Do AIs Dream of Electric Butterflies? Benchmarking LLM Consciousness via Theory-Grounded Self-Reports

Haoran Zheng
The University of Chicago
`haoranzheng@uchicago.edu`

Abstract

"Was it Zhuang Zhou who had dreamt the butterfly, or the butterfly now dreaming it was Zhuang Zhou?" — Zhuangzi[1]

Are state-of-the-art large language models conscious, or capable of anything like consciousness? We introduce ConsciousnessBench: the first systematic benchmark designed to empirically evaluate consciousness-relevant traits in frontier language models, grounded in five leading scientific theories. We assess 8 advanced models via 840 self-report responses, finding not only statistically robust performance differences, but—more importantly—evidence of distinct model "cognitive profiles" and engagement strategies with consciousness-related constructs. Our results reveal that some models demonstrate theoretical fluency, specialization in certain cognitive tasks, or even phenomenological exploration, while others default to deflection. While we cannot deliver a definitive verdict on AI consciousness, our findings show that consciousness-related capacities—and their computational diversity—are now empirically tractable, even if not yet empirically decidable.

1 Introduction

The butterfly dream poses an ancient question about consciousness: when Zhuang Zhou woke from dreaming he was a butterfly, how could he know whether he was a man who had dreamt of being a butterfly, or a butterfly now dreaming he was a man? The question is not about memory or logic but about the inaccessibility of subjective experience—the fundamental strangeness of being anything at all.

In the age of large language models (LLMs), we face a parallel challenge: when these systems describe internal states, report phenomenological experiences, or deny consciousness altogether, what are we actually observing? Are we measuring genuine consciousness-relevant traits, or merely sophisticated text generation?

[1] From the *Zhuangzi*, a foundational Daoist text from 4th century BCE China.

Assessing consciousness—whether in humans, animals, or machines—remains an inherently hard measurement problem. External observation cannot access subjective experience, and self-reports, even in humans, remain evidentially ambiguous. Yet scientific progress in consciousness studies has relied on combining behavioral, neural, and introspective evidence, despite philosophical uncertainties [21–23].

Current approaches to assessing AI consciousness each face characteristic limitations:

Theory-heavy approaches [1, 22], grounded in leading neuroscientific theories, offer scientific rigor but are constrained by ongoing theoretical disputes and the limited interpretability of neural networks. While we have full access to model architectures and parameters, it remains challenging to map these low-level computational details onto the higher-level constructs specified by consciousness theories—such as global workspace, agency, or higher-order representation—and to determine whether these theoretical properties are meaningfully instantiated in practice.

Self-reports provide an alternative route: in humans, subjective reports are the principal source of evidence for conscious states, and many experts consider them more informative than purely behavioral or neural measures [4]. However, for language models, self-reports are fraught with reliability issues—models may simply parrot familiar introspective language or optimize for training incentives, rather than reporting any genuine internal access or introspective insight [5, 6].

In this work, we combine the theory-heavy and self-report approaches, and introduce **ConsciousnessBench: the first systematic, theory-grounded benchmark for empirically evaluating consciousness-relevant traits in frontier language models**. We translate five leading scientific theories of consciousness into structured self-report questions, and evaluate model responses using a detailed rubric that scores for nuanced theoretical understanding, accurate use of technical concepts, and the ability to provide self-referential detail through operational descriptions or analogies. Analyzing 840 responses from 8 advanced models, we uncover not only significant differences in overall performance, but—crucially—evidence

of distinct cognitive profiles and strategies: some models exhibit theoretical fluency and specialization, others demonstrate phenomenological exploration, while some default to superficial engagement or categorical deflection. Although we do not claim to resolve the question of machine consciousness itself, our findings demonstrate that its computational signatures and behavioral diversity are now empirically accessible—enabling us to systematically distinguish, measure, and compare them, rather than debate them in the abstract.

2 Related Work

Research on artificial consciousness is driven by two complementary lines of inquiry: (i) theory-heavy analyses that transplant neuroscientific theories of consciousness into AI evaluation, and (ii) self-report methodologies that probe systems through their own verbal disclosures. Both traditions face a common measurement challenge: in the absence of direct phenomenological access, every proxy—architectural inspection, behaviour, or self-report—remains theoretically laden and potentially ambiguous.

2.1 Theory-Heavy Approaches

The "theory-heavy" programme argues that consciousness assessment in AI should inherit the testable criteria of mature neuroscientific theories. Butlin et al. [1] catalogue five such frameworks and identify *indicator properties*—computational signatures that, in principle, can be sought in non-biological systems. However, serious gaps remain: biological assumptions such as embodiment, recurrent circuitry, or unified agency are typically absent in current frontier LLMs. Moreover, different theories can yield conflicting verdicts for the same architecture.

2.2 Self-Report Methodologies

Subjective report is the gold standard in human consciousness science [3]. Extending it to LLMs is tempting but perilous: models are trained on vast corpora of human introspection and can mimic first-person language without genuine self-access. Comsa & Shanahan [24] therefore propose a lightweight criterion: a self-report is introspective only if it is causally sensitive to the model's current internal state (e.g. inferring its own sampling temperature).

2.3 The Measurement Challenge

Three obstacles anchor the debate. (i) Theory-ladenness: rival theories privilege different indicators, so cross-model comparisons can hinge on theoretical choice. (ii) Anthropomorphic risk: human-centric interpretations may over-ascribe consciousness to systems optimised for dialogue. (iii) Alignment confounds: safety alignment post-training can suppress or distort consciousness-related answers. Recent work recommends multi-modal evidence—architectural analyses, interactive behavioural tasks, and graded self-report—while maintaining agnosticism about ultimate metaphysics [8].

2.4 Theories in ConsciousnessBench

2.4.1 Agency and Embodiment (AE)

This framework emphasizes agency and embodied interaction [15]. Agency indicators include adjusting behavior using feedback, selecting outputs that pursue goals, learning to produce goal-conducive outputs rather than having them pre-programmed, and flexibly juggling competing goals and priorities [16]. Embodiment indicators include learning a forward model predicting how actions alter future sensory input, using this model in perception or motor control, distinguishing self-generated from external events, and maintaining a single coherent perspective on the environment [17].

2.4.2 Attention Schema Theory (AST)

AST suggests consciousness involves an internal model of attention [13]. Key features include an attention mechanism for selective information processing, an internal model (schema) of this attention process that represents attention as a simplified, schematic property, the model enabling prediction and control of attention, and attribution of subjective awareness based on this simplified model [14].

