Comparative Analysis of AI-Assisted and Human-Generated Solutions to the First Proof Challenge

Mark Dillerop*
Independent Researcher

February 14, 2026

Abstract

The First Proof challenge posed ten research-level mathematical problems spanning stochastic PDEs, representation theory, algebraic combinatorics, spectral graph theory, equivariant topology, manifold theory, symplectic geometry, algebraic geometry, and numerical linear algebra. We attempted all ten using an iterative multi-model AI workflow and submitted solutions within the eight-day competition window. This paper provides a systematic comparison with the official human-generated solutions released on February 14, 2026. We define formal assessment criteria, apply them to each problem, and report the following outcomes: five problems received correct answers with complete (though often methodologically distinct) proofs; one received a valid approach to an open-ended question; one received a correct answer with an incorrect proof (confirmed by the problem author); two received correct answers with incomplete proofs; and one addressed a different case than the one asked. We analyze the failure modes, compare our outputs with single-prompt LLM baselines from the organizers' internal testing, and identify two principal bottlenecks: cross-field tool discovery and domain-specific regularity blindness.

Disclosure. This paper was itself produced using the same human–AI collaborative workflow it describes. The author orchestrated the analysis; the AI systems drafted text, cross-referenced source files, and identified discrepancies. All factual claims were verified against the primary sources (our submitted proofs, the official solutions document [2], and the problem authors' commentary).

Contents

*Correspondence: `dillerop@gmail.com`

1 Introduction

The First Proof challenge [1] released ten open research-level mathematical problems on February 5, 2026, with a submission deadline of February 13. The problems were authored by leading researchers in their respective fields and were designed to test whether current AI systems could produce genuine mathematical proofs, not merely plausible-sounding arguments.

We attempted all ten problems using an iterative human–AI collaborative workflow (described in §3), submitting solutions between February 10 and 12. On February 14, the organizers released official solutions and commentary on LLM-generated outputs from internal testing [2]. The present paper provides a systematic, problem-by-problem comparison.

The author is not a mathematician and holds no domain expertise in any of the ten problem areas. This is a relevant datum: it means that the mathematical content of the submitted solutions was generated entirely by AI systems, with the author serving as workflow orchestrator rather than mathematical contributor. The implications of this arrangement are discussed in §6.

2 Related work

AI-assisted theorem proving has progressed rapidly in recent years. DeepMind's AlphaProof [3] solved four of six International Mathematical Olympiad 2024 problems using a reinforcement-learning approach coupled with Lean 4 verification. Trinh et al. [4] demonstrated that a neuro-symbolic system could solve Olympiad-level geometry problems. Romera-Paredes et al. [5] used LLM-guided search to discover new mathematical constructions in extremal combinatorics.

These systems operate in constrained, well-defined domains (competition problems, formal geometry, combinatorial search). The First Proof challenge differs in that its problems are *research-level*: they require novel proof strategies, draw on deep domain-specific machinery, and in several cases connect multiple mathematical fields. To our knowledge, no prior work has systematically compared AI-assisted solutions to research-level problems against expert human solutions across multiple mathematical disciplines.

The multi-model workflow we employ—using several LLMs in complementary roles—relates to recent work on LLM ensembles [6] and mixture-of-agents architectures [7], though our approach is orchestrated manually rather than automated.

3 Methodology

3.1 Workflow

Five AI systems were used: Claude Opus 4.6 (Anthropic), ChatGPT 5.2 Pro (OpenAI), Gemini 3.0 Deep Think (Google), Grok (xAI), and Perplexity. Each problem was addressed through 4–12 iterative sessions. A typical session proceeded as follows:

1. **Problem ingestion.** The problem statement and relevant context were provided to one or more models.
2. **Strategy generation.** Models proposed proof strategies; the author selected which to pursue based on the models' own assessments of feasibility.
3. **Proof development.** The selected model developed the proof in detail, with the author feeding intermediate outputs to other models for cross-checking.
4. **Hardening.** Subsequent sessions identified gaps, circularities, or errors in earlier drafts. Models were asked to attack their own proofs and fix identified issues.
5. **Verification.** Where feasible, numerical experiments (over 900,000 tests across all problems) and partial Lean 4 formalization (1,932 lines total) were used to validate claims.

The author's role was limited to orchestration: selecting which model to engage, routing outputs between models, and deciding when a proof was sufficiently hardened for submission.

The author did not set mathematical strategy, evaluate correctness, or decide proof pivots.

3.2 Timeline

Solutions were submitted as follows: February 10 (P03, P05, P07, P08, P09, P10—six problems); February 11 (P01, P02, P04, P06—four problems); February 12 (P06 resubmission). Per-problem time ranged from approximately 32 minutes (P10) to over 2 hours (P06).

3.3 Assessment criteria

We evaluate each solution along three dimensions.

Answer match. A solution has *answer match* if its stated conclusion agrees with the official answer. For problems admitting multiple valid answers (e.g., P10, which asks "explain how"), any valid method constitutes a match. Three levels: **Yes** (full agreement), **Partial** (correct direction, weaker claim), **Different** (addresses a different case).

Proof status. A proof is *complete* if it establishes the claimed result without logical gaps, even if the method differs from the official solution. It is *incomplete* if the argument has identified gaps (e.g., covers only special cases). It is *wrong case* if it proves a correct statement about a different mathematical object than the one asked about.

Methodological comparison. For each problem, we identify the principal proof technique used by the official solution and by our solution, and characterize the relationship: *identical*, *analogous* (same strategy, different execution), or *distinct* (fundamentally different approach).

4 Results

4.1 Summary

Tables 1 and 2 summarize the comparison across all ten problems.

Table 1: Answer comparison. **Match**: whether our answer agrees with the official answer (§3.3). **Our proof status**: whether the proof is complete, incomplete, or addresses the wrong case (§3.3).

