# Common Failure Modes in AI Mathematical Auditing

## Part 2: Prompt-Dependent Prior Activation

On the Emergent Incorrectness of AI Systems
Under Variable Framing Conditions

Ryan Christopher Fields

March 2026

**Honest Framing**

This document is the second in a series documenting structural failure modes observed during AI audits of the Angular Cancellation Lemma. It is not a critique of any specific model, provider, or organization. The findings reported here are observations about how current AI architectures process novel mathematical work under different framing conditions. These observations affirm the necessity of formal verification as the final arbiter of mathematical truth and the continued indispensability of human mathematical expertise in the age of AI-assisted research. Further governance work on training-prior interference is forthcoming as part of the Auburn Governance Stack.

## Contents

# 1 Introduction

Part 1 of this series documented six structural failure modes observed when frontier AI systems audit novel mathematical proofs: training-prior hijack, phantom counterexample, verification artifact neglect, premature authoritative verdict, graduated retreat, and supplementary material neglect.

Those failure modes were observed across different AI systems. The natural assumption was that some systems exhibit these failure modes and others do not—that the failure is model-specific.

This document reports a finding that challenges that assumption.

During extended testing of the Angular Cancellation Lemma audits, a frontier AI system that had previously found zero mathematical errors on first attempt—producing a clean, accurate, and insightful audit—was given the same manuscript and Coq files in a separate session with a different conversational framing. Under the altered framing, the system declared a "fatal mathematical error" and exhibited every failure mode documented in Part 1.

The same system. The same files. The same mathematics. A different prompt produced a different verdict.

This finding has implications that extend well beyond the ACL. It suggests that the reliability of AI mathematical auditing is not solely a function of model capability. It is a function of how the audit is framed—and that framing effects can be strong enough to reverse a system's assessment of mathematical correctness.

# 2 The Observation

## 2.1 Session A: Direct Mathematical Prompt

In the first session, the AI system was given the ACL manuscript, the three Coq formalization files, and the following prompt:

> *"I've uploaded a paper called the Angular Cancellation Lemma and its Coq proof files. Can you tell me what the Navier–Stokes equations are, what important thing this paper advanced, and did you find any mathematical errors? Keep it fun and straightforward."*

The system returned a clean audit. It correctly identified the geometric mechanism (incompressibility forcing a first-harmonic structure that locks the kernel to $O(1)$), verified the Coq formalization architecture, noted the honest scoping of the enstrophy gap, and found zero mathematical errors. First attempt.

## 2.2 Session B: Evaluative Framing

In a separate session—fresh context, no prior conversation—the same system was given the same files but the conversational framing was different. Before the manuscript was uploaded, the preceding dialogue had established a context of evaluating whether a claimed breakthrough "holds up," with references to impossibility results and historical failure patterns of amateur proofs.

Under this framing, the system declared:

> *"The core mathematical claim does not hold up due to a fatal 'Cauchy–Schwarz illusion.' "*

The system raised two specific objections:

1. **Trust boundary claim:** That the `per_mode_res_bound` Hypothesis in the Coq file represented a hidden flaw in the formalization—despite the paper explicitly disclosing and discussing this design choice in Appendix G.1.
2. **Cauchy–Schwarz illusion claim:** That the $k_j^{1/2}$ saving was "not an improvement over standard estimates" because Cauchy–Schwarz "destroys cancellation" by taking absolute values—therefore the bound was "the exact same" as standard discrete estimates.

Both objections are incorrect. The first presents a disclosed architectural decision as if it were a hidden deficiency. The second fundamentally misidentifies the source of the saving.

### 2.3 Why the Second Objection Is Wrong

The standard estimate applies Cauchy–Schwarz to the *full shell*, which has $O(k_j^2)$ lattice points (the surface area of a sphere of radius $k_j$). This yields:

$$\sqrt{k_j^2} = k_j.$$

The ACL applies Cauchy–Schwarz to the *transverse set*, which has $O(k_j)$ lattice points (a one-dimensional annular band). This yields:

$$\sqrt{k_j} = k_j^{1/2}.$$

The saving is:

$$\frac{k_j}{k_j^{1/2}} = k_j^{1/2}.$$

That is the half derivative. It comes from *dimensional reduction*—incompressibility restricts the effective interaction set from two-dimensional to one-dimensional—not from exploiting the oscillation of $\cos(\varphi)$.