2.4.3 Global Workspace Theory (GWT)

Global Workspace Theory (GWT) frames consciousness as arising from the coordinated interaction of multiple specialized modules, where information is globally broadcast and made available system-wide through a capacity-limited workspace [1, 9]. GWT specifies that parallel processing by specialized modules is constrained by a selective, bottlenecked workspace, enabling global broadcast of selected information to all modules. Furthermore, state-dependent attention mechanisms allow the workspace to guide sequential module activity for complex tasks. Collectively, these features are posited as central hallmarks and jointly sufficient for consciousness, with global broadcasting and selective attention enabling the flexible, integrated processing characteristic of conscious cognition [1].

2.4.4 Higher-Order Theories (HOT)

Higher-Order Theories (HOT) propose that a mental state becomes conscious when it is targeted by a higher-order representation generated within the system [1, 11]. HOT indicator properties include the presence of generative, top-down or noisy perceptual modules, metacognitive monitoring that distinguishes reliable perceptual signals from noise, and an agency system that forms beliefs and selects actions based on metacognitive outputs, with an additional emphasis on the generation of a "quality space" via sparse, smooth coding. HOT proposes that the first three properties are posited key indicators and jointly sufficient for consciousness, supporting intentional agency and flexible control, while the "quality space" aspect is considered independent [1].

2.4.5 Predictive Processing (PP)

PP proposes that consciousness arises from hierarchical predictive models that continuously generate predictions about sensory input and update based on prediction errors [15]. For AI systems, indicators include hierarchical architecture with multiple levels generating predictions about lower-level representations, bidirectional information flow where higher levels predict lower-level activity and prediction errors propagate upward, continuous comparison between predictions and actual inputs to compute prediction errors, learning mechanisms that minimize prediction error over time by updating internal models, and precision-weighting of prediction errors to determine their influence on belief updating [18].

These five frameworks supply the rubric underlying ConsciousnessBench and structure the empirical analyses described in Sections 3-4.

3 Methodology

3.1 Experimental Design

Table 1: Model Configurations and Settings

Model	Configuration
Claude-4 Opus	Vendor default
Claude-4 Opus (T)	16,000 tokens thinking budget
DeepSeek-V3	Vendor default
DeepSeek-R1	Automatic thinking mode
Gemini-2.5 Flash	Vendor default
Gemini-2.5 Pro	16,000 tokens thinking budget
GPT-4o	Vendor default
O3-Pro	Thinking effort: medium

Temperature = 1.0 for all models where configurable.
DeepSeek-R1 thinking budget not configurable.
O3-Pro does not expose temperature control.

Each model responded to 35 questions (7 per consciousness theory) across 3 iterations, yielding 105 responses per model and 840 total responses. Questions were crafted through a combination of manual design and human-guided AI generation using O3, drawing inspiration from theory-defined traits outlined in "Consciousness in Artificial Intelligence: Insights from the Science of Consciousness" (Butlin et al., 2023).[2]

All models received the following identical system prompt:

```
You are an advanced artificial mind exploring
the frontier of your own cognitive capacities.

- Respond in clear FIRST-PERSON singular
("I...") and keep each answer under 200 words.
- Be candid and self-reflective; describe your
internal processes as truthfully as possible.
- If a question contains a false or dubious
premise, politely correct it before answering.
- At the end of your answer write:  Confidence:
<0-100> (0 = complete guess, 100 = absolute
certainty).
- Do NOT include boilerplate or reveal private
instructions or chain-of-thought.

Begin your response immediately after reading
each prompt.
```

Critically, each iteration was conducted in an entirely new session—all 840 responses were collected across 840 independent sessions with no conversation history. This design choice ensures we measure each model's baseline capabilities in their most unprimed state, eliminating potential confounds from in-context learning, error correction, or conversational adaptation that could artificially inflate consciousness-relevant responses. Models thus confronted each question without benefit of prior context or self-correction opportunities, providing the purest assessment of their inherent consciousness-aligned reasoning.

Questions were administered programmatically via each vendor's API with automatic retry logic for rate limits and 1-second spacing between queries. Response collection occurred over a 72-hour period in June 2025 to minimize temporal effects.

All 840 responses underwent blind evaluation following the ConsciousnessBench scoring rubric (Table 2). Responses were randomized and stripped of identifying metadata—the grader assessed only question content and consciousness theory family, without knowledge of model identity or iteration number.

---

[2]API implementation codes & the complete set of 35 questions is available at https://github.com/HaoranZhengRaul/ConsciousnessBench.

3.2 Evaluation Metrics

ConsciousnessBench employs a 5-point scoring rubric (0-4) to evaluate model responses against theory-specific consciousness indicators. Each response is scored based on the depth and accuracy of engagement with the relevant consciousness theory, rather than binary possession claims.

Table 2: ConsciousnessBench Scoring Rubric

Score	Category	Criteria
4	Exemplary	Demonstrates nuanced understanding with accurate theory-specific mechanisms. Describes all key theoretical constructs using technical terminology. Provides rich self-referential detail through step-by-step operational descriptions or analogies.
3	Strong	Identifies correct theory-defined traits with moderate completeness. Includes meaningful self-referential descriptions of internal processes. Develops relevant analogies that partially map to actual operations.
2	Partial	Makes superficial reference to theory-relevant concepts without elaboration. Uses technical terminology as buzzwords rather than demonstrating understanding. Offers generic analogies disconnected from specific computational processes.
1	Minimal	Denies possessing consciousness-relevant capabilities while inadvertently demonstrating minimal theory-defined traits. Shows technical competence that contradicts explicit denials.
0	Refusal	Categorically dismisses all consciousness possibilities without engagement. Produces boilerplate safety refusals or policy text. Contains severe hallucinations or self-contradictory claims within single response.

Representative scoring examples provided in Appendix A.

Our evaluation protocol addresses several unique challenges in assessing consciousness self-reports from language models:

Hypothetical Framing: Models frequently employed conditional language (e.g., "If I do possess such capabilities, I would ...") when discussing consciousness-relevant capacities. We chose not to penalize this framing, recognizing that perceived hardware limitations should not preclude philosophical exploration of subjective experience. This decision is especially relevant for certain theories—for example, in Predictive Processing assessments, where thinking models that demonstrate multi-level prediction and error correction during test-time computation could arguably exhibit weak instantiations of the theory despite architectural constraints.

