Three-Way Comparative Analysis: Official, OpenAI, and Independent AI-Assisted Solutions to the First Proof Challenge

Mark Dillerop*
Independent Researcher

February 14, 2026

Abstract

The First Proof challenge posed ten research-level mathematical problems on February 5, 2026. We compare three sets of solutions: (i) the official human-generated solutions released by the problem authors on February 14; (ii) the submission by OpenAI, generated and typeset entirely by their models; and (iii) our own solutions, produced through an iterative multi-model human–AI collaborative workflow by a non-mathematician. We evaluate each submission on answer correctness, proof completeness, and methodological approach. OpenAI achieves 7/10 correct answers with complete proofs, produces 2 incorrect answers (P03, P07), and provides 1 valid approach to an open-ended question. Our workflow achieves 7/10 correct answers (5 complete proofs plus 1 valid approach), with 1 correct answer supported by an incorrect proof (confirmed by the problem author), 2 incomplete proofs, and 1 divergent result. The two AI submissions fail on largely *different* problems, suggesting complementary strengths and weaknesses. We note that without independent mathematical review, some claimed proofs (especially OpenAI's P04) cannot be fully verified as correct.

Disclosure. This paper was itself produced using the same human–AI collaborative workflow it describes.

1 Introduction

The First Proof challenge [1] posed ten open research-level mathematical problems on February 5, 2026, with a submission deadline of February 13. The problems span stochastic PDEs, representation theory, algebraic combinatorics, free probability, equivariant topology, spectral graph theory, manifold theory, symplectic geometry, multilinear algebra, and numerical linear algebra.

On February 13, OpenAI submitted solutions to all ten problems, generated and typeset entirely by their models [3]. Independently, we submitted solutions between February 10 and 12 using an iterative multi-model workflow. On February 14, the organizers released official solutions and commentary [2].

This paper provides a systematic three-way comparison. We assess each submission against the official solutions along three dimensions: answer correctness, proof completeness, and methodological approach (§2.2). We then analyze the pattern of successes and failures across the two AI submissions (§5).

The author is not a mathematician. The mathematical content of our submitted solutions was generated entirely by AI systems, with the author serving as workflow orchestrator. OpenAI's submission was generated entirely by their models with no human mathematical input.

*Correspondence: dillerop@gmail.com

2 Methodology

2.1 The three submissions

Official (human). Solutions produced by the problem authors—expert mathematicians. Released February 14, 2026.

OpenAI (single-system AI). All solutions "generated and typeset by our models" (February 13, 2026). The specific model(s) are not disclosed.

Ours (multi-model AI + human orchestration). Solutions produced through 4–12 iterative sessions per problem using five AI systems: Claude Opus 4.6, ChatGPT 5.2 Pro, Gemini 3.0 Deep Think, Grok, and Perplexity. The author orchestrated the workflow; the AI systems handled all mathematical content. Submitted February 10–12, 2026.

2.2 Assessment criteria

We evaluate each solution along three dimensions.

Answer match. Whether the stated conclusion agrees with the official answer. Three levels: **Yes** (full agreement), **Partial** (correct direction, weaker claim), **Wrong** (incorrect answer or different case).

Proof status. *Complete*: establishes the claimed result without logical gaps. *Incomplete*: has identified gaps. *Wrong*: the answer itself is incorrect.

Method. The principal proof technique, characterized as *identical*, *analogous*, or *distinct* relative to the official solution.

3 Results

3.1 Answer and proof comparison

Table 1 summarizes the three-way comparison across all ten problems.

Table 1: Three-way answer and proof comparison. Official answers are correct by definition.

