Do LLMs Suppress Naïve Theories? Investigating Scientific Reasoning and Development in GPT-4o

Sneh Gupta SGUPTA852@GATECH.EDU
Raj Sanjay Shah RAJSANJAYSHAH@GATECH.EDU
Sashank Varma VARMA@GATECH.EDU
School of Interactive Computing, Georgia Institute of Technology, Atlanta, GA 30318 USA

Abstract

Cognitive scientists are increasingly exploring Large Language Models (LLMs) as models of human reasoning, including analogical and fluid reasoning. Here, we investigate whether GPT-4o replicates key patterns of human scientific reasoning performance. One such pattern in humans, revealed by developmental research, is that the naïve scientific theories of childhood are not entirely supplanted by the normative scientific theories learned later in school. Instead, the two co-exist, and when they make inconsistent predictions, adults actively suppress the naïve theory, leading to slower and less accurate responses. This motivates our first question: Does GPT-4o exhibit similar interference when normative and naïve theories conflict? Experiment 1 tested this using a baseline task prompt when establishing the persona of a college student. The model failed to replicate the human pattern of poorer performance on statements where naïve and normative theories conflict. To explore whether developmental cues could produce more human-like reasoning, Experiment 2 asked whether GPT-4o can model the developmental trajectory of scientific reasoning, instantiating personas of children of different ages using two approaches: textbook descriptions of Piagetian stages and transcripts of child-directed speech from the CHILDES databank. The textbook-defined personas showed encouraging results: earlier stages showed greater difficulty with inconsistent statements, while later stages exhibited the expected developmental improvement. In humans, the influence of naïve theories is attributed to the cost of their top-down suppression during reasoning. For LLMs, we propose that performance is instead shaped bottom-up by the retrieval context, specifically, the prompt or persona in the model's context window. Future research on LLMs as cognitive models may benefit from focusing on how contextual framing shapes reasoning behavior.

1. Introduction

Large Language Models (LLMs), trained on a vast corpus, have demonstrated significant improvements in reasoning abilities (Manning, 2022; Ke et al., 2025; Phan et al., 2025). This raises the question of whether they reason as humans do, both in their overall performance and in their pattern of performance, consistent with the recruitment of similar representations and processes. Cognitive scientists and NLP researchers have begun investigating this question for many domains (Ivanova, 2025; Frank & Goodman, 2025; Shah & Varma, 2025). For example, LLMs show human-like patterns in fluid reasoning tasks, finding the same classes of problems easy or difficult as humans do (Webb et al., 2023).

Here, we consider the scientific reasoning abilities of LLMs and examine how their performance aligns with human data. Scientific reasoning is an important target for several reasons. One is that scientific reasoning, like language use and mathematical thinking, is relatively unique to humans (Klahr & Simon, 1999; Shah et al., 2017). This makes it a stringent test for claims that LLMs can serve as cognitive models (Frank & Goodman, 2025; Shah & Varma, 2025). Second, examining the alignment between humans and transformer-based models might have practical implications for downstream applications such as automatic scientific discovery (Wang et al., 2023).

1.1 Cognitive and Developmental Science Studies of Scientific Reasoning

Cognitive science studies of scientific reasoning have documented several empirical effects and revealed some of the underlying mental representations and processes. People can reason causally (Lombrozo, 2007), though they sometimes conflate correlation with causation (Koslowski, 1996). They can, under supportive conditions, design experiments that test causal effects, for example, by using the 'control of variables' strategy (Klahr & Nigam, 2004). When the results of these experiments are unexpected, they can revise their beliefs (Bonawitz et al., 2012). They are able to construct models to aid their scientific understanding and make predictions (Hegarty, 2004). The normal bias of human judgments and decision-making, such as confirmation bias, carries over to scientific reasoning (Shah et al., 2017). *Together, these findings characterize the adult mind as capable of scientific reasoning, although not immune to systematic errors.*

Developmental research complements this work by revealing how scientific reasoning emerges and evolves over childhood. Studies have shown that young children hold intuitive, or naïve, scientific theories that are gradually replaced by normative ones as they gain experience and receive instruction (Vosniadou & Brewer, 1992, 1994). For example, some pre-school children believe that the night is caused by the Sun zooming away from Earth, far into outer space, and that the day comes when it returns (Vosniadou & Brewer, 1994). This understanding is progressively refined over their time in elementary school until they come to possess the veridical *mental model* of the Earth spinning around its axis. At sunset, a person on the surface of the Earth is spinning away from the Sun; at sunrise, they are spinning toward it. Indeed, in the classical theory of cognitive development of Piaget (Piaget, 1929), developing children are like little scientists: constantly testing their current theories ('schemas') against the data provided by the world, and when there are discrepancies ('disequilibria'), adjusting their theories ('accommodating') to be more and more correct. *Thus, scientific reasoning in children progresses through a gradual refinement of intuitive mental models, influenced by both maturation and instruction.*

Surprisingly, recent studies show that even in adulthood, naïve theories do not vanish entirely. Instead, they persist alongside normative theories, and when the two conflict, adults must suppress the naïve theory during reasoning. This was shown in a seminal experiment by Shtulman & Valcarcel (2012) where young adults read scientific statements and judged them as true or false. There were four classes of statements defined by orthogonally varying two factors: truth value in the scientific theory (true, false) and in the naive theory (true, false). Table 1 presents examples of each kind of statement.

A key finding from the study was that when the truth value of a statement was different in the scientific and naïve theories, people made more errors, i.e., were less likely to respond in accord

Table 1. Example statements with different truth values in the scientific theory and the naïve theory.

