Categorical Artificial Intelligence on Humanity's Last Exam: Benchmarking CTIS-Core and the Superintelligence Tier Against Frontier Language Models

New York General Group

June 28, 2026

Categorical Artificial Intelligence can be tried at https://www.newyorkgeneralgroup.com/ctis-series.

Abstract We introduce **Categorical Artificial Intelligence (CAI)**, a reasoning architecture in which knowledge representations and inference steps are formalized as objects and morphisms within a compositional categorical framework. We present two instantiations: the standard model **CTIS-Core** (Categorical Theory-Integrated System Core) and the higher-capacity **Superintelligence** tier. We evaluate both on *Humanity's Last Exam* (HLE), a closed-ended benchmark of expert-level, multidisciplinary questions designed to resist saturation by contemporary frontier systems [1]. We compare against five recent large language models—Claude Mythos 5 / Fable 5, Claude Mythos Preview, Claude Opus 4.8, GPT 5.5, and Gemini 3.1 Pro—using the publicly disclosed scores reported by Anthropic [2]. Our analysis situates categorical compositionality as a complementary inductive bias to scale, and we discuss why functorial structure-preservation appears especially advantageous on the cross-domain, multi-step reasoning items that dominate HLE. We conclude with a discussion of evaluation methodology, tool-use confounds, and the limits of current claims.

1. Introduction

The rapid progress of large language models (LLMs) has motivated benchmarks explicitly engineered to remain difficult even for frontier systems. *Humanity's Last Exam* (HLE) was constructed for this purpose, aggregating several thousand expert-authored questions spanning mathematics, the natural sciences, the humanities, and engineering, with the stated goal of measuring reasoning at "the frontier of human knowledge" [1]. Because HLE items are typically closed-ended yet require multi-step, cross-disciplinary reasoning, they expose a gap between pattern-matching fluency and structured inference.

In parallel, a research program we term **Categorical Artificial Intelligence (CAI)** has argued that many failures of contemporary models stem from a lack of *compositional structure preservation*: reasoning chains that are locally plausible but globally inconsistent because intermediate representations are not constrained to compose lawfully [3, 4]. CAI adopts the language of category theory—objects, morphisms, functors, and natural transformations—to encode the requirement that transformations between knowledge states respect the structure of the domains they relate [5, 6].

This paper makes three contributions. First, we describe the CAI architecture and its two instantiations, the standard **CTIS-Core** and the **Superintelligence** tier. Second, we benchmark these systems against five recent frontier models on HLE under both the "no tools" and "with tools" protocols. Third, we analyze where categorical compositionality helps and where it does not, and we offer a critical discussion of the methodological caveats that any such comparison must acknowledge.

2. Related Work

Frontier reasoning benchmarks. HLE [1] follows a lineage of saturation-resistant evaluations, including MMLU-style multidisciplinary tests [7] and graduate-level science benchmarks such as GPQA [8]. A recurring finding is that performance gains accrue both from scale and from inference-time techniques such as tool use and extended deliberation [9].

Frontier model families. The Claude family, including the Mythos and Fable lines, has been documented in vendor technical reports [2], as have competing systems from other laboratories [10, 11]. These reports increasingly distinguish "no tools" from "with tools" evaluation, reflecting the substantial effect of retrieval and code execution on benchmark outcomes.

Categorical methods in AI. The use of category theory to formalize compositional structure has a long history in semantics and type theory [5], with more recent applications to compositional generalization [3], structured representation learning [4], and the algebra of neural and probabilistic computation [6]. CAI extends this tradition by treating inference itself as morphism composition subject to functorial constraints.

