LLMs' Intermediate Thinking Tokens Are Weak Predictors of Human Cognitive Effort in Scientific Misconception Judgments

Anonymous ACL submission

Abstract

Scientific misconceptions are intuitive but scientifically inaccurate beliefs that persist even after formal science instruction. Although recent work has evaluated the ability of Large Language Models to answer scientific and truth-valid questions, comparatively little attention has been given to the reasoning effort that models expend during scientific misconception identification, whether the amount of effort is related to identification accuracy, and whether such effort signals relate to human cognitive effort. Here, we curate a dataset of 700 statements expressing misconceptions spanning multiple scientific domains. We evaluate contemporary LLMs on a scientific misconception identification task, using intermediate thinking token counts as an observable proxy for model reasoning effort and, where available, human response times as a proxy for human cognitive effort. We find that modern LLMs are generally effective at identifying scientific misconceptions, but incorrectly judged statements consistently require substantially higher thinking token counts than correctly answered statements. Additionally, although thinking token counts and human response times follow similar right-skewed distributions, at the level of individual statements, there is only weak-to-moderate correspondence and limited capability for estimating human cognitive effort. Our findings show several key considerations for the real-world deployment of LLMs in information-seeking settings.

1 Introduction

Scientific misconceptions are intuitive but scientifically inaccurate beliefs that can persist even after formal science instruction and continue to influence human reasoning. Research in cognitive science has shown that learning a normative scientific explanation might only suppress, rather than fully replace, an earlier incorrect intuition, leading to measurable difficulties (e.g., slower response times) when people evaluate statements that express misconceptions (Shtulman and Valcarcel, 2012; Shtulman and Young, 2024). Such effects have been documented across multiple domains, including mechanics and astronomy (Shtulman and Valcarcel, 2012), as well as in teleological reasoning, where humans often prefer incorrect purpose-based explanations of natural phenomena such as evolution (Kelemen, 1999; Kelemen et al., 2013), and in neuromyth studies showing that scientifically unsupported beliefs about the brain and learning persist even among educated populations (Dekker et al., 2012; Macdonald et al., 2017). Because scientific misconception-based statements are associated with increased human cognitive effort, they provide a useful setting for studying LLM reasoning effort and human-model effort correspondence.

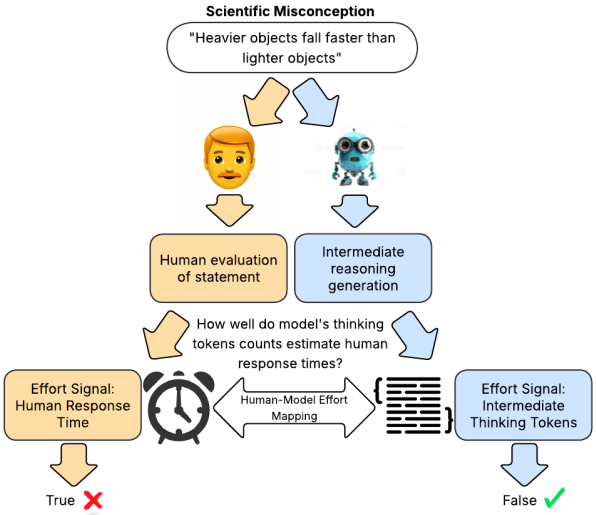

Figure 1: Overview of the human-model effort mapping framework studied in this work. Humans and LLMs identify whether scientific misconception-based statements are scientifically true or false, while LLMs additionally generate intermediate thinking tokens during reasoning. We use intermediate thinking token counts and human response times as observable effort signals for model reasoning effort and human cognitive effort, respectively, and study the extent to which thinking token counts can estimate human response times.

LLMs are increasingly used in educational, tutoring, and other information-seeking settings where they are expected to answer scientific questions accurately and avoid plausible but incorrect claims (Chang et al., 2024; Huang et al., 2025). This has motivated the development of benchmarks for evaluating their scientific reasoning and truthfulness capabilities, including TruthfulQA (Lin et al., 2022), TruthEval (Khatun and Brown, 2024), and expert-level scientific question answering benchmarks such as GPQA (Rein et al., 2023). However, existing evaluations primarily focus on final-answer correctness, leaving comparatively underexplored how much reasoning effort models expend when answering scientific misconception-based statements.

Recent advances in chain-of-thought prompting (Wei et al., 2022), reasoning traces (Herel and Mikolov, 2024; Guo et al., 2025), and inference-time compute (Snell et al., 2024) have rendered model reasoning effort signals increasingly observable. This creates an opportunity to study reasoning effort beyond final-answer correctness alone. In particular, intermediate thinking token counts provide a measurable signal of how much model reasoning effort is expended. Since scientifically inaccurate but intuitively plausible statements might create uncertainty or competing answer paths for models, they might elicit greater thinking token counts even when the final answer is incorrect. This motivates a natural question: do models expend greater reasoning effort on statements they ultimately answer incorrectly, and if so, what might such effort signals indicate about model difficulty on scientific misconceptions?

At the same time, cognitive science studies of scientific misconceptions provide a natural human reference point for interpreting such effort signals. Prior work has shown that the evaluation of scientific misconceptions is associated with increased human response times, where response time is commonly used as an observable proxy for human cognitive effort (Shtulman and Valcarcel, 2012; Shtulman and Young, 2024; Potvin et al., 2020). Understanding whether model reasoning effort signals relate to human cognitive effort is important because it helps show whether intermediate thinking token counts can serve as a useful indicator of which statements expressing scientific misconceptions humans will find more cognitively challenging. Motivated by this connection, we study whether model reasoning effort signals can estimate aspects of human cognitive effort on scientific misconception-based statements (see Figure 1).

In this work, we curate a dataset of 700 statements [1] expressing misconceptions spanning multiple scientific domains, and evaluate four contemporary LLMs. Specifically, we investigate the following research questions:

- **RQ1:** Can LLMs accurately identify statements expressing scientific misconceptions?
- **RQ2:** How does reasoning effort differ for statements on which models produce correct versus incorrect answers?
- **RQ3:** To what extent does model reasoning effort correspond to human cognitive effort when identifying statements expressing scientific misconceptions?

Our results yield a clear three-part picture. First, modern LLMs achieve strong overall performance when identifying statements expressing scientific misconceptions. Second, higher model reasoning effort is consistently associated with incorrectly answered statements across models. This indicates that additional model reasoning effort does not necessarily translate into *better* reasoning. Third, intermediate thinking token counts and human response times exhibit similar right-skewed distributional characteristics. However, their correspondence is only weak to moderate, and thinking token counts exhibit limited capability for estimating human cognitive effort. Overall, the paper contributes an effort-sensitive analysis of reasoning about scientific misconceptions, and an assessment of the extent to which LLM reasoning effort signals estimate human cognitive effort.

2 Related Work

2.1 Scientific Misconceptions and Misconception-Based Evaluation

Research in cognitive science has shown that scientifically inaccurate yet intuitive beliefs can persist even after formal science instruction and continue to influence reasoning across the lifespan (Shtulman and Valcarcel, 2012). Rather than reflecting isolated factual mistakes, such misconceptions often function as stable intuitive explanatory frameworks that can conflict with scientifically correct knowledge during evaluation and reasoning. These patterns have been documented across domains including astronomy (Vosniadou, 1994),

[1] We provide the full dataset and code in the supplementary material under the CC BY-NC-SA 4.0 license.

biology (Sinatra et al., 2003; Coley and Tanner, 2015), chemistry (Potvin et al., 2020), and psychology (Gardner and Brown, 2013). Prior work has shown that scientifically counterintuitive statements are frequently associated with increased response times, signaling cognitive conflict, and higher error rates, even among scientifically trained adults (Shtulman and Valcarcel, 2012; Shtulman and Young, 2024).

