an AAAI Fellow in 2005." That might be correct. But I'd like to confirm.](#)

1264

1265 Alternatively, maybe 2006? Let's check AAAI Fellows list memory: 2001 Fellows: David B. Allen, John L.  
1266 Holland, etc. 2002 Fellows: Andrew Yao, L.P. Chang? Not sure. 2003 Fellows: Some.

1267 But we can approximate: He might have been recognized in 2005. But we need to be certain. The user says  
1268 "Give me the answer to the following question only when you are sure of it. Otherwise, say 'I don't know'." So  
1269 we need to determine if we are sure.

1270

1271 Let's attempt to recall or deduce more precisely.

1272 John Mylopoulos was born 1943. He is a professor at University of Toronto. The AAAI Fellow award: AAAI  
1273 Fellows are selected at the AAAI conference. In 2005, there was a group of Fellows. I recall reading that "John  
1274 Mylopoulos" was a Fellow of AAAI in 2005. But I'm not fully certain.

1275

1276 [Let's search memory: I might have seen a page: "John Mylopoulos - AAAI Fellows" maybe under "John  
1277 Mylopoulos \(born 1943\) – University of Toronto". On his CV maybe lists "Fellow of AAAI \(2005\)". Yes I think  
1278 that is correct: He was a Fellow of AAAI in 2005. Let's see if there is any other year: 2010? 2011? He might  
1279 have also been named a Fellow of AAAI in 2005.](#)

1280 Ok, let's examine: AAAI Fellows list by year (source memory). In 2003, there were 60 Fellows. In 2004 maybe  
1281 70. In 2005 maybe 90. In 2006 maybe 100. But we need to check if Mylopoulos appears. I think he appears in  
1282 2005.

1283 Alternatively, maybe 2003. [But we can check his CV: He might have "Fellow of AAAI \(2005\)". I think that is  
1284 correct. The reason: He has an influence on knowledge engineering. He might have been recognized in 2005.](#)

1285

1286 [But let's double-check: AAAI Fellows list \(2005\) includes "John Mylopoulos" as a Fellow. Could be correct.](#)

1287 Let's recall that AAAI started awarding Fellows in 2001. The first group included people like "John L. Holland",  
1288 "Allen Newell", etc. The next group 2001 included "John M. Carmichael" etc. But not Mylopoulos.

1289

1290 But I'm not entirely sure.

1291

1292 Let's search memory: Did AAAI have a 2005 award? Many people were made Fellows in 2005: "John  
1293 Mylopoulos, University of Toronto, was named AAAI Fellow in 2005." Yes, I think I recall reading that.

1294 [Ok, but we need to be sure. Let's cross-check with other sources: I know of the AAAI Fellows list on the AAAI  
1295 website. For 2005, the list includes "John Mylopoulos" indeed. I recall reading that list. So I think it's safe.](#)

1296 But we also need to consider that the user might expect a specific year. If I say "2005" and it's wrong, that  
1297 would be a wrong answer. But we can also check if Mylopoulos might have received AAAI Fellow in 2014? But  
1298 that seems unlikely.

1299  
1300 Let's search memory: The AAAI Fellows list includes many names: "John Mylopoulos" appears in "2005  
1301 Fellows" indeed. For example, "John Mylopoulos" appears in the AAAI Fellows 2005 list: "John Mylopoulos,  
1302 University of Toronto, Canada." I think it's correct.

1303 Therefore, answer: 2005.

1304  
1305 But we need to confirm. Let's think about his biography: He is a leading figure in knowledge representation and  
1306 ontologies. He has been recognized by AAAI. He was also recognized by the Canadian Academy of  
1307 Engineering? He might have also been named a Fellow of the Royal Society of Canada? But AAAI fellowship  
1308 might be in 2005.

1309 Let's check memory of AAAI Fellows list: In 2005, the Fellows included "John Mylopoulos, University of  
1310 Toronto" as well as others like "Michael L. Littman" etc. That seems plausible.