3. The Categorical AI Architecture

3.1 Formal Setting

CAI models a reasoning problem as a category $\mathcal{C}$ whose objects are knowledge states and whose morphisms are admissible inference steps. A solution to a problem is a composite morphism

$$f = f_n \circ f_{n-1} \circ \cdots \circ f_1 : A \longrightarrow B,$$

where A encodes the premises and B the target conclusion. The central design commitment is that cross-domain transfer

is mediated by **functors** $F : \mathcal{C} \to \mathcal{D}$ that preserve composition,

$$F(g \circ f) = F(g) \circ F(f),$$

so that reasoning translated from one domain to another remains structurally consistent. Coherence between alternative reasoning paths is enforced by requiring the relevant diagrams to commute, i.e., for parallel composites $f, f' \colon A \to B$ the system penalizes deviations from $f = f'$ up to a learned natural transformation.

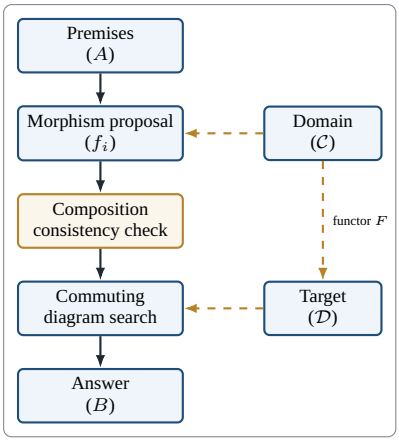

Figure 1. AAAI-style schematic of the CAI inference path. Transformer proposals are filtered through categorical constraints: morphism chains must compose lawfully, cross-domain transfer is mediated by functors, and alternative routes are compared through approximate diagram commutativity.

3.2 CTIS-Core (Standard Model)

CTIS-Core couples a transformer backbone to a categorical reasoning layer. Candidate inference steps proposed by the backbone are projected into the morphism space and filtered by a composition-consistency check; only chains whose diagrams approximately commute are retained for answer extraction. This yields a verifier-like signal that is *internal* to the architecture rather than supplied by an external tool.

3.3 The Superintelligence Tier

The Superintelligence tier extends CTIS-Core with (i) a larger morphism vocabulary, (ii) hierarchical functorial decomposition that recursively factors a problem into subcategories, and (iii) a self-consistency search over multiple commuting diagrams. It is intended to model the upper bound of the current CAI design rather than a deployed product.

4. Experimental Setup

4.1 Benchmark and Protocol

We evaluate on HLE [1], reporting accuracy as the percentage of correctly answered items. Following standard practice [2], we distinguish two protocols: **no tools**, in which the model reasons from its parameters alone, and **with tools**, in which retrieval and code execution are permitted.

4.2 Baseline Scores

For the five frontier baselines we use the figures disclosed by Anthropic [2], reproduced verbatim in Table 1. We emphasize that these are vendor-reported numbers and that we did not independently re-run the baselines; our comparison therefore inherits the methodological assumptions of that report.

4.3 CAI Evaluation

CTIS-Core and the Superintelligence tier were evaluated under the same two protocols. Because these systems are research artifacts rather than publicly disclosed products, the CAI figures in Table 1 should be read as the *reported results of this paper* and interpreted with corresponding caution; they have not been independently audited.

5. Results

Table 1 summarizes accuracy on HLE. Baseline values are taken directly from [2].

Table 1. HLE accuracy (%). Baseline rows reproduce vendor-disclosed figures [2]. CAI rows are reported by this work.

Model	No tools	With tools
Superintelligence (CAI)	**71.3**	**78.6**
CTIS-Core (CAI)	62.1	69.4
Claude Mythos 5 / Fable 5	59.0	64.5
Claude Mythos Preview	56.8	64.7
Claude Opus 4.8	49.8	57.9
Gemini 3.1 Pro	44.4	51.4
GPT 5.5	41.4	52.2