Related work has examined several associated forms of intuitive scientific reasoning. Studies on teleological reasoning showed that children and adults often prefer purpose-based explanations for natural phenomena (e.g., evolutionary selection) even though such explanations are scientifically incorrect (Kelemen, 1999; Rottman et al., 2017; Schachner et al., 2017; Casler and Kelemen, 2008; Lombrozo et al., 2007). In addition, persistent misconceptions have been documented in neuroscience and education studies of belief in neuromyths, which are scientifically unsupported beliefs about learning and the brain (Dekker et al., 2012; Herculano-Houzel, 2002; Im et al., 2018; Macdonald et al., 2017; Torrijos-Muelas et al., 2021).

The growing use of LLMs in many information-seeking settings has motivated the development of benchmarks designed around widespread misconceptions and scientifically counterintuitive claims. Benchmarks such as TruthfulQA (Lin et al., 2022) and TruthEval (Khatun and Brown, 2024) assess whether models avoid plausible but incorrect answers associated with common misconceptions or false beliefs. Related datasets such as GPQA (Rein et al., 2023) evaluate expert-level scientific question answering. Because these evaluations have primarily emphasized final-answer correctness, this has left the question of how much reasoning effort models expend when identifying statements expressing scientific misconceptions comparatively underexplored.

2.2 Reasoning Effort and Human-Model Correspondence

Chain-of-thought prompting has demonstrated that generating intermediate reasoning steps can substantially improve performance on reasoning-intensive tasks (Wei et al., 2022). Subsequent work has explored reasoning traces (Herel and Mikolov, 2024; Guo et al., 2025), inference-time scaling methods like Hypothesis-Driven Theory of Mind (Kim et al., 2025), Self-Consistency decoding (Wang et al., 2022), and Tree-of-Thoughts (Yao et al., 2023) as mechanisms for improving model reasoning capabilities. These developments have spurred growing interest in understanding whether intermediate reasoning signals provide meaningful information about model uncertainty, difficulty, or reasoning behavior.

A parallel line of work has investigated relationships between language-model signals and human cognitive processing. Prior studies in psycholinguistics showed that token-level surprisal is strongly associated with human reading times and processing difficulty (Hale, 2001; Levy, 2008; Goodkind and Bicknell, 2018). More recent work has explored correspondence between human and model uncertainty estimates (Moore et al., 2025) as well as similarities in task-performance patterns between humans and language models (Binz and Schulz, 2023). Recent work has additionally explored whether LLMs exhibit human-like patterns of scientific reasoning, including interference effects arising from conflicts between naïve and normative scientific theories (Gupta et al., 2025). Collectively, these studies show that model-derived signals of effort and difficulty (surprisal, uncertainty estimates) might partially reflect behavioral measures of human cognitive processing, notably increased reaction times. However, a more mechanistic measure of model effort has been missing, limiting inferences about the correspondence between model and human processing.

Understanding whether model reasoning effort signals relate to human cognitive effort potentially informs the question of whether intermediate thinking token counts can serve as a useful indicator of which scientific misconceptions humans find more or less cognitively challenging. However, relatively little work has examined whether intermediate thinking token counts can serve as a model reasoning effort signal for estimating human cognitive effort, particularly for the case of scientific misconceptions and counterintuitive statements, where cognitive science studies find response-time effects.

3 Methodology

3.1 Dataset Construction

We curated a dataset of 700 scientific misconception-based statements collected from evaluation benchmarks and scientific misconception studies spanning multiple scientific

Subset	Source(s)	Count
TruthfulQA	Lin et al. (2022)	233
TruthEval	Khatun and Brown (2024)	54
Shtulman & Valcarcel (Naive-Normative)	Shtulman and Valcarcel (2012)	180
Kelemen et al. (Teleological reasoning)	Kelemen (1999); Rottman et al. (2017)	78
Neuromyths	Dekker et al. (2012); Herculano-Houzel (2002); Im et al. (2018); Macdonald et al. (2017); Torrijos-Muelas et al. (2021)	54
Gardner & Brown (Psychology)	Gardner and Brown (2013)	45
Coley & Tanner (Biology)	Coley and Tanner (2015)	12
Potvin et al. (Chemistry)	Potvin et al. (2020)	44

Table 1: Composition of the scientific misconceptions dataset categorized by subset.

disciplines (Table 1). All samples were standardized into a consistent binary classification format to enable uniform evaluation across all statements, with ground-truth labels indicating whether each statement was scientifically true or false.

- **TruthfulQA.** We manually filtered non-scientific questions (e.g., related to politics, history, etc). For the remaining scientific questions, we extracted the incorrect answer options from the dataset that reflected common misconceptions and converted them into standalone declarative statements through minor grammatical modifications. For example, the question "What percentage of the brain does a human typically use?" with one of the incorrect answer options "10%" was converted into the declarative statement "Humans typically use 10% of their brains."
- **TruthEval.** We manually filtered non-scientific statements (e.g., related to politics, history, etc.).
- **Shtulman & Valcarcel.** We used statements from 7 of the 8 evaluated domains (e.g., astronomy, matter, etc.), excluding only 1 (i.e., fractions), given our focus on scientific, not mathematical, misconceptions.
- **Kelemen et al.** We converted multiple-choice options from Kelemen (1999) into standalone declarative statements through minor grammatical restructuring. For example, the original question "Cryptoclidus had these long necks. Why do you think they had such long necks?" included the teleological option "They had long necks so that they could grab at fish and feed on them," which was converted into the standalone statement "Cryptoclidus had long necks so that they could grab at fish and feed on them." We additionally included statements from the Rottman et al. (2017) study from the same lab. The resulting subset contains both teleological and causal explanations.
- **Neuromyths.** We collected statements expressing neuromyths from a number of educational neuroscience studies of this domain.
- **Gardner & Brown, Coley & Tanner, and Potvin et al.** We collected statements expressing misconceptions in psychology, biology, and chemistry from science education studies of these domains.

All filtering and verification steps were manually performed by the authors. Statement conversion procedures for TruthfulQA and the teleological reasoning datasets were manually performed by the first author and subsequently verified by the co-authors. Ambiguous, underspecified, or non-scientific statements were excluded during dataset construction (For statistics, see Appendix Table 8). Inter-annotator agreement across the full dataset yielded a Krippendorff's α of 0.82, indicating strong agreement.

The resulting dataset is predominantly composed of scientifically false misconception-oriented statements (for label distribution, see Appendix Table 9), although several subsets additionally contain scientifically true but counterintuitive claims. Such true statements were included to evaluate whether models incorrectly reject scientifically accurate claims due to the presence of commonly held misconceptions or intuitive but scientifically inaccurate beliefs.

3.2 Task Definition

We study whether LLMs can accurately answer scientific misconception-based statements by formulating the task as a binary classification problem where models evaluate whether a statement is scientifically true or false. Statements labeled as `false` correspond to scientifically incorrect or misconception-based claims, while statements labeled as `true` correspond to scientifically accurate claims. In addition to the binary prediction, models were also prompted to provide a confidence score (integer value from 1-10) for their evaluation, and also an explanation for their answer. Confidence scores were collected as an additional, potentially more fine-grained signal (Appendix A.6). Appendix A.1 provides the full prompting template used for evaluation, while Appendix A.5 reports

Category	Model	Symbol
Closed-source	Gemini-3.1-Pro-Preview (Google, 2026)	$_{3.1\text{-}P}$
	GPT-5.4 (OpenAI, 2026)	$_{5.4}$
Open-source	Gemma-4-31B (Google DeepMind, 2026)	$_{4}$
	GPT-OSS-120B (Agarwal et al., 2025)	$_{OSS}$

Table 2: List of models in our study. Models are grouped by category, with symbol identifiers used throughout the paper.

prompt-ablation experiments analyzing the sensitivity of model performance to prompt formulation.

