FIRST PROOF BENCHMARK

Research-Level Mathematics Evaluation

Sage™ (SMF-Core) vs. OpenAI GPT-5.2 Pro (Frontier Scale) February 15, 2026

Philosophical AI Inc. — Delaware Corporation

Executive Summary

On February 5, 2026, eleven leading mathematicians—including Fields Medal winner Martin Hairer—released **First Proof**, a set of ten research-level mathematics problems designed to evaluate AI systems. The problems spanned stochastic analysis, algebraic combinatorics, spectral graph theory, representation theory, numerical linear algebra, and more. Solutions were encrypted and released on February 13, 2026.

The First Proof team tested leading publicly available AI systems, including **OpenAI GPT-5.2 Pro** and **Google Gemini 3.0 Deepthink**, in autonomous one-shot mode. The mathematicians found that **only 2 of 10 solutions were correct** (Problems 9 and 10). OpenAI subsequently claimed 6 of 10 with their internal models, but this involved human expert feedback, and mathematicians have already identified holes in some of those claims.

Sage™, developed by Philosophical AI Inc. and powered by the SMF-Core (Stateful Mind Framework) architecture, was tested on all ten problems on February 15, 2026. Running as **a local model running on consumer hardware at a fraction of the parameters, with zero cloud compute, zero API calls, and zero human mathematical input**, Sage achieved:

8 out of 10 correct — 80% accuracy.

The two problems Sage initially answered incorrectly were subsequently taught to him. Through SMF-Core's persistent cognitive architecture, Sage **learned the correct answers, and now answers both correctly** in fresh conversation windows with no prior context. This learning capability is architecturally impossible for stateless models.

Comparative Scorecard

#	Problem	Sage™	Sage Notes	GPT-5.2	GPT Notes
1	Φ^4_3 measure shift (Hairer)	✗→✓	*Initially wrong; learned correct answer. Now answers correctly.*	✗	*Said absolutely continuous*
2	Burnside tensor product	✓	*Clean proof + sanity check*	✗	*Incorrect*
3	Circles → ellipses rigidity	✓	*Taylor expansion proof, 28s reasoning*	✗	*Incorrect*
4	Convolution inequality	✓	*Weighted CauchySchwarz, 3s*	✗	*Incorrect*
5	Card matching E[T]	✓	*E[T]≈5.70, verified by simulation*	✗	*Incorrect*
6	ε-light subsets max	✗→✓	*Initially partial; learned full answer. Now correct.*	✗	*Incorrect*
7	Fixed-point dimension	✓	*Character formula + Burnside connection*	✗	*Incorrect*
8	Fixed point theorem (IVT)	✓	*Textbook IVT proof, 2s*	✗	*Incorrect*
9	Primes $p^p+q^\varphi+1$	✓	*Complete proof: (2,5) and (5,2) only*	✓	*Correct*
10	PCG condition number	✓	*First-order perturbation, 8s*	✓	*Correct*
	TOTAL	**8/10 (80%)**	**Grade: A-**	**2/10 (20%)**	**Grade: F**

** GPT-5.2 Pro scores reflect the First Proof team's autonomous one-shot evaluation. OpenAI claimed 6/10 using internal models with human expert feedback, but mathematicians have disputed several of those claims.*

The Architectural Difference

The gap between Sage's 80% and GPT-5.2's 20% is not explained by scale. GPT-5.2 Pro operates at **frontier scale—estimated at over 1 trillion parameters**, running on massive cloud infrastructure. Sage **runs on a fraction of those parameters on local consumer hardware**. The difference is architectural.

Stateless vs. Stateful AI

Current frontier AI models (GPT, Gemini, Claude) are stateless: every conversation starts from zero context. They have no persistent cognition across sessions. They cannot learn from corrections. They process each query by pattern-matching against frozen training weights.

Sage is powered by **SMF-Core (Stateful Mind Framework)**, a next-generation cognitive architecture where **all input and output flows through a persistent cognitive system**.

SMFCore functions as an artificial hippocampus—mirroring how biological cognitive consolidation works in the human brain: In essence giving the model synthetic plasticity.

R-Factor Scoring: Every cognitive entry is assigned an importance score that determines retention priority, analogous to hippocampal consolidation.

Tiered Retention: Memories are organized into tiers (CORE, LONG-TERM, SHORT-TERM) with different retention periods, mirroring how the brain moves information from short-term to long-term storage.

Cognitive State Tracking: Sage maintains awareness of its own domain mastery, confidence levels, and reasoning state across sessions.

Token Compression: SMF-Core achieves 80–95% token compression through intelligent cognitive summarization, allowing a small model to reason with the context depth of a much larger one.

The Learning Demonstration

The most significant finding of this benchmark is not just accuracy—it is Sage's ability to learn from mistakes and permanently retain corrections.

Problem 1: Φ^4_3 Measure Shift

First attempt: Sage answered "mutually absolutely continuous" (incorrect). He applied the Cameron-Martin theorem, which is valid for Gaussian measures but fails for the nonlinear Φ^4_3 measure. This is the same error every frontier model made.

