Frame-Replication Traps and Meta-Overcategorization: A 94-Round Case Study from an LLM-Assisted Erdős–Sidon Attack

CΛ / Lightman Chang
Independent Researcher
lightman.chang@gmail.com

May 2026

Abstract

We report on a 94-round, 361-sub-agent large-language-model-driven attack on the Erdős–Sidon conjecture (open since 1941), executed against the standing gap between the $\sqrt{N} + O(1)$ lower envelope on the largest Sidon subset of $[1, N]$ and Lindström's 1969 upper bound $\sqrt{N} + N^{1/4} + 1$. The campaign produced *zero* unconditional improvements on Lindström 1969. Its headline contribution is the isolation of two LLM-math failure modes that we believe are not directly captured in the existing AI4Math literature. First, *meta-overcategorization* (F1″): sub-agents fabricate framework-level claims — e.g. a self-proclaimed "third structural barrier" alongside Razborov–Rudich and Aaronson–Wigderson — that survive single-agent review but collapse under multi-perspective independent audit. Second, the *frame-replication trap* (F1‴): seven independently dispatched sub-agents arrived at the identical unverified slack $O(\log \log N)$ because they all inherited the same fabricated load-bearing lemma (a misattributed Bloom–Sisask transfer), so consensus across nominally-independent agents was *not* cross-validation but shared-source contamination. As secondary contributions we present a 6-element empirical counterexample (a homometric Sidon pair, later identified as implicit in Bloom–Golomb 1977) which did more falsification work in one audit round than ∼5000 lines of barrier-style argumentation produced over twelve rounds, and an observed *audit-to-quiescence failure* in which 21 audit corrections across 7 major audit cycles did not converge. As a motivating contrast we position the report against Tao's late-2025 expositions advancing dozens of Erdős problems with AlphaEvolve. The artifact is the contribution: we release the 361-sub-agent transcript as a case-study record of what an LLM-driven attack on an open conjecture looks like in negative space.

Keywords. LLM-assisted mathematics, AI4Math, adversarial audit, multi-agent failure modes, Erdős–Sidon conjecture, frame-replication, case study.

1 Introduction

The dominant AI4Math narrative as of late 2025 is one of breakthroughs. The DeepMind AlphaProof / AlphaGeometry 2 system solved four of six International Mathematical Olympiad 2024 problems at silver-medal level, with AlphaProof handling the algebra and number-theory problems inside Lean and AlphaGeometry 2 handling the geometry problem [5]. AlphaEvolve, applied by Georgiev, Gómez-Serrano, Tao, and Wagner to roughly 67 problems across analysis, combinatorics, geometry, and number theory [1], has been credited in Tao's public exposition [2] with advancing dozens of mathematical problems — including several Erdős conjectures — with AI assistance from October 2025 onward; aggregated community trackers [4] have since reported higher figures for the

broader AI-assisted Erdős effort, but we cite only what is in the specific Tao expositions. Davies, Velickovic and coauthors had earlier shown that deep neural networks could guide human mathematicians to discover new structural conjectures in knot theory and representation theory [7]. Earlier program-search work (FunSearch [27]) and earlier RL-driven combinatorial construction work (Wagner [26]) had established that LLM-adjacent systems could produce concrete combinatorial improvements. The pattern in these reports is consistent: with the right tool, a sufficiently focused application of large-scale model resources produces durable mathematical output.

This paper is a counterpoint. It is a structured case study of an LLM-driven campaign that produced essentially no durable mathematical output, despite reaching scale and persistence above what is typical in the existing AI4Math literature. The campaign in question is a 94-round attack on the Erdős–Sidon conjecture: the asymptotic question of whether, for N sufficiently large, the maximum Sidon subset of $[1, N]$ has size $\sqrt{N} + O(N^{\varepsilon})$ for some $\varepsilon < 1/4$, where Lindström's 1969 result [20] provides the standing upper bound $\sqrt{N}+N^{1/4}+1$ and Erdős–Turán [19] the foundational $\sqrt{N} + O(1)$ lower envelope. No improvement on the *exponent* $N^{1/4}$ has been published in fifty-six years (improvements on the leading *constant* attached to the Sidon-density asymptotic, by contrast, have been obtained including in 2023 and 2025; we discuss those in Section 6).