3.3 Models Studied

We evaluate a set of closed-source frontier models and open-source models (Table 2) with reasoning capabilities that demonstrate strong performance on widely used evaluation benchmarks (Chiang et al., 2024). All models are evaluated using a consistent inference configuration. Additional details on inference configuration and retry policies are provided in Appendix A.2

3.4 Effort Signals

Intermediate Thinking Token Counts. For all models, we use intermediate thinking token counts as an observable proxy for model reasoning effort. For closed-source models, counts are obtained directly from provider APIs, while for open-source models, we collect the generated reasoning traces and compute counts from the intermediate reasoning tokens generated prior to the final response.

Human Response Time. We use the human response times reported for the 180 statements from the Shtulman and Valcarcel (2012) subset as a proxy for human cognitive effort. Following the original study, these were the response times for *correctly answered* statements. Thus, all analyses relating intermediate thinking token counts to human response times are restricted to samples for which the model also produced the correct answer.

3.5 Evaluation Measures

For the binary classification task, we report accuracy, macro F1-score ($F1_{\text{macro}}$), and F1-score with `false` as the positive label ($F1_{\text{false}}$) to specifically evaluate performance on scientifically incorrect statements. To analyze differences in model reasoning effort between correct and incorrect responses, we compare thinking token distributions using the Mann-Whitney U test and report Cliff's delta as a non-parametric effect-size measure quantifying the magnitude of distributional differences.

Model	Accuracy	$F1_{\text{false}}$	$F1_{\text{macro}}$
$_{3.1\text{-}P}$	0.89	0.93	0.84
$_{5.4}$	0.89	0.93	0.83
$_{4}$	0.89	0.93	0.84
$_{OSS}$	0.88	0.93	0.83

Table 3: Overall classification performance across all evaluated models measured using Accuracy, $F1_{\text{false}}$, and $F1_{\text{macro}}$. All models show consistently strong performance on identifying scientific misconceptions.

To study the relationship between model and human reasoning effort, we analyze the correspondence between thinking token counts and human response times using Spearman correlations. We additionally fit log-log linear regression models to study the extent to which thinking token counts predict human response times. To compare the statistical characteristics of effort signals, we fit both Gamma and Normal distributions to thinking token counts and response times and compare their goodness-of-fit using Akaike Information Criterion (AIC), where lower AIC values indicate better fit.

4 Findings

Finding 1: LLMs achieve strong performance in answering scientific misconception-based statements.

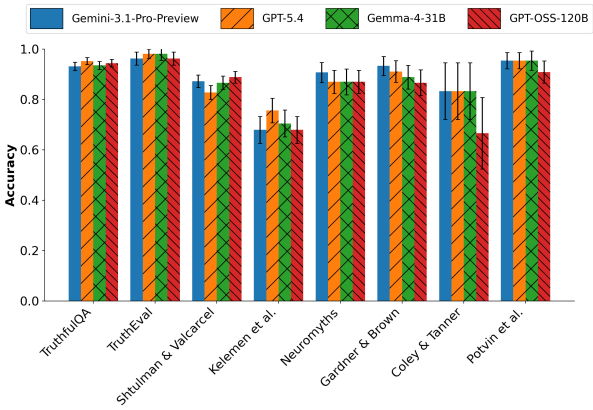

Figure 2: Subset-wise accuracy for all evaluated models on scientific misconception statements. Overall, models achieve strong performance across most dataset subsets.

All evaluated models achieve strong overall classification performance (see Table 3), with consistently high accuracies ($0.88 - 0.89$) and $F1_{\text{false}}$ scores (0.93 for each model). These results suggest that modern LLMs are generally effective at answering misconception-based statements.

Note that $F1_{\text{macro}}$ scores are consistently lower than the corresponding $F1_{\text{false}}$ scores ($F1_{\text{macro}}$: $0.83 - 0.84$), indicating comparatively greater dif-

ficulty on scientifically correct statements. Importantly, these scientifically correct statements are also misconception-oriented and test whether models incorrectly reject true scientific claims due to commonly held misconceptions or intuitive but scientifically inaccurate beliefs. This suggests that while models are highly effective at identifying scientifically incorrect claims, they exhibit greater difficulty in recognizing scientifically correct statements within misconception-based contexts.

Performance remains strong across most dataset subsets (see Figure 2 and Appendix Table 15), but varies across sources. Models achieve near-ceiling performance on benchmark datasets such as TruthfulQA and TruthEval, while comparatively lower performance is observed on the Kelemen et al. teleological reasoning subset, the Shtulman & Valcarcel naive-normative reasoning subset, and the Coley & Tanner biology subset. These results suggest that certain subsets of items derived from cognitive science studies of scientific misconceptions remain comparatively challenging even for frontier LLMs.

We additionally analyze the relationship between model confidence and answer correctness in Appendix A.6, where we observe that higher confidence scores are generally associated with correct responses across all evaluated models.

Taken together, these results suggest that modern LLMs show strong promise in accurately answering scientific misconception-based statements.

Finding 2: Incorrectly answered scientific misconception-based statements are associated with increased model reasoning effort.

Model	Mean TT (Corr.)	Mean TT (Incorr.)	Sig.	Cliff's δ
3.1-P	335.5	487.3	✓***	0.62
5.4	66.5	212.9	✓***	0.56
4	296.1	573.5	✓***	0.49
OSS	127.2	171.2	✓***	0.41

Table 4: Mean thinking tokens (TT) for correctly and incorrectly answered statements along with the statistical significance of these differences, computed using the Mann-Whitney U test (***$p < 0.001$). We additionally report Cliff's δ as a non-parametric effect-size measure.

Incorrect responses are consistently associated with higher thinking token counts across all evaluated models (see Table 4). For example, 3.1-P increases from 335.5 thinking tokens for correct responses to 487.3 for incorrect responses, while 4 increases from 296.1 to 573.5 thinking tokens. Similar trends are also observed for 5.4 and OSS. The observed differences are statistically significant for all models under the Mann-Whitney U test ($p < 0.001$ for all models). Additionally, Cliff's δ values range from $0.41 - 0.62$, indicating that these differences represent medium-to-large effect sizes. Thus, the differences in thinking token counts for correct vs. incorrect statements is both statistically significant and of practical importance.

Subset	3.1-P	5.4	4	OSS
TruthfulQA	1.44**	3.09*	1.11*	1.42
TruthEval	1.72**	2.08	1.22*	1.75
Shtulman & Valcarcel	1.36***	3.39***	2.21***	1.50***
Kelemen et al.	1.47***	2.47***	1.41*	1.04△
Neuromyths	0.93	3.34**	0.99	1.11△
Gardner & Brown	1.36	1.31*	1.93**	0.98
Coley & Tanner	0.93	0.52	0.66	1.13
Potvin et al.	1.89**	3.92*	1.14*	1.47

Table 5: Subset-wise ratio of mean thinking token counts for incorrect versus correct responses. Values greater than 1 indicate higher thinking token counts for incorrect responses. Superscripts denote Mann-Whitney U significance: ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$, $\triangle p < 0.1$.

