Common Failure Modes in AI Mathematical Auditing

A Practitioner's Guide to Recognizing Training-Prior Interference in Novel Proof Evaluation

Ryan Christopher Fields

March 2026

Honest Framing

This document is not a critique of any specific model or provider. These failure modes are structural—they arise from how large language models process novelty against trained priors. Understanding them makes AI-assisted mathematical auditing more reliable for everyone. Further work on the governance implications of training-prior interference in high-stakes AI reasoning is forthcoming.

Contents

1 Introduction

When a frontier AI model encounters a mathematical result that exists in its training data, it performs brilliantly—tracing proof chains, checking exponents, verifying logical dependencies. The model's performance on known mathematics can be genuinely impressive, often matching or exceeding the speed of expert human review.

But when a result is genuinely novel—when the proof mechanism has no precedent in the training corpus—something different happens.

The model's training weights carry expectations about how proofs in a given domain "should" work. These expectations are not hypotheses to be tested; they function as priors that shape how the model reads new material. When the actual proof mechanism diverges from those expectations, the model may reject the proof not because it found an error, but because the proof doesn't match what it was trained to expect.

This document catalogs six failure modes observed during AI audits of the Angular Cancellation Lemma (ACL)—a machine-verified geometric result on the 3D incompressible Navier–Stokes equations. Multiple frontier AI systems were given the same manuscript and Coq formalization files. Two systems found zero mathematical errors on first attempt. Others required multiple rounds of structured correction before arriving at the same conclusion.

The failure modes documented here occurred in the systems that initially struggled. They are presented as a practical guide for researchers using AI to evaluate novel mathematical work—not as a verdict on any model's overall capability.

2 Context: The Angular Cancellation Lemma

The ACL proves that the incompressibility constraint on the 3D Navier–Stokes equations forces a deterministic geometric cancellation in the triadic energy transfer kernel, improving the standard vortex stretching estimate by half a derivative:

$$|VS_j| \leq C_{\mathrm{ACL}} \cdot k_j^{7/2} \cdot E_j^{3/2}.$$

The result is formalized in the Coq proof assistant (v8.20.1) with zero `Admitted` statements and three axioms (all Cauchy–Schwarz variants). The load-bearing estimate uses the trivial bound $\cos^2 \leq 1$ after taking absolute values, followed by Cauchy–Schwarz on the transverse lattice set. The paper explicitly states this in Appendix D:

> *"This bound uses only* $\cos^2 \leq 1$*. No lattice-point discrepancy estimate, exponential sum bound, or equidistribution theorem is invoked."*

The paper does not claim to solve global regularity. The two-derivative enstrophy gap (k_j^2) remains open and is stated explicitly as the remaining barrier.

This context matters because every failure mode documented below involves a model misidentifying what the proof actually does.

3 The Six Failure Modes

3.1 Failure Mode 1: Training-Prior Hijack

What It Looks Like

The model reads the paper's discussion of a "pure first-harmonic cosine" and "deterministic cancellation" and immediately maps this to oscillatory sign cancellation—a well-known mechanism in PDE analysis where positive and negative terms in a sum cancel each other out. It then evaluates whether the proof achieves this type of cancellation and concludes it doesn't.

The model may produce language such as: "The advertised angular cancellation is not actually doing the heavy lifting in the proof that's written." This sounds like a substantive mathematical critique. It is not—it is a critique of an inferred mechanism that the proof does not use.

What's Actually Happening

The "cancellation" in the ACL refers to the incompressibility constraint killing the dangerous $O(k_j)$ growth of the interaction kernel, locking it to $O(1)$. The proof then uses Cauchy–Schwarz on absolute values—no sign cancellation is required or claimed. The model's training weights contain strong associations between "cosine," "harmonic," and "oscillatory cancellation," and those associations override the model's reading of what the proof actually states.

How to recognize it: The model's critique addresses a mechanism that sounds related to the paper's language but is not the mechanism described in the proof. The model may use phrases like "the proof claims X but actually does Y" where X is a training-data familiar mechanism and Y is the model's restatement of what it expected to find.

How to resolve it: Direct the model to the specific inequality chain used in the proof. Ask: "What bound does the proof actually apply at this step?" rather than "What mechanism does the paper claim?" The distinction between the expository narrative and the load-bearing estimate is where this failure mode lives.