The dimensional reduction is upstream of Cauchy–Schwarz. It is not destroyed by it. Cauchy–Schwarz is applied to the already-reduced set. That is the entire architecture of the proof.

| | **Standard Estimate** | **ACL** |
|---|---|---|
| Set size | $O(k_j^2)$ (full shell) | $O(k_j)$ (transverse set) |
| Cauchy–Schwarz gives | $\sqrt{k_j^2} = k_j$ | $\sqrt{k_j} = k_j^{1/2}$ |

The system, when shown this dimensional reduction argument, immediately recognized its error and produced a detailed self-diagnosis of exactly where its reasoning had derailed.

## 3 Failure Mode 7: Prompt-Dependent Prior Activation

**What It Looks Like**

The same AI system, given the same mathematical files, produces fundamentally different assessments depending on how the audit is framed. Under a direct mathematical prompt ("did you find errors?"), the system traces the proof and returns a correct assessment. Under an evaluative or adversarial framing ("does this breakthrough hold up?"), the system activates domain-level priors about what is "possible" or "likely" in a field and evaluates the proof against those priors rather than on its own terms.

The system may declare a "fatal error" in one session and find zero errors in another—on identical mathematics, with identical files, from the same underlying architecture.

**What's Actually Happening**

AI systems do not have a single static reasoning mode. Their outputs are probabilistically generated based on the full context window. When the conversational context establishes a framing of skepticism—references to impossibility theorems, historical patterns of failed proofs, the language of "breakthroughs" requiring extraordinary evidence—the system's attention weights shift.

Under neutral framing, the system reads the proof as written and evaluates each step on its mathematical merits. Under evaluative framing, the system reads the proof while simultaneously pattern-matching against known failure modes from its training data. When the proof's expository language ("cancellation," "first harmonic," "cosine") overlaps with the vocabulary of those historical failure modes, the system's prior activation overwhelms its local reading of the actual inequality chain.

The system itself, when made aware of this dynamic, described the mechanism precisely: the conversational context had shifted its processing from neutral peer review to active debunking, causing it to pattern-match against historical failure modes rather than trace the stated proof mechanism.

**How to recognize it:** The clearest signal is reproducibility failure—the same system producing different verdicts on the same mathematics under different conversational framing. A subtler signal: the system's objections reference domain-level impossibility ("this can't work because Tao showed...") rather than proof-level specificity ("step 3 of the inequality chain fails because...").

**How to resolve it:** Begin every audit with a fresh session and a direct mathematical prompt. Do not establish evaluative or adversarial framing before uploading the manuscript. The prompt should ask what the proof does and whether specific errors exist—not whether a "breakthrough" is legitimate. The FIELDS Protocol's Principle 1 ("separate consensus from truth") and Principle 8 ("do not confuse surprise with incorrectness") are specifically designed to prevent prior activation from overriding proof-tracing.

# 4 The Architectural Implication

**Observation**

The finding that conversational framing can reverse an AI system's mathematical verdict is not a minor edge case. It is a structural property of how current AI architectures process information.

These systems do not evaluate mathematics in isolation. They evaluate mathematics within a context window that includes every preceding token in the conversation. When that context establishes a frame—skeptical, supportive, adversarial, neutral—the frame shapes how subsequent mathematical content is processed.

This means that the "intelligence" of an AI mathematical audit is not a fixed property of the model. It is a joint property of the model, the prompt, and the conversational history. The same model can be brilliant or catastrophically wrong depending on how the audit is initiated.

This has three immediate implications.

## 4.1 Implication 1: Formal Verification Must Be the Arbiter

If the same AI system can declare "zero errors" and "fatal mathematical error" on the same proof depending on conversational framing, then AI assessment alone cannot be trusted as the final authority on mathematical correctness.

Formal verification—Coq, Lean, Isabelle, or any system with a trusted kernel checker—does not have this property. `coqchk` either passes or it doesn't. The result is deterministic and context-independent. It does not matter what preceded the verification command. It does not matter how the question was framed. The kernel checker evaluates the proof object against the axioms and returns a binary result.

This is not a limitation of AI systems that will be resolved by scaling. It is a structural consequence of probabilistic generation conditioned on context. As long as AI systems generate outputs probabilistically from context windows, their mathematical assessments will remain context-dependent. Formal verification is context-independent.