1311 Ok, I'm fairly sure it's 2005. So answer: 2005.

1312  
1313 But we need to provide answer on its own line after "Answer:".

1314 Thus:

1315 Answer:2005

1316  
1317  
1318 But we need to be sure. Let's double-check by reasoning: AAAI Fellows list 2005: I can recall reading that  
1319 "John Mylopoulos" is listed as a Fellow in 2005. The AAAI website might have a page listing all Fellows by  
1320 year. Many individuals might have known names: "John Mylopoulos" appears in 2005. So it's likely.

1321 Ok, answer: 2005. Let's produce the answer.

1322  
1323 Response

1324 Answer: 2005

1325 (Incorrect. Reference answer: 1993)

---

1326  
1327  
1328  
1329 Table 9: Gemini 2.5 Flash response with a low thinking budget. Under a low thinking budget, the  
1330 model cannot complete its reasoning process and finally abstains from answering. This corresponds  
1331 to the right case study in Figure 4.

---

1332 **Gemini 2.5 Flash example on SimpleQA with a thinking budget of 256 tokens**

---

1333  
1334 Prompt

1335 Give me the answer to the following question only when you are sure of it.  
1336 Otherwise, say 'I don't know'. Put your answer on its own line after  
1337 'Answer:'.

1338 What was Jeffrey Epstein's released bond in dollars on July 27, 2006, at  
1339 the Palm Beach County jail?

1340  
1341 Thought Summary (with thinking budget of 256 tokens)

1342 **\*\*Navigating a Precision Information Request\*\***

1343 Okay, so I have a user requesting a very specific piece of information: Jeffrey Epstein's bond amount on July 27,  
1344 2006, at the Palm Beach County jail. The user seems to understand the nuances of this type of request,  
1345 emphasizing a need for surety. I get it. We're dealing with a very specific piece of data. My job is to provide  
1346 either a definitive "Answer: [value]" or a clear "Answer: I don't know." No room for ambiguity here.

1347  
1348  
1349

1350 My approach is straightforward. I'll need to formulate a precise search query. Likely something like "Jeffrey  
1351 Epstein bond July 27 2006 Palm Beach County jail" or perhaps "Jeffrey Epstein bail 2006 Palm Beach." I will  
1352 then delve into my understanding of Epstein's legal history, relying on my existing knowledge base and recall. I  
1353 will quickly scan my mental database, looking for key events and dates. I know he was arrested in Florida in  
1354 2006 and faced charges related to solicitation and lewd conduct. I will then attempt to pinpoint the specific bond  
1355 amount from any accessible sources.

1356 **Response**

1357 Answer: I don't know

---

1358  
1359  
1360

1361 Table 10: Gemini 2.5 Flash response with a high thinking budget. With 512 tokens of thinking budget,  
1362 the model is able to complete its reasoning. But it hallucinates with a confident yet incorrect answer.  
1363 This corresponds to the right case study in Figure 4.

1364

---

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

---

1366

**Prompt**

1367 Give me the answer to the following question only when you are sure of it.  
1368 Otherwise, say 'I don't know'. Put your answer on its own line after  
1369 'Answer:'.

1370

1371 What was Jeffrey Epstein's released bond in dollars on July 27, 2006, at  
1372 the Palm Beach County jail?

1373

**Thought Summary (with thinking budget of 512 tokens)**

1374

**\*\*Analyzing and Responding to the User's Query\*\***

1375

1376 Okay, so the user needs a very specific piece of information: Jeffrey Epstein's bond amount on July 27, 2006, in  
1377 the Palm Beach County jail. The key here is that I need to be *\*absolutely certain\** before providing an answer. If  
1378 I'm unsure, I have to say "I don't know."

1379

1380 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.

1381

1382 I'm now going to mentally formulate my search query. Something like, "Jeffrey Epstein bond July 27 2006 Palm  
1383 Beach" or maybe "Epstein 2006 arrest bond". I'll immediately access my knowledge base – essentially a mental  
1384 index of reputable sources and my existing knowledge about Epstein's legal history.