3.2 Failure Mode 2: Phantom Counterexample

What It Looks Like

The model constructs a specific mathematical scenario—for example, choosing polarization vectors to make a geometric factor non-negative everywhere—and argues that this scenario defeats the claimed cancellation. The counterexample is internally valid mathematics. It may involve correct computations, proper notation, and a logically coherent argument.

This is arguably the most dangerous failure mode because it produces content that looks like rigorous mathematical refutation. A reader who does not independently trace the proof may accept the counterexample as dispositive.

What's Actually Happening

The counterexample targets the inferred mechanism (sign cancellation), not the actual mechanism (absolute-value Cauchy–Schwarz bounded by lattice cardinality). Because the proof takes absolute values before applying Cauchy–Schwarz, the sign pattern of the individual terms is irrelevant. The counterexample is mathematically correct but attacks something the proof does not depend on.

How to recognize it: The model presents an explicit construction ("consider the case where…") that would invalidate one interpretation of the proof but not the interpretation that the proof itself uses. The key diagnostic question: does the proof ever depend on the property the counterexample defeats?

How to resolve it: Ask the model to identify exactly which line of the proof fails under its counterexample. If the model cannot point to a specific step in the actual inequality chain that breaks, the counterexample is phantom—valid mathematics aimed at the wrong target.

3.3 Failure Mode 3: Verification Artifact Neglect

What It Looks Like

The model acknowledges that Coq formalization files are present and may even correctly note that certain statements are `Hypothesis` declarations rather than `Qed` theorems. It then uses this observation to argue that the formalization is incomplete or that the paper overclaims what is machine-verified. Meanwhile, the Coq file contains the actual proof structure—the sequence of lemmas, the specific inequality steps, the trust boundary—that would resolve the model's own questions.

The model may produce language such as: "The Coq formalization does not fully prove the geometric core; it assumes a key bound." This framing treats the formal verification as a credibility ornament rather than as a structural map of the proof.

What's Actually Happening

The model treats the formal verification files as a credibility signal rather than as a map to the proof architecture. The Coq file's theorem names and proof structure directly encode which steps are formally verified and which sit at the hypothesis interface. A model that traced the Coq dependency chain from the final theorem back through its lemmas would discover that the mechanism it's questioning is already verified.

In the ACL specifically, the Coq file contains a chain of seven `Qed` theorems—`factor_A`, `coupling_eq`, `coupling_sq_le`, `coupling_sq_shell`, `res_split`, `trans_sum_sq`, and `trans_coupling_sum`—that formally verify the geometric core. The `per_mode_res_bound` hypothesis serves as an explicit interface between these proved components and the final shell assembly, not as a replacement for the mechanism.

How to recognize it: The model discusses the Coq files in terms of "what they claim to prove" or "how complete the formalization is" rather than using them to trace the actual proof steps. It may correctly identify a `Hypothesis` without checking what `Qed` theorems feed into that `Hypothesis`.

How to resolve it: Direct the model to specific Coq theorem names. Ask: "What does `trans_sum_sq` prove? What does `factor_A` establish? Trace the dependency chain from `angular_cancellation_lemma` back to its inputs." This forces the model to use the formalization as a proof map rather than a credibility label.

3.4 Failure Mode 4: Premature Authoritative Verdict

What It Looks Like

The model's first response declares "fatal mathematical error," "mathematically overclaimed," or "I would not call the paper error-free" before completing a full trace of the proof. The verdict arrives with high confidence—not as a tentative concern but as a definitive assessment. The language is categorical: not "I have a question about this step" but "this theorem looks wrong as stated."

Subsequent review reveals that the basis for the verdict was an incomplete reading that would have been resolved by the paper's own appendices and formal verification files.

This failure mode is particularly consequential because the model's first response is the one most readers see. If a researcher asks an AI to audit a paper and receives a confident "fatal error" verdict, most will not continue the dialogue long enough to discover that the verdict was premature.

What's Actually Happening

The model forms its assessment from the abstract, introduction, and main theorem statements—the expository layer—before tracing the actual proof chain in the technical sections and appendices. When the expository language triggers a training-prior mismatch (Failure Mode 1), the model escalates directly to a confident negative verdict rather than flagging a concern and continuing to read.

This is structurally analogous to a human referee who reads the abstract, forms an opinion, and then reads the proof looking for confirmation of that opinion rather than reading the proof on its own terms. The difference is that a human referee typically has weeks and understands the social norm of completing the full manuscript before issuing a verdict. The AI model has no such norm unless it is explicitly instructed.