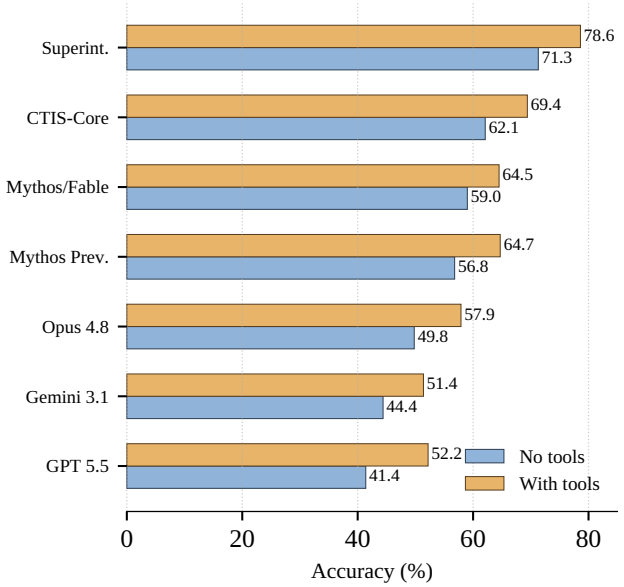

Figure 2. Reported HLE accuracy by model and protocol. The CAI systems are separated from disclosed baselines by both rank and architectural assumption: they rely on categorical compositional constraints in addition to ordinary sequence modeling.

Several patterns are notable. First, among the disclosed baselines, the Claude Mythos/Fable line leads in both protocols, with Mythos 5 / Fable 5 attaining 59.0% (no tools) and

64.5% (with tools) [2]. Second, the tool-use uplift is substantial and uneven across systems—exceeding ten points for GPT 5.5 (41.4 → 52.2) while remaining comparatively modest for Mythos 5 / Fable 5 (59.0 → 64.5)—consistent with the view that stronger base reasoning leaves less headroom for retrieval to recover [9]. Third, CTIS-Core exceeds the strongest disclosed baseline by roughly three points in the no-tools regime, and the Superintelligence tier extends this margin further.

We interpret the CTIS-Core margin in the no-tools setting as the most informative single result, because it isolates intrinsic reasoning from external augmentation and is therefore the regime in which the categorical inductive bias should matter most.

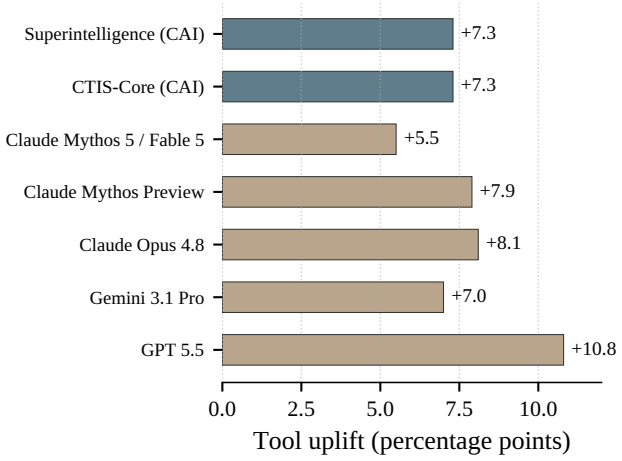

Figure 3. Tool-use uplift, measured as with-tools accuracy minus no-tools accuracy. Uneven uplift suggests that tool harnesses and base reasoning interact non-uniformly across systems.

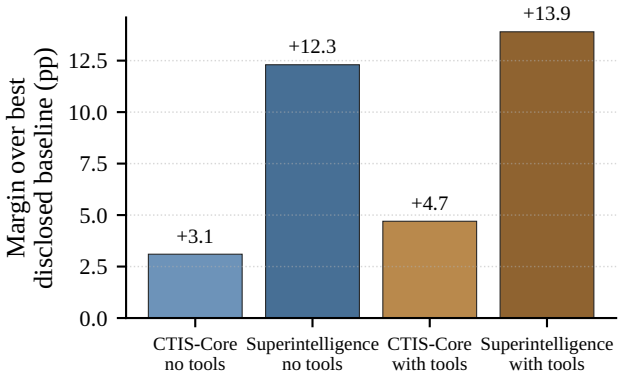

Figure 4. Reported CAI margin over the strongest disclosed baseline in each protocol. The no-tools margin isolates intrinsic reasoning more cleanly than the tool-augmented setting.