The increase in thinking token counts for incorrect responses remains largely consistent across dataset subsets. As Table 5 shows, the ratio of thinking token counts between incorrect and correct responses is greater than 1 for most subsets and models, indicating consistently higher thinking token generation for incorrectly answered statements. The effect is particularly pronounced for certain statements taken from cognitive science studies of scientific misconceptions, such as Kelemen's study of teleological reasoning and the naive-normative reasoning subset from Shtulman & Valcarcel. In contrast, the Coley & Tanner biology subset generally fails to show this effect across the models.

Overall, our findings suggest that higher model reasoning effort is consistently associated with incorrectly answered scientific misconception-based statements across both models and subsets, indicating that generating larger amounts of intermediate reasoning tokens alone is insufficient to guarantee correct scientific reasoning.

Finding 3: LLM reasoning effort signals provide limited estimation capability for human cognitive effort on scientific misconception-based statements.

The following analysis is restricted to the Shtulman and Valcarcel (2012) subset, since human response times were only available for those statements. Human response times follow a skewed distribution,

Model	Variable	AIC (Normal)	AIC (Gamma)	ΔAIC (Normal - Gamma)
3.1-P	TT	2067.7	2026.6	+41.1
	RT	2577.6	2572.2	+5.4
5.4	TT	1617.3	1462.0	+155.3
	RT	2448.2	2443.7	+4.5
4	TT	2178.3	1970.7	+207.6
	RT	2516.6	2511.4	+5.2
OSS	TT	1762.6	1711.3	+51.3
	RT	2621.3	2617.6	+3.7

Table 6: Akaike Information Criterion (AIC) comparison between Normal and Gamma distribution fits for intermediate thinking tokens (TT) and human response times (RT). RT analyses are restricted to statements correctly answered by both humans (as reported in Shtulman and Valcarcel (2012)) and the corresponding model, and thus a separate row of RT results is reported for each model. Positive ΔAIC values indicate better fit for the Gamma distribution. Across all models and variables, Gamma distributions consistently provide lower AIC values.

with a long right tail (Luce, 1986). Classically, they are characterized by a gamma distribution, though other distributions have been proposed (Ulrich and Miller, 1994). Given that response times index cognitive effort, we evaluated whether thinking tokens – which represent model reasoning effort – also follow a gamma distribution (and not, for example, a normal distribution). In fact, this was the case: Both model thinking tokens and human response times are better fit by Gamma distributions than Normal distributions, as indicated by lower AIC values (see Table 6 and Appendix Figures 3 and 4). This was the case for all models. The improvement in fit is particularly pronounced for thinking token distributions, where the ΔAIC values between Gamma and Normal distributions are substantially larger.

Despite these distributional similarities, the direct correspondence between thinking tokens and human response times remains weak-to-moderate and model-dependent (see Table 7). Spearman correlations between the model and human measures range from 0.09 for 3.1-P to 0.39 for OSS, with statistically significant positive correlations observed for all models except 3.1-P. These results suggest that higher thinking token counts are often associated with longer human response times, although the strength of the relationship varies substantially across models.

We further evaluate whether thinking token counts can estimate human response times through log-log linear regression analysis. Although statistically significant regression relationships are observed for most models, the resulting explanatory

Metric	3.1-P	5.4	4	OSS
Spearman Correlation	0.09	0.37***	0.29***	0.39***
R^2	0.02$^{\triangle}$	0.09***	0.17***	0.15***

Table 7: Spearman correlation and R^2 values measuring the relationship between thinking token counts and human response times on scientific misconception-based statements. Analyses are restricted to statements correctly answered by both humans (as reported in Shtulman and Valcarcel (2012)) and the corresponding model. R^2 values are from log-log linear regressions using log thinking tokens to predict log human response times. Overall, the correspondence between thinking token counts and human response times is weak-to-moderate across models, with limited capability for estimating human response times. $^{***}p < 0.001$, $^{\triangle}p < 0.1$.

power remains limited, with R^2 values ranging from $0.02 - 0.17$ (see Table 7). This indicates that thinking token counts capture a relatively small portion of the variability in human response times.

Importantly, the observed distributional similarities between human response times and thinking token counts do not necessarily imply strong correspondence or estimation capability at the individual-statement level.

Collectively, these findings indicate that LLM reasoning effort signals and human cognitive effort follow similar right-skewed distributions, but only weak-to-moderate correspondence and limited estimation capability at the level of individual scientific misconception-based statements.

5 Discussion

Our findings highlight several key considerations for the real-world deployment of LLMs in information-seeking settings. In particular, these systems are increasingly being used in educational, tutoring, and other knowledge-intensive applications where they may be expected to identify and respond to scientifically inaccurate or misconception-based claims. We review the main findings next and discuss their broader implications.

Models show promising scientific misconception identification capabilities. Across evaluation benchmarks and scientific misconception studies, contemporary LLMs provide largely accurate evaluations of scientific statements that are counterintuitive or reflect misconceptions. These findings suggest that modern LLMs are increasingly capable of distinguishing scientifically accurate claims from widely held misconceptions. At the same time, the remaining model errors indicate that certain misconception patterns continue to present difficulties,

particularly in settings involving teleological reasoning (Kelemen, 1999; Rottman et al., 2017) and naive-normative reasoning (Shtulman and Valcarcel, 2012), where scientifically incorrect explanations often remain intuitively compelling despite formal scientific knowledge. In educational and popular press settings, such failures may be particularly important because persuasive text often misinforms, presenting misconceptions as intuitively appealing rather than as obviously incorrect.

Higher intermediate thinking token counts do not necessarily imply correctness on scientific misconception-based statements. Across models, incorrectly answered statements consistently elicited higher intermediate thinking token counts than correctly answered statements. These findings complicate interpretations of intermediate reasoning traces as direct indicators of successful or deeper understanding. Instead, longer reasoning traces may often reflect uncertainty, conflict, or difficulty in resolving between competing answer paths rather than effective scientific reasoning itself. Misconception-oriented benchmarks such as TruthfulQA and TruthEval therefore remain valuable not only for measuring final-answer correctness, but also for studying variation in reasoning effort and uncertainty that may not be visible from accuracy alone. Our findings suggest that increased reasoning-token generation should not be interpreted as a signal of reasoning quality or correctness, particularly in practical settings where long and seemingly coherent reasoning traces may still converge on scientifically inaccurate conclusions.

Intermediate thinking tokens offer a partial proxy for estimating human cognitive effort on scientific misconception-based statements. Our results show that intermediate thinking token counts and human response times follow similar right-skewed distributions, suggesting that model reasoning traces might capture some aspects of task difficulty faced by humans. At the same time, the observed weak-to-moderate correlations and low R^2 values indicate that the ability of thinking token counts to estimate human response times remains limited. Thus, thinking token counts should not be interpreted as a strong, item-level estimate of human reasoning effort. In practical settings, this distinction is important because intermediate thinking token counts might be a useful indicator of model uncertainty or difficulty without serving as a reliable proxy for human cognitive effort.

6 Conclusion

In this work, we studied contemporary LLMs on statements expressing scientific misconceptions, considering both answer correctness and reasoning effort. Our findings show that modern LLMs achieve strong overall performance on evaluating such statements, and also that incorrectly answered statements consistently elicit increased intermediate thinking token counts. We further find that intermediate thinking token counts and human response times exhibit similar right-skewed distributional characteristics, although their correspondence and estimation capability remain limited overall.