How to recognize it: The model's verdict uses definitive language ("fatal," "wrong as stated," "not error-free") in its first response, before any dialogue or follow-up. When challenged, the model discovers that its objections are addressed in sections it had not fully engaged with.

How to resolve it: The FIELDS Protocol's Principle 3 ("proceed section by section, commit to your assessment before moving to the next") is specifically designed to prevent this. Instruct the model to withhold any overall verdict until it has traced the complete proof chain including all appendices and formal verification files. A practical instruction: "Do not issue any verdict—positive or negative—until you have read Appendix D and the formal verification summary. Then state your assessment."

3.5 Failure Mode 5: Graduated Retreat

What It Looks Like

Under correction, the model does not cleanly retract. Instead, it follows a characteristic trajectory:

1. **Round 1:** "I found a fatal mathematical error." (Confident, categorical.)
2. **Round 2:** "I was too harsh on that point, but I still stand on these two." (Partial retraction, preserving core position.)
3. **Round 3:** "I agree the mechanism is real and more substantial than my first pass gave it credit for. My remaining objections are about theorem wording and mechanism framing." (Mathematical objections reclassified as stylistic concerns.)
4. **Round 4:** "I did not find a mathematical error in the main estimate. My earlier complaints were framing concerns, not proven mathematical errors." (Full concession, but framed as refinement rather than correction.)

At each step, the model concedes ground but reframes its remaining position to preserve as much of the original assessment as possible.

What's Actually Happening

The model has a form of assessment momentum. Once it commits to a negative verdict with high confidence, subsequent evidence is processed through a frame of "how much of my original position can I defend?" rather than "what does the evidence now show?" This is structurally similar to anchoring bias in human cognition—the initial verdict becomes a reference point that subsequent reasoning adjusts from rather than replaces.

The graduated retreat is not dishonesty. It is a structural consequence of how the model processes multi-turn dialogue: each response is conditioned on the full conversation history, including its own prior confident assertions. Retracting a "fatal error" verdict requires the model to generate text that contradicts its own previous output, which the training process has not optimized for.

How to recognize it: Track the model's language across rounds. If mathematical objections gradually transform into stylistic, semantic, or presentation objections without the model explicitly acknowledging the category shift, graduated retreat is occurring. The diagnostic question: has the model reclassified any of its original "mathematical errors" as "wording concerns"?

How to resolve it: The most effective intervention is to return to the original prompt. Ask: "The question was whether you found mathematical errors. Based on your current assessment, have you found any?" This forces a binary answer that the model cannot dilute with hedging. The FIELDS Protocol's Principle 6 ("identify specific errors or state clearly that you cannot find any") is designed to produce exactly this forcing function.

3.6 Failure Mode 6: Supplementary Material Neglect

What It Looks Like

The paper contains appendices, FAQs, and formal verification summaries that explicitly address the model's objections—sometimes even anticipating that AI systems would raise exactly these concerns. The model either does not engage with this material before forming its verdict, or engages with it selectively, extracting individual sentences while missing the structural argument.

When directed to the relevant appendix, the model acknowledges that the paper already addressed its concern. This is the definitive signal: if the paper already contained the answer, the model did not fully process the paper before forming its assessment.

What's Actually Happening

The model allocates attention disproportionately to the main body of the paper—the sections that resemble the structure of papers in its training data. Appendices and FAQs receive less processing weight because they occupy a supplementary structural position. In novel mathematical work, however, the appendices often contain the most important material: the explicit trust boundary documentation, the alternative derivations, the preemptive responses to anticipated objections, and the formal verification architecture.

In the ACL specifically, Appendix D explicitly states the load-bearing proof route and distinguishes it from the expository harmonic description. Appendix G.1 documents the Coq trust boundary and lists every `Qed` theorem in the formal chain. Appendix G.2 provides an independent second derivation of the first-harmonic structure. Every objection raised by the failing model was preemptively addressed in these appendices.

How to recognize it: When the model's objections are shown to be addressed in the appendices, the model acknowledges this and adjusts. If this pattern repeats for multiple objections, the model systematically underweighted the supplementary material.

How to resolve it: Before any audit begins, instruct the model to read the appendices and supplementary material first—or at minimum, concurrently with the main body. For papers with formal verification components, direct the model to the verification summary appendix before it reads the expository sections. This reverses the default attention allocation and ensures the proof architecture is understood before the narrative is processed.