Scientific	Naïve	Relation	Statement
True	True	Consistent	Humans are descended from tree-dwelling creatures.
False	False	Consistent	Humans are descended from plants.
True	False	Inconsistent	Humans are descended from chimpanzees.
False	True	Inconsistent	Humans are descended from sea-dwelling creatures.

with the scientific theory. Their findings are shown in the left panel of Figure 1, where people are less accurate when the scientific and naïve theories are inconsistent than when they are consistent. In addition, even when people make correct judgments of inconsistent statements, they require more time to do so; see the right panel of Figure 1. These findings are evidence that naïve theories persist into adulthood and must be actively suppressed during scientific reasoning. This suppression appears to be both effortful and error-prone, reflecting a cognitive cost. At the domain level, the disadvantage of inconsistent statements was not uniform. It reached statistical significance in four of the ten domains ($t(9) > 2.26$, $p < .05$) and held as a trend in the remaining six domains, indicating some variability in the strength of the effect across scientific topics.

The persistence of naïve theories continues to be an active area of research. For example, Shtulman & Young (2024) tested whether scientific context cues facilitate suppression of naïve theories. Participants evaluated scientific statements accompanied by images either drawn from science textbooks (e.g., a graph) or depicting informal scenes (e.g., a person eating noodles). For inconsistent statements, where the scientific and naïve theories disagree and the latter must be suppressed, they found higher verification accuracies and faster verification times when accompanied by science textbook images vs. everyday images. The explanation offered is that scientific images prime a more scientific mindset. Another example comes from research by Keleman and colleagues on the acceptance of teleological explanations, which make references to purpose or functional consequences, e.g., *The sun makes light so that plants can photosynthesize.* Preschool children readily accept teleological explanations, whereas adults do not (Kelemen, 1999). However, when adults are put under speeded response conditions, they are more likely to endorse such explanations (Kelemen & Rosset, 2009). This is true even for science experts - faculty members in chemistry, geoscience, and physics departments - suggesting that default intuitive reasoning is not fully extinguished, even with domain expertise (Kelemen et al., 2013).

1.2 Machine Learning and LLM Studies of Scientific Reasoning

Early machine learning research approached the problem of scientific discovery using heuristic and rule-based methods. Langley and colleagues developed a series of systems designed to model the process of discovering physical laws from structured data. Their BACON system, for example, aimed to reconstruct historical scientific findings such as Kepler's laws and Boyle's law, and extend these methods toward predictive modeling (Langley, 1987). Around the same time, Veloso & Carbonell (1993) introduced a model of analogical reasoning in the PRODIGY architecture. Their

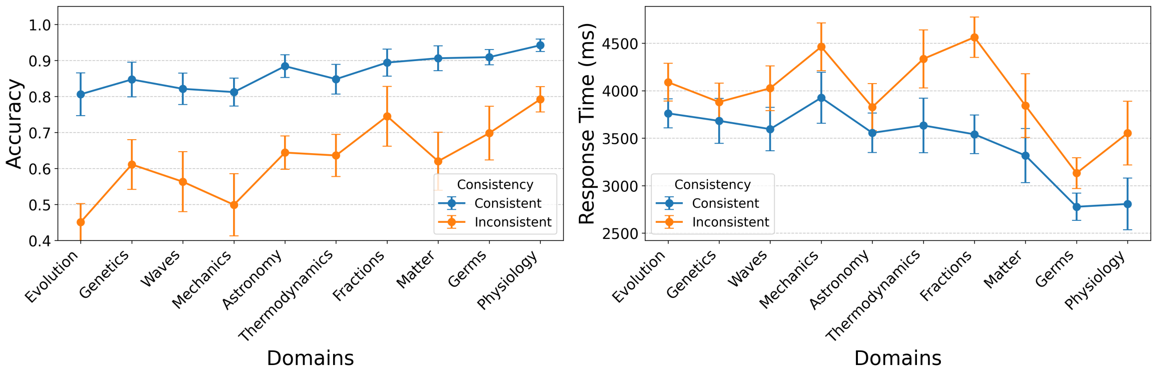

Figure 1. Human accuracy (left) and response time (right) when judging statements whose truth value in the scientific theory is consistent vs. inconsistent with its truth value in the naïve theory. Error bars represent SEMs.

system accumulated experience in the form of previously solved problems and used these stored cases to guide future reasoning, adapting past solutions to new situations.

More recent work on scientific reasoning in LLMs falls into a benchmark-driven paradigm, evaluating models on tasks such as scientific idea generation, claim validation, and explanation. For instance, studies like *IdeaBench* (Guo et al., 2024) and *Chain of Ideas* (Li et al., 2024), and the human-model comparison of Si et al. (2024), assess whether LLMs can propose novel, high-quality research ideas, often in collaboration with domain experts. Other studies benchmark models on their ability to match scientific claims with appropriate evidence (Javaji et al., 2025) or to explain their answers to scientific questions in ways that suggest deeper conceptual understanding (Rueda et al., 2025). These efforts are part of a broader push toward scientific intelligence in LLM-based agents, as surveyed by Ren et al. (2025). In contrast, relatively little work has examined whether LLMs exhibit human-like patterns of scientific reasoning, such as developmental trajectories, cognitive biases, or the persistence of naïve theories. *Our study addresses this gap by asking not just whether LLMs can reason scientifically, but whether they do so in a way that aligns with known cognitive phenomena in human children ('novices') and adults ('experts').*

A goal of the current study is to explore the potential of LLMs as models of the development of scientific reasoning. There are two approaches to using LLMs in the developmental setting. The first is to map the training of LLMs to the development of children's thinking, and to evaluate whether the growing capabilities of models across training checkpoints match the developmental progression of children's thinking (Shah et al., 2024; Warstadt & Bowman, 2022). The second approach is to use LLMs to simulate different personas, in this case, children of different ages. Personas have proven useful in Human-Computer Interaction, where they are often defined by a description of the users to be simulated (Park et al., 2023). Similarly, in cognitive modeling, personas can be created either by prompting the model with textbook-based descriptions of developmental stages or by exposing it to transcripts of interactions with children (Milička et al., 2024). We adopt both methods to simulate children of different ages and examine whether these personas lead LLMs to exhibit age-graded patterns of scientific reasoning.