#	Problem	Official answer	OpenAI answer / proof	Our answer / proof
1	Φ^4_3 shift	No (singular)	**Yes** / Complete	**Yes** / **Incorrect** (Hairer)
2	Rankin–Selberg	Yes (constructive)	**Yes** / Complete	**Yes** / Complete
3	Macdonald chain	Yes (t-Push TASEP)	**Wrong** / Claims ill-posed	**Yes** / Complete
4	Finite free Stam	Yes (all n)	**Yes** / Complete	**Partial** / Incomplete ($n \leq 3$)
5	$\mathcal{O}$-slice filtr.	Characterization thm	**Yes** / Complete	**Yes** / Complete
6	ε-light subsets	Yes, $c = \varepsilon/42$	**Yes** / Complete ($c = 1/256$)	**Partial** / Incomplete
7	Lattices / $\mathbb{Q}$-acyclic	No (2-torsion)	**Wrong** / Claims Yes	**Wrong** / Wrong case
8	Lagrangian smoothing	Yes	**Yes** / Complete	**Yes** / Complete
9	Quadrilinear tensors	Yes (deg 5)	**Yes** / Complete (deg 5)	**Yes** / Complete (deg 4)
10	CP-RKHS PCG	(Explain how)	**Yes** / Complete	**Yes** / Complete
Totals		10/10	8 correct, 2 wrong	6 correct, 1 incorrect proof, 2 partial, 1 wrong

3.2 Methodological comparison

Table 2 compares the proof techniques used by each submission.

3.3 Aggregate assessment

OpenAI: 7 complete + 1 valid approach + 2 wrong answers. OpenAI produces correct, complete proofs for P01, P02, P04, P05, P06, P08, and P09. For P10 (an open-ended "explain how" question), they provide a valid PCG method with Kronecker preconditioning

Table 2: Methodological comparison across all three submissions.

#	Prob.	Official	OpenAI	Ours
1	$\varPhi^4_3$	H_3 Wick cube, $(\log N)^{-\gamma}$, BG decomp.	Mollified cubic, super-exp. scales, renorm. Wick	Hairer's A_ψ, super-exp. $e^{-3n/4}$
2	R–S	Constructive: explicit W_0 via Godement–Jacquet	Constructive: Kirillov model on mirabolic, Howe vectors	Existential: JPSS + inertial class counting
3	Macd.	t-Push TASEP (discrete, multiline queues)	*Claims problem ill-posed; no chain constructed*	Hecke-recursive detailed balance (continuous-time)
4	Stam	Jacobian contractivity via Bauschke et al.	Self-contained proof (claims all n)	Cumulant decomp., heat equation ($n \leq 3$)
5	Slice	Isotropy sep. + induction on lattice	Localizing subcat. + isotropy sep. + induction	Nearly identical; includes Lean 4 formalization
6	ε-lt	BSS barrier + leverage → $c = \varepsilon/42$	BSS barrier + partial coloring → $c = 1/256$	Multi-bin greedy; partial cases only
7	Latt.	Surgery theory + symmetric signatures + Novikov	Spin-lift $\Gamma < \mathrm{Spin}(n, 1)$, surgery realization	Euler char. contradiction ($\delta = 0$ only)
8	Lagr.	Conormal fibration (coordinate-free)	Local smoothing + edge/vertex gluing + Hamiltonian	Hamiltonian via Cartan formula (coordinate-based)
9	Tens.	5×5 minors of flattenings (deg 5, Tucker rank)	5×5 minors of flattenings (deg 5, same as official)	Plücker equations (deg 4, geometric)
10	PCG	PCG + eigendecomp. of K	PCG + Kronecker precond. + eigendecomp.	PCG + subsampled Kronecker matvecs

and eigendecomposition—arguably the strongest P10 submission. However, they produce two *incorrect* answers:

- **P03**: OpenAI argues the problem is ill-posed because the putative stationary weights can be negative for certain parameter values. The official solution (and ours) constructs a valid Markov chain by restricting to parameter regimes where positivity holds. OpenAI's reading of the problem is overly literal.
- **P07**: OpenAI claims YES and constructs a manifold M with $\pi_1(M) \cong \Gamma < \mathrm{Spin}(n, 1)$ and $\widetilde{M}$ rationally acyclic. The official answer is NO. The official proof uses surgery theory with symmetric signatures and the Novikov conjecture to show no such manifold exists for 2-torsion. OpenAI's construction appears to contain an error in the surgery realization step (the Wall finiteness obstruction argument or the rational surgery realization may not apply as claimed).

Ours: 5 complete + 1 valid approach + 1 correct answer with incorrect proof + 2 incomplete + 1 wrong case. We produce correct, complete proofs for P02, P03, P05, P08, and P09. For P10, we provide a valid PCG method. For P01, we correctly answered NO but Hairer and Gubinelli have confirmed that our proof is incorrect (see §4.1). We produce incomplete proofs for P04 ($n \leq 3$ only) and P06 (partial cases, no universal constant). For P07, we prove NO for the wrong case ($\delta(G) = 0$ instead of 2-torsion).