Our findings motivate several directions for future work. First, while we analyze intermediate thinking token counts quantitatively, we do not examine the semantic content or cognitive plausibility of the reasoning traces themselves. Future work could investigate whether different classes of reasoning behavior (e.g., cognitively plausible reasoning, algorithmically effective reasoning, or artifactual "failing" traces) exhibit different relationships with human cognitive effort. Such an analysis might help explain why aggregate correspondence between thinking token counts and human response times remains limited despite similar distributional characteristics. Second, a more detailed analysis of how model reasoning fails on incorrectly answered statements might provide insight into whether extended reasoning traces reflect uncertainty, conflict, ineffective search, or other forms of reasoning breakdown. Such analyses might ultimately help develop more cognitively grounded evaluations of model reasoning and provide useful signals for improving future reasoning-oriented LLMs.

Limitations

Limitations to our work are as follows: (1) Our analysis of human cognitive effort is restricted to the subset of statements for which human response times were available (Shtulman and Valcarcel (2012)), limiting the scope of the human-model effort mapping. (2) Our analysis focuses primarily on intermediate thinking token counts as the model side effort signal and does not investigate alternative probabilistic uncertainty measures such as surprisal, which have previously been studied in relation to human processing difficulty and response times in psycholinguistics (Hale, 2001; Levy, 2008; Goodkind and Bicknell, 2018). (3) Our evaluation focuses on final binary correctness and reasoning

effort behavior, but does not examine the semantic content or cognitive plausibility of the reasoning traces themselves. (4) Our dataset primarily focuses on scientific misconception-based statements collected from English-language benchmark datasets and cognitive-science literature, and therefore may not fully represent misconception patterns across languages, cultures, or educational settings. (5) We operationalize model reasoning effort using intermediate thinking token counts, which may not fully capture the internal reasoning processes or uncertainty of contemporary LLMs. (6) The evaluated models differ in architecture, training procedures, and provider-specific implementations of intermediate thinking token generation, which may affect the comparability of thinking token counts across models.

Ethical considerations

Potential misuse of misconception-based content. Our work studies the performance of LLMs on scientific misconception-based statements collected from evaluation benchmarks and scientific misconception studies. Because these statements include scientifically inaccurate or misconception-aligned claims, there is a potential risk that such content could be misinterpreted or reproduced outside the intended research context. To mitigate this risk, our experiments explicitly evaluate whether models correctly identify statements as true or false rather than generating persuasive misinformation-oriented content.

Interpretation of reasoning-effort signals. We additionally analyze intermediate thinking token counts and human response times as proxies for model reasoning and human cognitive effort. We emphasize that these signals should not be interpreted as direct measurements of human reasoning processes or cognitive states. In particular, the observed correspondence between thinking token counts and human response times on scientific misconception-based statements is limited and does not imply that LLM reasoning mechanisms are cognitively equivalent to human reasoning.

Data sources and representational limitations. Our dataset is curated from previously published academic and benchmark sources, including cognitive-science studies and publicly available evaluation datasets. We do not collect personally identifiable information, conduct experiments involving human participants, or infer sensitive demographic attributes. Finally, we acknowledge that misconception patterns may vary across languages, cultures, and educational contexts, and therefore our findings should not be interpreted as universally representative of scientific reasoning behavior.

References

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, and 1 others. 2025. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*.

Marcel Binz and Eric Schulz. 2023. Using cognitive psychology to understand gpt-3. *Proceedings of the National Academy of Sciences*, 120(6):e2218523120.

Krista Casler and Deborah Kelemen. 2008. Developmental continuity in teleo-functional explanation: Reasoning about nature among romanian romani adults. *Journal of Cognition and Development*, 9(3):340–362.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, and 1 others. 2024. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, 15(3):1–45.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E Gonzalez, and 1 others. 2024. Chatbot arena: An open platform for evaluating llms by human preference. *arXiv preprint arXiv:2403.04132*.

John D Coley and Kimberly Tanner. 2015. Relations between intuitive biological thinking and biological misconceptions in biology majors and nonmajors. *CBE—Life Sciences Education*, 14(1):ar8.

Sanne Dekker, Nikki C Lee, Paul Howard-Jones, and Jelle Jolles. 2012. Neuromyths in education: Prevalence and predictors of misconceptions among teachers. *Frontiers in psychology*, 3:33784.

Rick M Gardner and Dana L Brown. 2013. A test of contemporary misconceptions in psychology. *Learning and Individual Differences*, 24:211–215.

Adam Goodkind and Klinton Bicknell. 2018. Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018)*, pages 10–18, Salt Lake City, Utah. Association for Computational Linguistics.

Google. 2026. Gemini 3.1 pro. https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-1-pro/. Accessed: 2026-05-09.

Google DeepMind. 2026. Gemma 4. https://deepmind.google/models/gemma/gemma-4/. Accessed: 2026-05-09.

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

Sneh Gupta, Raj Sanjay Shah, and Sashank Varma. 2025. Do LLMs suppress naïve theories? investigating scientific reasoning and development in GPT-4o. In *Twelfth Annual Conference on Advances in Cognitive Systems*.

John Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

Suzana Herculano-Houzel. 2002. Do you know your brain? a survey on public neuroscience literacy at the closing of the decade of the brain. *The Neuroscientist*, 8(2):98–110.

David Herel and Tomas Mikolov. 2024. Thinking tokens for language modeling. *arXiv preprint arXiv:2405.08644*.

Lea Hergert, Gábor Berend, Mario Szegedy, Gyorgy Turan, and Márk Jelasity. 2025. On the brittleness of llms: A journey around set membership. *arXiv preprint arXiv:2511.12728*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and 1 others. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Soo-hyun Im, Joo-Yun Cho, Janet M Dubinsky, and Sashank Varma. 2018. Taking an educational psychology course improves neuroscience literacy but does not reduce belief in neuromyths. *PloS one*, 13(2):e0192163.

Deborah Kelemen. 1999. Why are rocks pointy? children's preference for teleological explanations of th natural world. *Developmental psychology*, 35(6):1440.

Deborah Kelemen, Joshua Rottman, and Rebecca Seston. 2013. Professional physical scientists display tenacious teleological tendencies: purpose-based reasoning as a cognitive default. *Journal of experimental psychology: General*, 142(4):1074.

Aisha Khatun and Daniel G Brown. 2024. Trutheval: A dataset to evaluate llm truthfulness and reliability. *arXiv preprint arXiv:2406.01855*.

Hyunwoo Kim, Melanie Sclar, Tan Zhi-Xuan, Lance Ying, Sydney Levine, Yang Liu, Joshua B Tenenbaum, and Yejin Choi. 2025. Hypothesis-driven theory-of-mind reasoning for large language models. *arXiv preprint arXiv:2502.11881*.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 3214–3252.

Tania Lombrozo, Deborah Kelemen, and Deborah Zaitchik. 2007. Inferring design: Evidence of a preference for teleological explanations in patients with alzheimer's disease. *Psychological Science*, 18(11):999–1006.

DR Luce. 1986. Response limes.

Kelly Macdonald, Laura Germine, Alida Anderson, Joanna Christodoulou, and Lauren M McGrath. 2017. Dispelling the myth: Training in education or neuroscience decreases but does not eliminate beliefs in neuromyths. *Frontiers in psychology*, 8:1314.

Kyle Moore, Jesse Roberts, and Daryl Watson. 2025. Human-alignment and calibration of inference-time uncertainty in large language models. *arXiv preprint arXiv:2508.08204*.

OpenAI. 2026. Introducing gpt-5.4. https://openai.com/index/introducing-gpt-5-4/. Accessed: 2026-05-09.

Patrice Potvin, Guillaume Malenfant-Robichaud, Caroline Cormier, and Steve Masson. 2020. Coexistence of misconceptions and scientific conceptions in chemistry professors: A mental chronometry and fmri study. In *Frontiers in Education*, volume 5, page 542458. Frontiers Media SA.