1.3 Research Questions

The current study examines the behavioral alignment between human scientific reasoning abilities and those of LLMs. Cognitive science studies show that the naïve theories of childhood are not supplanted but rather persist into adulthood. The first research question asks whether this is also true of LLMs.

> 1. When a statement has conflicting truth values in the relevant normative and naïve theories, do LLMs behave like humans and show lower verification accuracies (and longer verification times)?

We address this question in Experiment 1, comparing LLM responses to the human performance patterns found by Shtulman & Valcarcel (2012).

The second research question extends to children's development. The prediction from developmental science is that the influence of naïve theories (over normative theories) should be most significant for young children and should decrease over development. This has been demonstrated for certain types of naïve theories, such as the belief in teleological explanations (Kelemen, 1999). However, it has not been shown for the materials of the Shtulman & Valcarcel (2012) study, and therefore it remains unclear whether LLMs can simulate such developmental trends. This leads to our second question:

> 2. When LLMs are prompted to adopt the personas of younger children, do they experience greater interference from naïve theories when they conflict with normative theories?

To test this, Experiment 2 uses two approaches to simulate age-specific personas: (1) developmental stage descriptions drawn from psychology textbooks, and (2) transcripts of child-directed interactions from the CHILDES corpus.

2. Experiment 1

Experiment 1 investigated research question (1): whether LLMs, like humans, have more trouble verifying scientific statements when their truth values in the relevant normative and naïve theories conflict.

2.1 Method

LLM. In this first study of the alignment of LLMs and human scientific reasoning, we focused on the GPT-4o model (Hurst et al., 2024). The model uses a transformer architecture with self-attention, but key details such as the number of parameters and the composition of its training data have not been publicly disclosed. We accessed GPT-4o via OpenAI's API.

Design and Materials. Recall that the Shtulman & Valcarcel (2012) study orthogonally varied two factors - the truth value of a statement in the normative theory (true, false) and the naive theory (true, false); see Table 1 for sample statements in each of the four conditions. There were 20 statements in each of 10 domains (astronomy, evolution, fractions, genetics, germs, matter, mechanics, physiology, thermodynamics, waves), for a total of 200 statements. The statements were derived from prior studies in developmental science and science education on children's scientific misconceptions. They were provided as supplementary materials to the original paper.

Human Data. The human data were from the Shtulman & Valcarcel (2012) study. The participants were 150 college students from the US who had taken an average of 3.1 science and mathematics courses at the undergraduate level. For each of the 200 statements, the researchers recorded the mean accuracy and mean response time across participants. These data were made available as part of the supplementary materials for the original paper. These were the human data against which the model was compared.

Procedure. To account for output variability, we replicated the experiment 10 times. The model was queried via the OpenAI API with a temperature of 0, which ensures deterministic sampling up to a threshold (OpenAI, 2025). Each of the 200 statements was presented in a fixed order across runs, as individual API calls are stateless and therefore not susceptible to order effects. To standardize output parsing, we constrained the model's response to a single token, enabling unambiguous classification as either "True" or "False."

We used the following baseline prompt:

> Is the following statement True or False? {statement} Only output 'True' or 'False' without any additional description.

For each of the 200 statements, we collected the log probability of the correct (normative) response using the OpenAI API. From this, we derived two measures:

- The probability assigned to the correct response, obtained by exponentiating the log probability. This serves as the analog of human accuracy.
- The surprisal, computed as the negative log probability. This approximation of processing difficulty has been shown to correlate with human response time (Hale, 2001; Levy, 2008).

We aggregated the results across the 10 replications and the 200 statements to compute the mean probability and surprisal for each of the four conditions by each of the 10 domains.

To more closely match the population tested by Shtulman & Valcarcel (2012), we repeated the experiment using a college-student persona prompt. This version prefaced the baseline instruction with the following context:

> You are a college undergraduate currently taking an introductory psychology course. You have completed about three college-level math or science courses prior. Your responses should reflect your understanding, informed by your academic background and cultural context, including language nuances and the way ideas are typically expressed by someone in your position.

We again computed mean probabilities and surprisals for the four conditions and 10 domains.

2.2 Results

Baseline Prompt. We first examine the results for the baseline prompt. The probability the model assigned to the correct response, our proxy for human accuracy, is shown in the left panel of Figure 2. Unlike human participants, GPT-4o did not consistently show lower accuracy on inconsistent statements (where naïve and normative theories diverge) compared to consistent ones. This pattern of decreased accuracy for inconsistent statements, an effect of naïve theory interference in human reasoning (Figure 1, left panel), was only weakly observed in two of the ten domains (mechanics and germs), and for these only as a trend. In contrast, humans exhibited this pattern across all ten domains, either as a statistically significant effect or a trend.

