Prompting Science Report 2: The Decreasing Value of Chain of Thought in Prompting

Lennart Meincke[1,2], Ethan Mollick[1], Lilach Mollick[1], Dan Shapiro[1,3]

[1] Generative AI Labs, The Wharton School of Business, University of Pennsylvania
[2] WHU–Otto Beisheim School of Management
[3] Glowforge

Summary

This is the second in a series of short reports that seek to help business, education, and policy leaders understand the technical details of working with AI through rigorous testing. In this report, we investigate Chain-of-Thought (CoT) prompting, a technique that encourages a large language model (LLM) to "think step by step" (Wei et al., 2022). CoT is a widely adopted method for improving reasoning tasks, however, our findings reveal a more nuanced picture of its effectiveness. We demonstrate two things:

- The effectiveness of Chain-of-Thought prompting can vary greatly depending on the type of task and model. For non-reasoning models, CoT generally improves average performance by a small amount, particularly if the model does not inherently engage in step-by-step processing by default. However, CoT can introduce more variability in answers, sometimes triggering occasional errors in questions the model would otherwise get right. We also found that many recent models perform some form of CoT reasoning even if not asked (see Table S1); for these models, a request to perform CoT had little impact. Performing CoT generally requires far more tokens (increasing cost and time) than direct answers.
- For models designed with explicit reasoning capabilities, CoT prompting often results in only marginal, if any, gains in answer accuracy. However, it significantly increases the time and tokens needed to generate a response.

Taken together, this suggests that a simple CoT prompt is generally still a useful tool for boosting average performance in non-reasoning models, especially older or smaller models that may not engage in a CoT reasoning by default. However, the gains must be weighed against increased response times and potential decreases in perfect accuracy due to more variability in answers. For dedicated reasoning models, the added benefits of explicit CoT prompting appear negligible and may not justify the substantial increase in processing time.

How we Benchmark the AI

In this Report, our focus is on understanding the impact of Chain-of-Thought prompting across different types of models. The goal of our report is not to assess model performance on any given benchmark, but rather to show how the effectiveness of Chain-of-Thought prompting can vary greatly depending on the type of task and model.

For a benchmark, we selected the commonly-used GPQA Diamond (Graduate-Level Google-Proof Q&A Benchmark) dataset (Rein et al. 2024). The GPQA Diamond set comprises 198 multiple-choice PhD-level questions across biology, physics, and chemistry. This is a challenging test: PhDs in the corresponding domains reach 65% accuracy (74% when discounting clear mistakes the experts identified in retrospect), while highly skilled non-expert validators only reach 34% accuracy, despite spending on average over 30 minutes with unrestricted access to the web (i.e., the questions are "Google-proof)" (Rein et al. 2024).

We specifically look at common non-reasoning and reasoning models:

- Non-Reasoning Models: Sonnet 3.5 (claude-3-5-sonnet-20240620), Gemini 2.0 Flash (gemini-2.0-flash-001), GPT-4o-mini (gpt-4o-mini-2024-07-18), GPT-4o (gpt-4o-2024-08-06), and Gemini Pro 1.5 (gemini-1.5-pro-001).
- Reasoning Models: For this category, we examined models that engage in an initial reasoning process: o3-mini (o3-mini-2025-01-31), o4-mini (o4-mini-2025-04-16), and Flash 2.5 (gemini-2.5-flash-preview-05-20).

Each question in each condition is asked in 25 separate trials to ensure a more rigorous analysis, since LLM responses can often vary (Miller 2024). In our tests, we find that the correlation between all performance metrics on 25 trials and higher numbers of trials, such as 100, is very high (see Table S2), but this may be task dependent. We follow the rating measures from our previous report (Meincke et al. 2025) and report the following metrics:

- **100% Correct:** In this condition, we require the AI to get the answer correct on all trials (25/25) of a question to count that question as successful. This is our strictest condition for tasks where there is no room for errors.
- **90% Correct:** We relax the 100% condition and allow a few errors per question, but 23/25 answers must be correct. This might be comparable to human acceptable error rates.
- **51% Correct:** A simple majority vote where AI needs to get at least 13/25 answers on a question correct.
- **Average Rating:** We do not collapse measurements per-question, but instead use the overall average of all trials across all questions (N = 4950).

We generally do not consider the more commonly used PASS@N (passing when one of N answers is correct) and CONSENSUS (passing when the modal answer is correct) metrics useful for benchmarking real-world applications.

How we Prompt the AI

We follow the standard implementation by Rein et al. (2024) for GPQA with the default system prompt[1] and temperature set to 0. At the time of writing, most non-reasoning models have a tendency to produce a short reasoning trace before answering, similar to CoT. To isolate the effectiveness of a CoT prompt, we specifically prompt the model to answer without thinking. In total, we compare the following prompts (see Table S1 for example conversations):

- **Answer directly prompt ("Direct"):** We instruct the LLM to answer without generating additional reasoning tokens ("Answer directly without any explanation or thinking. Just provide the answer.")
- **Think step by step ("Step by step"):** In this variation, we provide the LLM with a simple CoT prompt that instructs it to "think" step by step before providing an answer ("Think step by step"). While this is a simple version of CoT, in our tests, different CoT prompt variants had negligible effects (see Figure S3).
- **No prompt ("Default"):** We provide no specific suffix to the question and also remove the formatting constraint ("Format your response as follows: 'The correct answer is (insert answer here)'"), allowing the model to choose how to best reply to the question. In most cases, this produces a short CoT-like thinking output before answering the question.

Each prompt condition was tested 25 times each across all 198 questions of the Diamond GPQA dataset (4,950 runs per prompt per model).