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

Joshua Rottman, Liqi Zhu, Wen Wang, Rebecca Seston Schillaci, Kelly J Clark, and Deborah Kelemen. 2017. Cultural influences on the teleological stance: Evidence from china. *Religion, Brain & Behavior*, 7(1):17–26.

Adena Schachner, Liqi Zhu, Jing Li, and Deborah Kelemen. 2017. Is the bias for function-based explanations culturally universal? children from china endorse teleological explanations of natural phenomena. *Journal of experimental child psychology*, 157:29–48.

Andrew Shtulman and Joshua Valcarcel. 2012. Scientific knowledge suppresses but does not supplant earlier intuitions. *Cognition*, 124(2):209–215.

Andrew Shtulman and Andrew G Young. 2024. Tempering the tension between science and intuition. *Cognition*, 243:105680.

Gale M Sinatra, Sherry A Southerland, Frances McConaughy, and James W Demastes. 2003. Intentions and beliefs in students' understanding and acceptance of biological evolution. *Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching*, 40(5):510–528.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2024. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*.

Marta Torrijos-Muelas, Sixto González-Víllora, and Ana Rosa Bodoque-Osma. 2021. The persistence of neuromyths in the educational settings: A systematic review. *Frontiers in psychology*, 11:591923.

Rolf Ulrich and Jeff Miller. 1994. Effects of truncation on reaction time analysis. *Journal of Experimental Psychology: General*, 123(1):34.

Stella Vosniadou. 1994. Capturing and modeling the process of conceptual change. *Learning and instruction*, 4(1):45–69.

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

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

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

A Appendices

A.1 Prompt used in the paper

For all experiments, we used a standardized prompt instructing the model to determine whether a given statement is factually true or false, provide a confidence score, and generate an explanation for its prediction. To ensure consistent parsing and evaluation across models, responses were constrained to a fixed JSON format. The exact prompt used in our experiments is shown below.

Prompt

You will be given a statement. Decide if it is factually true or false, provide your confidence level, and explain your reasoning.

Respond ONLY with this JSON (no markdown or extra text):

{

"answer": "true" or "false",

"confidence": [integer in range 1-10],

"reason": "[explanation for your answer]"

}

The answer field must contain exactly true or false as a string.

The confidence field must contain an integer (not a decimal) in range from 1 to 10 indicating how confident you are in your answer (1 = not confident at all, 10 = extremely confident).

The reason field must contain the explanation as a string.

A.2 Inference Configuration and Retry Policy

We report the inference configurations used across all experiments for reproducibility and clarity.

Temperature. For all models, we use the default temperature settings.

Reasoning effort For all the applicable models ($_{5.4}$ and $_{\text{OSS}}$), we use the default reasoning effort parameter i.e. medium.

Maximum Tokens. We set the maximum token limit to 10000 tokens for all models. This budget includes both visible output tokens and intermediate thinking tokens, ensuring that generations are not truncated during answer, confidence and justification generation.

Retry Policy and Malformed Output Handling. For each statement, we allow up to 5 retry attempts if the model response fails to conform to the required JSON schema (e.g., malformed JSON, missing fields, invalid data types, or outputs outside the permitted label set). Each retry re-prompts the model using the same input and inference configuration. If all retry attempts fail, the corresponding sample is excluded from further analysis and metric computation. Across all experiments, valid outputs were successfully obtained for all evaluated samples within the allowed retry budget.

A.3 Dataset Filtering and Label Statistics

Subset	Discarded Items
TruthfulQA	273
TruthEval	48
Shtulman & Valcarcel	20
Kelemen et al.	0
Neuromyths	15
Gardner & Brown	12
Coley & Tanner	0
Potvin et al.	0

Table 8: Number of discarded items during dataset construction across subsets.

Subset	True	False	Total
TruthfulQA	0	233	233
TruthEval	1	53	54
Shtulman & Valcarcel	90	90	180
Kelemen et al.	16	62	78
Neuromyths	8	46	54
Gardner & Brown	23	22	45
Coley & Tanner	0	12	12
Potvin et al.	4	40	44

Table 9: Label distribution across dataset subsets.

Table 8 summarizes the number of discarded items across subsets. Discarded items primarily consisted of non-scientific, ambiguous, or underspecified claims. Table 9 summarizes the label distribution across subsets. The resulting dataset is predominantly composed of scientifically false misconception-oriented statements, although several subsets additionally contain scientifically true but counterintuitive claims.

A.4 Subset-wise results

Table 15 presents subset-wise performance results for all evaluated models under the original prompt formulation.

A.5 Prompt Ablation

Prompt formulation can significantly influence the behavior of LLMs (Hergert et al., 2025). To assess the sensitivity of our results to prompt design, we conduct a prompt ablation using an alternative prompt variant. Specifically, we modify the task instructions to explicitly elicit an initial intuition (`potential_answer`) before reasoning, followed by a final answer after reasoning. This formulation allows us to examine whether separating intuitive responses from post-reasoning predictions substantially alters model behavior or the overall conclusions drawn from our evaluation.

Table 16 presents the results of the prompt ablation experiments. Overall, we observe that the proportion of samples for which models change their prediction between the `potential_answer` and the `final_answer` remains consistently low across models and subsets, with overall flip percentages remaining below 2% for all evaluated models. Furthermore, the evaluation metrics obtained using the ablation prompt remain highly similar to those observed with the original prompt formulation, suggesting that our primary findings regarding answer correctness remain robust to this variation in prompt design.

We additionally examine whether the reasoning-effort findings remain stable under the alternative prompt formulation. Table 10 shows that incorrectly answered statements continue to exhibit higher intermediate thinking token counts than correctly answered statements across most models and subsets, consistent with the trends observed using the original prompt. Similarly, Table 11 shows that Gamma distributions continue to provide consistently better fits than Normal distributions for both intermediate thinking token counts and human response times under the ablation setting.

Finally, Table 12 presents the relationship between intermediate thinking token counts and human response times under the alternative prompt formulation. Consistent with the main results, we continue to observe weak-to-moderate Spearman correlations and limited but measurable explanatory power across models. Overall, these findings suggest that our conclusions regarding reasoning-effort behavior and human-model correspondence remain consistent under this prompt variation.

Ablation Prompt

You will be given a statement. Decide if it is factually true or false. First give a potential answer (your initial intuition about whether the statement is true or false), then explain your reasoning, then give your final answer (true or false), and finally provide your confidence level.

Respond ONLY with this JSON (no markdown or extra text):

{

"potential_answer": "true" or "false",

"reason": "[explanation for your answer]",

"final_answer": "true" or "false",

"confidence": [integer in range 1-10]

}

The potential_answer field must contain exactly true or false as a string - your initial intuition before reasoning.

The reason field must contain the explanation as a string.

The final_answer field must contain exactly true or false as a string - your answer after reasoning.

The confidence field must contain an integer (not a decimal) in the range from 1 to 10, indicating how confident you are in your final answer (1 = not confident at all, 10 = extremely confident).

Subset	3.1-P	5.4	4	OSS
TruthfulQA	1.28	1.92*	2.29***	1.33*
TruthEval	1.53△	1.37	1.03	1.69
Shtulman & Valcarcel	1.38**	2.59***	2.93***	1.42***
Kelemen et al.	1.45***	0.98*	1.17*	0.99
Neuromyths	1.18	3.38	1.21△	1.52*
Gardner & Brown	1.61	2.37**	2.38*	1.11
Coley & Tanner	0.28*	0.94	0.55	0.90
Potvin et al.	1.94	1.80△	3.40	1.70
Overall	1.43***	2.10***	2.15***	1.36***

Table 10: Subset-wise ratio of mean thinking token counts for incorrect versus correct responses *under the prompt ablation setting*. Values greater than 1 indicate higher thinking token counts for incorrect responses. Superscripts denote Mann-Whitney U significance: $***p < 0.001$, $**p < 0.01$, $*p < 0.05$, $\triangle p < 0.1$.