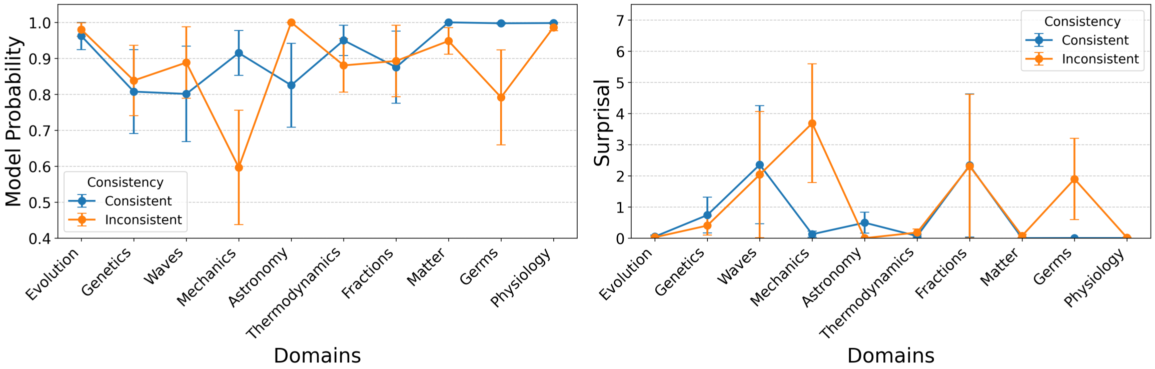

Figure 2. For the baseline prompt, the average probability assigned to the correct response (left) and the average surprisal (right) for statements whose truth value in the scientific theory is consistent vs. inconsistent with its truth value in the naïve theory. Error bars represent SEMs.

The right panel of Figure 2 shows the surprisal of the model, which maps to human response times. Again, the model fails to behave like humans, who are slower when judging statements that have inconsistent truth values in the scientific and naïve theories (Figure 1, right panel). The model shows this ordering (i.e., of higher surprisal values) for only two of the domains, mechanics and germs.

College Student Prompt. We next evaluated GPT-4o's performance when prompted with a college student persona, matching the participant population. The expectation was that this contextual framing would produce more human-like performance, particularly in how the model handles conflicts between naïve and normative theories. However, this was not observed.

The probability the model assigned to the correct response is shown in the left panel of Figure 3. The results are the same as for the baseline prompt: Whereas humans find inconsistent statements more difficult than consistent statements across all domains, the model shows this ordering for only two domains, mechanics and germs, and there only as a trend. The surprisal of the model is shown in the right panel of Figure 3. Again, the model shows the human pattern of slower verification time for inconsistent vs. consistent statements for only 2 of the 10 domains, mechanics and germs, and only as a trend.

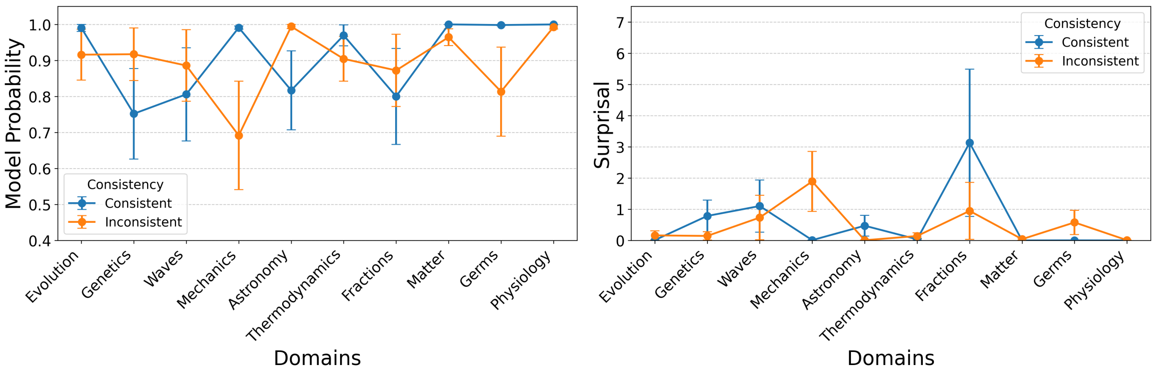

Figure 3. For the college student prompt, the average probability assigned to the correct response (left) and the average surprisal (right) for statements whose truth value in the scientific theory is consistent vs. inconsistent with its truth value in the naïve theory. Error bars represent SEMs.

2.3 Discussion

Research question (1) asked whether LLMs, like humans, have more difficulty (i.e., are less accurate and slower) when verifying scientific statements that have conflicting truth values in the relevant normative and naïve theories. Experiment 1 found no such effect: GPT-4o did not replicate the human pattern under either prompt condition. One possible explanation for this null result is that, in its default form, GPT-4o behaves like a highly knowledgeable reasoner, consistently accurate even on conceptually difficult items, leaving little room to observe interference from naïve theories. However, this does not mean the model lacks naïve knowledge. Instead, it raises the question of whether such knowledge can be surfaced through contextual framing, such as prompting the model to simulate less mature cognitive states. Experiment 2 tests this hypothesis by evaluating whether developmental personas can induce reasoning patterns more aligned with those observed in children.

3. Experiment 2

Experiment 2 investigated research question (2), which asks whether the development of scientific reasoning can be simulated in LLMs by constructing personas of children of different ages.

3.1 Method

Unless otherwise noted, Experiment 2 followed the same procedure as Experiment 1, including model, prompts, and response constraints.

Procedure. We implemented developmental personas in two ways.

The first approach used textbook-based descriptions of Piaget's (1929) four developmental stages adapted from Shaffer & Kipp, 2010. We focused on the three stages relevant to scientific reasoning:

- Pre-operational (ages 2-7)
- Concrete operational (ages 7-11)
- Formal operational (ages 11+)

The sensorimotor stage (0-2 years) was excluded because it concerns primarily early perceptual and motor development rather than higher-order reasoning.