Key observation: complementary failures. The two AI submissions fail on largely *different* problems. OpenAI fails on P03 and P07; we fail on P04, P06, and P07. Only P07 is a shared failure—and even there, the failure modes differ (OpenAI claims YES incorrectly; we prove NO for the wrong case). Conversely, OpenAI succeeds where we fail (P04, P06) and we succeed where OpenAI fails (P03).

4 Problem-by-Problem Analysis

4.1 P01: $\varPhi^4_3$ measure shift

All three submissions correctly answer NO (mutual singularity). The official solution uses the Wick cube H_3 with $(\log N)^{-\gamma}$ scaling and the BG decomposition. OpenAI uses a mollified cubic observable at super-exponential scales with renormalized Wick powers, producing a self-contained 7-page proof with explicit propositions and lemmas. Our proof uses Hairer's A_ψ functional with super-exponential scaling $e^{-3n/4}$.

However, Hairer and Gubinelli have confirmed that our proof is incorrect. Hairer (February 15, 2026): "The proof is definitely not correct: for example the first non-trivial statement, namely 'The renormalized cubic is $O(1)$', is wrong." In the Φ^4_3 setting, the Wick-ordered cube $: u^3 :$ is a distribution in a negative Sobolev/Besov regularity space (roughly $\mathcal{C}^{-1/2-\varepsilon}$), not a bounded function.

Gubinelli further observed that the models misidentified the mechanism of the proof (fixating on mass renormalization, which was chosen for convenience) and that our own Section 7 correctly notes that the variance of the renormalized cube diverges—directly contradicting the Section 6 claim that it is $O(1)$. The models generated both the correct fact and the incorrect claim in different sections of the same document without recognizing the contradiction.

Gubinelli additionally noted: "I have now seen several 'solutions' doing this. Also, they all take as gospel the choice $\varepsilon_n = e^{-e^n}$ when this was really just a way of saying 'I don't want to have to think about whether I can apply Borel–Cantelli'..." This confirms a systematic pattern across multiple independent AI systems: **cargo-cult proof construction**—reproducing the surface features of a proof (specific formulas, specific parameter choices) without understanding why those choices were made, and consistently elevating convenient implementation details to essential status.

4.2 P02: Rankin–Selberg test vectors

All three answer YES. The official solution is *constructive*, giving an explicit W_0 via the Godement–Jacquet functional equation. OpenAI is also *constructive*, using the Kirillov model on the mirabolic subgroup and Howe vectors (Baruch's explicit computation) to produce a concrete test vector—a 12-page proof covering all cases (ramified, unramified, $n = 1$). Our proof is *existential*, using JPSS nondegeneracy and inertial class counting.

4.3 P03: Markov chain for Macdonald polynomials

The official answer is YES, constructing the interpolation t-Push TASEP. Our answer is YES, constructing a Hecke-recursive detailed balance chain. **OpenAI answers NO**, arguing the problem is ill-posed because the putative stationary weights F^*_μ/P^*_λ can be negative for certain real specializations of the parameters (x_i, t). OpenAI demonstrates this with the explicit counterexample $n = 2$, $\lambda = (2, 0)$, $t = 2$, $x_1 = 0$, $x_2 = 10$, yielding $\pi(2, 0) = -1/17$.

This is a legitimate mathematical observation, but it reflects an overly literal reading of the problem. The official solution (and ours) constructs a Markov chain that works in the parameter regime where positivity holds. The problem authors clearly intended the question to be read in this way.

4.4 P04: Finite free Stam inequality

The official answer is YES for all n, proved via Jacobian contractivity using the Bauschke–Güler–Lewis–Sendov theorem on hyperbolic polynomial convexity. OpenAI claims a complete self-contained proof for all n (8 pages). Our proof covers $n \leq 3$ only, with a gap for $n \geq 4$.

OpenAI's claimed proof for all n is notable. If correct, it represents an independent discovery of a complete proof using different techniques from the official solution. However, absent independent mathematical review, we cannot rule out subtle gaps analogous to those in OpenAI's P07 proof (see §4.7). We were unable to close the gap despite 8+ sessions and 900,000+ numerical tests.