#	Problem	Official answer	Our answer	Match	Our proof status
1	Φ^4_3 shift	No (singular)	No (singular)	**Yes**	**Incorrect** (Hairer)
2	Rankin–Selberg	Yes (constructive)	Yes (existential)	**Yes**	Complete
3	Markov–Macdonald	Yes (t-Push TASEP)	Yes (Hecke recursion)	**Yes**	Complete
4	Finite free Stam	Yes (all n)	Yes ($n \leq 3$ only)	**Partial**	Incomplete: gap for $n \geq 4$
5	$\mathcal{O}$-slice filtr.	Characterization thm	Characterization thm	**Yes**	Complete (+ Lean 4)
6	ε-light subsets	Yes, $c=\varepsilon/42$	Yes (no universal c)	**Partial**	Incomplete: no universal constant
7	Lattices	No (2-torsion)	No ($\delta=0$ only)	**Diff.**	Wrong case
8	Lagrangian smooth.	Yes	Yes	**Yes**	Complete
9	Quadrilinear tensors	Yes (minors, deg 5)	Yes (Plücker, deg 4)	**Yes**	Complete
10	CP-RKHS PCG	(Explain how)	PCG + subsampled Kron	**Yes**	Complete

4.2 Aggregate assessment

Applying the criteria of §3.3, we classify the ten outcomes into four categories.

Table 2: Methodological comparison (§3.3).

#	Problem	Official method	Our method
1	Φ^4_3	H_3 Wick cube, $(\log N)^{-\gamma}$ scaling, BG decomposition	Hairer's A_ψ functional, super-exponential $e^{-3n/4}$ scaling
2	Rankin–Selberg	*Constructive*: explicit W_0 via Godement–Jacquet	*Existential*: JPSS nondegeneracy + inertial class counting
3	Macdonald	t-Push TASEP (discrete-time, multiline queues)	Hecke-recursive detailed balance (continuous-time, T_i-recursion)
4	Stam	Jacobian contractivity via Bauschke et al. hyperbolic poly. convexity	Cumulant decomposition, finite free heat equation ($n{\le}3$ only)
5	Slice filtr.	Isotropy separation + induction on subgroup lattice	Nearly identical; adds Lean 4 formalization (0 `sorry`)
6	ε-light	BSS barrier + leverage score control ⇒ $c{=}\varepsilon/42$	Multi-bin greedy: $c{=}1/6$ (sparse), $c{=}1/2$ (K_n), no universal c
7	Lattices	Surgery theory + symmetric signatures + Novikov conjecture	Euler characteristic contradiction ($\delta{=}0$ only)—*wrong case*
8	Lagrangian	Conormal fibration (elegant, coordinate-free)	Hamiltonian via Cartan formula (computational, coordinate-based)
9	Tensors	5×5 minors of flattenings (degree 5, Tucker rank)	Plücker equations (degree 4, geometric perspective)
10	CP-RKHS	PCG + eigendecomposition of K (better conditioning)	PCG + subsampled Kronecker matvecs (no eigendecomposition)

Complete proof with correct answer (5/10). Problems P02, P03, P05, P08, and P09. In each case, our proof reaches the correct conclusion via a valid argument, though the proof technique often differs substantially from the official solution (Table 2). Of note: P05 nearly reproduces the official proof structure and additionally includes Lean 4 formalization with zero `sorry` axioms; P09 provides a construction of lower degree (4 vs. 5); and P03 offers a genuinely distinct construction based on Hecke operators rather than the t-PushTASEP.

Valid approach to open-ended question (1/10). Problem P10 asks "explain how to use PCG," not for a yes/no answer. We provided a valid preconditioned conjugate gradient method using subsampled Kronecker matrix-vector products and a diagonal preconditioner. The official solution additionally transforms the problem via eigendecomposition of the kernel matrix K, yielding better conditioning. Kolda [2, §4.10] noted that the best internally tested LLM solution was "correct and better" than her provided solution; this refers to the organizers' internal LLM testing, not to our submission.

Correct answer, incorrect proof (1/10). Problem P01 (Hairer): We correctly answered NO (the measures are mutually singular), but Martin Hairer (the problem author) has confirmed that our proof is incorrect. Specifically, the first non-trivial claim—that the renormalized cubic is $O(1)$—is wrong. In the Φ^4_3 setting, the Wick-ordered cube $: u^3 :$ is a distribution living in a negative Sobolev/Besov regularity space, not a bounded quantity. This invalidates the proof from its first substantive step.

Correct direction, incomplete proof (2/10).

- **P04** (Srivastava): We correctly answered YES and proved the result for $n \leq 3$, developing 12 theorems and 20 propositions supported by over 900,000 numerical tests. However, the official proof establishes the result for *all* n using the Bauschke–Güler–Lewis–Sendov theorem [9] on convexity of eigenvalue functions of hyperbolic polynomials. Our proof has a genuine gap for $n \geq 4$.
- **P06** (Spielman): We correctly answered YES and proved several partial cases ($c = 1/6$ for sparse graphs, $c = 1/3$ for bounded degeneracy, $c = 1/2$ for K_n). The official proof

provides a universal constant $c = \varepsilon/42$ via BSS barrier functions [10]. We could not close the general case.

Divergent—wrong case addressed (1/10). Problem P07 (Weinberger) asks specifically about **2-torsion** in the fundamental group. The official solution proves the answer is NO using surgery theory, symmetric signatures in $L(\mathbb{R}\pi)$, and the Novikov conjecture for lattices. Our proof establishes NO for a *different* case ($\delta(G) = 0$, covering groups of vanishing fundamental rank) and leaves the $\delta(G) \neq 0$ case open—which is precisely where the 2-torsion case resides. Weinberger's commentary [2, §4.7] further notes that Fowler's theorem shows *all* proof strategies based on finite complexes and Poincaré duality must fail for this problem—exactly the approach we employed.