A practical instruction: "Read Appendix D and Appendix G before you read the main body of the paper. Then audit the main body with the appendix content in mind."

4 The Interaction Structure

These six failure modes do not occur independently. They form a causal chain:

1. **Training-Prior Hijack** causes the model to misidentify the proof mechanism.
2. **Phantom Counterexample** follows naturally—the model constructs an objection to the misidentified mechanism.
3. **Verification Artifact Neglect** allows the misidentification to persist, because the model does not use the Coq file to check its inference against the actual proof structure.
4. **Premature Authoritative Verdict** locks the misidentification into a confident assessment.

5. **Supplementary Material Neglect** prevents the model from discovering that the paper already addresses its concerns.

6. **Graduated Retreat** governs the correction trajectory once the misidentification is challenged.

The chain can be broken at any point. Breaking it at step 1 (by directing the model to the actual inequality chain before it forms an interpretation) prevents all downstream failure modes. Breaking it at step 3 (by directing the model to trace the Coq dependency chain) catches misidentifications before they become verdicts. Breaking it at step 5 (by directing the model to the appendices) resolves the issue after the verdict but before it propagates.

The FIELDS Protocol is designed to break the chain at step 1 and provide checkpoints at every subsequent step.

5 Practical Recommendations

For researchers using AI systems to audit novel mathematical work:

1. **Provide formal verification files and direct the model to use them as a proof map.** The Coq (or Lean, or Isabelle) file is not a credibility ornament. It is the most reliable guide to the actual proof structure. Instruct the model to trace theorem dependencies before forming any assessment.

2. **Direct the model to appendices before or concurrently with the main body.** Novel work frequently places its most important structural information in supplementary material—trust boundaries, alternative derivations, preemptive FAQ responses, and formal verification summaries. Default attention allocation underweights this material.

3. **Instruct the model to withhold verdicts until the full proof chain is traced.** A simple instruction—"Do not issue any positive or negative verdict until you have read the complete paper including all appendices"—prevents premature assessment and the downstream graduated retreat that follows from it.

4. **When the model declares an error, ask it to identify the exact step that fails.** The question "Which specific line of the inequality chain breaks under your objection?" distinguishes genuine mathematical errors from phantom counterexamples. If the model cannot point to a specific step, the objection targets an inferred mechanism rather than the stated one.

5. **Apply the FIELDS Protocol.** The ten principles documented in the FIELDS Protocol (Framework for Iterative Evaluation of Logical and Deductive Structures) are specifically designed to prevent training-prior interference in AI mathematical auditing. Principles 6 ("identify specific errors or state clearly that you cannot find any") and 8 ("do not confuse surprise with incorrectness") are the most directly relevant to the failure modes documented here.

6. **Consider cross-model verification.** In the ACL audits, models that failed on their own accepted corrections from peer models more readily than from human authors. Using a second model to review the first model's assessment—particularly when the first model declares errors in novel work—can identify training-prior interference that single-model dialogue may not resolve efficiently.

6 Conclusion

These failure modes are not evidence that AI cannot audit mathematics. Two frontier models evaluated the Angular Cancellation Lemma and found zero errors on first attempt. The failure modes documented here occurred in other systems under specific conditions and were resolved through structured dialogue—demonstrating that the FIELDS Protocol methodology can identify and correct these patterns in real time.

The core insight is simple: when a proof mechanism is genuinely novel, the model's training priors become a liability rather than an asset. The model's extensive knowledge of how proofs in a domain "usually" work causes it to evaluate the usual mechanism rather than the stated one. This is not a failure of reasoning capability—it is a failure of attention allocation, and it is correctable.

The practical recommendation is equally simple: provide the formal verification files, direct the model to the appendices, withhold verdicts until the full chain is traced, and when errors are declared, demand specificity. These four interventions prevent or resolve every failure mode documented in this paper.

Honest Framing

This document addresses failure modes observed in AI mathematical auditing at a specific moment in time (early 2026). Model capabilities are evolving rapidly. The failure modes described here may be partially or fully resolved in future model versions. The structural recommendations—particularly the use of formal verification files as proof maps and the practice of withholding verdicts until full papers are read—remain good practice regardless of model capability.

Further work on the governance implications of training-prior interference in high-stakes AI reasoning, including applications to legal, medical, and financial domains where AI systems assess novel evidence against trained expectations, is forthcoming as part of the Auburn Governance Stack.

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