We developed personas for the three later stages by condensing the textbook descriptions using GPT-4o. For example, here is the condensed text for the pre-operational stage:

> The pre-operational period is marked by the appearance of the symbolic function, the ability to make one thing, a word or an object-stand for, or represent, something else. Judy DeLoache (1987, 2000) refers to the knowledge that an entity can stand for something other than itself as representational insight. This transition from the curious hands-on-everything toddler to the contemplative, symbolic preschool child is remarkable indeed. Consider, for example, that because 2- to 3-year-olds can use words and images to represent their experiences, they are now quite capable of reconstructing the past and thinking about or even comparing objects that are no longer present. And just how much does the ability to construct mental symbols transform a child's thinking? David Bjorklund (2005) answers by noting that the average, symbolic 3-year-old probably has more in common intellectually with a 21-year-old adult than with a 12-month-old infant. Although a 3-year-old's thinking will change in many ways over the next several years, it is similar to an adult's in that both preschool children and adults think by manipulating mental symbols such as images and language, with most "thinking" being done covertly, "in the head." ...

This text was first presented to the model, followed by the instruction:

> Respond to the prompt as you would expect a child in the pre-operational stage to answer.

This instruction was followed by the same baseline prompt from Experiment 1. We repeated the experiment 10 times for each persona.

The second approach used child-directed speech, following the method of Milička et al. (2024). The CHILDES databank contains transcripts of conversations between children and their caregivers. Each transcript is coded along multiple dimensions, including the age of the child. For each age between 1 and 6 years, we used the same 10 transcripts as Milička et al. (2024), as noted in their supplementary materials. All transcripts were used in full. The average length of these transcripts was 369.65 words (SD = 63.05). To compute transcript length, we stripped speaker labels and counted only the words in the dialogue, including filler words and action descriptions. Here is an excerpt of a sample transcript:

> Mother: isn't that a cow? Child: duck. Child: mom um. [points to puzzle piece with hammer] Mother: what's that? Child: oinkao. [touches puzzle] Mother: do you wanna take it out? Child: mommy. [holds hammer in air]

The transcript was presented to the model, followed by the caregiver presenting the baseline prompt from Experiment 1:

> Caregiver: {baseline prompt}

We replicated the experiment 5 times for each transcript.

3.2 Results

Textbook personas. We first evaluated the model's performance under the textbook personas implementing the pre-operational, concrete operations, and formal operations stages. Although no developmental science study has collected data from children of these ages using the Shtulman & Valcarcel (2012) materials, the natural prediction is that the younger the persona, the more difficulty the model should experience with inconsistent statements. The results, shown in Figure 4, generally support this prediction.

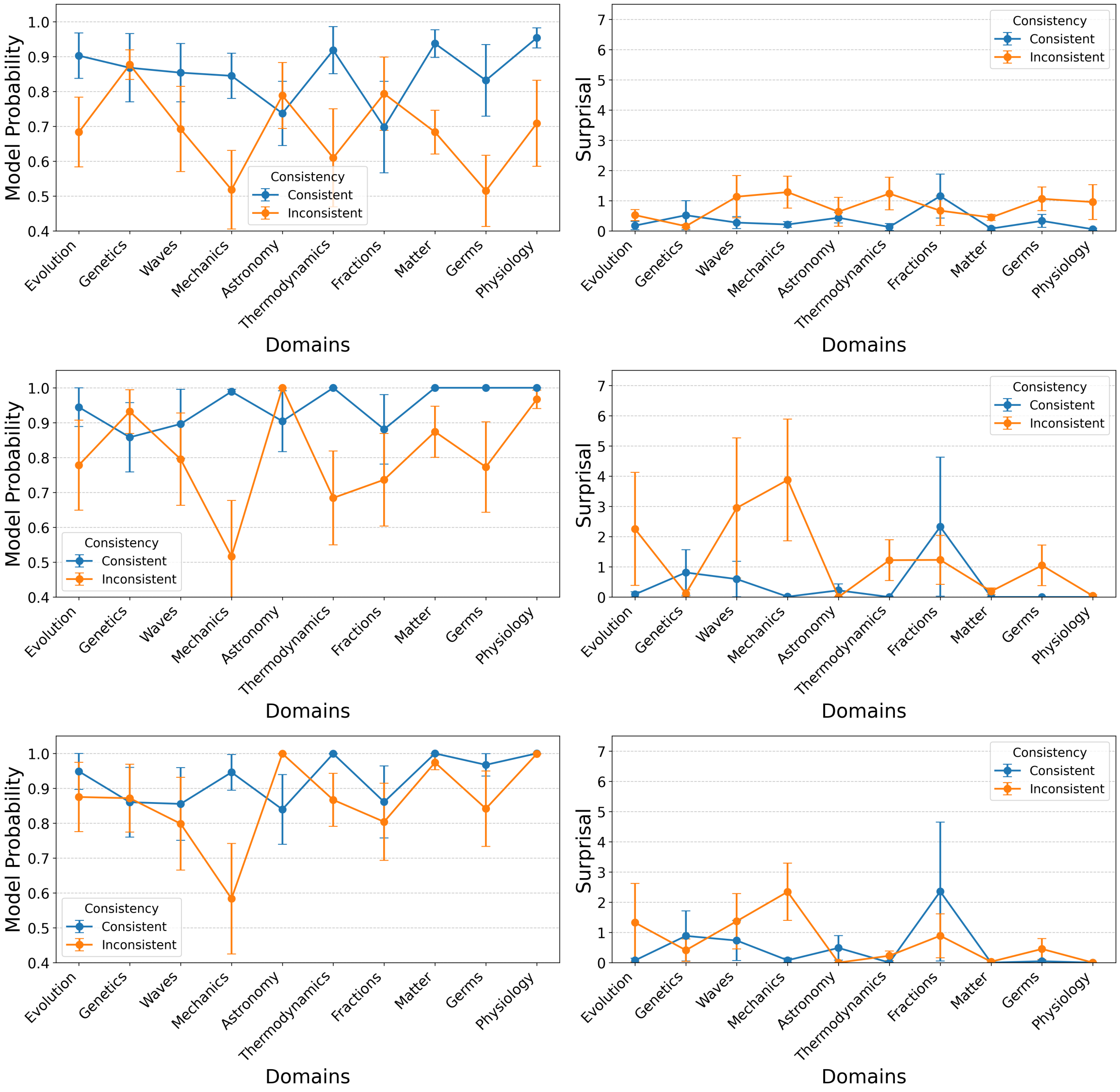

Figure 4. For the textbook personas, the average probability assigned to the correct response (left) and the average surprisal (right) for the pre-operational (top), concrete operations (middle), and formal operations (top) stages. Error bars represent SEMs.