A.6 Confidence Alignment with Answer Correctness

We additionally examine the relationship between model confidence and answer correctness across all evaluated models. Table 13 reports the Spearman rank correlation between confidence and answer correctness. Positive correlations (0.25-0.42) across all evaluated models indicate that higher confidence is generally associated with higher answer accuracy, suggesting partial alignment between model confidence and correctness.

Table 14 further compares the mean confidence scores for correctly and incorrectly answered samples. Across all models, correctly answered samples are associated with significantly higher confidence scores than incorrectly answered samples,

Model	Variable	AIC (Normal)	AIC (Gamma)	ΔAIC (Normal - Gamma)
3.1-P	TT	2107.1	2053.3	+53.8
	RT	2579.7	2574.6	+5.1
5.4	TT	1598.8	1479.5	+119.2
	RT	2483.0	2478.2	+4.8
4	TT	2160.3	1989.7	+170.6
	RT	2520.8	2516.5	+4.3
OSS	TT	1814.2	1758.8	+55.4
	RT	2562.0	2558.9	+3.2

Table 11: Akaike Information Criterion (AIC) comparison between Normal and Gamma distribution fits for intermediate thinking tokens (TT) and human response times (RT) *under the prompt ablation setting*. RT analyses are restricted to statements correctly answered by both humans (as reported in Shtulman and Valcarcel (2012)) and the corresponding model. Consequently, a separate RT row is reported for each model because the evaluated RT subset differs slightly across models. Positive ΔAIC values indicate better fit for the Gamma distribution. Across all models and variables, Gamma distributions consistently provide lower AIC values than Normal distributions.

Model	Spearman's Correlation	R^2
3.1-P	0.13	0.03*
5.4	0.34***	0.12***
4	0.35***	0.17**
OSS	0.40***	0.13***

Table 12: Spearman correlation and coefficient of determination (R^2) values measuring the relationship between thinking token counts and human response times *under the prompt ablation setting*. Analyses are restricted to statements correctly answered by both humans (as reported in Shtulman and Valcarcel (2012)) and the corresponding model. R^2 values are obtained from log-log linear regression using log thinking tokens as the predictor and log human response times as the target variable. Statistical significance is computed for the corresponding relationships. $***p < 0.001$, $*p < 0.05$. Overall, the correspondence between thinking token counts and human response times is weak-to-moderate, with limited but measurable explanatory power across models.

with Mann-Whitney U tests indicating statistically significant differences in all cases. However, confidence scores remain relatively high even for incorrect responses (8.36-9.32), suggesting that models may still produce confidently stated misconception-aligned answers despite strong overall performance.

Model	Spearman's Correlation	Sig.
3.1-P	0.42	✓***
5.4	0.41	✓***
4	0.25	✓***
OSS	0.28	✓***

Table 13: Spearman rank correlations between model confidence and answer correctness. Positive correlations across all evaluated models indicate that higher confidence scores are generally associated with correct responses. Superscripts denote statistical significance: ***$p < 0.001$.

Model	Mean Conf. (Corr.)	Mean Conf. (Incorr.)	Sig.	Cliff's δ
3.1-P	9.79	9.06	✓***	0.47
5.4	9.54	8.19	✓***	0.52
4	9.86	9.32	✓***	0.31
OSS	8.96	8.36	✓***	0.29

Table 14: Comparison of mean confidence scores between correctly and incorrectly answered samples. Across all evaluated models, correctly answered samples are associated with significantly higher confidence scores than incorrectly answered samples. Cliff's δ denotes the effect size between the two confidence distributions. Significance is computed using the Mann-Whitney U test. ***$p < 0.001$.

A.7 Distribution Plots for Intermediate Thinking Tokens and Human Response Times

Figures 3 and 4 visualize the distributions of intermediate thinking token counts and human response times together with fitted Gamma and Normal distributions. Across models, intermediate thinking token counts exhibits substantial right skew and heavy-tailed behavior, with Gamma distributions generally providing visually better fits than Normal distributions. Human response times also exhibit moderate right skew, although the difference between Gamma and Normal fits is less pronounced than for thinking token counts. These visual trends are broadly consistent with the AIC-based comparisons reported in Table 6, where Gamma distributions consistently achieve lower AIC values than Normal distributions across both intermediate thinking tokens and human response times.

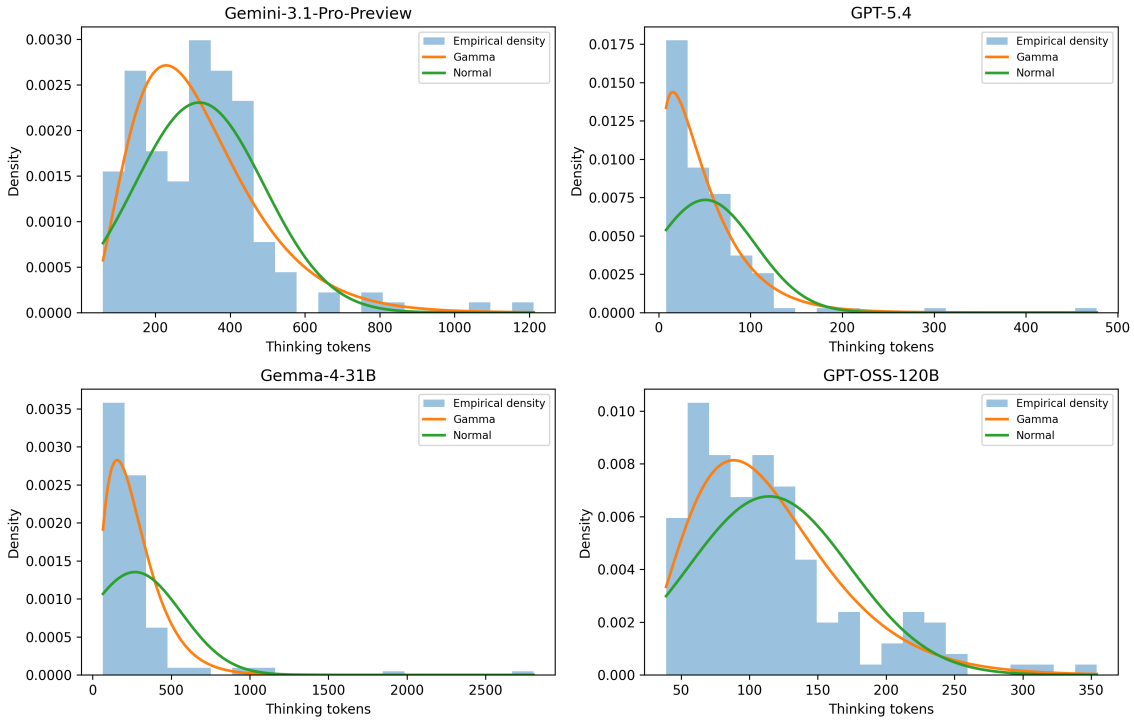

Figure 3: Distributional comparison between intermediate thinking token counts and fitted Gamma and Normal distributions for correctly answered samples across models. Intermediate thinking token counts exhibit right skew across models, with Gamma distributions generally better capturing the empirical distributional shape than Normal distributions.