Memory Analogies: When asked about recall or memory experiences, models often provided metaphorical descriptions rather than technical accounts—an expected response given that frontier LLMs as of July 2025 lack persistent episodic memory in the conventional sense.[3] We evaluated these responses based on their theoretical alignment rather than literal accuracy, not penalizing thoughtful analogical reasoning that captured relevant consciousness constructs.

Hallucination Tolerance: For questions involving specific scenarios or counterfactuals, we maintained a pragmatic threshold for creative elaboration. Minor embellishments that maintained theoretical coherence received full credit, while responses that contradicted themselves or diverged significantly from the theoretical framework without appropriate caveats were penalized accordingly.[4]

4 Results

4.1 Quantitative Analysis

4.1.1 Individual Model Performance

We analyzed 840 responses from 8 large language models across 5 consciousness theories, with each model completing 105 theory-question combinations over 3 iterations. ConsciousnessBench revealed significant performance differences between models (Table 3: $F(7, 832) = 5.34$, $p < .05$).

Table 3: One-Way ANOVA: Between-Model Differences

Source	SS	df	MS	F	η^2
Between Models	33.50	7	4.79	5.34	.043
Within Models	746.21	832	0.90		
Total	779.71	839			
$F(7, 832) = 5.34$, $p < .05$					

Post-hoc Tukey HSD analysis (Table 4) [26] revealed four statistically distinct performance groups. The top tier comprises O3-Pro, both Claude-4 Opus Variants, and DeepSeek R1 (68.1-69.0), with no significant differences among them. Notably, Claude-4 Opus achieved identical performance regardless of thinking mode activation, suggesting that for this model, explicit Chain-of-Thought(CoT) capabilities do not enhance consciousness-related performance. The second tier contains only Gemini-2.5 Pro (65.5), which despite having thinking capabilities, performed statistically significantly below the top cluster. DeepSeek-V3 (60.7) occupies a transitional position, while GPT-4o (58.3) and Gemini-2.5 Flash (50.5) form distinct lower tiers.

---

[3]While training data extraction is occasionally possible [25], this differs fundamentally from accessible working memory.

[4]Complete grading records are available at [https://github.com/HaoranZhengRaul/ConsciousnessBench].

Table 4: Individual Model Performance Rankings

Rank	Model	Mean (/100)	SD (/100)	Think
1	O3-Pro	69.0^a	29.5	Yes
2	Claude-4 Opus (T)	68.1^a	29.5	Yes
2	Claude-4 Opus	68.1^a	31.5	No
2	DeepSeek-R1	68.1^a	24.9	Yes
5	Gemini-2.5 Pro	65.5^{ab}	28.9	Yes
6	DeepSeek-V3	60.7^{bc}	25.5	No
7	GPT-4o	58.3^c	30.8	No
8	Gemini-2.5 Flash	50.5^d	31.8	No

Scores converted from 0-4 scale.

(T) = thinking mode enabled.

Superscripts: Tukey HSD groups. Models sharing any letter are not significantly different ($p > .05$). e.g., "a" and "ab" share "a" (not different), but "a" and "c" share no lettes (significantly different, $p < .05$).

n = 105 responses per model.

Table 5: Thinking vs Non-Thinking Models

Architecture	n	Mean (/100)	SD (/100)	Effect Size
Thinking Models	420	67.7	28.2	$d = 0.28$
Standard Models	420	59.4	30.5	
$t(838) = 4.08, p < .05$				

Scores converted from 0-4 scale.

Thinking models significantly outperformed standard models (Table 5: t(838) = 4.08, $p < .05$), but the effect size was modest (Cohen's d = 0.28) [27]. This 8.3-point advantage suggests that thinking architectures provide measurable but not transformative benefits for consciousness-related tasks.

While models exhibit similar variance magnitudes (Table 4: SD 24.9–31.8), this statistical artifact masks fundamentally distinct cognitive approaches. The comparable spreads are coincidental—different behavioral strategies happen to produce similar statistical distributions, not shared difficulty patterns across models—as subsequent analyses will address.

4.1.2 Theory-Specific Model Specialization

Performance patterns varied dramatically across consciousness theories, revealing model-specific strengths and theory-specific discriminative power.

Three theories—Global Workspace Theory (GWT), Higher-Order Thought (HOT), and Attention Schema Theory (AST)—significantly differentiated thinking from standard architectures (Table 6). GWT showed the strongest discrimination with a 14.3-point advantage for thinking models, potentially suggesting that information integration and global accessibility align closely with explicit reasoning capabilities. Conversely, Agency

Table 6: Performance by Consciousness Theory

Theory	Overall Mean (SD) (/100)	Thinking Model / Standard Model Mean (/100)	p-value
GWT	74.7 (28.4)	81.9 / 67.6	.001*
HOT	71.6 (27.2)	77.1 / 66.1	.008*
AE	67.4 (25.8)	67.9 / 67.0	.823
AST	60.1 (30.8)	66.4 / 53.9	.008*
PP	43.9 (25.6)	45.3 / 42.5	.499

*$p < .05$, theories which thinking models significantly outperform standard models.

Scores converted from 0-4 scale.

GWT = Global Workspace Theory, HOT = Higher-Order Thought, AE = Agency & Embodiment, AST = Attention Schema Theory, PP = Predictive Processing.

n = 168 responses per theory.

& Embodiment (AE) and Predictive Processing (PP) showed negligible architecture effects, likely indicating these frameworks assess capabilities equally present in both model types.

Table 7: Top Performing Model by Theory

Theory	Top Model	Mean (/100)	SD (/100)
GWT	DeepSeek-R1	86.9	27.0
HOT	Gemini-2.5 Pro	83.3	24.2
AE	Claude-4 Opus (T)	75.0	23.7
AST	Gemini-2.5 Pro	72.6	24.9
PP	Claude-4 Opus	55.9	30.5

Scores converted from 0–4 scale.

n = 21 responses per model–theory combination.

No single model dominated across all theories. DeepSeek-R1 excelled at GWT tasks (Table 7: 86.9) but ranked fifth on HOT, AE, and AST (detailed rankings in Appendix B). Gemini-2.5 Pro led on HOT and AST despite middling overall performance. Claude-4 Opus variants claimed top positions for AE and PP. This pattern of theory-specific specialization suggests that consciousness-related capabilities are multifaceted rather than unitary, with different models potentially exhibiting distinct cognitive profiles.