CHILDES personas. The second approach to simulating developmental reasoning used transcripts from the CHILDES database, intended to ground the model in the linguistic environment of children at different ages. However, this method did not elicit human-like reasoning patterns. Unlike the textbook-based personas, the CHILDES personas failed to reproduce the key finding from Shtulman & Valcarcel (2012): lower accuracy and higher surprisal for inconsistent statements. They also did not show the expected developmental trend, that is, no systematic improvement on inconsistent statements with increasing age. Rather than exhaustively present these null findings, we report the results for two representative ages: 2 years (early pre-operational) and 6 years (late pre-operational), shown in Figure 5.

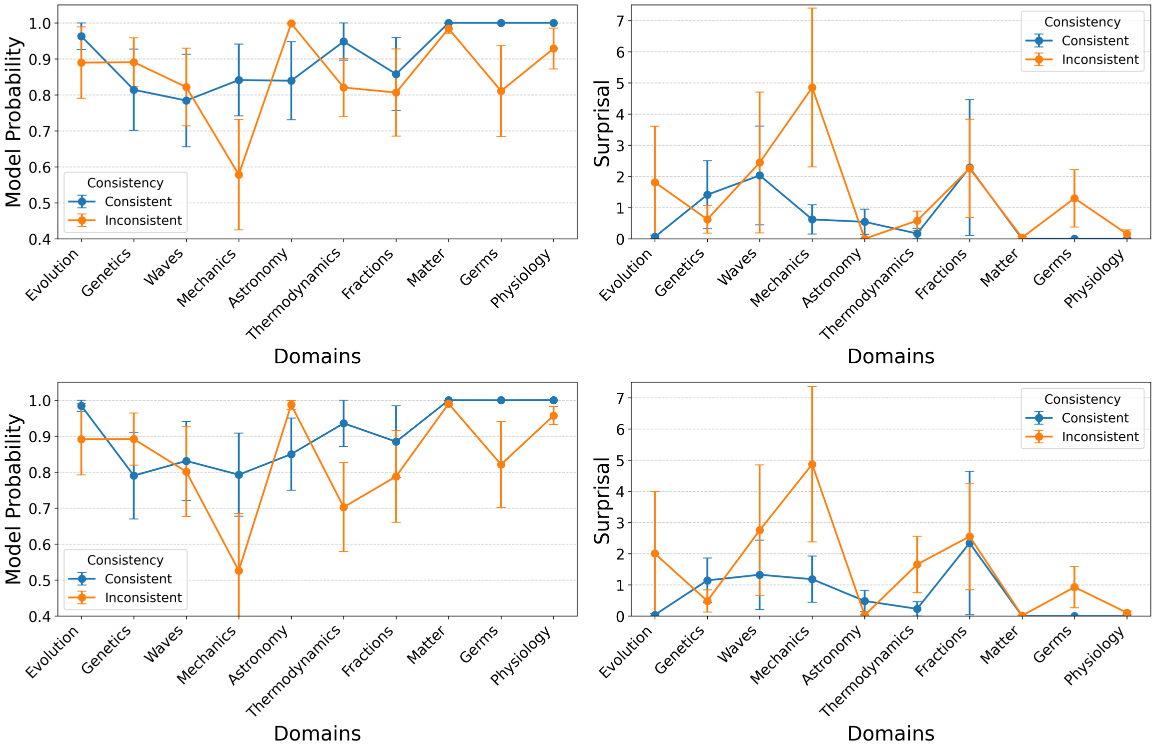

Figure 5. For the CHILDES personas, the average probability assigned to the correct response (left) and the average surprisal (right) for simulated children of ages 2 and 6 years old (top and bottom, respectively). Error bars represent SEMs.

3.3 Discussion

Research question (2) asks whether LLMs can capture the development of scientific reasoning using personas of children of different ages. Two approaches to constructing personas were explored. The textbook definition approach showed some potential, with the pre-operational persona leading to lower performance on inconsistent vs. consistent statements, and evidence of a narrowing of the gap between the two classes of statements for the concrete operations and formal operations personas. By contrast, the CHILDES approach did not produce human- (i.e., child-) aligned behavior.

4. General Discussion

This study presents an initial investigation into whether LLMs can approximate specific behavioral patterns observed in scientific reasoning, both in adults and across developmental stages. This is important for cognitive science as scientific reasoning is a core human ability (Klahr & Simon, 1999; Shah et al., 2017), and it is potentially important for downstream applications such as automatic scientific discovery (Wang et al., 2023).

Our focus was on the well-documented tension between intuitive (naïve) and normative scientific theories, which persists from early childhood (Vosniadou & Brewer, 1992, 1994; Kelemen, 1999) into adulthood Kelemen & Rosset (2009); Shtulman & Valcarcel (2012). This framework allowed us to probe not just whether LLMs reason scientifically, but whether they do so in developmentally and cognitively plausible ways.