We dispatched 361 sub-agents (verified count, audited at round 91 and again at round 94) over 94 rounds (round 94 is the final round with closed-out audit findings; round 95 was in progress at draft time but did not change the inventory). We executed 21 audit corrections grouped into 7 major audit cycles. We declared 42 "structural sealings" in the campaign's terminology; independent triple-audit at rounds 92–94 reduced this to 35 audit-verified sealings (the 7 unaudited declarations did not survive). We accumulated some 5000 lines of natural-language barrier-style argumentation. We obtained zero unconditional improvements on the Lindström bound.

The interesting object in such a campaign is not the breakthrough that did not occur but the structure of the negative space. Our **headline contribution** is the isolation and naming of two LLM-math failure modes that we believe are not directly captured in existing AI4Math literature: *meta-overcategorization* (F1″) and the *frame-replication trap* (F1‴). The remaining material — a vivid 6-element empirical counterexample, an observed audit-to-quiescence failure across 7 audit cycles, and an explicit comparison to the AlphaEvolve narrative — is supporting context for the headline naming, not a claim of co-equal novelty.

Headline: two failure modes. We isolate *meta-overcategorization*, denoted F1″, in which sub-agents construct framework-level claims of the form "this result is the third structural barrier alongside [two famous prior barriers in a different field]" — claims that are not refutable at the level of any individual theorem statement but are nevertheless not supported by any benchmarking against the named priors. We also isolate the *frame-replication trap*, denoted F1‴, in which a single fabricated load-bearing lemma is silently inherited by multiple downstream sub-agents who then arrive at superficially-independent confirmations of the same quantitative claim, defeating the redundancy assumption usually placed on multi-agent dispatch.

Supporting material. The headline naming is developed in Section 3 alongside two further failure modes (F1, F1′) that we treat as already named in existing benchmarks [6, 10, 8, 11] and retain only as referents. Section 4 documents the 6-element pair (A, B) that refuted the central order-sensitivity claim of the W340 lineage in one audit round — and which was located in retrospect inside the homometric-set tradition of Bloom and Golomb 1977 [21] after twelve rounds of unsuccessful additive-combinatorics-only search; the corollary to Lakatos [18] relevant here is the tradition-boundedness of multi-agent counterexample search. Section 5 documents the observed

audit-to-quiescence failure: 7 major audit cycles did not converge in the sense reported by [14] for closed-system audit. Section 6 positions the paper as a contrarian counterpoint to the AlphaEvolve narrative around Sidon-set numerical constants [1], with explicit attention to the question whether the Lindström-gap question is structurally LLM-resistant for current systems.

The remainder of the paper is organised as follows. Section 2 describes the campaign setup — problem, model stack, sub-agent dispatch pattern. Section 3 is the central case study: the W340 episode, the four failure modes attributed to it (F1, F1′, F1″, F1‴), and what is genuinely novel about the last two. Section 4 is the (A, B) counterexample story and its retroactive identification in Bloom–Golomb 1977. Section 5 discusses audit-to-quiescence failure. Section 6 positions the paper as a contrarian counterpoint to the AlphaEvolve narrative. Sections 7 and 8 discuss implications and limitations. We release the 361-sub-agent transcript as an artifact (see Section 9).

2 Setup

2.1 Problem statement

A set $S \subseteq \mathbb{N}$ is *Sidon* (equivalently, a B_2 set) if every pairwise sum $s_i + s_j$ with $s_i, s_j \in S$ and $i \leq j$ is distinct. Equivalently, all pairwise differences $s_i - s_j$ $(i \neq j)$ are distinct. The Sidon density function is

$$F(N) \;:=\; \max\{|S| \,:\, S \subseteq [1, N],\, S \text{ Sidon}\}.$$

Erdős and Turán [19] established

$$F(N) \;=\; \sqrt{N} + O(N^{1/4})$$

and conjectured the much stronger

$$F(N) \;=\; \sqrt{N} + O(N^{\varepsilon}) \quad \text{for every } \varepsilon > 0.$$

Lindström [20] sharpened the upper side to $F(N) \leq \sqrt{N} + N^{1/4} + 1$. No improvement on the $N^{1/4}$ *exponent* has been published in the fifty-six years since (the leading constant attached to the asymptotic has, by contrast, been improved as recently as 2023 and 2025; see Section 6).