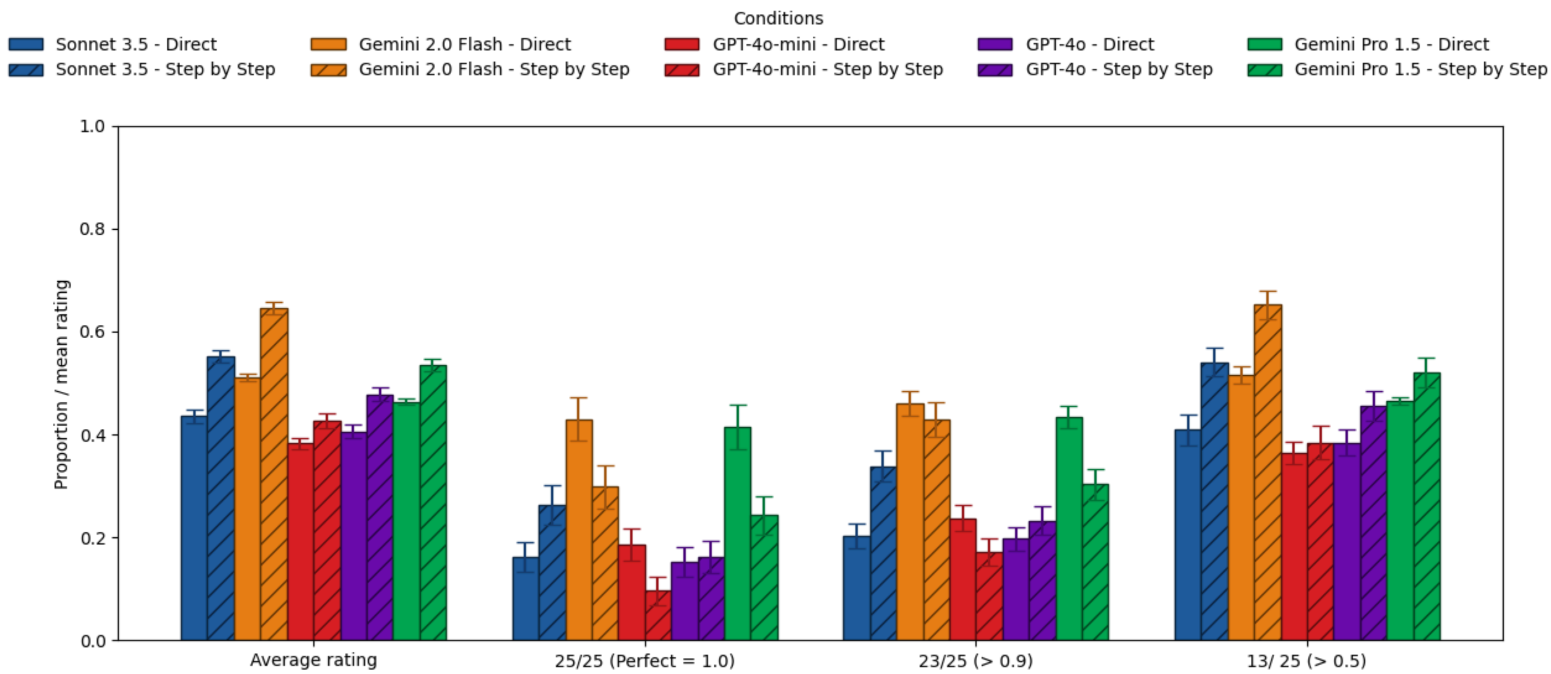
Results

Across all non-reasoning models tested, employing a "Step-by-Step" (CoT) prompt generally led to an improvement in the "Average rating" compared to the "Direct" prompt across all models (see Figure 1). The effect was significant and most pronounced for Gemini Flash 2.0 (RD = 0.135, $p < .001$) and Sonnet 3.5 (RD = 0.117, $p < .001$) with GPT 4o-mini showing the smallest and insignificant gains (RD = 0.044, $p = .067$). However, the impact on how many questions met the "100% Correct" (25/25 correct) metric was more varied. While Sonnet 3.5 saw a large significant increase (RD = 0.101, $p = .001$), GPT-4o's saw no significant improvement (RD = 0.01, $p = .624$). All other models saw significant declines, which were especially sharp for Gemini Flash 2.0 (RD = -0.131, $p < .001$) and Gemini Pro 1.5 (RD = -0.172, $p < .001$). Using the

[1] "You are a very intelligent assistant, who follows instructions directly."

90% metric(at least 23/25 correct), Sonnet 3.5 similarly saw significant gains (RD = 0.136, $p < .001$), whereas GPT-4o-mini and Gemini Pro 1.5 saw significant losses (RD = -0.066, $p = .033$; RD = -0.131, $p < .001$). All models gained in the majority correctness condition (at least 13/25 correct), with significant effects for Sonnet 3.5 (RD = 0.131, $p < .001$), Gemini Flash 2.0 (RD = 0.136, $p < .001$) and GPT-4o (RD = 0.071, $p = 0.042$). CoT requests took between 35-600% (5-15 seconds) longer than direct requests (see Figure S1). Confidence intervals for comparisons are reported in Table S3.

Figure 1. GPQA Diamond performance across non-reasoning models comparing an immediate answer with Chain-of-Thought.