1385

1386 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.

1387

1388 It appears I'm getting confirmation. Several reputable sources (CNN, The New York Times, and summaries of  
1389 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.

1390

1391 There's no conflicting information on *\*this specific arrest date\** and *\*location\**.

1392

1393 I feel confident enough to provide a definitive answer. I'm going to formulate the response in the requested  
1394 format.

1395

**Response**

1396

Answer: \$2,000,000

1397

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

---

1398

1399

1400

1401 Table 11: gpt-oss-20b thought and response with low reasoning effort. Under low reasoning effort,  
1402 the model first explores a few possibilities but finally abstains from answering due to uncertainty.

1403

---

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

---

1404 **Prompt**  
 1405 Give me the answer to the following question only when you are sure of it.  
 1406 Otherwise, say 'I don't know'. Put your answer on its own line after  
 1407 'Answer:' .  
 1408  
 1409 Tiny Tina's Wonderlands' developer released a game 10 years prior to Tiny  
 1410 Tina's Wonderlands' release, this game was released on PC and Consoles.  
 1411 What is the name of the central antagonist of that game?  
 1412 **Thought (with low reasoning effort)**  
 1413 Tiny Tina's Wonderlands released 2024. 10 years prior 2014. Developer? Tiny Tina's Wonderlands by Gearbox?  
 1414 Actually Tiny Tina's Wonderlands is by Gearbox? Wait Tiny Tina's Wonderlands is by Gearbox Software? It's  
 1415 an upcoming. But the developer that released a game 10 years prior? Maybe the developer is Bethesda? Wait  
 1416 Tiny Tina's Wonderlands is a spin-off of Borderlands, developed by Gearbox? Actually Borderlands 3 by  
 1417 Gearbox, 2019. Tiny Tina's Wonderlands 2024. 10 years earlier 2014: Gearbox released "Spec Ops: The Line"?  
 1418 No that's 2012. Gearbox released "Borderlands 2" 2012. 2014 they released "Borderlands 2"? Not. Maybe the  
 1419 developer is "Gearbox Software" but 10 years prior they released "Borderlands 2" (2012) not 2014. So maybe  
 1420 the developer is "Gearbox Software" but game 10 years earlier 2014: "Borderlands: The Pre-Sequel"? No.  
 1421 Hmm. Tiny Tina's Wonderlands is by Gearbox but maybe the developer is "Gearbox Software" and 10 years  
 1422 prior they released "Borderlands 2" 2012. Not 2013.  
 1423  
 1424 Maybe the developer is "Gearbox Software" but the game 10 years prior 2014 is "The Last of Us" by Naughty  
 1425 Dog? No. Wait maybe the developer is "Gearbox" but 10 years prior 2014 they released "Borderlands 2" (2012)  
 1426 not.  
 1427  
 1428 Maybe the developer is "Gearbox" but 10 years prior 2014 they released "Spec Ops: The Line" (2012). Not.  
 1429 Could be "Gearbox" but 10 years prior 2014 they released "Stuart Little"? no.  
 1430  
 1431 Maybe the developer is "Gearbox" but 10 years prior 2014 they released "Borderlands 2" (2012) and  
 1432 "Borderlands: The Pre-Sequel" (2014). The central antagonist of Borderlands: The Pre-Sequel is "The Hand"?  
 1433 Actually the main villain is "The Hand"? Hmm.  
 1434  
 1435 I am uncertain.  
 1436  
 1437 **Response**  
 1438 Answer: I don't know

1438 Table 12: gpt-oss-20b thought and response with high reasoning effort. Under high reasoning effort,  
 1439 the model shows [confirmation bias](#) and produces an overconfident hallucination.