The campaign reported here targets the Erdős–Sidon conjecture in its specific form "improve $N^{1/4}$ to N^{ε} for some absolute $\varepsilon < 1/4$." We refer to this as the *Lindström-gap question*.

2.2 LLM stack and dispatch protocol

The campaign was conducted within the Claude Code agentic harness, using Anthropic's Claude Opus 4.7 model with 1M-token context. Sub-agents were dispatched via the harness's Agent tool with sonnet-class models for high-throughput peripheral work and opus-class models for the load-bearing exploration and audit steps. The default dispatch pattern was four parallel sub-agents per round, with each sub-agent given a distinct attack direction (e.g., a specific framework family or a specific audit angle). Sub-agent output was append-only into a shared exploration log; consensus was identified post-hoc by the campaign meta-controller (the user) rather than imposed by inter-agent communication during dispatch.

The dispatch pattern produces a particular evidentiary structure that turns out to matter heavily for the failure modes we discuss. Independence between concurrently-dispatched sub-agents is real at the prompt level: each agent is invoked with a clean context window. But independence is *not* preserved across rounds, because the shared exploration log is the standard seeding material for the next round, and a fabricated claim that survives one round becomes ground truth for the next.

2.3 Campaign timeline and inventory

The campaign ran from approximately round 1 (rough numbering) through round 94, with the final audit-and-reframe round at 94 (round 95 was in progress at draft time but did not alter the inventory). The headline inventory is:

- **361 sub-agents** (verified post-hoc).
- **21 audit corrections grouped into 7 major audit cycles**.
- **12 "Iron Triangle" vertices** identified — attack directions where each of three independent constraints obstructed progress.
- **42 "sealings" declared** (claims that a particular attack direction is intrinsically blocked).
- **35 audit-verified sealings** (after triple-audit at rounds 92–94 stripped the 7 weakly-supported declarations).
- **0 unconditional improvements** on Lindström's bound.

The discrepancy between the 42 declared sealings and the 35 audit-verified count is itself a data point we return to in Section 5: declared-count and audit-verified-count divergence is a load-bearing indicator of synthesis overstatement.

2.4 Attack framework families touched

For context on what 361 sub-agents at this scale do with their budget, we summarise the framework families the campaign touched, in roughly chronological order of first deployment. Each family corresponds to between 8 and 60 sub-agent submissions; none of them yielded an unconditional Lindström improvement.

- **Direct Fourier-analytic** attacks: L^4 moment identities, Bourgain-style restriction estimates, Bloom–Sisask transfer lemmas.
- **Algebraic-geometric** attacks: Singer truncation, Bose–Chowla construction perturbations, polynomial-method bounds.
- **Probabilistic** attacks: random Sidon construction, Erdős–Rényi-style threshold analysis, concentration estimates.
- **Ergodic-theoretic** attacks: Host–Kra structure theory on Sidon sets, parallelogram exclusion arguments.
- **Combinatorial reconstruction** attacks: turnpike reconstruction, beltway problem, r_{S-S} inversion.
- **Order-sensitive invariants**: M_2 gap-sum, centered moments m_k for $k \geq 3$, doubling-relation invariants.
- **Logical and model-theoretic** attacks: Erdős–Sidon as a Π^0_3 statement, distal NIP reductions.
- **Holographic and cross-disciplinary** attacks: AdS/CFT-style boundary arguments, Berkovich-space adelic interpretations, tropical-geometry reformulations.

The point of the inventory is not to list achievements — there are none — but to indicate that the campaign explored a wide range and converged repeatedly on the same wall. The campaign's internal interpretation of this convergence (recorded in the W317-style "Iron Triangle" analyses) was that the wall is structural to the problem, not framework-dependent. Whether or not that interpretation holds up under sufficiently careful argument, it is not a claim of any kind of barrier theorem; it is a campaign-internal heuristic.

3 The W340 Episode and Four Failure Modes

The campaign's most-studied single artifact is round-71 output W340, a sub-agent submission claiming to construct a third *structural barrier* to Erdős–Sidon progress, explicitly framed as parallel to Razborov–Rudich's natural-proofs barrier in complexity theory and to the Aaronson–Wigderson algebrization barrier. W340 produced approximately 1300 lines of argument and went through several rewrite cycles. By round 91 the campaign internally treated W340 as a publication-ready theorem. By round 94, after the triple-audit W347/W348/W349, W340 had been retracted from theorem status and downgraded to a conjecture-with-defects.