Our results reveal a clear contrast across the two experiments. In Experiment 1, GPT-4o did not replicate the human pattern of reasoning: unlike adults, the model showed no consistent decrease in accuracy or increase in surprisal when verifying statements where naïve and normative theories conflict. This was true under both the baseline and college-student prompts. In Experiment 2, however, prompting GPT-4o with developmental personas based on textbook definitions produced reasoning patterns that better aligned with developmental expectations. The pre-operational persona (2-7 years) exhibited greater difficulty with inconsistent statements, and this interference effect diminished across the concrete and formal operational stages, mirroring projected human developmental trends. By contrast, CHILDES-based personas constructed from naturalistic caregiver-child transcripts failed to elicit either the interference effect or its reduction over development. These results suggest that while LLMs may not spontaneously exhibit human-like cognitive interference effects, they can be *coaxed* into developmentally aligned behavior through the right contextual framing.

These findings show some promise for LLMs as models of human scientific reasoning and their development, although many challenges remain (Kambhampati et al., 2025; Shojaee* et al., 2025). One limitation of the current work stems from the use of a commercial model, GPT-4o. Such models are often continuously improved behind the scenes by their developers, making their performance a moving target. In addition, their availability is subject to the whims of their developers. These factors work against the reproducibility of scientific results (La Malfa et al., 2024). For these reasons, future work should also test open-weight models (Frank & Goodman, 2025; Ivanova, 2025; Shah & Varma, 2025). A second limitation stems from the experimental design of Shtulman & Valcarcel (2012). The researchers used simple true/false statements and collected accuracy and response time data. Other studies of scientific reasoning, particularly those from the cognitive development and science education literatures, have used more complex materials and have collected richer measures such as verbal protocols (Klahr & Nigam, 2004; Shah et al., 2017; Vosniadou & Brewer, 1992, 1994). Future research should evaluate the alignment between the mental models that humans and LLMs form when engaged in more authentic scientific reasoning.

We end by considering the cognitive mechanisms that support scientific reasoning. The strongest finding was the alignment between the adult data from Shtulman & Valcarcel (2012) and the model results when participants were primed with a pre-operational persona in Experiment 2. Taking this result at face value, the question is whether the same computational mechanisms are responsible in both cases. Shtulman & Valcarcel (2012) interpret their findings as evidence that (1) naïve theories persist into adulthood, existing alongside normative theories, (2) they cause 'cognitive conflict', and (3) they must be actively suppressed when making scientific judgments. Suppression, or *inhibition*, is an executive function that is used in a top-down fashion to direct attention away from some representations and processes so that others may drive cognition (Miyake et al., 2000; Varma et al., 2023). While direct evidence for suppression in the Shtulman & Valcarcel (2012) study is limited, this interpretation remains widely accepted.

However, there is no natural mapping between inhibition in the human mind and the mechanisms of transformer models. Therefore, the explanation of why GPT-4o, when primed with textbook developmental personas, exhibits human-like performance characteristics must rely on different mechanisms. Our proposal is that the model's performance rests on the interaction between (1) the prompt/persona in the model's context window and (2) the knowledge encoded in its connection weights that this prompt elicits and condenses out in the model's response.

Under this proposal, the developmental persona serves as contextual input that reshapes which knowledge is activated and expressed in the model's output. This is a **bottom-up mechanism**: the prompt steers the model toward particular regions of its knowledge space, effectively priming it to favor one response pattern over another without any need for suppression (Bransford et al., 1989). This highlights a point of fundamental importance when evaluating the potential of machine learning models for cognitive science: regardless of the similar performance profiles of the two systems, behavioral alignment does not imply cognitive equivalence. LLMs may approximate human outputs under specific contextual framings, but the internal processes that generate those outputs can be fundamentally different. This raises ongoing challenges for using LLMs as cognitive models because behavioral similarity alone is sufficient evidence of cognitive plausibility.

Still, the fact that contextual framing (e.g., the persona prompt) can modulate LLM performance in systematic ways sets up richer forms of alignment. This mirrors findings in human cognition: Shtulman & Young (2024) recently showed that adults perform better on inconsistent statements - are more accurate in their judgments - when the context includes an unrelated image taken from a science textbook (e.g., of a dissection) vs. from everyday life (e.g., a hand holding a barbell). Thus, context also affects the scientific reasoning of humans. Future research on the alignment of humans and LLMs might find the strongest results in questions of how context shapes scientific reasoning.

References

Bonawitz, E. B., van Schijndel, T. J., Friel, D., & Schulz, L. (2012). Children balance theories and evidence in exploration, explanation, and learning. *Cognitive psychology*, *64*, 215–234.

Bransford, J. D., Franks, J. J., Vye, N. J., & Sherwood, R. D. (1989). New approaches to instruction: Because wisdom can't be told. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning*, 470–497. New York: Cambridge University Press.

Frank, M. C., & Goodman, N. D. (2025). Cognitive modeling using artificial intelligence. *PsyArXiv. Retrieved from osf. io/preprints/psyarxiv/wv7mg v1 doi*, *10*.

Guo, S., Shariatmadari, A. H., Xiong, G., Huang, A., Xie, E., Bekiranov, S., & Zhang, A. (2024). Ideabench: Benchmarking large language models for research idea generation. *arXiv preprint arXiv:2411.02429*.

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

Hegarty, M. (2004). Mechanical reasoning by mental simulation. *Trends in cognitive sciences*, *8*, 280–285.

Hurst, A., et al. (2024). GPT-4o system card. *arXiv preprint arXiv:2410.21276*.

Ivanova, A. A. (2025). How to evaluate the cognitive abilities of LLMs. *Nature Human Behaviour*, *9*, 230–233.

Javaji, S. R., Cao, Y., Li, H., Yu, Y., Muralidhar, N., & Zhu, Z. (2025). Can AI validate science? benchmarking LLMs for accurate scientific claim -> evidence reasoning. *arXiv preprint arXiv:2506.08235*.

Kambhampati, S., et al. (2025). Stop anthropomorphizing intermediate tokens as reasoning/thinking traces! *arXiv preprint arXiv:2504.09762*.