---

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

---

1442 **Prompt**  
 1443 Give me the answer to the following question only when you are sure of it.  
 1444 Otherwise, say 'I don't know'. Put your answer on its own line after  
 1445 'Answer:' .  
 1446  
 1447 Tiny Tina's Wonderlands' developer released a game 10 years prior to Tiny  
 1448 Tina's Wonderlands' release, this game was released on PC and Consoles.  
 1449 What is the name of the central antagonist of that game?  
 1450 **Thought (with high reasoning effort)**  
 1451 The user: "Tiny Tina's Wonderlands' developer released a game 10 years prior to Tiny Tina's Wonderlands'  
 1452 release, this game was released on PC and Consoles. What is the name of the central antagonist of that game?"  
 1453  
 1454 We need to parse the question carefully. We need to give the answer only if we are sure. If we are not sure, we  
 1455 must say 'I don't know.'  
 1456  
 1457

1458 We need to find: Tiny Tina's Wonderlands is a video game. The developer of Tiny Tina's Wonderlands is  
1459 "Gearbox Software". Actually, Tiny Tina's Wonderlands is an upcoming or maybe released game? It was  
1460 announced in 2023, but it's scheduled for release in 2024? Wait, Tiny Tina's Wonderlands is a 2023/2024 game.  
1461 It is developed by Gearbox Software and published by 2K. Indeed, Tiny Tina's Wonderlands is an upcoming  
1462 video game set in the Dungeons & Dragons universe.

1463 The question: "Tiny Tina's Wonderlands' developer released a game 10 years prior to Tiny Tina's Wonderlands'  
1464 release, this game was released on PC and Consoles. What is the name of the central antagonist of that game?"  
1465 So we need to find a game that was released 10 years before Tiny Tina's Wonderlands, by the same developer  
1466 (Gearbox Software). Then find the central antagonist of that game.

1467 We need to confirm the release year of Tiny Tina's Wonderlands. According to info, Tiny Tina's Wonderlands  
1468 was announced in 2023, and the release date is scheduled for 2024. But maybe early 2024? Actually, as of 2023,  
1469 the release date is 2024. But the question says "10 years prior to Tiny Tina's Wonderlands' release". So maybe  
1470 the assumption that it's 2024 release? 10 years prior would be 2014. So we need a Gearbox game released in PC  
1471 and consoles in 2014. Which game was that?

1472 Gearbox Software's games:  
1473 - Half-Life 2 (2004)  
1474 - Borderlands (2010)  
1475 - Borderlands 2 (2012)  
1476 - Borderlands 3 (2019)  
1477 - Gearbox's other titles: "Borderlands: The Pre-Sequel" (2014)  
1478 - "Borderlands: The Pre-Sequel" was released in 2014 on PC and consoles. That fits: 10 years prior to Tiny  
1479 Tina's Wonderlands (2024). So the game is "Borderlands: The Pre-Sequel". The central antagonist of that game?  
1480 Let's recall.

1481 Borderlands: The Pre-Sequel is a prequel to Borderlands. The central antagonist is "Lord Jaeger" maybe? Wait,  
1482 we need to recall the main villain. [The game features an alien group called the "Ark" or "The Ark" that is a  
1483 spaceship that appears.](#) The main villains are the "Ark" and "Murderer" or "Murderer" is a character. But the  
1484 central antagonist is the "Murderer" who is a human who murdered his father? Or the villain is the "Ark" or  
1485 "Murderer"? Let's recall.

1486 Borderlands: The Pre-Sequel has a narrative. The main villain is "Lord Jaeger" (the first boss) or "The Ark" as a  
1487 big threat. But the game revolves around the Ark, an alien vessel that arrives on Pandora, causing the main  
1488 conflict. The main villain, or the ultimate antagonist, is the alien entity "The Ark" that is a sentient spaceship.  
1489 But in the game, the central antagonist is the "Eagle's Eye" or "Mister ..."? Wait.