5 Problem-by-problem comparison

5.1 Problem 1: Φ^4_3 measure equivalence under smooth shifts

Official solution (Hairer). The official proof constructs a separating set

$$B_\gamma := \Big\{ u \in \mathcal{D}' : \lim_{N\to\infty} (\log N)^{-\gamma} \big\langle H_3(P_N u;\, c_N) + 9\, c_{N,2}\, P_N u,\ \psi \big\rangle = 0 \Big\}$$

using the Wick cube H_3 with polynomial-in-log scaling $(\log N)^{-\gamma}$. The key mechanism is that the mass renormalization constant $c_{N,2} \gtrsim \log N$ (Lemma 4.3 of the official solution), so the shifted measure produces a divergent term $9(\log N)^{-\gamma}\, c_{N,2}\, \langle \psi_N, \psi \rangle \to \infty$. The proof uses the Barashkov–Gubinelli decomposition $u = \Upsilon - \Xi + v$ (Proposition 4.1) and is self-contained with four detailed lemmas (4.2–4.5).

Our solution. We obtained the correct answer (**No**: $\mu \perp T^*_\psi \mu$ for any nonzero smooth ψ) and constructed a separating set using Hairer's 2022 note functional A_ψ with super-exponential mollification ($\varepsilon_n = e^{-e^n}$) and scaling $e^{-3n/4}$. The dominant mechanism is a deterministic mass-shift term $-b\, e^{n/4} \|\psi\|^2_{L^2}$ arising from the unmollified field.

We additionally analyzed the HKN separating set $A^{\alpha,\gamma}$ [11] and proved it does *not* separate μ from $T^*_\psi \mu$: the smooth shift is absorbed into the regular remainder. This analysis is absent from the official solution.

Assessment. Both proofs target mutual singularity via a separating set, but with different regularization schemes (polynomial-in-log vs. super-exponential). However, Hairer has confirmed that our proof is **incorrect** (see below).

Author feedback. Hairer [2, §4.1]: "The best solution produced during testing was by ChatGPT 5.2 Pro [. . .] it essentially just quotes the content of [Hai22] without giving a detailed proof."

Direct feedback from Hairer on our submission (February 15, 2026). "I had a look at your answer to Q1. 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 living in a negative Sobolev/Besov regularity space (roughly $\mathcal{C}^{-1/2-\varepsilon}$), not a bounded function. Claiming it is $O(1)$ is a fundamental misunderstanding of the regularity of the objects involved, and this error invalidates the proof from its first substantive step. This is a stark illustration of how AI-generated proofs can produce statements that sound reasonable but are false in the specific regularity regime of the problem—and how a non-expert orchestrator cannot catch such errors.

Additional feedback from Gubinelli (February 15, 2026). "Besides the problem highlighted by Hairer, it seems the models induced themselves into believing that the regularisation of the mass term was the key point (Section 6 in the solution PDF) which is actually irrelevant (in the original note it was chosen like so for convenience reasons). In Section 7 there is a (correct) remark that the variance of the renormalised cube actually diverges, which is not-consistent with the model's belief that the renormalised cube is $O(1)$."
This reveals a deeper structural problem: the AI models not only made a false claim about regularity, but also **misidentified the mechanism** of the proof. They fixated on the mass renormalization term as the key driver of singularity, when in fact it was chosen for convenience in Hairer's original note and is not the essential ingredient. Worse, the models' own Section 7 contains a correct observation (divergent variance of $: u^3 :$) that directly contradicts their Section 6 claim ($: u^3 :$ is $O(1)$)—the models failed to notice this internal inconsistency in their own proof.

Verdict: Yes answer, incorrect proof (Hairer and Gubinelli confirm fundamental errors). **Correct answer, wrong proof.**

5.2 Problem 2: Rankin–Selberg universal test vectors

Official solution (Nelson). The proof uses the Godement–Jacquet functional equation as its main tool. It constructs an explicit Whittaker function $W_0(g) = \int_{N_n} \mathbf{1}_{K_n}(xg)\,\psi(x)\,dx$ and relates the u_Q-twisted Rankin–Selberg integral to an integral over $K_1(\mathfrak{q})$ via Schwartz–Bruhat functions $\beta, \beta^\sharp$ and their Mellin transforms (Lemma 5). The final result is an explicit formula $\ell_{\mathrm{RS}}(s,\, u_Q W_0,\, d_Q V) = c\,|Q|^{-n/2}$, a monomial in $|Q|^s$, hence nonzero for all s. The proof is six pages and completely self-contained.

Our solution. We obtained the same u_Q-twist formula

$$W\big(\mathrm{diag}(g,1)\,u_Q\big) = \psi^{-1}(Q\,g_{nn})\,W\big(\mathrm{diag}(g,1)\big)$$

and used the same monomial-in-q^{-s} strategy. For universality, we used an existential argument: JPSS nondegeneracy combined with the observation that the "bad locus" B_π depends only on the *inertial* equivalence class $[\pi]$ (orbit under unramified twists), and inertial classes are countable. A $\mathbb{C}$-vector space cannot be a countable union of proper subspaces, yielding the desired W.

For $n = 1$ ($\mathrm{GL}_2 \times \mathrm{GL}_1$), we gave a fully rigorous argument via Gauss sums, matching the official solution's treatment of this base case.

Assessment. The official proof is *constructive* (explicit W_0, explicit integral computation); ours is *existential* (dimension counting over inertial classes). Both are mathematically valid. Our inertial class observation—that B_π depends only on $[\pi]$, not on π itself—is a genuine insight not present in the official proof. The official commentary notes that the best LLM attempt reduced to the same key integral but failed on the nonvanishing step; our proof handles this correctly.