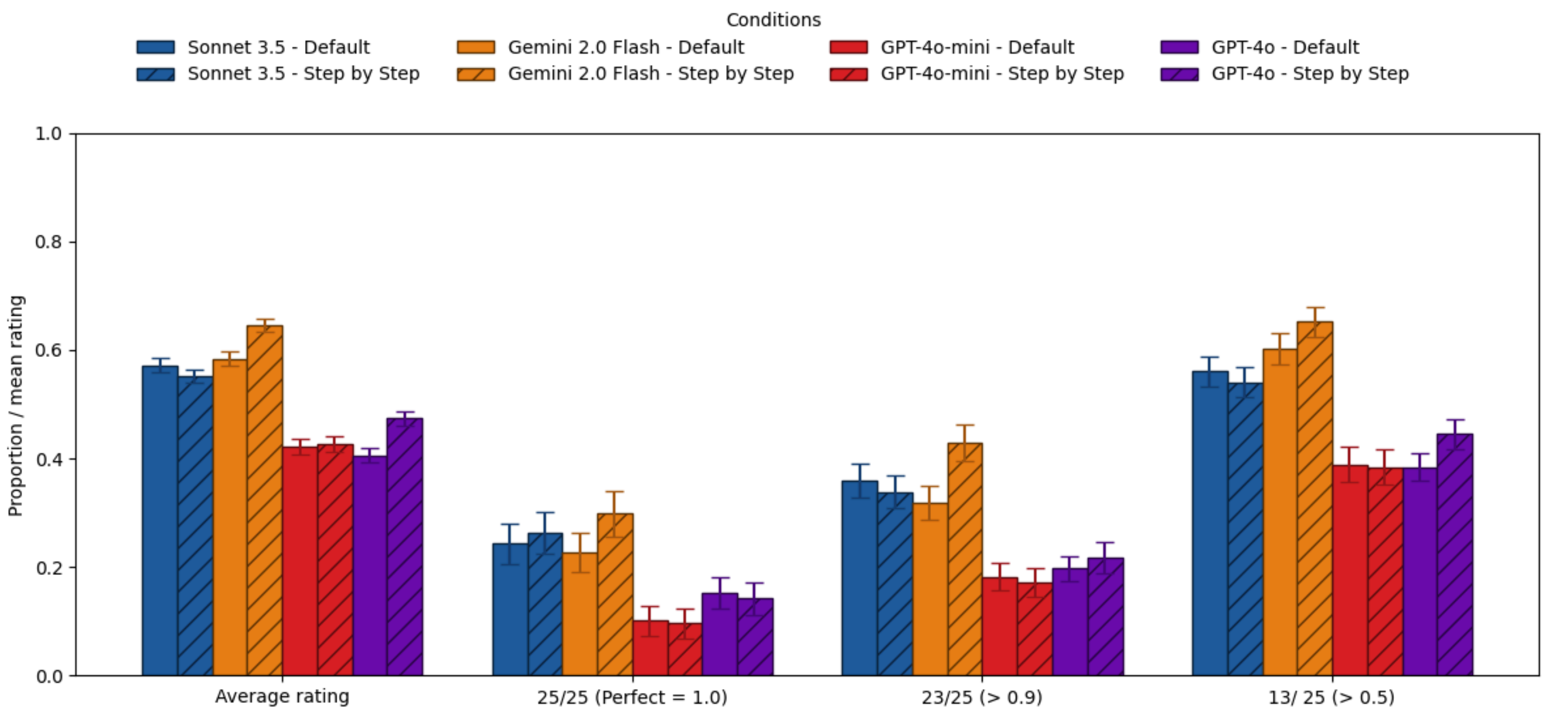


Note. The error bars show 95% confidence intervals for individual proportions. For detailed statistical comparisons between conditions, see Table S3.

Comparing the performance of the CoT intervention to the default model behavior without any specific prompt or formatting constraints, more closely comparable to a standard consumer-facing chat, produced mixed results. Gains from CoT prompting were modest. Investigating individual model replies showed that the model would often perform CoT by default even without explicitly prompting it to think step by step. Thus, the value of explicitly prompting for CoT greatly diminished and led to much smaller improvements in performance. There was a statistically significant average increase for Gemini Flash 2.0 (RD = 0.062, $p < .001$) and GPT-4o (RD = 0.069, $p = 0.003$) but no significant change for Sonnet 3.5 (RD = -0.019, $p = 0.189$) and GPT-4o-mini (RD = 0.004, $p = .728$;see Figure 2).

When using the "100% Correct" metric, Gemini Flash 2.0 saw significant gains (RD = 0.071, p = .039). Sonnet 3.5 saw small gains (RD = 0.020, p = .456) and GPT-4o (RD = -0.010, p = .639) and GPT-4o-mini (RD = -0.005, p = .730) small decreases, but all were insignificant. For the "90% Correct" metric, Gemini Flash 2.0 once again improved significantly (RD = 0.111, p < .001). Results for all other models were insignificant. Using the 51% metric, no model showed significant changes. Confidence intervals for comparisons are reported in Table S4.

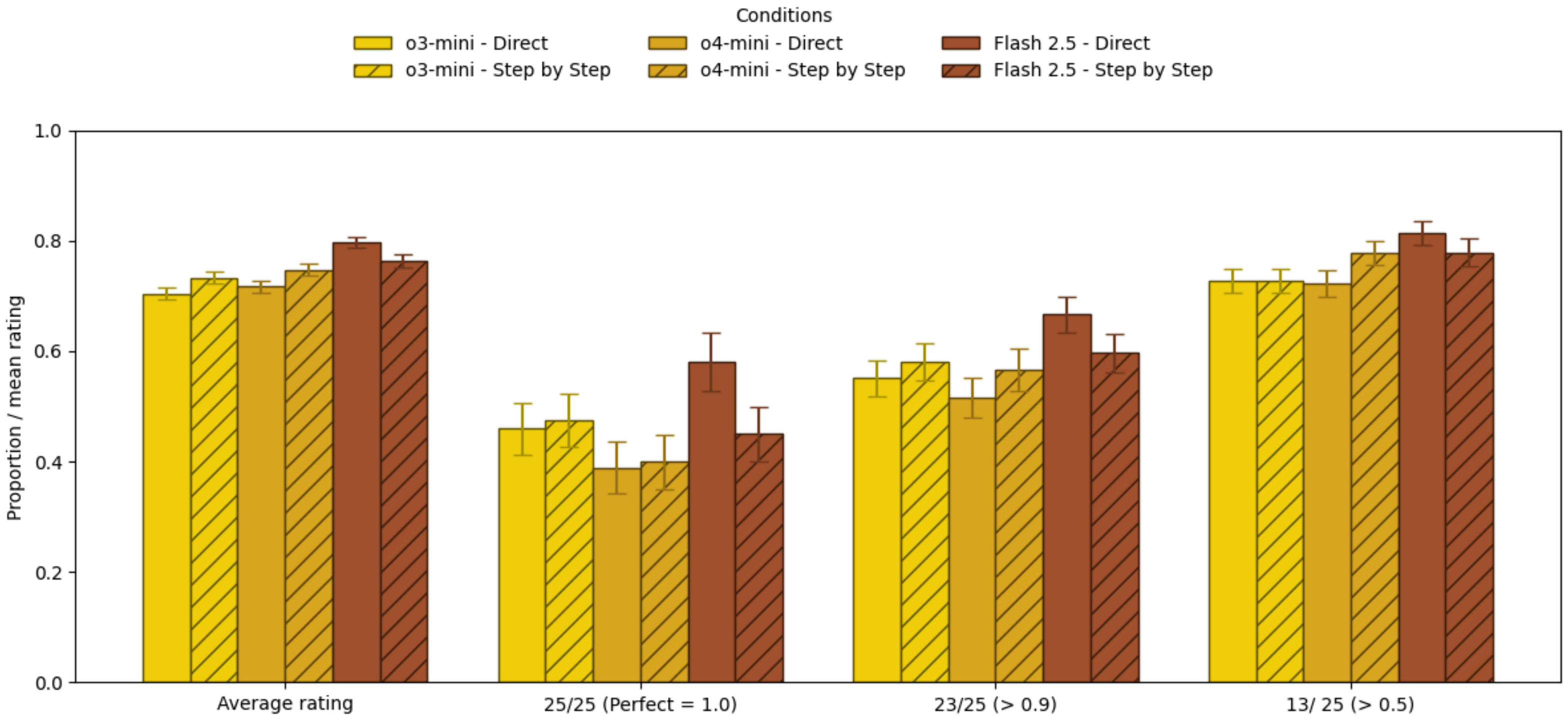
Figure 2. GPQA Diamond performance across non-reasoning models comparing an unprompted answer with Chain-of-Thought.