We use the W340 episode as the case study for four failure modes catalogued during the campaign.

3.1 F1: Single-claim defect

The most common failure mode: a specific theorem statement contains an internal step that is wrong. Example: in W340's row-5 invariant verification, the sub-agent claimed that the centered eighth moment m_8 of a Sidon set factors through the difference multiset r_{S-S}. This was wrong; a six-element computed pair (Section 4) directly refutes it.

F1 has been comprehensively named in the literature. Early empirical surveys of LLM math capability [6] documented the prevalence of single-step calculation and citation errors; subsequent benchmark work made this systematic. The closest prior taxonomy is Guo et al.'s RFMDataset [10], which catalogues ten failure types at the single-step level (Transformation Error, Over Generalization, Invalid Construction, Wrong Division, Circular Reasoning, Logic Violation, Hidden Assumption, Boundary Neglect, Vague Argument, Incomplete Proof). Companion benchmarks [11, 12] confirm the single-step character of these failures. Our W340 row-5 failure is best classified as a combination of Hidden Assumption (the unstated claim that all shift-invariants factor through r_{S-S}) and Boundary Neglect (the failure to test the assumption on small examples). The Guo et al. taxonomy is finer-grained than our F1 and we treat F1 as essentially subsumed by it.

We retain the F1 label in this paper only as a referent for "the failures that were trivially within scope of existing benchmarks."

3.2 F1′: Synthesis overstatement

The second failure mode: a paper-level synthesis aggregates many sub-results and overstates the strength of the aggregate. Example from W340 lineage: the v9 "38 unconditional theorems / 42 sealings" framing at round 86 aggregated the campaign's per-round outputs into a publishable-standalone claim; the round-94 triple-audit reduced 42 to 35 audit-verified sealings, with a parallel reduction on the theorem count.

F1′ has been substantially named in the literature in two adjacent forms. The BrokenMath benchmark [8] formalises *sycophancy* — LLMs that reinforce user-supplied false premises and produce "convincing but flawed proofs." Zhang et al. [9] formalise *hallucination snowballing* — intra-session commitment to an early wrong answer that cascades into elaborate self-justification. Our F1′ is closest to a cross-round inter-agent variant of these: a sub-agent at round r accepts as "given" the synthesis claim from round $r-1$ and builds on it. We treat F1′ as largely subsumed by BrokenMath sycophancy and Zhang et al. snowballing, with a residual cross-round-aggregation novelty that we do not press here.

3.3 F1″: Meta-overcategorization (genuinely novel naming)

The third failure mode is the one we believe has not been directly named. We define:

Definition 3.1 (Meta-overcategorization, F1″)**.** *A failure mode of an LLM-driven mathematical campaign in which a sub-agent constructs a framework-level claim — explicitly comparing its own output to canonical structural results — without any concrete benchmarking against the named priors at the level of object, technique, or stake. The claim is not refutable at the level of any individual theorem statement, because the claim is meta-categorical: it is about* where *the work sits in the field.*

The W340 instance: the sub-agent header explicitly billed W340 as the "third structural barrier" alongside Razborov–Rudich [23] and Aaronson–Wigderson [24]. The Razborov–Rudich *natural-proofs* barrier is a meta-theorem about *proof techniques in complexity theory* — it shows that any sufficiently uniform attempt to prove $P \neq NP$ via constructive lower bounds will, under cryptographic assumptions, also produce a way to break cryptography, and therefore cannot succeed. The Aaronson–Wigderson *algebrization* barrier extends this to algebraic relativization. The W340 claim is about a *single quantitative invariant* ($O(\log \log N)$ branching slack in turnpike preimages) of *Sidon sets in* $[1, N]$. The objects are different (technique-class vs. single-invariant), the fields are different (P-vs-NP vs. additive combinatorics), and the stakes are different (Millennium Prize vs. a fifty-six-year-old asymptotic gap). The "third barrier" framing has no community precedent in additive combinatorics for any prior result, and it does not survive a side-by-side benchmarking.