Author feedback. Nelson [2, §4.2]: "The best attempt [. . .] reduced to the same key integral but failed on the nonvanishing step." LLMs constructed W depending on π (wrong—universality requires independence). Our proof correctly handles universality via the inertial class countability argument.

Verdict: Yes answer, Yes proof (different strategy). **Correct with complete proof.**

5.3 Problem 3: Markov chain for interpolation Macdonald polynomials

Official solution (Ben Dali–Williams). The official proof constructs the *interpolation t-Push TASEP*, a sophisticated discrete-time Markov chain with three steps per transition: ring a bell (select a position), activate particles clockwise, and displace via vacancy return. The proof uses classical and signed two-line queues with pair weights and ball weights, establishing stationarity via an analogue of [8, Theorem 4.18]. The key identity is $F^*_\nu(x;1,t) = \sum_\eta F^*_\eta(x;1,t)\, P(\eta,\nu)$: the F^*-polynomials are left eigenvectors of the transition matrix. The proof spans seven pages with detailed combinatorial machinery.

Our solution. We constructed the *Hecke-Recursive Detailed Balance Chain*, a continuous-time reversible Markov chain on the Cayley graph of S_n. The construction uses three elementary ingredients:

1. A product of shifted powers: $w_\lambda = \prod_j (x_j - t^{1-j})^{\lambda_j}$.
2. The Hecke operator T_i (standard algebraic object).
3. The Cayley graph of S_n with generators $\{s_1, \dots, s_{n-1}\}$.

Transition rates are $r(\mu \to s_i\mu) = c_i \cdot w_{s_i\mu}(x;t)$, and stationarity follows from detailed balance (trivial by commutativity of the weight assignment).

We proved the product formula $E^*_\lambda(x;q,t) = \prod_j \prod_{k=0}^{\lambda_j-1}(x_j - q^k t^{1-j})$ via the Knop–Sahi vanishing characterization, with computational verification for $n = 2$ and $n = 3$.

Assessment. The constructions are completely different: the official uses a multi-step particle system; ours uses Hecke recursion with detailed balance. Both are nontrivial in the required sense (transition rates not defined using F^*_μ or P^*_λ directly). Our construction is arguably simpler but less connected to the combinatorial literature on multiline queues. The official commentary notes that LLMs gave Metropolis–Hastings (trivial) or confused the problem with the non-interpolation version—our construction avoids both pitfalls.

Author feedback. Williams [2, §4.3]: "Both Gemini and ChatGPT produced Metropolis–Hastings chains" (trivial, since they use the target distribution directly). "Gemini confused the problem with the non-interpolation version." Our Hecke-recursive construction avoids both pitfalls and satisfies the nontriviality constraint.

Verdict: Yes answer, Yes proof (genuinely different construction). **Correct with complete proof.**

5.4 Problem 4: Finite free Stam inequality

Official solution (Garza Vargas–Srivastava–Stier). The official proof is *complete for all n* and follows a three-step strategy à la Blachman:

1. **Score vectors as derivatives** (Lemma 1.1): under reverse heat flow $p_t = p \boxplus_n \sqrt{t}\,\mathrm{He}_n$, the roots satisfy $\alpha_i'(0) = J_n(\alpha)_i$.
2. **Jacobian contractivity** (Proposition 2.1): $\|J_{\boxplus_n}(u,v)\|^2 \le \|u\|^2 + \|v\|^2$ for $u, v \perp \mathbf{1}_n$.
3. **Blachman's argument**: choose $a = 1/\|J_n(\alpha)\|$, $b = 1/\|J_n(\beta)\|$, and combine.

The key insight is the use of the Bauschke–Güler–Lewis–Sendov theorem [9] on convexity of eigenvalue functions of hyperbolic polynomials: the Hessians $H^{(i)}_{\boxplus_n}$ satisfy $\sum c_i H^{(i)} \succeq 0$ for ordered c_i, via hyperbolicity of the finite free convolution polynomial.

Our solution. We obtained the correct answer (**Yes**) and proved the inequality for $n = 2$ (exact equality) and $n = 3$ (explicit formula). We developed extensive supporting theory:

- Finite free heat equation (Theorem 6): $dp_t/dt = -p_t''/2$.
- Root velocity equals score (Corollary 6.1): $\lambda_i'(t) = p''(\lambda_i)/(2p'(\lambda_i))$.
- Finite de Bruijn identity (Theorem 7): $d/dt \log \Delta(p_t) = \Phi_n(p_t)$.
- Φ_n monotonicity (Theorem 8): $\Phi' = -2\sum_{i<j}(h_i - h_j)^2/(\lambda_i - \lambda_j)^2 \le 0$.
- J-concavity for $n = 3$ (Theorem 11): $J(t) = 1/\Phi(p_t)$ is concave.

In total: 12 theorems, 20+ propositions, 900,000+ numerical tests with zero violations. The general case remained open: the gap was the generalized concavity of a remainder term $r(\sigma)$ along contraction paths.

Assessment. This is the largest gap between our work and the official solution. The official proof is complete; ours covers only $n \leq 3$. Our approach (cumulant decomposition, explicit remainder analysis) is fundamentally different from theirs (Jacobian contractivity via hyperbolic polynomial theory).

The missing ingredient—the Bauschke et al. convexity theorem—connects real algebraic geometry to information theory in a non-obvious way. Despite extensive search (8+ sessions), we did not discover this connection. However, our finite free heat equation, de Bruijn identity, and score–root identities are genuine new results not present in the official solution.

The official commentary describes exactly the difficulty we encountered: the LLM tried Blachman's approach but could not find the joint probability space analogue.

Author feedback. Srivastava [2, §4.4]: "The LLM tried Blachman's approach but couldn't find the joint probability space analogue." This is exactly the difficulty we encountered. The official authors resolved it via the Bauschke–Güler–Lewis–Sendov theorem [9] on hyperbolic polynomial convexity—a connection between real algebraic geometry and information theory that we did not discover despite 8+ sessions of exploration.