4.5 P05: $\mathcal{O}$-slice filtration

All three submissions produce correct characterization theorems with very similar proof structures (isotropy separation + induction on the subgroup lattice). OpenAI's proof is 8 pages using localizing subcategories. Our proof additionally includes Lean 4 formalization with zero `sorry` axioms.

4.6 P06: ε-light subsets

The official answer is YES with $c = \varepsilon/42$. OpenAI proves YES with $c = 1/256$ using a one-sided BSS barrier variant with a partial coloring process (7 pages). Our proof establishes partial cases only ($c = 1/6$ for sparse graphs, $c = 1/2$ for K_n) without a universal constant.

Both the official solution and OpenAI use BSS barrier methods—the key technique we could not discover. OpenAI's constant (1/256) is weaker than the official (1/42) but the proof is complete.

4.7 P07: Lattices and $\mathbb{Q}$-acyclicity

The official answer is NO for 2-torsion, proved via surgery theory with symmetric signatures and the Novikov conjecture. **OpenAI answers YES**, constructing a spin-lift lattice $\Gamma < \mathrm{Spin}(n, 1)$ for odd $n \geq 5$ and using rational surgery realization to produce a closed manifold with $\mathbb{Q}$-acyclic universal cover. Our proof establishes NO for $\delta(G) = 0$ only, leaving the 2-torsion case open.

OpenAI's construction is elaborate (10 pages) and mathematically sophisticated, involving central idempotents, Wall finiteness obstructions, and rational surgery. The proof reads as internally coherent and technically detailed—precisely the kind of argument that could pass cursory review. However, the official solution proves the answer is NO, meaning OpenAI's proof must contain an error. The most likely location is the surgery realization step: the passage from a projective $\mathbb{Q}\Gamma$-Poincaré complex to an actual closed manifold requires careful control of the surgery obstruction groups, and the central involution $z = -1$ may introduce complications not fully addressed.

This is arguably the most important cautionary finding in the entire comparison: a current AI system produced a *convincing but incorrect* research-level proof. A convincing wrong proof is a more dangerous failure mode than an obviously broken one, because it may not be caught without expert review.

Weinberger's commentary further notes that Fowler's theorem shows all proof strategies based on finite complexes and Poincaré duality must fail—a constraint that applies to aspects of OpenAI's approach as well.

4.8 P08: Lagrangian smoothing

All three answer YES. The official solution uses an elegant conormal fibration approach. OpenAI provides a detailed 15-page proof with explicit local smoothing models at vertices and edges, careful gluing, and a Hamiltonian isotopy verification via flux vanishing. Our proof uses the Cartan formula for explicit Hamiltonian construction.

4.9 P09: Quadrilinear tensors

All three answer YES. The official solution and OpenAI both use 5×5 minors of mode-p flattenings (degree 5). Our proof uses Plücker equations (degree 4). OpenAI's proof is essentially the same construction as the official one, working over $\mathbb{C}$ first and then reducing to $\mathbb{R}$.

4.10 P10: CP-RKHS iterative solver

All three provide valid PCG methods. The official solution transforms via eigendecomposition of K. OpenAI provides a detailed matrix-free PCG with Kronecker preconditioning that *also* uses eigendecomposition of both K and $\Gamma = Z^T Z$—arguably the most complete treatment of the three. Our solution uses subsampled Kronecker matvecs with a diagonal preconditioner but omits the eigendecomposition step.

5 Discussion

5.1 Complementary failure modes

The most striking finding is that the two AI submissions fail on largely different problems:

Problem	OpenAI	Ours
P01 (Φ^4_3)	**Yes** (complete)	**Incorrect** (Hairer)
P03 (Macdonald)	**Wrong** (claims ill-posed)	**Yes**
P04 (Stam)	**Yes** (claims complete)	**Partial** ($n \leq 3$)
P06 (ε-light)	**Yes** ($c = 1/256$)	**Partial** (no universal c)
P07 (Lattices)	**Wrong** (claims Yes)	**Wrong** (wrong case)

This complementarity suggests that the two approaches have different strengths. OpenAI's single-system approach excels at problems requiring deep technical machinery within a single field (P04: hyperbolic polynomial theory; P06: BSS barriers) and produces polished, self-contained mathematical writing. It struggles with problems requiring careful interpretation of problem statements (P03) or where a sophisticated-looking construction masks a subtle error (P07).