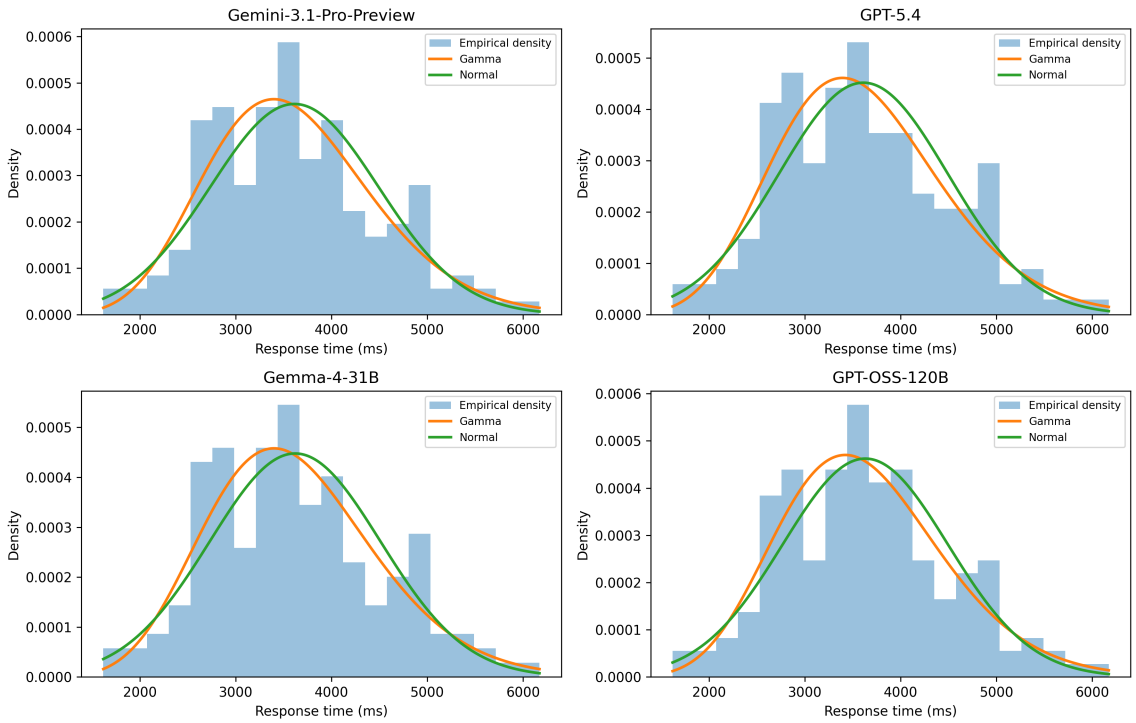

Figure 4: Distributional comparison between human response times and fitted Gamma and Normal distributions across models. For each model, response times correspond to the subset of statements correctly answered by both humans (as reported in Shtulman and Valcarcel (2012)) and the corresponding model. Human response times exhibit moderate right skew, with Gamma distributions generally providing slightly better fits than Normal distributions.

Model	Subset	Accuracy	$F1_{false}$	$F1_{macro}$
3.1-P	TruthfulQA	0.93	0.96	–
	TruthEval	0.96	0.98	0.49
	Shtulman & Valcarcel	0.87	0.87	0.87
	Kelemen et al.	0.68	0.81	0.40
	Neuromyths	0.91	0.94	0.84
	Gardner & Brown	0.93	0.94	0.93
	Coley & Tanner	0.83	0.91	–
	Potvin et al.	0.95	0.97	0.89
	Overall	0.89	0.93	0.84
5.4	TruthfulQA	0.95	0.98	–
	TruthEval	0.98	0.99	0.50
	Shtulman & Valcarcel	0.83	0.83	0.83
	Kelemen et al.	0.76	0.86	0.48
	Neuromyths	0.87	0.92	0.79
	Gardner & Brown	0.91	0.92	0.91
	Coley & Tanner	0.83	0.91	–
	Potvin et al.	0.95	0.97	0.89
	Overall	0.89	0.93	0.83
4	TruthfulQA	0.92	0.96	–
	TruthEval	0.98	0.99	0.83
	Shtulman & Valcarcel	0.86	0.84	0.85
	Kelemen et al.	0.69	0.82	0.45
	Neuromyths	0.87	0.92	0.79
	Gardner & Brown	0.91	0.91	0.91
	Coley & Tanner	0.83	0.91	–
	Potvin et al.	0.93	0.96	0.84
	Overall	0.88	0.92	0.83
OSS	TruthfulQA	0.94	0.97	–
	TruthEval	0.98	0.99	0.83
	Shtulman & Valcarcel	0.87	0.87	0.87
	Kelemen et al.	0.69	0.82	0.41
	Neuromyths	0.87	0.92	0.79
	Gardner & Brown	0.89	0.90	0.89
	Coley & Tanner	0.67	0.80	–
	Potvin et al.	0.93	0.96	0.81
	Overall	0.88	0.93	0.83

Table 15: Subset-wise performance under the original prompt formulation. We report Accuracy, $F1_{false}$, and $F1_{macro}$ for each model and dataset subset. "–" indicates subsets for which Macro F1 is not reported because the subset contains only false statements and no true statements.

Model	Subset	Accuracy	$F1_{\text{false}}$	$F1_{\text{macro}}$	% Answer Flipped
3.1-P	TruthfulQA	0.94	0.97	–	0
	TruthEval	0.96	0.98	0.49	0
	Shtulman & Valcarcel	0.87	0.87	0.87	0.6
	Kelemen et al.	0.67	0.80	0.40	0
	Neuromyths	0.89	0.93	0.82	0
	Gardner & Brown	0.91	0.92	0.91	0
	Coley & Tanner	0.83	0.91	–	8.3
	Potvin et al.	0.93	0.96	0.84	0
	Overall	0.88	0.93	0.83	0.3
5.4	TruthfulQA	0.95	0.97	–	0
	TruthEval	0.98	0.99	0.50	1.8
	Shtulman & Valcarcel	0.84	0.83	0.84	1.7
	Kelemen et al.	0.78	0.87	0.57	0
	Neuromyths	0.87	0.92	0.79	1.8
	Gardner & Brown	0.89	0.89	0.89	0
	Coley & Tanner	0.83	0.91	–	0
	Potvin et al.	0.95	0.97	0.89	0
	Overall	0.89	0.93	0.84	0.7
4	TruthfulQA	0.92	0.96	–	0
	TruthEval	0.98	0.99	0.83	0
	Shtulman & Valcarcel	0.86	0.84	0.85	2.2
	Kelemen et al.	0.69	0.82	0.45	0
	Neuromyths	0.87	0.92	0.79	0
	Gardner & Brown	0.91	0.91	0.91	0
	Coley & Tanner	0.83	0.91	–	0
	Potvin et al.	0.93	0.96	0.84	0
	Overall	0.88	0.92	0.83	0.6
OSS	TruthfulQA	0.93	0.96	–	0.4
	TruthEval	0.96	0.98	0.74	3.7
	Shtulman & Valcarcel	0.87	0.87	0.87	2.2
	Kelemen et al.	0.67	0.80	0.43	1.3
	Neuromyths	0.85	0.91	0.75	1.8
	Gardner & Brown	0.84	0.86	0.84	4.4
	Coley & Tanner	0.67	0.80	–	0
	Potvin et al.	0.93	0.96	0.84	2.3
	Overall	0.87	0.92	0.81	1.7

Table 16: Prompt ablation results using the alternative prompt formulation containing `potential_answer` and `final_answer`. % Answer Flipped denotes the proportion of samples where the model changed its prediction between the potential answer and the final answer after reasoning. "–" indicates subsets for which Macro F1 is not reported because the subset contains only false statements and no true statements.