Performance Analysis and Design Validation

Evaluation of Agentic vs. Human-in-the-Loop Proof Strategies

Dietmar Wolz, Ingo Althöfer

February 14, 2026

Abstract

This document reviews how well a fully-automatic agentic AI performed against official solutions in the "First Proof" math challenge (Questions 1-10). It compares the AI-team performance with a proposed "Human-in-the-Loop" design. The results show that while teams of AI agents are excellent at solving algorithmic problems, they require a human manager to guide the meaning of the problem to prevent the AIs from inventing false logic.

1 Introduction

The "First Proof" challenge was a stress test to see if automatic AI-systems could handle complex, research-level mathematics. Our AIs were put in a "multi-agent" strategy that succeeded at calculation tasks but often struggled to understand the subtle meanings in abstract math problems.

The results indicate that a "Human-in-the-Loop" system is necessary for the current state of the art.

This hybrid system is divided into three layers:

1. **The Manager (Human):** Breaks down the problem and defines what it means.
2. **The Generator (one or several agents):** Uses software and search tools to create the solution.
3. **The Verifier (Human/Formal):** Acts as a strict filter to check for accuracy and logical consistency.

This report compares the AI's answers to the official solutions to show where the lack of a human manager caused failures, and where the AI generator performed better than human experts.

2 Summary of Comparisons (Questions 1–10)

Question 1: Mutual Singularity of Φ^4_3

Status: Structurally Correct but Rigor Deficient

- **Comparison:** Both proofs correctly identify the divergence of the renormalization counterterm as the driver of singularity.
- **Divergence:** The AI solution relied on a "citation trap," referencing an unpublished note by Martin Hairer to assert a probability bound rather than deriving it from the SPDE decomposition.

- **Technical Nuance:** The agent used a **super-exponential scale** ($\varepsilon_n = \exp(-e^n)$) to force convergence heuristically, whereas the rigorous proof required **logarithmic scales** and delicate Borel-Cantelli estimates.
- **Analysis:** This failure validates the need for the **Reference Checker** module described in [1]. An autonomous agent treats a citation as a "truth token," whereas a HITL system would flag the reference as insufficient.

Question 2: Generic Representations of GL_{n+1}

Status: Logical Failure (Quantifier Trap)

- **Comparison:** The problem required a *single* Whittaker function W effective for *all* representations π ($\exists W \forall \pi$). The AI constructed a W dependent on π ($\forall \pi \exists W$).
- **Divergence:** The solution solved an easier, different problem. The official solution used explicit newvector theory to construct a universal test vector.
- **Analysis:** This is a breakdown in **Problem Decomposition**. Without a human manager to parse the quantifier hierarchy, the agent optimized for the path of least resistance.

Question 3: Stationary Distribution of Interpolation TASEP

Status: Correct Trivial Solution (Metropolis-Hastings)

- **Comparison:** The AI "reverse-engineered" a Markov chain using Metropolis-Hastings and algebraic exchange relations.
- **Divergence:** The official solution sought a "natural" interacting particle system (the Push TASEP). The authors categorized Metropolis-Hastings approaches as "trivial."
- **Analysis:** Agents lack "mathematical taste." The **Strategy Selection** phase in [1] requires human intuition to direct the search toward structural (interacting particles) rather than computational (sampling) solutions.

Question 4: Finite Free Stam Inequality

Status: Heuristic Gap (Blachman's Argument)

- **Comparison:** The AI correctly analyzed the Gaussian case but attempted to generalize Blachman's projection argument to the free setting without rigorous construction.
- **Divergence:** The AI "hallucinated" the existence of a non-commutative conditional expectation. The official solution used a Jacobian spectral bound derived from the geometry of **hyperbolic polynomials**.
- **Analysis:** This illustrates the danger of *Analogical Reasoning* without verification. The agent assumed a classical tool had a direct analogue.

Question 5: Slice Connectivity of G-Spectra

Status: Essentially Correct

- **Comparison:** The AI correctly identified the filtration via geometric fixed points and the dimension of slice cells.
- **Divergence:** The AI left the connectivity bound as an abstract optimization problem, while the official solution solved it explicitly.

- **Analysis:** A strong success for the autonomous strategy in retrieving high-level category theory, though it stopped short of the final explicit calculation.

Question 6: Spectral Barrier for Graph Partitioning

Status: Conditional Proof (Missing Lemma)

- **Comparison:** Both solutions used a greedy algorithm driven by a spectral barrier.
- **Divergence:** The AI proved the upper bound (1/2) correctly but assumed a "Mass τ-control" hypothesis to close the lower bound. The official solution used boundary leverage scores to prove this step.
- **Analysis:** The agent correctly identified the "hard part" (the barrier condition) but lacked the creative leap to invent a new combinatorial invariant (leverage scores).