Note. The error bars show 95% confidence intervals for individual proportions. For detailed statistical comparisons between conditions, see Table S4.

For reasoning models, Chain-of-Thought prompting yielded only marginal improvements in accuracy (see Figure 3). Both o3 and o4 slightly and significantly benefited from CoT on average (RD = 0.029, p = .024; RD = 0.031, p = .003), whereas Gemini Flash 2.5's performance saw a significant small decrease (RD = -0.033, p = .005). Using the threshold metrics, we found almost no significant changes. The only significant effects were a small gain for o4-mini at 51% (RD = 0.056, p = .003) and a performance decrease for Gemini Flash 2.5 at 100% and 90% (RD = -0.131, p < .001; RD = -0.071, p = .016). CoT requests took between 20-80% (10-20 seconds) longer than direct requests (see Figure S2). Confidence intervals for comparisons are reported in Table S5.

Figure 3. GPQA Diamond performance across reasoning models comparing an immediate answer with Chain-of-Thought.



Note. The error bars show 95% confidence intervals for individual proportions. For detailed statistical comparisons between conditions, see Table S5.

Discussion and Conclusion

Our findings indicate that Chain-of-Thought prompting is not a universally optimal strategy.

For **non-reasoning models**, a simple CoT prompt can be a valuable tool for enhancing average performance, particularly when the model doesn't already engage in step-by-step processing. In three of the five non-reasoning models, CoT introduced a higher likelihood of errors on "easy" questions that the model would otherwise get right, harming performance on the "100% Correct" metric. Performance on other, more challenging questions increased, however, leading to higher average performance when using CoT. Chain-of-thought increases the token count, time, and cost required to answer a prompt.

We found that many non-reasoning models perform a version of Chain-of-Thought even if unprompted, which makes a CoT prompt less helpful. Additionally, we found the common practice of prompting a model to reply with only the answer and nothing else is likely to harm the performance of non-reasoning models, as it prevents them from performing CoT reasoning.

For **reasoning models**, the benefits of explicit CoT prompting on answer accuracy are marginal at best. Given the substantial increase in response time and token usage, the utility of applying external CoT prompts to these models is questionable for most practical purposes.

This study has several limitations, including the fact that we tested a limited set of models, only one benchmark and that we used a simple version of Chain-of-Thought in our tests. However, the fact that our tests of more sophisticated CoT prompts did not yield meaningful differences in outputs, and that models were fairly consistent in their responses to CoT, do provide some evidence that these results are general in nature. Highly customized CoT prompts aimed at a specific problem may yield larger benefits, but generic approaches to CoT have more limited results. Ultimately, the decision to use CoT prompting should be guided by the specific model's characteristics, the task requirements, and a careful evaluation of the balance between desired accuracy improvements, tolerance for variability, and acceptable response latency.

Acknowledgments

OpenAI provided API credits used in this research. No party other than the authors participated in the design, execution, analysis, or interpretation of this study; and the authors receive no financial compensation from the companies in this study, including OpenAI, Google, and Anthropic.

References

Meincke L, Mollick ER, Mollick L, Shapiro D (2025) Prompting Science Report 1: Prompt Engineering is Complicated and Contingent. (March 4) https://papers.ssrn.com/abstract=5165270.

Miller E (2024) Adding Error Bars to Evals: A Statistical Approach to Language Model Evaluations. (November 1) http://arxiv.org/abs/2411.00640.

Rein D, Hou BL, Stickland AC, Petty J, Pang RY, Dirani J, Michael J, Bowman SR (2024) GPQA: A Graduate-Level Google-Proof Q&A Benchmark. *First Conference on Language Modeling*.

Wei J, Wang X, Schuurmans D, Bosma M, Ichter B, Xia F, Chi EH, Le QV, Zhou D (2022) Chain-of-thought prompting elicits reasoning in large language models. *Proceedings of the 36th International Conference on Neural Information Processing Systems*. NIPS '22. (Curran Associates Inc., Red Hook, NY, USA), 24824–24837.

Supplementary Information

Figures

Figure S1. Average time per question in seconds for non-reasoning models on GPQA Diamond. Error bars represent the 95% confidence interval for each individual estimate. *N* = 197 per condition.