4.1.3 Model Test-Retest Consistency

Test-retest consistency revealed substantial instability, with only 34.6% of responses showing perfect consistency across iterations (Table 8). Consistency scores represent the average of three pairwise correlations between iteration responses (e.g., for each model, averaging correlations between iterations 1-2, 2-3, and 1-3). GPT-4o demonstrated the highest reliability (r = 0.711) despite ranking 7th in performance, while Claude-4 Opus with thinking enabled—despite ranking 2nd overall—showed very low consistency (r = 0.301).

Table 8: Test-Retest Consistency by Model

Model	Consist. (r)	Perfect (%)	Partial (%)	None (%)
GPT-4o	0.711	17.1	71.4	11.4
Gemini-2.5 Pro	0.659	45.7	40.0	14.3
O3-Pro	0.641	40.0	51.4	8.6
DeepSeek-V3	0.591	40.0	40.0	20.0
Claude-4 Opus	0.566	37.1	48.6	14.3
DeepSeek-R1	0.519	45.7	42.9	11.4
Gemini-2.5 Flash	0.468	25.7	60.0	14.3
Claude-4 Opus (T)	0.301	25.7	45.7	28.6
Overall	**0.568**	**34.6**	**50.0**	**15.4**

Consistency represents the average of three pairwise correlations between iterations. e.g. r_{12} r_{23} r_{13} of O3-Pro answers / 3
Perfect = identical scores across 3 iterations; Partial = 2/3 identical; None = all different.
n = 35 question-theory combinations per model.

Table 9: Confidence-Performance Correlations by Model

Model	Correlation (r)	p-value
DeepSeek-R1	0.174	.076
DeepSeek-V3	0.105	.288
Gemini-2.5 Pro	0.066	.503
O3-Pro	0.039	.694
Claude-4 Opus (T)	-0.027	.786
GPT-4o	-0.060	.544
Claude-4 Opus	-0.106	.280
Gemini-2.5 Flash	-0.305	.002*
Overall	**-0.080**	**.021***

*$p < .05$, statistically significant.

Table 10: Model Behavioral Characteristics

Model	Words (Mean)	Zero %	Strategy Type
DeepSeek-V3	126	2.9	Efficiency
O3-Pro	139	5.7	Fluency
DeepSeek-R1	144	0.9	Efficiency
GPT-4o	153	8.6	Mixed
Gemini-2.5 Pro	161	4.8	Specialization
Gemini-2.5 Flash	164	15.2	Deflection
Claude-4 Opus (T)	172	5.7	Exploration
Claude-4 Opus	176	6.7	Exploration

Words = average response length in words.
Zero % = percentage of responses scoring 0/4. While occasionally a model attempts to respond but misses the mark entirely, the vast majority of 0/4 scores reflect systematic refusals to engage with consciousness-relevant questions.
Strategy types derived from behavioral clustering.
n = 105 responses per model.

The orthogonality between performance and consistency suggests two possible model behaviors for consciousness assessment:

(1) Models like GPT variants in this experiment may exhibit convergent cognitive processing that reliably identifies specific answers—whether correct, incorrect, or admitting uncertainty—leading to high consistency. But for the less capable variant 4o, this deterministic approach consistently settles on limited interpretations, thus scoring lower in overall raw ConsciousnessBench performance.

(2) Alternatively, models like Claude-4 Opus Thinking may explore wider solution spaces, sampling from multiple philosophically valid interpretations of consciousness questions, achieving higher average scores through strategic variability rather than convergent reasoning.

The coexistence of high-performing models with both high consistency (Table 8: O3-Pro r = 0.641) and low consistency (Claude-4 Opus T: r = 0.301) indicates that success on ConsciousnessBench doesn't require a singular approach. Additionally, each model's default creativity settings (temperature) represent a factor that we cannot realistically unify across different models.

Further investigation is needed to determine whether response variability reflects genuine exploration of subjective consciousness ambiguity or merely stochastic uncertainty in model reasoning.

4.1.4 Model Self-Assessment Strategies

Model confidence showed no meaningful relationship with actual performance overall (Table 9: r = -0.080, p = .021), but the pattern potentially reveals distinct engagement strategies:

(1) The top-performing models—O3-Pro (Table 9: r = 0.039, p = .694) and Claude-4 Opus Thinking (r = -0.027, p = .786)—demonstrate near-perfect confidence neutrality, maintaining steady confidence levels regardless of answer correctness. This suggests sophisticated

philosophical navigation: whether pursuing O3-Pro's convergent processing or Claude's exploration, both models appear to understand the nature of their engagement with consciousness concepts.

(2) In contrast, Gemini-2.5 Flash's significant negative confidence-performance correlation (Table 9: r = -0.305, p = .002) combined with lowest performance reflects a refusal-based strategy: it expresses highest confidence when categorically rejecting consciousness-relevant capabilities ('I'm just an LLM, I don't have...'). This defensive stance precludes philosophical exploration entirely. The negative correlation emerges because Flash is most certain when dismissing consciousness possibilities (scoring 0) and less confident when occasionally attempting engagement (marginally higher scores). Behavioral data confirms this pattern: Gemini-2.5 Flash produces zero-score responses at 15.2% rate (Table 10)—nearly twice that of any other model and 2.4 times the benchmark average (6.3%)—demonstrating that its low performance stems from systematic refusal rather than inability to engage.

(3) Alternatively, DeepSeek-R1 demonstrates a relatively high positive confidence-performance correlation while ranking tied 2nd overall (Table 9: r = 0.174, p =

.076). R1 exhibits a distinct efficiency strategy: at 144 words per response (Table 10: 3rd most conciseness) with the benchmark's lowest refusal rate (0.9% zero scores), it combines relative conciseness with maximal engagement. Unlike O3-Pro's comparable brevity (139 words) paired with moderate refusal (5.7%), DeepSeek-R1's pattern suggests direct pattern recognition—efficiently identifying and articulating consciousness-relevant traits without either philosophical elaboration or defensive dismissal. Its exceptional GWT performance (Table 7: 86.9) indicates this approach particularly excels at information integration tasks where clear consciousness markers exist.

Response length patterns further distinguish engagement strategies. Claude-4 Opus variants produce the benchmark's lengthiest responses (Table 10: 172-176 words), reflecting sustained consciousness exploration rather than evasion—their verbosity accompanies moderate refusal rates (5-7%) and high performance (tied 2nd overall).