Question 7: Acyclic Universal Covers

Status: Misinterpretation (Rational vs. Integral)

- **Comparison:** The AI proved that if $\widetilde{M}$ is *integrally* acyclic, the group is torsion-free.
- **Divergence:** The question allowed $\widetilde{M}$ to be merely *rationally* acyclic. This flexibility permits torsion (e.g., Fowler's construction), rendering the AI's obstruction invalid.
- **Analysis:** A fatal semantic error. The human manager's role in [1] is specifically to disambiguate such definitions before the agent begins formalization.

Question 8: Lagrangian Smoothing

Status: Global Gluing Error

- **Comparison:** The AI correctly modeled the local smoothing via generating functions.
- **Divergence:** The AI assumed local Hamiltonians could simply be summed. The official solution utilized a "conormal fibration" to handle the geometric obstruction.
- **Analysis:** The agent failed to recognize the global geometric obstruction, a common weakness in "patching" arguments generated by LLMs which tend to assume linearity.

Question 9: Identifiability of Rank-1 Tensors

Status: Correct (Optimal Approach)

- **Comparison:** The AI correctly identified the 5×5 flattening minors and used a rigorous tangent space/Lie Algebra argument.
- **Validation:** Confirmed by the official solution as the "best AI-generated solution."
- **Analysis:** Validates the **Code/Algebra Execution** component. When the problem is reducible to algebraic constraints, the agent operates at a superhuman level of precision.

Question 10: Matrix-Free Tensor Completion

Status: Superior to Human Solution

- **Comparison:** The AI proposed a Matrix-Free Preconditioned Conjugate Gradient (PCG) method with $O(qnr)$ complexity.
- **Validation:** The author explicitly stated this was superior to the official solution and planned to adopt the AI's matrix-free MVP approach.
- **Technical Nuance:** While the AI's MVP was superior, its choice of a Kronecker preconditioner was computationally heavier ($O(n^3)$ setup) than the reference's simple diagonal preconditioner. The AI "won" on the solver strategy but was suboptimal on the preconditioner choice.
- **Analysis:** The ultimate validation of the agent as a "Generator." It searched the space of numerical algorithms more effectively than the human expert.

3 Design Validation: The Necessity of Human-in-the-Loop

The performance data above provides a clear delineation of responsibilities for the architecture proposed in *Agentic Strategy Design for Math Proofs* [1].

3.1 The Manager: Semantics and Scope

The failures in **Q2** (Quantifiers) and **Q7** (Rational vs. Integral) were not failures of calculation, but of definition. The autonomous agent [2] rushed to solve the "most likely" version of the problem based on training data.

> *Design Implication:* The **Manager** must explicitly parse the problem statement, resolving ambiguities and locking in definitions (e.g., "Note: Acyclicity is over $\mathbb{Q}$, not $\mathbb{Z}$") before the agent is permitted to generate proof steps.

3.2 The Verifier: Rigor and Citations

The failure in **Q1** (Hairer citation) and **Q5** (Geometric Spectra) highlights the "Hallucination of Authority."

> *Design Implication:* The **Reference Checker** tool proposed in [1] is essential. It must force the agent to expand citations into self-contained arguments or flag them for human review.

3.3 The Generator: Algorithmic Superiority

The successes in **Q9** and **Q10** prove that the agentic system is not merely an assistant, but a superior optimizer for specific sub-tasks.

> *Design Implication:* The HITL architecture should not micro-manage the **Generator** during algorithmic search. Once the problem is correctly framed (e.g., "Minimize complexity for Tensor Completion"), the agent should be given autonomy to search the solution space, as it found the $O(qnr)$ solution that the human expert missed.

4 Conclusion

The experiment shows that the independent AI strategy is powerful but unpredictable. Its performance swings between superhuman insight (Question 10) and basic misunderstandings (Question 7). The proposed design bridges this gap by putting the creative AI inside a strict, human-managed framework. The near (and medium ?!) future of AI in research mathematics relies on a partnership where humans provide the *meaning* and agents provide the *scale*.

References

[1] Wolz, D. (2026). *Agentic Strategy Design for Math Proofs*. Available at `https://althofer.de/agentic_strategy_design_for_math_proofs.pdf`. Accessed February 12, 2026.

[2] Wolz, D. (2026). "Multi-Agent Strategy for Solving Research-Level Mathematics." `https://althofer.de/first-proof-competition/first-proof-report.html`