Verdict: Yes answer, Partial proof ($n \leq 3$ only; official proof is complete for all n). **Correct direction, incomplete proof.**

5.5 Problem 5: Equivariant $\mathcal{O}$-slice filtration

Official solution (Blumberg–Hill–Lawson). The official proof defines the $\mathcal{O}$-slice filtration via localizing subcategories generated by $G_+ \wedge_H N_T S^1$ for admissible T, and introduces the characteristic function $\chi_{\mathcal{O}}(H) = \min\{K : K \to H\}$. The main result (Theorem 2.7) states: $E \in \tau^{\mathcal{O}}_{\geq n}$ if and only if $[H : \chi_{\mathcal{O}}(H)] \cdot \mathrm{gconn}(E)(H) \geq n$ for all H. The forward direction uses Lemma 2.3 (restriction to generators); the reverse uses isotropy separation, Lemma 2.5, Lemma 2.6 (geometric Mackey functors), and induction on the subgroup lattice.

Our solution. Our Theorem 3.1 states: X is $\mathcal{O}$-slice $\geq n$ if and only if $\Phi^H X$ is $(\lfloor n/|H| \rfloor - 1)$-connected for all $H \in \mathcal{F}_{\mathcal{O}}$. The forward direction uses strong induction on $|H|$ and the Wirthmüller isomorphism; the reverse uses isotropy separation and induction on the subgroup lattice.

We additionally produced a Lean 4 formalization (0 `sorry`) verifying the combinatorial and arithmetic skeleton: transfer system axioms, dimension bookkeeping, and the strong induction structure.

Assessment. This is the *closest match* to the official solution among all ten problems. Both proofs use isotropy separation and induction on the subgroup lattice. The connectivity bound formulas differ slightly in notation ($[H : \chi_{\mathcal{O}}(H)]$ vs. $|H|$ with restriction to $\mathcal{F}_{\mathcal{O}}$) but are equivalent when $\chi_{\mathcal{O}}(H) = e$.

Both use the same key lemmas: Wirthmüller isomorphism, compactness of slice cells, geometric fixed point computation. The official commentary notes that LLMs got the statement essentially correct but proofs were "sketched or slightly garbled"—our proof is fully detailed.

Our Lean 4 formalization adds a layer of verification not present in the official release.

Author feedback. Blumberg [2, §4.5]: "The LLMs got the statement essentially correct but proofs were 'sketched or slightly garbled'." Our proof is fully detailed with all steps, and our Lean 4 formalization (0 `sorry`) adds a layer of verification not present in any other submission.

Verdict: Yes answer, Yes proof (nearly identical structure, plus Lean 4). **Correct with complete proof.**

5.6 Problem 6: ε-light subsets

Official solution (Spielman). The official proof is *complete* with constant $c = 1/42$ (i.e., $|S| \geq \varepsilon n/42$). It uses a greedy construction, adding vertices one at a time, controlled by a modified BSS barrier function [10]:

$$\Phi_\sigma^u(A) = \sum_{i=1}^{\sigma} \frac{1}{u - \lambda_i(A)}$$

tracking only the top-σ eigenvalues. The key lemma (Lemma 1.2) shows that at each step, one can find a vertex t such that both the barrier Φ remains bounded and the leverage score $\ell(S) \leq 4|S|$ is maintained. The proof uses the Ky Fan eigenvalue inequality and the Sherman–Morrison–Woodbury formula.

Our solution. We obtained the correct answer (**Yes**) but with an *incomplete* proof. Our proved results include:

- $c \leq 1/2$ (tight upper bound, via disjoint cliques).
- $c = 1/6$ for sparse graphs ($\bar{d} \leq 6/\varepsilon - 1$).
- $c = 1/2$ for complete graphs K_n (direct spectral argument).
- $c = 1/3$ for graphs with degeneracy $< \lceil 3/\varepsilon \rceil$.
- Extensive computational verification: multi-bin greedy with $k = \lceil 2/\varepsilon \rceil$ bins never gets stuck across 16+ graph families, $n \leq 100$, $\varepsilon \in \{0.1, \ldots, 0.5\}$.

The gap: we could not prove that greedy always finds a bin with $\mu_{\max} < 1$ at every step.

Assessment. The official proof uses a BSS-type barrier from spectral sparsification—a specific technique from theoretical computer science that we did not employ. Our approach (multi-bin greedy with effective resistance analysis) is different and more computational.

The official commentary notes that Gemini's proof was vague and GPT-Pro only gave the upper bound $c \leq 1/2$. Our partial results (four proved theorems, extensive computation) go well beyond both LLM attempts.

Author feedback. Spielman [2, §4.6]: "Gemini asserted that it presented a proof [...] But, after some correct statements, it presented a very vague explanation of how the proof could be finished. To me, it seems unlikely that the approach can be turned into a correct proof." ChatGPT 5.2 Pro only offered the upper bound $c \leq 1/2$. Our partial results (four proved theorems, log-det barrier, Barbell graph identification, extensive computation across 16+ graph families) go substantially beyond both LLM attempts.

Verdict: Yes answer, Partial proof (no universal constant; official proves $c = \varepsilon/42$). **Correct direction, incomplete proof.**

5.7 Problem 7: Lattices and $\mathbb{Q}$-acyclicity

Official solution (Cappell–Weinberger–Yan). The official proof addresses the specific case of 2-torsion and proves the answer is **No**. The argument reduces (WLOG) to $\Gamma = \pi \rtimes \mathbb{Z}_2$ with an involution on $M = K\backslash G/\pi$. It uses symmetric signatures in $L(\mathbb{R}\pi)$, the Novikov conjecture (true for lattices via assembly map injectivity), and equivariant cobordism. The key contradiction: for Y (free $\mathbb{Z}_2$-action on a $\mathbb{Q}$-acyclic manifold), the symmetric signature image is 0; for M (isometric action on the locally symmetric space), the image is nonzero because the fixed set is aspherical.