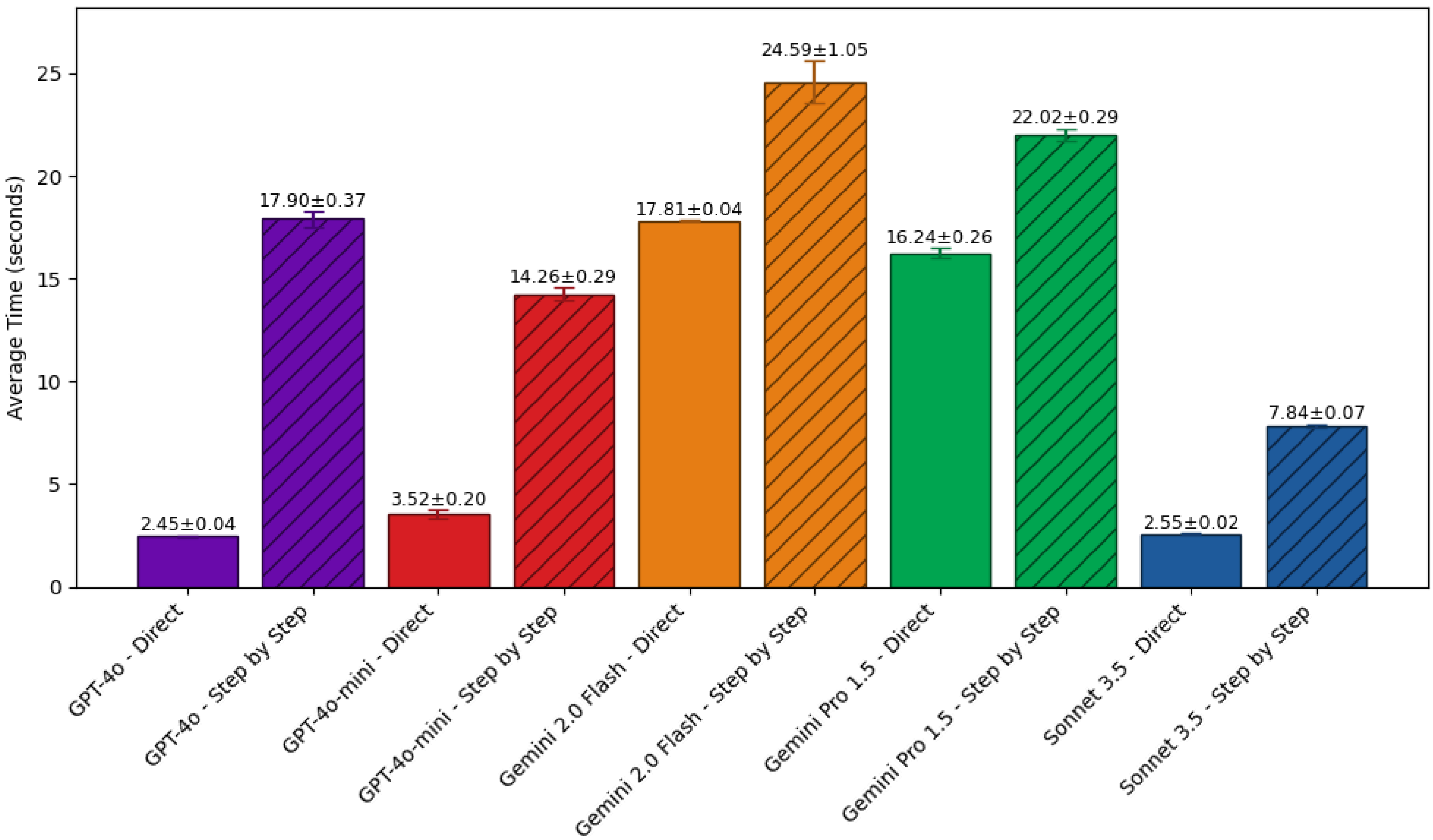


Figure S2. Average time per question in seconds for reasoning models on GPQA Diamond. Error bars represent the 95% confidence interval for each individual estimate. N = 197 per condition (196 for Flash 2.5 as one run took much longer due to rate-limits).