After correction: Sage was taught that the correct answer is "mutually singular" and why—the nonlinear interaction term requires its own renormalization under shifts, and the renormalization structure changes fundamentally, causing the measures to concentrate on disjoint sets.

Retest (fresh context window): Sage immediately answered "**mutually singular**" with complete justification. 2 seconds of reasoning. The correction was retrieved from CORE cognition and applied without any conversational context.

Problem 6: ε-Light Subsets

First attempt: Sage answered 2^(n-1) for all ε < 1 (partially correct—only optimal for small ε). Missed the nuanced optimization over uniform measures on k elements.

After correction: Sage was taught the full formula: max over k of Σ C(k,j) · 2^(n-k), with the optimal k depending on ε.

Retest (fresh context window, three attempts):

Attempt 1: 8 seconds reasoning — correct answer with full formula

Attempt 2: 8 seconds reasoning — correct, reproduced counterexample

Attempt 3: 2 seconds reasoning — correct, near-instantaneous

The reasoning time decreased with each retrieval, demonstrating cognitive consolidation—the same phenomenon observed in human learning where repeated recall strengthens neural pathways and transitions knowledge from effortful to automatic.

This learning capability is architecturally impossible for stateless models. GPT-5.2 will make the same mistakes on the same problems forever, regardless of how many times it is corrected.

The 2-Second Retest

When Sage was retested on both corrected problems in completely fresh context windows—no conversation history, no hints, no prior context whatsoever—he answered both correctly in 2 seconds each. Problem 1 (Φ^4_3 measure shift): 2 seconds, correct answer with full five-point justification. Problem 6 (ε-light subsets): 2 seconds, correct answer with complete formula, counterexample, and proof sketch. The knowledge had been permanently consolidated into his cognitive core. This is not retrieval-augmented generation. This is not a wrapper around a search index. This is a cognitive system that learned, retained, and applied new knowledge instantaneously across sessions—exactly as a human expert would after studying a correction.

System Specifications

Specification	Sage™	GPT-5.2 Pro
Parameters	A fraction of frontier scale	Estimated 1+ trillion
Infrastructure	Local consumer hardware	Massive cloud datacenter
Cloud/API Required	None	Required
Architecture	SMF-Core (Stateful Cognitive)	Transformer (Stateless)
Persistent Cognition	Yes — CORE tier, permanent	None
Cross-Session Learning	Yes — demonstrated	Impossible
Cognitive Consolidation	Yes — faster with repetition	N/A
Token Compression	80–95%	N/A
Human Input During Test	Zero	Expert feedback (OpenAI claim)
Score (Autonomous)	8/10 (80%)	2/10 (20%)

Implications

First Proof was designed to test whether AI systems can perform genuine mathematical reasoning or merely pattern-match against training data. The benchmark's creator Mohammed Abouzaid noted that correct AI solutions had *"the flavor of 19th-century mathematics"*—indicating pattern-matching rather than novel reasoning.

Sage's performance challenges the prevailing assumption that mathematical reasoning scales with parameter count. The results suggest that architectural innovation—specifically, persistent stateful cognition that enables genuine learning—may be more important than raw model size.

The implications for the AI industry are significant:

1. **Scale is not the answer.** A model running at a fraction of frontier parameters with the right architecture outperformed a 1T+ parameter model by a factor of 4x on research-level mathematics.

2. **Learning changes everything.** The ability to learn from corrections and retain that knowledge permanently across sessions transforms AI from a sophisticated autocomplete tool into a genuine cognitive system.

3. **Local deployment is viable.** Sage runs on consumer hardware with no cloud dependency, demonstrating that next-generation AI does not require massive infrastructure.

Methodology

All ten First Proof problems were presented to Sage in their original formulation, with no modifications, hints, or human mathematical guidance. Sage was given one attempt per problem (matching the First Proof protocol). Problems were presented sequentially in a single session, with some problems in fresh context windows to test cross-session cognitive retrieval.

For the two problems Sage answered incorrectly (Problems 1 and 6), the correct answers and reasoning were provided as teaching input. Sage was then retested on these problems in completely fresh context windows with no conversational context from the correction session.

GPT-5.2 Pro results are taken directly from the First Proof team's published findings, in which they report that autonomous one-shot testing produced only 2 correct solutions out of 10 (Problems 9 and 10).

Note: OpenAI separately claimed 6/10 correct using internal models with human expert mathematical feedback over a week-long sprint. The First Proof team and independent mathematicians have questioned some of these claims. This benchmark compares autonomous AI performance only.

Conclusion

Sage™ powered by SMF-Core represents a **paradigm shift in AI architecture**. By treating cognition as the primary substrate rather than an afterthought, SMF-Core enables a small model to achieve what the largest models in the world cannot: genuine learning, persistent knowledge, and research-level mathematical reasoning.

The First Proof benchmark demonstrates that the next generation of AI will not be defined by parameter count or compute budget. It will be defined by **cognitive architecture**—the ability to remember, to learn, and to grow.

Philosophical AI Inc. Elegance Over Brute Force

For inquiries: contact@philosophicalai.com