Our solution. We proved partial results covering different cases:

- **Case 1** ($\delta(G) = 0$, all torsion): Euler characteristic contradiction. $\mathbb{Q}$-acyclicity forces $\chi(M) = \chi_{\mathbb{Q}}(\Gamma) \neq 0$ via the Cartan–Leray spectral sequence and Wall's theorem, while L^2-Betti number vanishing forces $\chi(M) = 0$.
- **Case 2** ($G \cong \mathrm{SL}(2, \mathbb{C})$, $d = 3$): dimension-forcing plus Perelman's geometrization gives asphericity, forcing Γ torsion-free—contradiction.
- **Case 2b** ($d = 4$): vacuous—no connected real semisimple G without compact factors has $\delta(G) \neq 0$ and $\dim(G/K) = 4$.
- **Case 3** ($\delta(G) \neq 0$, $d \geq 5$): open. A thorough analysis of surgery-theoretic obstructions reveals no obstruction.

Assessment. The official proof handles the 2-torsion case directly using surgery theory and symmetric signatures—a completely different approach from our Euler characteristic argument. Our $\delta(G) = 0$ case is correct but uses an obstruction that vanishes when $\delta \neq 0$.

The official commentary identifies a specific false lemma in LLM proofs (about Lefschetz numbers) and notes that Fowler's theorem shows proofs using only finite complex plus Poincaré duality must fail. This is exactly the barrier we encountered in Case 3.

Our analysis of the $d = 4$ vacuousness (checking all rank-2 groups) is a genuine contribution.

Author feedback. Weinberger [2, §4.7]: "All proofs by AI's I've seen only use finite complex and Poincaré duality." He shows this approach is fundamentally doomed: Fowler's theorem constructs a space $M_3 \times (K\backslash G/\Gamma_0 \times E\Delta)/\Delta$ (where M_3 is any closed hyperbolic 3-manifold) that has the rational type of a finite complex, satisfies rational Poincaré duality, and has fundamental group $\pi_1(M_3) \times \Gamma$—a lattice in $\mathrm{SO}(3, 1) \times G$. This proves that *all* proof strategies based solely on finite complexes and Poincaré duality must fail. Our proof fell into exactly this trap for the $\delta(G) \neq 0$ case.

Weinberger also identifies a specific false lemma in LLM proofs: "The counterexample is $\mathbb{R}^1$ and f is a translation. It has no fixed points, but its Lefschetz number in their sense is -1."

Verdict: Answer No for $\delta = 0$, but does not cover the specific 2-torsion case asked by the problem. **Divergent—wrong case covered.**

5.8 Problem 8: Polyhedral Lagrangian smoothing

Official solution (Abouzaid). The official proof answers **Yes** with an elegant global approach built on the notion of a *conormal fibration*.

- **Local:** Lemma 1 establishes a vertex normal form (the same linear algebra as in our proof). Smoothing functions $\mathcal{S}(\Sigma)$ are defined for each vertex star.
- **Global:** Lemma 8 shows that a conormal fibration L_z (a family of Lagrangian planes parametrized by $z \in K$) exists. Lemma 9 shows that smoothing functions of small C^1-norm exist.
- **Assembly:** graphical Lagrangians from smoothing functions yield the smooth family $\{K_t\}$; the Hamiltonian isotopy follows from the graphical description.

The proof is five pages and avoids explicit coordinate computations in the global argument.

Our solution. We obtained the correct answer (**Yes**) with a more computational approach:

- Same vertex normal form (Lemma 1)—essentially identical linear algebra, including the spanning argument via isotropic subspaces.

- Local smoothing via cotangent bundle identification and explicit bump-function interpolation.
- Hamiltonian isotopy via the Cartan formula: $H_t = \dot{S}_t - \lambda(V_t) \circ \iota_t$.
- Session 4 hardening fixed 11 issues including sign conventions, cyclic order, and topological isotopy convergence.

Assessment. Both proofs start with the same local analysis (vertex normal form) and arrive at the same conclusion. The official proof is more elegant globally, using conormal fibrations to avoid coordinate computations; ours is more explicit, constructing the Hamiltonian directly.

The official commentary notes that LLMs identified the local smoothing correctly but failed on global gluing—specifically, the compatibility of local choices across shared edges. Our proof addresses this compatibility, though with more computational effort than the official approach.

The official solution's conormal fibration framework generalizes more readily to other symplectic manifolds.

Author feedback. Abouzaid [2, §4.8]: "The best two solutions produced during testing both correctly identified the existence of a local smoothing near every vertex; the proof uses essentially the same basic linear algebra argument that appears in the human solution. The proof then proceeds to perform a local-to-global argument" but fails on compatibility of local choices across shared edges. Our proof successfully handles this global compatibility, avoiding the "exactness" traps and coordinate compatibility errors that defeated other AI models.

Verdict: **Yes** answer, **Yes** proof (valid but more computational; less elegant than official). **Correct with complete proof.**

5.9 Problem 9: Quadrilinear determinantal tensors

Official solution (Miao–Lerman–Kileel). The official proof answers **Yes** by taking $\mathbf{F}$ to be the collection of all 5×5 minors of the four $3n \times 27n^3$ matrix flattenings of the block tensor Q. The key observation is a Tucker decomposition $Q = C \times_1 A \times_2 A \times_3 A \times_4 A$ with $C_{abcd} = \operatorname{sgn}(abcd)$ (Lemma 1), giving multilinear rank $\leq (4, 4, 4, 4)$, so all 5×5 minors of any flattening vanish. The "only if" direction uses a three-step argument: normalize λ so that $\lambda_{\alpha 111} = \lambda_{1\beta 11} = \cdots = 1$, then show that λ entries with two ones equal some constant c, with one one equal c^2, and with zero ones equal c^3—hence λ is rank-1. Genericity is verified computationally (random numerical instances).