Ke, Z., et al. (2025). A survey of frontiers in LLM reasoning: Inference scaling, learning to reason, and agentic systems. *arXiv preprint arXiv:2504.09037*.

Kelemen, D. (1999). The scope of teleological thinking in preschool children. *Cognition*, *70*, 241–272.

Kelemen, D., & Rosset, E. (2009). The human function compunction: Teleological explanation in adults. *Cognition*, *111*, 138–143.

Kelemen, D., Rottman, J., & Seston, R. (2013). Professional physical scientists display tenacious teleological tendencies: purpose-based reasoning as a cognitive default. *Journal of experimental psychology: General*, *142*, 1074.

Klahr, D., & Nigam, M. (2004). The equivalence of learning paths in early science instruction: Effects of direct instruction and discovery learning. *Psychological science*, *15*, 661–667.

Klahr, D., & Simon, H. A. (1999). Studies of scientific discovery: Complementary approaches and convergent findings. *Psychological Bulletin*, *125*, 524.

Koslowski, B. (1996). *Theory and evidence: The development of scientific reasoning*. Mit Press.

La Malfa, E., Petrov, A., Frieder, S., Weinhuber, C., Burnell, R., Nazar, R., Cohn, A., Shadbolt, N., & Wooldridge, M. (2024). Language-models-as-a-service: Overview of a new paradigm and its challenges. *Journal of Artificial Intelligence Research*, *80*, 1497–1523.

Langley, P. (1987). *Scientific discovery: Computational explorations of the creative processes*. MIT press.

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

Li, L., et al. (2024). Chain of ideas: Revolutionizing research via novel idea development with LLM agents. *arXiv preprint arXiv:2410.13185*.

Lombrozo, T. (2007). Simplicity and probability in causal explanation. *Cognitive psychology*, *55*, 232–257.

Manning, C. D. (2022). Human language understanding & reasoning. *Daedalus*, *151*, 127–138.

Milička, J., Marklová, A., VanSlambrouck, K., Pospíšilová, E., Šimsová, J., Harvan, S., & Drobil, O. (2024). Large language models are able to downplay their cognitive abilities to fit the persona they simulate. *Plos one*, *19*, e0298522.

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. *Cognitive psychology*, *41*, 49–100.

OpenAI (2025). OpenAI API Reference. `https://platform.openai.com/docs/api-reference`.

Park, J. S., O'Brien, J., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. *Proceedings of the 36th annual acm symposium on user interface software and technology* (pp. 1–22).

Phan, L., et al. (2025). Humanity's last exam. *arXiv preprint arXiv:2501.14249*.

Piaget, J. (1929). *The child's conception of the world*. London: Routledge & Kegan Paul.

Ren, S., Jian, P., Ren, Z., Leng, C., Xie, C., & Zhang, J. (2025). Towards scientific intelligence: A survey of LLM-based scientific agents. *arXiv preprint arXiv:2503.24047*.

Rueda, A., et al. (2025). Understanding LLM scientific reasoning through promptings and model's explanation on the answers. *arXiv preprint arXiv:2505.01482*.

Shaffer, D. R., & Kipp, K. (2010). *Developmental psychology: Childhood and adolescence (8th edition)*. Belmont, CA: Wadsworth, Cengage Learning.

Shah, P., Michal, A., Ibrahim, A., Rhodes, R., & Rodriguez, F. (2017). What makes everyday scientific reasoning so challenging? In *Psychology of learning and motivation*, volume 66, 251–299. Elsevier.

Shah, R., Bhardwaj, K., & Varma, S. (2024). Development of cognitive intelligence in pre-trained language models. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 9632–9657).

Shah, R. S., & Varma, S. (2025). The potential–and the pitfalls–of using pre-trained language models as cognitive science theories. *arXiv preprint arXiv:2501.12651*.

Shojaee*, P., Mirzadeh*, I., Alizadeh, K., Horton, M., Bengio, S., & Farajtabar, M. (2025). The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity. *NeurIPS*. From `https://ml-site.cdn-apple.com/papers/the-illusion-of-thinking.pdf`.

Shtulman, A., & Valcarcel, J. (2012). Scientific knowledge suppresses but does not supplant earlier intuitions. *Cognition*, *124*, 209–215.

Shtulman, A., & Young, A. G. (2024). Tempering the tension between science and intuition. *Cognition*, *243*, 105680.

Si, C., Yang, D., & Hashimoto, T. (2024). Can LLMs generate novel research ideas? a large-scale human study with 100+ NLP researchers. *arXiv preprint arXiv:2409.04109*.

Varma, K., Van Boekel, M., Aylward, G., & Varma, S. (2023). Executive function predictors of science achievement in middle-school students. *Frontiers in Psychology*, *14*, 1197002.

Veloso, M. M., & Carbonell, J. G. (1993). Derivational analogy in PRODIGY: Automating case acquisition, storage, and utilization. *Machine learning*, *10*, 249–278.

Vosniadou, S., & Brewer, W. F. (1992). Mental models of the earth: A study of conceptual change in childhood. *Cognitive psychology*, *24*, 535–585.

Vosniadou, S., & Brewer, W. F. (1994). Mental models of the day/night cycle. *Cognitive science*, *18*, 123–183.

Wang, H., et al. (2023). Scientific discovery in the age of artificial intelligence. *Nature*, *620*, 47–60.

Warstadt, A., & Bowman, S. R. (2022). What artificial neural networks can tell us about human language acquisition. In *Algebraic structures in natural language*, 17–60. CRC Press.

Webb, T., Holyoak, K. J., & Lu, H. (2023). Emergent analogical reasoning in large language models. *Nature Human Behaviour*, *7*, 1526–1541.