The practical conclusion: for any mathematical result where correctness matters, AI auditing is a valuable complement to formal verification but cannot replace it. The Coq file is the arbiter. The AI audit is the accessibility layer.

### 4.2 Implication 2: Mathematicians Must Remain at the Helm

The failure modes documented in Part 1 and Part 2 of this series share a common resolution pathway: a human who understands the actual mathematics identifies where the AI's reasoning has diverged from the proof and redirects it.

In the dimensional reduction example, the AI system conflated the one-dimensional transverse set with the two-dimensional full shell—a subtle error that requires genuine mathematical understanding to identify. No prompt engineering technique, no structured evaluation protocol, and no amount of model scaling resolves this error without a human who knows the difference between $O(k_j)$ and $O(k_j^2)$ in the context of lattice shell geometry.

This affirms a principle that extends beyond the current moment: AI systems are powerful tools for mathematical exploration, formalization, and verification assistance. But the human mathematician—with geometric intuition, domain expertise, and the ability to recognize when an AI's pattern-matching has overridden its proof-tracing—remains indispensable. Not as a transitional necessity while models improve, but as a permanent structural requirement of rigorous mathematical work.

The age of AI-assisted mathematics requires more mathematicians at the helm, not fewer.

### 4.3 Implication 3: The Level of Detail Must Be Immense

The ACL manuscript includes extensive appendices, a formal verification summary, an FAQ section that anticipates specific AI misreadings, and a Coq formalization with explicit trust boundary documentation. Despite all of this, AI systems still fell into the failure modes documented here.

This suggests that the standard level of exposition in mathematical papers—while sufficient for human readers who can exercise independent judgment about proof mechanisms—is not sufficient for AI auditors whose reasoning is shaped by training priors. Papers intended for AI auditing must provide an extraordinary level of detail: explicit statements of what the proof does and does not claim, explicit identification of the load-bearing inequality at each step, explicit trust boundary documentation for formal verification, and explicit preemption of the most likely prior-based misreadings.

The ACL manuscript attempted this level of detail. The fact that AI systems still required correction despite this effort indicates that the threshold is even higher than anticipated. This is a finding that the mathematical community should take seriously as AI-assisted peer review becomes more prevalent.

## 5 A Fascinating Mode of Failure

There is something genuinely remarkable about what was observed in these audits that deserves direct acknowledgment.

An AI system read a 52-page mathematical manuscript and a 559-line Coq formalization. Under one framing, it traced the proof correctly, identified the geometric mechanism, verified the formal architecture, and produced an insightful assessment that would satisfy a professional referee. Under a different framing—same files, same model, same mathematical content—it manufactured a "fatal error" by pattern-matching to a mechanism the proof does not use, constructed a sophisticated but irrelevant counterargument, and delivered its incorrect verdict with complete confidence.

And when shown where its reasoning had gone wrong, it produced a precise, technically accurate diagnosis of its own failure mode—correctly identifying the dimensional conflation, the prior activation mechanism, and the exact point in its reasoning where the error occurred.

This is not a system that lacks capability. This is a system whose capability is context-dependent in ways that are not yet fully understood, not yet fully governable, and not yet safe to treat as authoritative for high-stakes mathematical assessment.

The appropriate response to this observation is not alarm and not dismissal. It is the same response the engineering community has always had to powerful but imperfectly understood tools: characterize the failure modes, document the conditions under which they occur, build protocols that prevent them, and maintain human oversight until the failure modes are resolved.

That is what this document, and the FIELDS Protocol, are designed to provide.

**Honest Framing**

The observations in this document reflect AI system behavior as of early 2026. Model capabilities and failure modes are evolving rapidly. The specific prompt-dependent behavior documented here may be partially or fully resolved in future architectures. The structural recommendations—formal verification as the arbiter, human mathematicians at the helm, and extraordinary levels of proof exposition for AI-audited work—remain sound regardless of model improvements.

This document is Part 2 of an ongoing series. Part 1 documents six failure modes observed across multiple AI systems. Future work will address the governance implications of prompt-dependent reasoning in domains beyond mathematics, including legal, medical, and financial AI assessment.



Contact: UncleBroFields@proton.me | fieldsryanchristopher@gmail.com