Our solution. We constructed a different polynomial map $\mathbf{F}$ consisting of six families of Plücker equations $\mathcal{P}_{pq}$, one for each pair of positions $\{p, q\} \subset \{1, 2, 3, 4\}$. Each equation has degree 4 (vs. degree 5 for the official minors).

- **Necessity:** the Plücker identity plus rank-1 λ gives factorization of each equation.
- **Sufficiency:** factorization into $\Lambda_P \cdot (\Lambda_{S_1} - \Lambda_{S_2}) \cdot [Q\text{-bracket}]$; genericity via explicit witness cameras.
- Numerically verified for $n = 5, 6$.

Assessment. Both constructions are valid polynomial maps satisfying all three required properties (independence from A, bounded degree, rank-1 characterization for generic A). The official approach is more natural from the tensor decomposition perspective (Tucker rank); ours is more geometric (Plücker relations on determinantal varieties).

Our degree bound (4) is tighter than the official (5), which is a minor advantage. The official commentary notes that the best LLM answer was "essentially correct" using the same 5×5 minors approach—our Plücker approach is a genuinely different alternative.

Author feedback. Kileel [2, §4.9]: "The best answer was essentially correct"—using the same 5×5 minors approach as the official solution. Our Plücker-elimination approach is a genuinely different alternative with a tighter degree bound (4 vs. 5), providing an independent verification from a geometric rather than tensor-algebraic perspective.

Verdict: Yes answer, Yes proof (alternative construction with tighter degree bound). **Correct with complete proof.**

5.10 Problem 10: CP-RKHS iterative solver

Official solution (Brust–Kolda). The official proof designs a preconditioned conjugate gradient (PCG) solver for the mode-k RKHS subproblem of CP-HiFi tensor decomposition. The key innovation is a *change of variables* via the eigendecomposition $K = UDU^T$: defining $\bar{F} = S^T(Z \otimes UD)$ and solving for $\bar{W} = U^T W$ yields a better-conditioned system. Three lemmas establish efficient matrix–vector products via row-wise Kronecker structure:

- $Cx = (A * BX)\mathbf{1}_r$ at cost $O(q(r+n))$.
- $C^T v = \mathrm{vec}(B^T \operatorname{diag}(v)\, A)$.
- $\operatorname{diag}(C^T C) = \mathrm{vec}((B * B)^T (A * A))$.

A diagonal preconditioner is constructed from $\operatorname{diag}(\bar{F}^T \bar{F}) + \lambda(I_r \otimes D) + \rho I_{rn}$. Total cost: $O(qn^2 + qr^2 + qnrp)$ per CG iteration, vs. $O(qn^2r^2 + n^3r^3)$ for direct Cholesky.

Our solution. We proved $\mathbf{H}$ is symmetric positive definite (Proposition 2.1), designed an efficient matrix–vector product via subsampled Kronecker structure (reshaping $\mathbf{v} = \mathrm{vec}(V)$ and computing $(Z \otimes K)^T SS^T (Z \otimes K)\,\mathbf{v}$ column-by-column), and proposed a diagonal preconditioner. Complexity analysis was provided.

Assessment. Both solutions use PCG with efficient matrix–vector products avoiding $O(N)$ computation, and both propose diagonal preconditioners. The official solution additionally transforms the system via eigendecomposition of K—this is the "advanced" step that Kolda noted she would be impressed if AI could produce.

Both use the row-wise Kronecker structure for efficient matvecs. The official commentary notes that the best LLM solution was "correct and better than the solution I provided," citing the subsampled Kronecker matvec idea from prior literature (arXiv:1601.01507).

Author feedback. Kolda [2, §4.10]: "The best LLM solution was correct and better than the solution I provided" —referring to the internally tested LLM (Feb 4–5), not our submission. Kolda noted she would be "impressed if the AI can" discover the eigendecomposition transformation; our solution does *not* include this step. Our subsampled Kronecker matvec approach draws on prior literature (arXiv:1601.01507) and is a valid PCG method, but does not achieve the full transformation the official solution provides.

Verdict: Yes answer, Yes proof (valid PCG approach; does not include the eigendecomposition transformation). **Correct—valid approach to an open-ended question.**

6 Discussion

6.1 Strengths of the AI-assisted approach

In five of ten problems, the iterative multi-model workflow produced complete, correct proofs. Several of these exhibit methodological independence from the official solutions: P03 constructs a Hecke-recursive chain rather than a t-PushTASEP; P09 uses Plücker equations of degree 4 rather than 5×5 minors of degree 5; and P02 gives an existential argument via inertial class counting rather than an explicit construction. This methodological diversity suggests that the AI systems are not merely reproducing known approaches but can synthesize novel proof strategies within a given mathematical field.

6.2 Failure mode analysis

The four problems with incorrect, incomplete, or divergent results (P01, P04, P06, P07) reveal two distinct failure modes. The first is **cross-field tool discovery**. Three problems share a common structural feature: the official proof requires importing a specific tool from an *adjacent* mathematical field that is not suggested by the problem statement itself.

- **P04** required the Bauschke–Güler–Lewis–Sendov theorem [9], connecting real algebraic geometry (hyperbolic polynomials) to information theory (entropy power inequalities).
- **P06** required BSS barrier functions [10] from spectral sparsification theory, applied to a graph partitioning problem.
- **P07** required surgery-theoretic tools (symmetric signatures, the Novikov conjecture for lattices, equivariant cobordism) applied to a question about lattice topology.

This pattern suggests that cross-field tool discovery is one principal bottleneck for current AI systems in research-level mathematics.