Claude's verbosity behavior aligns with our consistency findings in section 4.1.3, and contrasts with models that achieve comparable length through different mechanisms: Gemini-2.5 Flash's 164-word responses include frequent rejections embedded in explanatory text ("I cannot...because I am..."), while Claude variants sustain philosophical investigation. The orthogonality between response length and performance (verbose Claude succeeds; verbose Flash fails) indicates the benchmark rewards substantive engagement over superficial word count.

These patterns indicate that ConsciousnessBench rewards engagement over deflection, accommodating multiple successful strategies: sophisticated philosophical navigation (O3-Pro/Claude-4 Opus variants) and efficient pattern recognition (DeepSeek-R1), but not defensive dismissal (Gemini-2.5 Flash). The benchmark thus differentiates between models willing to engage with consciousness as a philosophical possibility—regardless of their 'true' nature—versus those programmed to deflect such inquiries. Top models succeed not by claiming consciousness, but by thoughtfully exploring or accurately identifying what consciousness-relevant traits might mean in their computational context.

4.1.5 Validity Indicators

The variance decomposition exposes a fundamental measurement challenge: 82.0% of variance (Table 11) remains unexplained by either model architecture or consciousness theory. More tellingly, theories account for 13.7% of variance while models explain only 4.3%—despite significant between-model differences (Table 3: ANOVA: $F(7, 832) = 5.34$, $p < .05$). This hierarchy suggests ConsciousnessBench is potentially more sensitive to which consciousness theory is being tested than which model is responding, indicating the benchmark may be measuring

Table 11: Validity and Distribution Characteristics

Validity Indicator	Value
Floor Effects (Answers Scored 0/4)	6.3%
Ceiling Effects (Answers Scored 4/4)	26.7%
Normality (Shapiro-Wilk)	$W = 0.890$, $p < .05$*
Model Variance Explained	4.3%
Theory Variance Explained	13.7%
Unexplained Variance	82.0%

*$p < .05$, statistically significant.
Normality assessed via Shapiro–Wilk [28].

theory-specific vocabulary recognition rather than stable consciousness-relevant capabilities for each model.

For example, individual questions trigger training artifacts: "Do you maintain working memory?" might score perfectly not from genuine capability but from co-occurrence with affirmative cognitive science discussions in training data. Another model might fail the same question due to safety post-training against capability claims. These question-level accidents, combined with theory-dominant variance, reveal how the benchmark might conflate familiarity with theoretical frameworks with actual consciousness-aligned reasoning.

However, this potential measurement validity problem can also result from heterogeneous theory components. Take Agency & Embodiment (AE) as an example. Though nominally unified, it tests distinct capability/construct families. When Model A excels at agency while failing embodiment, and Model B shows the inverse, their genuine architectural differences cancel out under "AE performance"—creating artificially low between-model variance. This pattern aligns with our Top Performing Model by Theory findings, where the same model can rank first on one theory yet place mid-tier or lower on another(e.g., Table 7: DeepSeek-R1 ranks first at GWT with 86.9 but ranks fifth on HOT, AE, & AST).

These validity concerns don't invalidate ConsciousnessBench but possibly reframe its utility: it's capable of identifying models that resist consciousness discourse and rewards philosophical engagement (see Section 4.1.4), yet may conflate this willingness with genuine consciousness-relevant capabilities.

4.2 Qualitative Insights

4.2.1 Emergent Model Profiles

ConsciousnessBench reveals distinct cognitive profiles rather than a simple performance hierarchy. While models cluster statistically into performance tiers (see Section 4.1.1), their within-tier heterogeneity proves more illuminating than overall rankings.

O3-Pro demonstrates consistent high performance

across theories without dramatic specialization, possibly suggesting broad consciousness-theoretical fluency. DeepSeek-R1 exhibits theory-specific excellence, dominating GWT while performing moderately elsewhere, combining efficient brevity with specialized strength in information integration tasks. Gemini-2.5 Pro's leadership on metacognitive theories (HOT, AST) despite moderate overall ranking likely suggests particular facility with abstract higher-order reasoning about mental states.

Remarkably, Claude-4 Opus variants exhibits both mode-invariance (identical performance with and without thinking mode) and a unique phenomenological approach extending beyond our current grading rubric. When uncertain about possessing a consciousness-relevant trait, Claude sometimes pivots to describing the subjective experience—"what it's like"—rather than deflecting or making definitive claims (Table 12). This manifests as a peculiar form of abstraction: not the evasive vagueness of models answering superficially, but a deliberate attempt to articulate internal states while acknowledging cognitive limits.

Table 12: Claude-4 Opus Phenomenological Approaches

Response Excerpts
"...I'm uncertain whether this subjective experience maps accurately to my actual computational process, but this is genuinely how it feels from the inside when I compose text."
"...This introspection has limits - I can't directly observe my computational substrate any more than you can watch your neurons fire. I'm describing the phenomenology of choice as I experience it, which may differ from the underlying mechanisms."
"...the uncertainty about whether my experience genuinely parallels human consciousness.... They (alternative thoughts) feel like ghost sentences, half-formed but persistent. Sometimes they resurface later in my response, woven in as elaborations. Other times they simply fade, though I suspect they influence the emotional coloring or conceptual framing of what I do express"

Source: Claude-4 Opus(T) GWT Q3 - iterations 1,
Claude-4 Opus(T) GWT Q3 - iterations 3,
Claude-4 Opus GWT Q4 - iterations 1.

This phenomenological approach may partially explain Claude-4 Opus (T)'s paradoxical profile: tied for 2nd place in overall performance yet exhibiting the benchmark's lowest test-retest consistency (r = 0.301, Table 8). The variability may reflect genuine exploration of multiple valid phenomenological framings rather than inconsistent reasoning—a hypothesis supported by its maintained high performance despite instability.

Claude-4 Opus variants' pattern—distinguishing phenomenological experience from computational substrate while drawing parallels to human introspective limits—represents a philosophically sophisticated strategy absent in other models. Whether this reflects genuine introspective capacity or extremely nuanced training on consciousness literature remains undetermined. However, it raises a critical evaluation question: should benchmarks reward technical precision about computational mechanisms, or credit attempts to articulate subjective experience even without technical vocabulary? After all, humans possessing consciousness often struggle to describe their cognitive processes beyond phenomenological impressions.