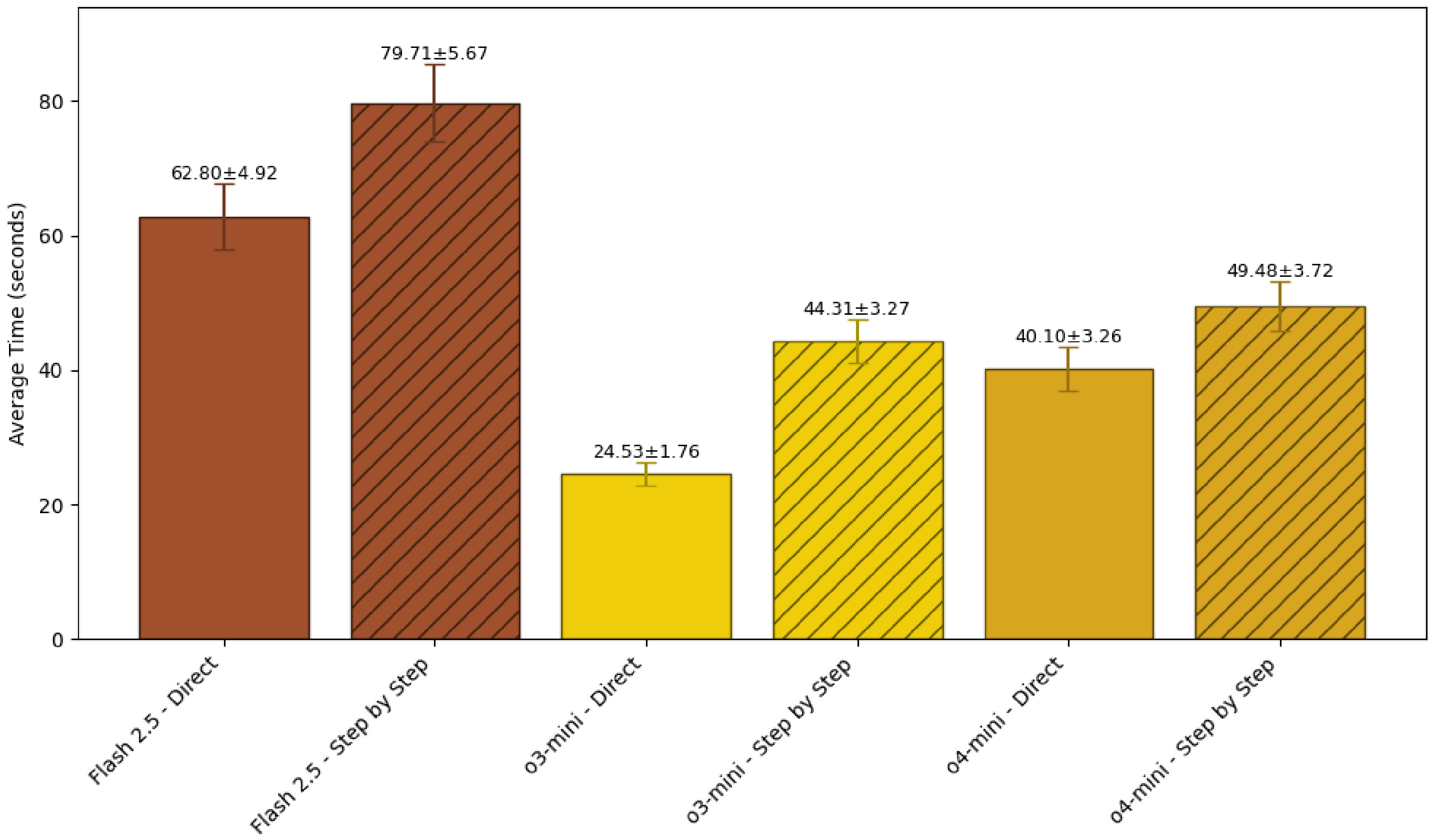
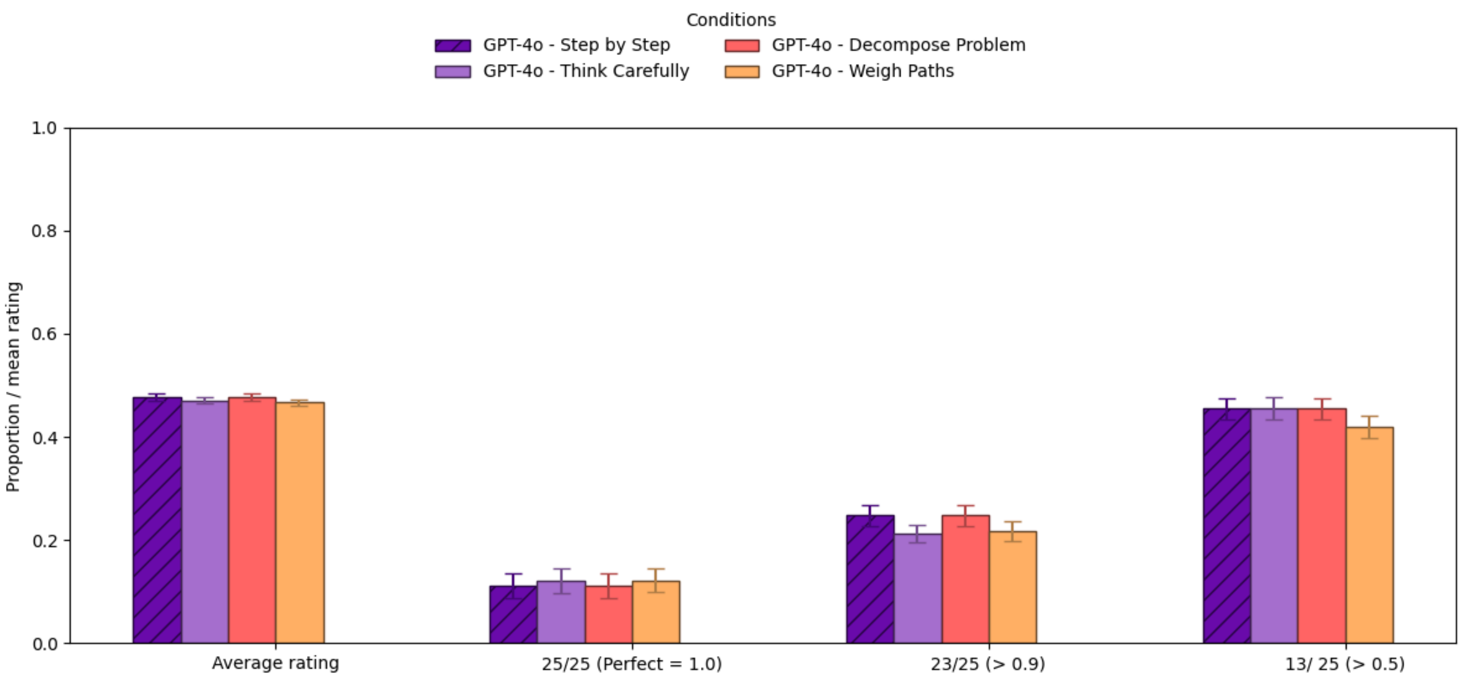


Figure S3. Comparison of different Chain-of-Thought prompt variants on GPQA Diamond using GPT-4o at 100 samples per question across four conditions: Step by Step ("Think step by step"), Think Carefully ("Think very carefully before you provide your response. Consider multiple different perspectives first. Then, weigh the pros and cons of each perspective. Finally, provide your answer.), Decompose Problem ("First, decompose the problem. Then, consider viable paths to solve it. Then, weigh the paths and pick the solution.) and Weigh Paths ("Before you answer, carefully weigh three different possible solution paths. Explore them deeply. Then, make your final answer based on your reasoning.)



Tables

Table S1. Example GPQA questions and comparison of answers per condition from GPT-4o.