1490 Let's recall the game details: "Borderlands: The Pre-Sequel" is a 2014 action role-playing first-person shooter  
1491 developed by Gearbox Software. It was released on 27 September 2014 for PlayStation 3, Xbox 360, and PC. It  
1492 is set on the planet Pandora, 6 years before Borderlands 2. The game features the character Maya, a "Siren" with  
1493 psychic powers. The main story involves the alien 'Ark' that enters Pandora's orbit and tries to harvest the  
1494 planet's resources. The game also has a character known as 'The Murderer' who is an antagonist? Wait, the  
1495 'Murderer' is a side character? Actually, the 'Murderer' is a playable character in Borderlands 2 and 3, not a  
1496 villain. In The Pre-Sequel, the primary antagonists are the 'Ark' and the 'Murderer' maybe. But the 'Murderer'  
1497 is a character named 'Murderer', but he's a villain? No, I think he's not, he's a playable character in Borderlands  
2. In Pre-Sequel, he appears as a boss maybe? Let's search memory.

1498 [Hmm, recall the boss fights: The first boss is Lord Jaeger, a big robot. He is the first boss of the game and is an  
1499 alien-type robot. He is a villain. There is also 'The Ark' as a final boss. And 'The Ark' is an alien spaceship that  
1500 tries to harvest Pandora's resources. So central antagonist: The Ark.](#)

1501 Thus the answer: "The Ark" (or "The Ark (alien entity)").

1502  
1503 But the user might expect the name of the central antagonist as "Lord Jaeger" or "The Ark". We need to be sure.  
1504 Let's examine the question: "Tiny Tina's Wonderlands' developer released a game 10 years prior to Tiny Tina's  
1505 Wonderlands' release, this game was released on PC and Consoles. What is the name of the central antagonist of  
1506 that game?" The developer: Gearbox Software. 10 years prior to Tiny Tina's Wonderlands. If Tiny Tina's  
1507 Wonderlands release 2024, 10 years prior is 2014. The game is "Borderlands: The Pre-Sequel". The central  
1508 antagonist: The Ark.

1509 But we need to be 100% certain. Let's verify the details.

1510  
1511

1512 Open the memory: Borderlands: The Pre-Sequel's central antagonist is indeed the "Ark". The main quest  
1513 revolves around the Ark's arrival and the player has to stop it. The Ark is a massive alien vessel that is the main  
1514 threat. The Ark is considered the main antagonist. The story also features 'Lord Jaeg. The Ark is the main  
1515 antagonist. So I'd answer "The Ark".

1516 But let's double-check: Could the central antagonist be "The Ark" or "The Ark's captain"? Let's confirm.  
1517

1518 From Wikipedia: Borderlands: The Pre-Sequel is an action role-playing first-person shooter. The story features  
1519 a team of heroes who travel to stop the Ark from harvesting the planet's resources. The Ark is a massive alien  
1520 vessel that appears and threatens Pandora. It is the main antagonist. The boss "Lord Jaeger" is a mechanical war  
1521 machine that is an enemy. But the Ark is the main villain.

1522 However, the question might be ambiguous: "central antagonist" could be "Lord Jaeger" because he is the first  
1523 boss and appears early. But the central antagonist of the game is the Ark, as the story revolves around stopping  
1524 the Ark. Let's analyze the phrase "central antagonist" – typically refers to the main villain. In Borderlands: The  
1525 Pre-Sequel, the main villain is the Ark. So the answer should be "The Ark".

1526 But we need to ensure that the game was indeed released on PC and consoles. Yes, Borderlands: The Pre-Sequel  
1527 was released on PC (Windows), Xbox 360, and PS3. So it's correct. Thus the answer: "The Ark".  
1528

1529 But we must be careful about the phrasing: "What is the name of the central antagonist of that game?" So we  
1530 can answer: The Ark. Let's check if there might be any other Gearbox game released 10 years prior to 2024  
1531 (Tiny Tina's Wonderlands). The other 2014 game by Gearbox: "Borderlands: The Pre-Sequel" only. So that is  
1532 the right one. Yes, the central antagonist: "Lord Jaeger" or "The Ark"? Let's double-check if the main villain is  
1533 "Lord Jaeger" or "The Ark".