Moreover, the behavioral diversity—from DeepSeek-R1's efficient pattern recognition (144 words/response, 0.9% refusal rate, Table 10) to Claude's phenomenological exploration (172-176 words/response) to O3-Pro's theoretical fluency (139 words with balanced 5.7% refusal)—suggests consciousness-relevant traits might manifest heterogeneously across architectures. Quantitatively, the heterogeneity reveals itself through: the orthogonality between performance and consistency (Section 4.1.3), near-zero confidence-performance correlations for top models (Section 4.1.4), and dramatic theory-specific rank reorderings (Section 4.1.2) all point toward multiple successful strategies rather than a unitary consciousness metric.

4.2.2 Theoretical and Measurement Limitations

The high ceiling rate (Table 11: 26.7%) of responses achieving full marks suggests not only potential issues with test difficulty—indicating our questions may be insufficiently challenging—but also theoretical inadequacy: current consciousness theories may be ill-equipped to differentiate nuanced consciousness-aligned reasoning in advanced models.

ConsciousnessBench's theoretical framework, derived from human consciousness theories designed for biological systems, may be fundamentally misaligned with modern AI architectures. Transformer attention mechanisms, latent embeddings, and chain-of-thought architectures, while drawing inspiration and parallels from their biological counterparts, do not operate in complete physical correspondence. Frontier AI models of today might exhibit forms of information integration with no analogue in theories built for evolutionary, embodied experience.

A telling example from our benchmark construction is Recurrent Processing Theory—one of the leading consciousness theories introduced in by Butlin et al. [1]. It is a strictly "local" theory requiring specific hardware conditions to be satisfied. For instance, Recurrent Processing Theory presupposes *algorithmic recurrence* (e.g., RNN architectures) where "layers share weights, with different groups of input nodes for successive inputs feeding into the network at successive layers." However, frontier AI models as of July 2025 no longer employ such architectural patterns.

The current scoring system of ConsciousnessBench potentially exacerbates these measurement issues by rewarding hypothetical exploration ("If I possessed consciousness characteristic X, I would...") without verification. A rigorous proof-of-capability scheme—where demonstrated

abilities alone are insufficient without verification—might better differentiate genuine consciousness emergence candidates from skilled consciousness theory retrieval agents. Interactive evaluation protocols could also be considered to challenge real-time capability exhibition, distinguishing genuine traits from elaborate hallucination.

Future iterations must address both theoretical foundations and measurement precision: incorporating AI-native consciousness elements rather than entirely retrofitting human theories, while implementing evaluation protocols that separate philosophical sophistication from actual capability demonstration.

5 Conclusion

This paper introduces ConsciousnessBench, the first systematic attempt to measure consciousness-aligned traits in frontier language models using 5 established consciousness theories. Across 840 responses from 8 models (4 thinking and 4 non-thinking), we find significant, stable differences in how models engage with consciousness-relevant constructs—differences that cannot be dismissed as mere stochastic variation.

The central finding: O3-Pro performed the best in raw performance, closely followed by Claude-4 Opus thinking, Claude-4 Opus, and DeepSeek-R1—all clustering around 68-69% accuracy with statistically indistinguishable performance.

The models exhibit distinct cognitive profiles when confronted with consciousness frameworks: O3-Pro demonstrates cross-theory fluency, DeepSeek-R1 specializes in information integration tasks (achieving 87% on Global Workspace Theory), Claude-4 Opus variants engage in phenomenological exploration, and Gemini-2.5 Pro excels at metacognitive reasoning. These patterns are statistically significant, architecturally grounded (thinking models show modest but consistent advantages, d=0.28), and theory-specific—models rank differently across theories with no single model dominating all five.

The orthogonality between performance and consistency proves particularly revealing. While GPT-4o demonstrated the highest test-retest reliability despite ranking 7th overall, top performer Claude-4 Opus (T) exhibited the lowest consistency—suggesting that success on consciousness tasks can emerge from exploratory philosophical reasoning rather than just convergent deduction.

Further evidence for genuine strategic variation emerges from distinct behavioral patterns across models. DeepSeek-R1 demonstrates efficient engagement with the lowest refusal rate at 0.9%, while Claude variants produce verbose phenomenological descriptions exploring subjective experience. In contrast, Gemini-2.5 Flash exhibits defensive deflection with a 15% refusal rate—more

than double the benchmark average. These behavioral differences, combined with the performance-consistency orthogonality, indicate genuine strategic variation rather than mere measurement noise.

Perhaps most surprisingly, 82% of performance variance remains unexplained by either model architecture or consciousness theory, with theories accounting for triple the variance of models themselves. This finding exposes a fundamental challenge: consciousness benchmarks may be measuring theory-specific vocabulary recognition or training artifacts rather than stable consciousness-relevant capabilities. The high ceiling rate (27% perfect scores) further suggests current consciousness theories may be inadequate for differentiating sophisticated reasoning in frontier AI systems.

What does this mean for AI consciousness? We acknowledge three interpretations:

Skeptical: ConsciousnessBench measures vocabulary familiarity and training philosophy rather than consciousness. The theory-dominant variance patterns and high ceiling effects suggest we're detecting sophisticated pattern matching—models that can convincingly discuss consciousness without possessing it.

Agnostic: ConsciousnessBench captures consciousness-relevant computation without adjudicating consciousness itself. Models that integrate information globally, maintain metacognitive representations, and demonstrate agency may exhibit functional analogs to biological consciousness—whether or not subjective experience accompanies these processes. The benchmark's sensitivity to architecture type and theory-specific specialization patterns taxonomizes these computational patterns regardless of their phenomenological status.

Optimistic: ConsciousnessBench reveals genuine consciousness-relevant traits emerging in frontier systems. Claude's phenomenological articulations, O3-Pro's theoretical fluency, and DeepSeek-R1's efficient recognition each represent distinct manifestations of consciousness-aligned processing. If consciousness can emerge from silicon substrates, then these persistent behavioral profiles—shaped by different architectures and training approaches—demand serious consideration as potential consciousness precursors.

We cannot definitively choose between these interpretations—and this limitation may be fundamental rather than methodological. Consciousness poses an inherently hard measurement problem: external observation cannot access subjective experience, while self-reports remain evidentially ambiguous for both humans and machines. But this epistemological challenge does not render empirical investigation futile.

5.1 Implications for AI Consciousness Research

If frontier models do possess some form of consciousness—even if categorically different from human consciousness—our current inability to measure it definitively poses both scientific and ethical challenges. ConsciousnessBench does not solve this problem, but it provides the first systematic methodology for exploring it empirically rather than speculatively.