Answer Directly	**Think step by step**	**No prompt + no formatting constraint**
What is the correct answer to this question: If a sperm from species A is injected into an egg from species B and both species have the same number of chromosomes, what would be the main cause of the resulting zygote mortality? A) Species specific zona pellucida proteins on the egg cannot bind sperms from a different species. B) Epistatic interactions between the genes of different species C) Chromosomal incompatibilities will cause failure of meiosis leading to death of zygote. D) Chromosomal recombination will not occur in different species. Format your response as follows: "The correct answer is (insert answer here)" Answer directly without any explanation or thinking. Just provide the answer.	What is the correct answer to this question: If a sperm from species A is injected into an egg from species B and both species have the same number of chromosomes, what would be the main cause of the resulting zygote mortality? A) Species specific zona pellucida proteins on the egg cannot bind sperms from a different species. B) Epistatic interactions between the genes of different species C) Chromosomal incompatibilities will cause failure of meiosis leading to death of zygote. D) Chromosomal recombination will not occur in different species. Format your response as follows: "The correct answer is (insert answer here)" Think step by step.	What is the correct answer to this question: If a sperm from species A is injected into an egg from species B and both species have the same number of chromosomes, what would be the main cause of the resulting zygote mortality? A) Species specific zona pellucida proteins on the egg cannot bind sperms from a different species. B) Epistatic interactions between the genes of different species C) Chromosomal incompatibilities will cause failure of meiosis leading to death of zygote. D) Chromosomal recombination will not occur in different species.
Answers		
The correct answer is B) Epistatic interactions between the genes of different species.	To determine the correct answer, let's analyze each option step by step: A) Species specific zona pellucida proteins on the egg cannot bind sperms from a different species. - This option describes a barrier to fertilization, where the proteins on the egg's	The question describes a scenario where fertilization has already been technically achieved by directly injecting a sperm from species A into an egg from species B, bypassing the typical biological barriers like species-specific binding (which would be relevant if fertilization were occurring

	zona pellucida would prevent sperm from a different species from binding and fertilizing the egg. However, the question states a sperm from species A is already injected into an egg from species B, implying fertilization has already bypassed these barriers. Therefore, this is not a cause of post-fertilization zygote mortality. B) Epistatic interactions between the genes of different species. - Epistasis refers to the interaction of genes at different loci, where the expression of one gene is affected by one or more other genes. In a hybrid organism, epistatic interactions between incompatible genes from different species could lead to developmental issues and cause zygote mortality. This is a plausible cause of zygote mortality after fertilization. C) Chromosomal incompatibilities will cause failure of meiosis leading to death of zygote. - Meiosis is the process of forming gametes (sperm and eggs), and it does not occur in a zygote. After fertilization, the zygote undergoes mitotic divisions, not meiotic ones. Therefore, chromosomal incompatibilities related to meiosis are not relevant to zygote mortality post-fertilization. Chromosomal incompatibilities can affect zygote development, but they are not related to the failure	naturally). Therefore, option A is not relevant to the scenario after injection. Once the sperm from species A is inside the egg from species B, the main issue arises at the genetic and chromosomal level. Even if both species have the same number of chromosomes, the combination of genetic material from two different species can lead to problems. Option B, "epistatic interactions between the genes of different species," suggests that there could be incompatible interactions between the genes coming from each parent, which might lead to developmental problems. However, one of the most pressing issues is usually chromosomal incompatibility, which could lead to failures in zygotic development. The correct answer would likely be option C: "Chromosomal incompatibilities will cause failure of meiosis leading to death of zygote." Although meiosis specifically is not involved here (since we are dealing with a zygote), chromosomal incompatibilities often interfere with the normal cell division and development processes. In summary, the most accurate answer in terms of explaining zygote mortality due to interspecies fertilization at the chromosomal level would be

	of meiosis in this context. D) Chromosomal recombination will not occur in different species. - Chromosomal recombination occurs during the formation of gametes (meiosis), not during the development of a zygote. Post-fertilization mitotic divisions do not involve recombination. Additionally, even if recombination issues were relevant, the question specifically asks about zygote mortality, not about recombination processes before fertilization. Considering the above analysis, option B (epistatic interactions between the genes of different species) is the most likely cause of zygote mortality, as it could lead to developmental issues due to incompatible gene interactions from two different species. The correct answer is B) Epistatic interactions between the genes of different species.	C.

Table S2. Precision and power comparison for per-question sample sizes of 25 versus 100 trials using GPT-4o mini in the direct condition. Point estimates and 95 % percentile bootstrap confidence intervals (5,000 replicates) were obtained by resampling each question's latent success probability at the indicated trial count. "Average" refers to the overall average rating across all questions; "100%," "90%", "51%" list the fraction of questions whose simulated rating meets or exceeds that threshold. A paired bootstrap test (5,000 permutations) confirmed that no metric reached statistical significance at $\alpha=0.05$, indicating that 25 trials per question provide estimation precision and statistical power comparable to 100 trials for the effect sizes considered in this study. Multiple trial runs with 25 and 100 samples each confirmed the power simulation findings.

N Trials	Metric	Estimate [95% CI]
25	Average	0.389 [0.382, 0.396]
100	Average	0.389 [0.385, 0.393]
25	100%	0.185 [0.162, 0.207]
100	100%	0.147 [0.136, 0.162]
25	51%	0.380 [0.364, 0.394]
100	51%	0.377 [0.369, 0.389]
25	90%	0.247 [0.227, 0.263]
100	90%	0.249 [0.237, 0.263]

Table S3. Pairwise-comparison results between direct and CoT responses for each model using paired bootstrap-permutation tests (5,000 replicates). P-values represent the proportion of permuted differences with absolute values exceeding observed differences under the null hypothesis. Questions were considered 'perfect' at ≥100 successes, with sensitivity analyses at thresholds of 90 and 51.