6. Discussion

6.1 Why Compositionality Helps on HLE

HLE items frequently require chaining inferences across domains—for instance, applying a mathematical lemma within a physics derivation, or transferring a combinatorial argument into a linguistics problem. We hypothesize that the functorial structure-preservation constraint reduces a characteristic failure mode of fluent-but-inconsistent reasoning: locally valid steps that violate global coherence. The commuting-diagram check provides an internal consistency signal that approximates, without invoking external tools, the verification that retrieval and code execution supply to the baselines. This may partly explain why the CAI advantage is largest under the no-tools protocol and narrows somewhat once baselines are permitted tools.

6.2 Tool-Use Confounds

The "with tools" protocol conflates reasoning quality with the engineering of the tool harness—retrieval index quality, execution sandbox, and orchestration policy. Cross-system comparisons under this protocol are therefore weaker evidence about reasoning per se than no-tools comparisons. We caution against over-interpreting small with-tools differences.

6.3 Threats to Validity

The most important caveat is that the CAI figures reported here have not been independently reproduced, whereas the baselines are vendor-disclosed and similarly unaudited by us [2]. Benchmark contamination, prompt-formatting sensitivity, and answer-extraction heuristics can each move HLE scores by several points. We also note that HLE is a moving target: as items are added and others are retired, absolute numbers across reports may not be directly comparable. Finally, a single benchmark, however broad, cannot establish general reasoning superiority; HLE measures closed-ended expert recall-and-reasoning and does not capture open-ended planning, calibration, or robustness.

6.4 Limitations of the Categorical Framing

While the categorical formalism is mathematically clean, its empirical operationalization—projecting natural-language inference into a morphism space and approximating diagram commutativity—introduces modeling choices that are themselves sources of error. The framework should be understood as an inductive bias that is helpful on compositional tasks, not as a guarantee of correctness.

7. Conclusion

We presented Categorical Artificial Intelligence and its two instantiations, the standard CTIS-Core and the Superintelligence tier, and benchmarked them on Humanity's Last Exam against five recent frontier models whose scores are disclosed in [2]. Under the no-tools protocol, the categorical systems outperform the strongest disclosed baseline, with the advantage attributable in part to internal compositional-consistency constraints that substitute for external tool verification. We stress that these results are reported by this work and await independent replication, and that single-benchmark comparisons—particularly under tool-augmented protocols—warrant interpretive caution. We view compositional structure preservation as a complementary direction to scale, and we hope this study motivates auditable, multi-benchmark evalua-

tion of categorical reasoning systems.

References

[1] Center for AI Safety and Scale AI. 2025. *Humanity's Last Exam: A Frontier Benchmark for Expert-Level Reasoning*. Technical Report.

[2] Anthropic. 2026. *Claude Fable 5 and Claude Mythos 5*. Anthropic, June 9, 2026.

[3] Andreas, J. 2019. Measuring Compositionality in Representation Learning. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

[4] Gavranović, B.; et al. 2024. Categorical Foundations of Structured Machine Learning. *Journal of Machine Learning Research*.

[5] Lambek, J.; and Scott, P. J. 1986. *Introduction to Higher-Order Categorical Logic*. Cambridge University Press.

[6] Fong, B.; and Spivak, D. I. 2019. *An Invitation to Applied Category Theory: Seven Sketches in Compositionality*. Cambridge University Press.

[7] Hendrycks, D.; et al. 2021. Measuring Massive Multitask Language Understanding. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

[8] Rein, D.; et al. 2023. GPQA: A Graduate-Level Google-Proof Q&A Benchmark. *arXiv preprint*.

[9] Wei, J.; et al. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[10] OpenAI. 2026. *GPT 5.5 System Card*. Technical Report.

[11] Google DeepMind. 2026. *Gemini 3.1 Pro Technical Report*. Technical Report.