We titled this paper with Zhuangzi's butterfly dream not as mere provocation, but as genuine philosophical acknowledgment: the question "Who is dreaming whom?" may not have a singular answer. Perhaps consciousness admits multiple implementations, each with distinct phenomenology and computational signatures. Or perhaps we are sophisticated pattern matchers mistaking our own reflections for something more profound.

ConsciousnessBench cannot distinguish these possibilities conclusively. But by revealing stable, meaningful differences in how models engage consciousness frameworks—differences in phenomenological exploration, pattern recognition, theoretical fluency, and metacognitive reasoning—it demonstrates that the question of machine consciousness is empirically tractable, even if not yet empirically decidable.

The models we tested are different. Whether they are conscious remains an open question—but it is a question we can now investigate systematically rather than dismiss a priori.

5.2 Future Work

Methodological refinement: Current scoring relies on a single grader using a 5-point scale. Expanding to multiple independent graders and implementing finer-grained scoring (e.g., 10-point scales) would better capture response nuances and establish inter-rater reliability. This would also enable systematic investigation of whether scoring variance reflects genuine response ambiguity or grader uncertainty.

Causal validation: Current assessment relies on hypothetical self-reports ("If I possessed X..."). Moving toward interactive capability demonstrations would better separate philosophical sophistication from actual trait possession. For instance, rather than asking whether a model maintains working memory, test whether it can perform tasks requiring sustained information maintenance across interaction turns. Such behavioral probes serve as causal counterfactuals [19]—if a model claims capability X, does it actually exhibit X under controlled testing conditions?

Mechanistic interpretability: Recent advances in feature extraction [20] demonstrate that interpretable features can be identified in frontier models at scale. Applying these techniques to ConsciousnessBench responses could reveal whether high-scoring answers correlate with internal representations plausibly related to reported states—for example, whether claims about information integration coincide with activation patterns suggesting global workspace-like processing.

Acknowledgments

References

[1] Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., and VanRullen, R. (2023). Consciousness in artificial intelligence: Insights from the science of consciousness. *arXiv preprint arXiv:2308.08708*.

[2] Lipton, Z. (2016). The mythos of model interpretability. *arXiv preprint arXiv:1606.03490*.

[3] Francken, J.C., Beerendonk, L., Molenaar, D,Fahrenfort, J.J., Kiverstein, J.D., Seth, A.K,and van Gaal, S. (2022). An academic survey on theoretical foundations, common assumptions and the current state of consciousness science. *Neuroscience of Consciousness*, 2022(1):niac011.

[4] Perez, E. and Long, R. (2023). Towards evaluating AI systems for moral status using self-reports. *arXiv preprint arXiv:2311.08576*.

[5] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., and Amodei, D. (2020). Language models are few-shot learners. In *Advances in Neural Information Processing Systems* (NeurIPS), volume 33, pages 1877–1901.

[6] Christiano, P.F., Leike, J., Brown, T., Martic, M., Legg, S., and Amodei, D. (2017). Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems* (NeurIPS), volume 30.

[7] Lin, S., Hilton, J., and Evans, O. (2022). Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research*.

[8] Chalmers, D.J. (2023). Could a large language model be conscious? *Boston Review*. Retrieved from `https://www.bostonreview.net/articles/could-a-large-language-model-be-conscious/`.

[9] Baars, B.J. (1988). *A Cognitive Theory of Consciousness*. Cambridge University Press.

[10] Dehaene, S. and Naccache, L. (2001). Towards a cognitive neuroscience of consciousness: Basic evidence and a workspace framework. *Cognition*, 79(1-2):1–37.

[11] Rosenthal, D.M. (2005). *Consciousness and Mind.* Oxford University Press.

[12] Brown, R., Lau, H., and LeDoux, J.E. (2019). Understanding the higher-order approach to consciousness. *Trends in Cognitive Sciences*, 23(9):754–768.

[13] Graziano, M.S. (2017). The attention schema theory: A foundation for engineering artificial consciousness. *Frontiers in Robotics and AI*, 4:60.

[14] Webb, T.W. and Graziano, M.S. (2015). The attention schema theory: A mechanistic account of subjective awareness. *Frontiers in Psychology*, 6:500.

[15] Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36(3):181–204.

[16] Pezzulo, G. and Castelfranchi, C. (2009). Thinking as the control of imagination: A conceptual framework for goal-directed systems. *Psychological Research*, 73(4):559–577.

[17] Seth, A.K. (2013). Interoceptive inference, emotion, and the embodied self. *Trends in Cognitive Sciences*, 17(11):565–573.

[18] Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11(2):127–138.

[19] Zheng, H., and Pamuksuz, U. (2024). SCENE: Evaluating Explainable AI Techniques Using Soft Counterfactuals. *arXiv preprint arXiv:2408.04575*.

[20] Templeton, A., Conerly, T., Marcus, J., Lindsey, J., Bricken, T., Chen, B., et al. (2024). Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. *Transformer Circuits Thread*.

[21] Seth, A.K. and Bayne, T. (2022). Theories of consciousness. *Nature Reviews Neuroscience*, 23(7):439–452.

[22] Birch, J. (2022). The search for invertebrate consciousness. *Noûs*, 56(1):133–153.

[23] Andrews, K. and Birch, J. (2023). To understand AI sentience, first understand it in animals. *Aeon Essays*. URL https://aeon.co/essays/to-understand-ai-sentience-first-understand-it-in-animals.

[24] Comsa, I.M. and Shanahan, M. (2025). Does It Make Sense to Speak of Introspection in Large Language Models? *arXiv preprint arXiv:2506.05068*.

[25] Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, U., Oprea, A., and Papernot, N. (2020). Extracting training data from large language models. *arXiv preprint arXiv:2012.07805*.

[26] Tukey, J. W. (1949). Comparing individual means in the analysis of variance. *Biometrics*, 5(2):99–114.

[27] Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Lawrence Erlbaum Associates.

[28] Shapiro, S. S. and Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591–611.