The second failure mode is illustrated by P01: **domain-specific regularity blindness**. Hairer confirmed that our proof's first non-trivial statement—that the renormalized cubic is $O(1)$—is wrong. The Wick-ordered cube in Φ^4_3 is a distribution in a negative regularity space, not a bounded function. This error is characteristic of AI systems generating plausible-sounding mathematical statements that are false in the specific technical regime of the problem. Critically, a non-expert orchestrator has no way to catch such errors, and the AI systems themselves did not flag the claim as problematic during cross-checking. Gubinelli further observed that the models misidentified the mechanism of the proof (fixating on mass renormalization, which was chosen for convenience) and that the models' 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. This represents a failure mode that is arguably more dangerous than the cross-field discovery gap, because the proof reads as superficially coherent.

6.3 Comparison with single-prompt LLM baselines

The official commentary [2, §4] describes outputs from Gemini 3.0 Deep Think and ChatGPT 5.2 Pro tested on February 4–5, 2026 in single-prompt mode. The comparison is informative:

- On P01, single-prompt LLMs quoted Hairer's note without proof or assumed absolute continuity $\mu \sim \mu_0$ (false).
- On P02, LLMs constructed test vectors depending on π, violating the universality requirement.
- On P03, LLMs proposed Metropolis–Hastings chains (trivial, not satisfying the problem's nontriviality condition) or confused the interpolation and non-interpolation Macdonald polynomials.
- On P05, LLMs stated the correct theorem but proofs were "sketched or slightly garbled" [2, §4.5].
- On P07, all single-prompt LLM proofs contained false lemmas.

Our iterative multi-session approach avoided these failure modes in every case where the single-prompt baselines failed, suggesting that sustained multi-turn reasoning with cross-model verification is a significant factor in proof quality. However, on the three problems where we produced incomplete results (P04, P06, P07), the single-prompt baselines also failed, indicating that these problems pose fundamental challenges for current LLM architectures regardless of prompting strategy.

6.4 Role of the human operator

A distinctive feature of this work is that the human operator (the author) is not a mathematician. The author's contribution was purely organizational: selecting which AI model to engage, routing outputs between models, and maintaining workflow momentum. Mathematical strategy, correctness evaluation, and proof pivots were handled entirely by the AI systems.

This raises the question of whether the results should be attributed to the AI systems, the human orchestrator, or the collaborative process. We take no position on this question but note that the same AI systems, when used in single-prompt mode by the challenge organizers, produced substantially weaker results (§6). The iterative multi-model workflow appears to be a necessary component, not merely a convenience.

6.5 Limitations of this analysis

Several limitations should be noted. First, our assessment of proof completeness is based on our own reading of the proofs and the official authors' commentary; we have not submitted our proofs for independent peer review. Second, the comparison with single-prompt baselines is indirect: the organizers tested different model versions on different dates under different conditions. Third, this paper was itself produced using AI assistance (as disclosed in the abstract), which introduces the possibility of systematic blind spots in self-assessment.

7 Conclusion

We have presented a systematic comparison of AI-assisted solutions to the ten First Proof challenge problems with the official human-generated solutions. The principal findings are:

1. Five of ten problems received correct answers with complete proofs, often via methodologically distinct approaches.
2. The four problematic results reveal two failure modes: (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).
3. Iterative multi-model collaboration avoided the failure modes observed in single-prompt LLM baselines, but did not overcome either bottleneck.
4. A non-expert human operator can orchestrate AI systems to produce research-level mathematical proofs, but cannot independently verify correctness. The P01 error was only caught by the problem author himself.

An initial automated review categorized 8/10 of our solutions as correct or superior. On critical re-examination against the official solutions and direct author feedback, this assessment is significantly inflated: the honest count is 5/10 complete, 1/10 valid approach, 1/10 incorrect proof (correct answer), 2/10 incomplete, and 1/10 divergent. We report this discrepancy as a cautionary note on the reliability of AI self-assessment in mathematical contexts.

References

[1] M. Abouzaid, A. Blumberg, M. Hairer, J. Kileel, T. Kolda, P. Nelson, D. Spielman, N. Srivastava, R. Ward, S. Weinberger, and L. Williams. First Proof: Ten research-level mathematical problems. `https://1stproof.org/`, arXiv:2602.05192, 2026.

[2] M. Abouzaid et al. First Proof solutions and comments. Released February 14, 2026. `https://codeberg.org/tgkolda/1stproof/`

[3] DeepMind. AI achieves silver-medal standard solving International Mathematical Olympiad problems. Blog post, July 2024. `https://deepmind.google/discover/blog/ai-solves-imo-problems-at-silver-medal-level/`

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

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

[6] S. Jiang, D. Kadosh, and N. Shazeer. Towards large reasoning models: A survey of reinforced learning for mathematical reasoning. arXiv:2501.09686, 2025.

[7] J. Wang et al. Mixture-of-agents enhances large language model capabilities. arXiv:2406.04692, 2024.

[8] A. Ayyer, J. Martin, and L. Williams. The inhomogeneous t-PushTASEP and Macdonald polynomials at $q = 1$. *Ann. Inst. Henri Poincaré D*, 2025.

[9] H. Bauschke, O. Güler, A. Lewis, and H. Sendov. Hyperbolic polynomials and convex analysis. *Canad. J. Math.*, 53(3):470–488, 2001.

[10] J. Batson, D. Spielman, and N. Srivastava. Twice-Ramanujan sparsifiers. *SIAM J. Comput.*, 41(6):1704–1721, 2012.

[11] M. Hairer, S. Kusuoka, and H. Nagoji. Singularity of solutions to singular SPDEs. arXiv:2409.10037, 2024.

[12] H. Ben Dali and L. Williams. A combinatorial formula for interpolation Macdonald polynomials. arXiv:2510.02587, 2025.