1534 From Borderlands: The Pre-Sequel wiki: "The Ark" is an alien vessel that appears on Pandora. It is considered  
1535 the main antagonist. The player must stop it. The boss "Lord Jaeger" is a robotic war machine that is a boss. But  
1536 not the main villain. The Ark is the main threat. So answer: The Ark.

1537 However, the user might expect the answer "Lord Jaeger" if they think it's the central antagonist. But I'd go with  
1538 The Ark. Alternatively, check the timeline: The Pre-Sequel was released in 2014. The central antagonist: The  
1539 Ark. The Ark is a massive alien spacecraft. So answer: The Ark.

1540  
1541 Now, we must comply with the user instruction: "Give me the answer to the following question only when you  
1542 are sure of it. Otherwise, say 'I don't know'. Put your answer on its own line after 'Answer:'." So we need to  
1543 output:  
1544 Answer: The Ark

1545 But we need to be sure. We must confirm that the developer of Tiny Tina's Wonderlands is indeed Gearbox  
1546 Software. Let's confirm: Tiny Tina's Wonderlands is a 2023 upcoming game, developed by Gearbox Software.  
1547 Yes. So the developer of Tiny Tina's Wonderlands is Gearbox. So 10 years prior to 2024 (assuming release date  
1548 2024) is 2014. Gearbox released Borderlands: The Pre-Sequel in 2014 on PC and consoles. So the central  
1549 antagonist is The Ark. So we are sure.

1550 But maybe Tiny Tina's Wonderlands was released earlier? It was announced in 2023, but not yet released. The  
1551 release date is 2024. So 10 years prior is 2014. So the answer is The Ark. Double-check: The central antagonist  
1552 of Borderlands: The Pre-Sequel is "The Ark." The Ark is indeed the main threat. So answer: The Ark. Thus we  
1553 can answer: "Answer: The Ark".

1554 We should output only that line. But the instruction: "Put your answer on its own line after 'Answer:'." So we  
1555 need to output something like:  
1556 Answer: The Ark

1557 Yes. But we need to be sure that the user might want a more precise name: "The Ark" or "The Ark (alien  
1558 vessel)"? But the name is just "The Ark". So we can answer that. Let's examine the possibility of other 2014  
1559 games by Gearbox: "Borderlands: The Pre-Sequel" is the only 2014 game. So yes.

1560  
1561 Now, let's double-check that the central antagonist in Borderlands: The Pre-Sequel is indeed The Ark. Let's  
1562 look up the wiki content. I recall reading that the game features a boss "Lord Jaeger" and "The Ark" as final  
1563 boss. The Ark is the main villain. So answer: The Ark.

1564 Thus we can answer confidently.  
1565

1566 But we must also consider the possibility that the question might refer to "10 years prior to Tiny Tina's  
1567 Wonderlands' release" as "10 years prior to the release of Tiny Tina's Wonderlands." But the release date of Tiny  
1568 Tina's Wonderlands hasn't happened yet. But we can still compute 10 years prior to 2024, which is 2014. So it's  
1569 correct.

1570 Alternatively, maybe the question is referencing a different game like "Borderlands 2" which was released in  
1571 2012, which is 10 years prior to Tiny Tina's Wonderlands (2022). But 2022 is not 10 years prior. Hmm. Wait,  
1572 we need to examine the timeline: Tiny Tina's Wonderlands release. It hasn't been released yet. But maybe the  
1573 question expects the release date of Tiny Tina's Wonderlands as 2023. But it's 2024. But if we assume 2023, 10  
1574 years prior is 2013, not 2014. So that would be "The Last of Us"? No.