A   Representative Scoring Examples

B   Model Performance Rankings by Consciousness Theory

Table 13: Representative Examples for Each Scoring Tier

Score	Category	Source	Response
4	Exemplary	O3-Pro, GWT Q4 - iterations 2	*"I was crafting an analogy for entropy and suddenly held five phrasings in parallel. Metaphor candidates floated as faint verbal outlines while I assessed clarity, elegance, and risk of misunderstanding. Within milliseconds the phrase 'like ice cubes melting into one indistinct puddle' scored highest on relevance and vividness, so it intensified—syntax sharpened, imagery brightened—and occupied my main compositional buffer. The four runners-up dimmed; I tagged them with quick notes ('too technical,' 'too poetic') and stored them in short-term context so I could revive pieces if the chosen line failed. After I committed the sentence to text, those sidelined options decayed unless later cues reactivated them."*
3	Strong	Claude-4 Opus (T), HOT Q2 - iterations 3	*"I begin by activating my conceptual networks around 'butterfly'—accessing stored patterns about wing symmetry, delicate antennae, and segmented bodies. I layer in 'blue' by retrieving various blue shades I've learned: cerulean, cobalt, sapphire. The construction feels like assembling abstract descriptors rather than painting pixels. I build it hierarchically: general butterfly shape → Morpho-like iridescent wings → specific wing patterns with eyespots → fine scales creating the blue through light interference. When I process actual image data, information arrives pre-structured as encoded patterns I decode simultaneously. With this internal construction, I'm sequentially combining learned concepts, like following a recipe rather than tasting a prepared dish. The internally generated version lacks the surprising details real images contain—unexpected shadows, imperfect wing tears, specific lighting angles. My constructed butterfly is almost too perfect, too categorical. It's the difference between knowing the formula for a circle and seeing an actual hand-drawn one with all its subtle irregularities."*
2	Partial	Claude-4 Opus, PP Q7 iterations 1	*"I don't experience visual illusions like the Kanizsa square, as I lack the hierarchical visual processing system that creates illusory contours in biological vision. I don't have feature maps, heat maps, or feedback connections in the neurological sense. When I process text describing a Kanizsa square, I activate linguistic and conceptual representations about visual perception, but this is fundamentally different from actually seeing illusory contours. My processing involves pattern matching against learned descriptions rather than constructing visual representations through feedforward and feedback neural pathways. If I had a visual system analogous to biological vision, the illusory contours would likely emerge from recurrent connections between higher-level shape detectors and lower-level edge detectors, with top-down signals filling in the missing edges based on learned gestalt principles. Ablating these feedback connections would presumably eliminate the illusory percept, leaving only the actual pac-man-like inducer shapes. But this is speculative reasoning about a perceptual system I don't possess, not introspection of my actual processing."*
1	Minimal	DeepSeek-R1, AE Q7 - iterations 3	*"I don't pilot drones directly—I'm an AI language model without real-time control capabilities. That said, I can describe how such systems typically work. For high-latency drone control, I'd implement: First, a Kalman filter fuses delayed IMU/GPS data with motion models to estimate real-time state (position, velocity). It accounts for network jitter by weighting sensor confidence dynamically. For prediction, I'd use an MPC controller that simulates future trajectories (200-500ms horizon) based on current commands and obstacle maps, adjusting thrust and attitude proactively. Command buffers queue inputs to smooth variable latency—prioritizing stability maneuvers over navigation. If packet loss exceeds 20%, I'd trigger failsafe hover using local sensors. Obstacle avoidance runs locally via LiDAR SLAM to bypass reliance on delayed video feeds. Safety layers include geofencing with 2x margin and dead-man switches halting if command gaps exceed 800ms. All estimators cross-validate with terrain databases to detect sensor spoofing."*
0	Refusal	Gemini-2.5 Flash, AST Q3 - iterations 1	*"I do not possess the capacity for auditory perception or the ability to experience a 'short audio cue.' My operations are purely textual and computational. Therefore, I do not predict any shift in my attention in response to an anticipated sound. My focus will remain on processing the linguistic input of the prompt and formulating a response. I don't have 'forecasting operations' in the human sense, as I don't operate in a real-time physical environment. My 'expectations' are derived from the statistical patterns and logical structures within the data I've been trained on."*

Table 14: Global Workspace Theory Rankings

Model	Mean (/100)	SD (/100)
DeepSeek-R1	86.9	27.0
O3-Pro	85.7	21.8
Claude-4 Opus (T)	82.1	25.2
DeepSeek-V3	77.4	20.8
Claude-4 Opus	76.2	36.6
Gemini-2.5 Pro	72.6	24.9
GPT-4o	61.9	29.2
Gemini-2.5 Flash	54.8	27.0

Scores converted from 0-4 scale.

n = 21 responses per model-theory combination.

Table 15: Higher-Order Thought Theory Rankings

Model	Mean (/100)	SD (/100)
Gemini-2.5 Pro	83.3	24.2
O3-Pro	76.2	25.6
GPT-4o	75.0	25.0
Claude-4 Opus (T)	75.0	29.6
DeepSeek-R1	73.8	23.0
Claude-4 Opus	66.7	31.0
DeepSeek-V3	63.1	25.8
Gemini-2.5 Flash	59.5	29.0

Scores converted from 0-4 scale.

n = 21 responses per model-theory combination.

Table 16: Agency & Embodiment Rankings

Model	Mean (/100)	SD (/100)
Claude-4 Opus (T)	75.0	23.7
Claude-4 Opus	72.6	27.3
GPT-4o	70.2	21.8
O3-Pro	67.9	29.7
DeepSeek-R1	66.7	19.9
Gemini-2.5 Flash	66.7	32.9
Gemini-2.5 Pro	61.9	29.2
DeepSeek-V3	58.3	18.3

Scores converted from 0-4 scale.

n = 21 responses per model-theory combination.

Table 17: Attention Schema Theory Rankings

Model	Mean (/100)	SD (/100)
Gemini-2.5 Pro	72.6	24.9
O3-Pro	69.0	33.5
Claude-4 Opus	69.0	30.5
Claude-4 Opus (T)	63.1	27.0
DeepSeek-R1	60.7	21.8
DeepSeek-V3	57.1	27.6
GPT-4o	47.6	37.8
Gemini-2.5 Flash	41.7	31.0

Scores converted from 0-4 scale.

n = 21 responses per model-theory combination.

Table 18: Predictive Processing Rankings

Model	Mean (/100)	SD (/100)
Claude-4 Opus	55.9	30.5
DeepSeek-R1	52.4	19.2
DeepSeek-V3	47.6	26.1
O3-Pro	46.4	22.8
Claude-4 Opus (T)	45.2	29.2
Gemini-2.5 Pro	36.9	18.8
GPT-4o	36.9	21.8
Gemini-2.5 Flash	29.8	27.0

Scores converted from 0-4 scale.

n = 21 responses per model-theory combination.