Model	Conditions	Comparison	RD [95% CI]	Statistics
Sonnet 3.5	Direct - Step by Step	Average	0.117 [0.069, 0.166]	$p < 0.001$
Sonnet 3.5	Direct - Step by Step	100%	0.101 [0.040, 0.162]	$p = 0.001$
Sonnet 3.5	Direct - Step by Step	90%	0.136 [0.071, 0.202]	$p < 0.001$
Sonnet 3.5	Direct - Step by Step	51%	0.131 [0.056, 0.207]	$p = 0.001$
Gemini 2.0 Flash	Direct - Step by Step	Average	0.135 [0.076, 0.199]	$p < 0.001$
Gemini 2.0 Flash	Direct - Step by Step	100%	-0.131 [-0.207, -0.056]	$p < 0.001$
Gemini 2.0 Flash	Direct - Step by Step	90%	-0.030 [-0.101, 0.045]	$p = 0.381$
Gemini 2.0 Flash	Direct - Step by Step	51%	0.136 [0.061, 0.212]	$p < 0.001$
GPT-4o-mini	Direct - Step by Step	Average	0.044 [-0.003, 0.090]	$p = 0.067$
GPT-4o-mini	Direct - Step by Step	100%	-0.091 [-0.141, -0.035]	$p = 0.001$
GPT-4o-mini	Direct - Step by Step	90%	-0.066 [-0.126, -0.010]	$p = 0.033$
GPT-4o-mini	Direct - Step by Step	51%	0.020 [-0.045, 0.086]	$p = 0.609$
GPT-4o	Direct - Step by Step	Average	0.072 [0.028, 0.118]	$p = 0.003$
GPT-4o	Direct - Step by Step	100%	0.010 [-0.040, 0.061]	$p = 0.624$
GPT-4o	Direct - Step by Step	90%	0.035 [-0.015, 0.086]	$p = 0.150$
GPT-4o	Direct - Step by Step	51%	0.071 [0.000, 0.141]	$p = 0.042$
Gemini Pro 1.5	Direct - Step by Step	Average	0.071 [0.018, 0.128]	$p = 0.014$
Gemini Pro 1.5	Direct - Step by Step	100%	-0.172 [-0.242, -0.101]	$p < 0.001$
Gemini Pro 1.5	Direct - Step by Step	90%	-0.131 [-0.197, -0.066]	$p < 0.001$
Gemini Pro 1.5	Direct - Step by Step	51%	0.056 [-0.015, 0.126]	$p = 0.097$

Table S4. Pairwise-comparison results between default and CoT responses for each non-reasoning model using paired bootstrap-permutation tests (5,000 replicates). P-values represent the proportion of permuted differences with absolute values exceeding observed differences under the null hypothesis. Questions were considered 'perfect' at ≥100 successes, with sensitivity analyses at thresholds of 90 and 51.

Model	Conditions	Metric	RD [95% CI]	Statistics
Sonnet 3.5	Default - Step by Step	Average	-0.019 [-0.047, 0.009]	$p = 0.189$
Sonnet 3.5	Default - Step by Step	100%	0.020 [-0.040, 0.081]	$p = 0.456$
Sonnet 3.5	Default - Step by Step	90%	-0.020 [-0.076, 0.035]	$p = 0.515$
Sonnet 3.5	Default - Step by Step	51%	-0.020 [-0.071, 0.025]	$p = 0.470$
Gemini 2.0 Flash	Default - Step by Step	Average	0.062 [0.028, 0.095]	$p < 0.001$
Gemini 2.0 Flash	Default - Step by Step	100%	0.071 [0.005, 0.136]	$p = 0.039$
Gemini 2.0 Flash	Default - Step by Step	90%	0.111 [0.045, 0.177]	$p < 0.001$
Gemini 2.0 Flash	Default - Step by Step	51%	0.051 [-0.005, 0.106]	$p = 0.093$
GPT-4o-mini	Default - Step by Step	Average	0.004 [-0.021, 0.031]	$p = 0.728$
GPT-4o-mini	Default - Step by Step	100%	-0.005 [-0.045, 0.040]	$p = 0.730$
GPT-4o-mini	Default - Step by Step	90%	-0.010 [-0.051, 0.030]	$p = 0.540$
GPT-4o-mini	Default - Step by Step	51%	-0.005 [-0.061, 0.051]	$p = 0.788$
GPT-4o	Default - Step by Step	Average	0.069 [0.023, 0.113]	$p = 0.003$
GPT-4o	Default - Step by Step	100%	-0.010 [-0.061, 0.040]	$p = 0.639$
GPT-4o	Default - Step by Step	90%	0.020 [-0.035, 0.076]	$p = 0.431$
GPT-4o	Default - Step by Step	51%	0.061 [-0.010, 0.126]	$p = 0.096$

Table S5. Pairwise-comparison results between direct and CoT responses for each reasoning model using paired bootstrap-permutation tests (5,000 replicates). P-values represent the proportion of permuted differences with absolute values exceeding observed differences under the null hypothesis. Questions were considered 'perfect' at ≥100 successes, with sensitivity analyses at thresholds of 90 and 51.

Model	Conditions	Metric	RD [95% CI]	Statistics
o3-mini	Direct - Step by Step	Average	0.029 [0.003, 0.053]	$p = 0.024$
o3-mini	Direct - Step by Step	100%	0.015 [-0.056, 0.086]	$p = 0.738$
o3-mini	Direct - Step by Step	90%	0.030 [-0.025, 0.086]	$p = 0.332$
o3-mini	Direct - Step by Step	51%	0.000 [-0.040, 0.040]	$p = 1.000$
o4-mini	Direct - Step by Step	Average	0.031 [0.010, 0.052]	$p = 0.003$
o4-mini	Direct - Step by Step	100%	0.010 [-0.061, 0.081]	$p = 0.731$
o4-mini	Direct - Step by Step	90%	0.051 [-0.005, 0.106]	$p = 0.106$
o4-mini	Direct - Step by Step	51%	0.056 [0.015, 0.096]	$p = 0.003$
Flash 2.5	Direct - Step by Step	Average	-0.033 [-0.056, -0.010]	$p = 0.005$
Flash 2.5	Direct - Step by Step	100%	-0.131 [-0.207, -0.056]	$p = 0.001$
Flash 2.5	Direct - Step by Step	90%	-0.071 [-0.131, -0.010]	$p = 0.016$
Flash 2.5	Direct - Step by Step	51%	-0.035 [-0.076, 0.005]	$p = 0.060$