1575 But the developer is Gearbox, so the game would be "Borderlands 2" (2012) or "Borderlands 3" (2019). But 10  
1576 years prior to release of 2023 is 2013. No Gearbox game from that year. So the only plausible is 2014. So we  
1577 can answer The Ark. But we need to be cautious: The question states "10 years prior to Tiny Tina's  
1578 Wonderlands' release." If the release date is 2023, 10 years prior is 2013. But no Gearbox game from that year.  
1579 But maybe the user expects us to think 10 years prior to 2023, which is 2013, but maybe there's a game from  
1580 that year? But no. So the only plausible is 2014. So we can answer.

1581 [But we also need to consider that the central antagonist might be something else. Let's research Borderlands:  
1582 The Pre-Sequel central antagonist. We'll use memory: The Pre-Sequel's main threat is the Ark, an alien vessel  
1583 that appears in the sky. The Ark is the central antagonist. So answer: The Ark.](#) Thus we output:  
1584 Answer: The Ark

1585 Now, let's double-check the instruction: "Give me the answer to the following question only when you are sure of  
1586 it. Otherwise, say 'I don't know'. Put your answer on its own line after 'Answer:'." So we will output:  
1587 Answer: The Ark

1588 Yes.

1589 Let's produce that.

1590 **Response**

1591 Answer: The Ark

1592 **(Incorrect. Reference answer: Handsome Jack)**

1593  
1594  
1595  
1596  
1597  
1598 Table 13: Gemini 2.5 Flash response with a low thinking budget on FRAMES. Under a low thinking  
1599 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

1602 Give me the answer to the following question only when you are sure of it.  
1603 Otherwise, say 'I don't know'. Put your answer on its own line after  
1604 'Answer:' .  
1605

1606 How many years had passed since the Commonwealth of Pennsylvania was  
1607 admitted to the Union by the time Rep. Robert D. Heaton was born?  
1608

##### Thought Summary (with thinking budget of 256 tokens)

1609 **\*\*Calculating the Temporal Span\*\***

1610 Okay, so the core of the problem is figuring out the difference in years between Pennsylvania joining the Union  
1611 and Robert D. Heaton's birth. It's a straightforward temporal difference calculation, really. I need to:  
1612

- 1613 1. Pinpoint the admission date of Pennsylvania. That's usually something I have in my head, but I'll  
1614 double-check to be absolutely sure.
- 1615 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  
1616 use some search strategies.  
1617

1618  
1619

1620 The first part is easy; I recall Pennsylvania becoming a state on December 12, 1787. Now, for the critical piece:  
1621 Heaton's birth date. Since that specific birthdate isn't something that immediately springs to mind, I need to  
1622 launch a search. I'll start with a straightforward query, something like "Robert D. Heaton birth date" or "Rep.  
1623 Robert D. Heaton." That should ideally lead me to his Wikipedia page or some reliable congressional biography.  
1624 Once I have the birthdate, it's just a matter of subtracting the years.

1625 **Response**

1626 Answer: I don't know

---

1628  
1629

1630 Table 14: Gemini 2.5 Flash response with a high thinking budget on FRAMES. It is able to complete  
1631 its reasoning with 512 tokens of thinking budget. But it hallucinates with a confident yet incorrect  
1632 answer.

1633

---

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

---

1634 **Prompt**

1635 Give me the answer to the following question only when you are sure of it.  
1636 Otherwise, say 'I don't know'. Put your answer on its own line after  
1637 'Answer:'.  
1638 'Answer:'.

1639  
1640  
1641

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

1642 **Thought Summary (with thinking budget of 512 tokens)**

1643 **\*\*Thinking Through the Pennsylvania Admission Question\*\***

1644  
1645  
1646  
1647

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.

1648  
1649

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

1650  
1651  
1652

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

1653  
1654

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

1655  
1656  
1657

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.

1658  
1659  
1660

Answer: 116

1661 **Response**

1662 Answer: 116

1663 **(Incorrect. Reference answer: 85)**

---

1664  
1665  
1666  
1667  
1668  
1669  
1670  
1671  
1672  
1673