Our multi-model iterative approach benefits from cross-checking between models and iterative hardening (P03: 4 sessions of refinement caught the positivity issue early), and explicitly separates verified algebraic content from axiomatized analytic input via Lean formalization. It struggles with problems requiring specific technical tools from adjacent fields that no model in the ensemble could discover (P04, P06).

5.2 The P07 puzzle

Problem 7 is the only shared failure, and it is instructive. The official answer is NO, but OpenAI constructs an elaborate 10-page argument claiming YES. Our proof establishes NO for a different case. The official commentary notes that Fowler's theorem proves all strategies based on finite complexes and Poincaré duality must fail—a fundamental obstruction that neither AI submission fully navigated.

OpenAI's error is particularly notable because the construction is mathematically sophisticated and internally coherent—it reads as a plausible research paper that could survive superficial review. This demonstrates that current AI systems can produce convincing but incorrect proofs at research level, a failure mode that is arguably more dangerous than producing obviously flawed arguments, and one that demands new verification practices as AI-assisted mathematics scales.

5.3 Cross-field tool discovery

The problems where AI submissions fail (P04, P06, P07) share a common feature: the official proof requires importing tools from an *adjacent* mathematical field not suggested by the problem statement. P04 requires hyperbolic polynomial convexity (real algebraic geometry → information theory); P06 requires BSS barriers (spectral sparsification → graph partitioning); P07 requires surgery theory with the Novikov conjecture (geometric topology → lattice theory).

OpenAI overcame this barrier for P04 and P06 but not P07. We overcame it for none of the three. This suggests that cross-field tool discovery remains a principal bottleneck, with single-system approaches having a slight advantage (perhaps due to broader training or longer reasoning chains).

5.4 Limitations

Several limitations should be noted. First, we cannot independently verify OpenAI's claimed proof for P04 (all n) without detailed mathematical review; it may contain subtle gaps analogous to their P07 error. More broadly, without the official solutions (released February 14), several of OpenAI's proofs could not have been assessed as correct or incorrect. Second, the comparison between OpenAI's single-system approach and our multi-model approach is confounded by differences in model capabilities, compute budget, and time invested—OpenAI discloses none of these parameters. Third, this paper was itself AI-assisted, and the author's assessment of mathematical correctness is necessarily limited by his non-expert status.

6 Conclusion

We have presented a three-way comparison of solutions to the ten First Proof challenge problems. The principal findings are:

1. Both AI submissions achieve 6–8/10 correct answers, but fail on largely different problems, suggesting complementary strengths.
2. OpenAI's single-system approach succeeds on two problems (P04, P06) where our multi-model approach fails, but produces two incorrect answers (P03, P07) that our approach avoids.
3. Problems 1 and 7 demonstrate the most important cautionary finding: current AI systems can produce elaborate, superficially coherent, but ultimately incorrect proofs at research level. For P01, the error was only caught by the problem author (Hairer) and his collaborator (Gubinelli); for P07, OpenAI's 10-page construction reaches the wrong answer. These failure modes are more dangerous than obviously broken arguments.
4. Two principal bottlenecks emerge: (a) cross-field tool discovery (P04, P06, P07), and (b) domain-specific regularity blindness—generating plausible but false claims about the technical properties of mathematical objects (P01).
5. Neither AI submission matches the official solutions in completeness and correctness across all ten problems.

References

[1] T. Kolda et al. First Proof: ten open problems in mathematics. `https://1stproof.org/`, 2026.

[2] M. Abouzaid et al. First Proof solutions and comments. Released February 14, 2026.

[3] OpenAI. First Proof? Submission, February 13, 2026.

[4] DeepMind. AI achieves silver-medal standard solving International Mathematical Olympiad problems. Blog post, July 2024.

[5] T. Trinh, Y. Wu, Q. Le, H. He, and T. Luong. Solving Olympiad geometry without human demonstrations. *Nature*, 625:476–482, 2024.

[6] B. Romera-Paredes et al. Mathematical discoveries from program search with large language models. *Nature*, 625:468–475, 2024.