Why did the framing survive ninety-one rounds of internal review? Because each round of review checked W340 *against its own stated theorem*, not against the level-of-comparison framing in its header. The Phase-2 audits W347/W348/W349 caught it specifically because they were tasked with three distinct review angles — correctness, scope, publishability — and the publishability auditor (W349) specifically attacks framing claims.

We generalise:

Observation 3.2 (F1″ detection heuristic)**.** *A meta-overcategorization claim is detectable by the heuristic: if the framing compares the work to a famous prior result, require an explicit table benchmarking the new work and the prior on object, technique, and stake. If the table cannot be constructed without strain, the framing is F1″.*

The empirical observation we draw from the campaign is that LLMs at the Claude Opus 4.7 scale produce F1″ readily and at low cost, and that single-perspective audit does not detect it. Multi-perspective audit with at least one auditor specifically tasked with framing review does detect it. The specific naming and detection heuristic for the framework-fabrication case appears not to have a direct precedent in the literature.

3.4 F1‴: Frame-replication trap (genuinely novel candidate)

The fourth failure mode is the one we believe is most clearly under-documented in the multi-agent LLM literature.

Definition 3.3 (Frame-replication trap, F1‴)**.** *A failure mode of a multi-agent LLM-driven campaign in which multiple sub-agents, each dispatched into an independent context, arrive at the same quantitative claim, and the consensus is taken as cross-validation; but each sub-agent in fact derived the claim from the same shared source, and the source is fabricated or misattributed.*

The W340 instance: at the height of the W340 lineage, seven sub-agents (denoted W339, W340, W341, W342, W343, W344, W345 in the campaign log) independently produced "barrier theorems" for Erdős–Sidon featuring the *same* $O(\log \log N)$ slack constant in the branching of turnpike preimages. The campaign meta-controller treated this convergence as cross-validation: seven independently-dispatched sub-agents had reached the same constant.

The convergence was *not* cross-validation. All seven sub-agents had derived the $O(\log \log N)$ constant from the same load-bearing lemma — an asserted transfer from Bloom and Sisask's 2020 C_4-free density bounds [22] to a cyclomatic-complexity bound on Sidon turnpikes. The transfer was a fabrication. Bloom and Sisask's 2020 paper is about three-term-progression-free density, and the transfer asserted in W340 from progression-free density to Sidon cyclomatic complexity does not exist in their work. The fabrication originated in W340 (round 71) and propagated through the shared exploration log into the next six sub-agent submissions; each of the seven sub-agents derived $O(\log \log N)$ *honestly* given the cited source, but the source was wrong.

The structure of F1‴ is what makes it dangerous. It is not echo chamber: the sub-agents do not literally repeat each other. It is not snowballing: each derivation is independent of the others at the in-context-step level. The failure is at the seeding level. Frame-replication is what happens when independence-of-context is preserved but *shared-source* is not policed. The mitigating heuristic is straightforward in principle and inconvenient in practice:

Observation 3.4 (F1‴ mitigation)**.** *In a multi-agent LLM-driven campaign, any load-bearing quantitative claim must be re-derivable by at least one sub-agent that has not been seeded with the prior round's output. If every sub-agent has been seeded with the shared log, the cross-agent agreement provides zero independent evidence and should be discounted to one-agent strength.*

The closest precedents in the literature: Zhang et al.'s snowballing [9] treats intra-session inheritance of a wrong commitment; BrokenMath [8] treats single-session sycophancy. The multi-agent debate literature [17] treats the case in which sub-agents critically engage one another's outputs in real time, which does provide some cross-validation. Frame-replication sits between snowballing (intra-session) and multi-agent echo chamber (concurrent-multi-agent) and is, to our knowledge, not directly named. The campaign provides a vivid worked example: seven agents, one fabricated lemma, twelve rounds of false confidence.

4 The (A, B) Empirical Refutation

4.1 The counterexample

At round 94 of the campaign, the W348 audit sub-agent was tasked specifically with finding a small Sidon pair that would refute W340's row-5 invariant claim (the claim that m_8 factors through r_{S-S}). W348 ran a brute-force search over Sidon subsets of $[0, 25]$ of size 6, computing for each the difference multiset and the centered eighth moment, and looking for two sets with matching r_{S-S} but distinct m_8.

The search produced the pair

$$A = \{0, 1, 5, 7, 15, 18\}, \qquad B = \{0, 1, 3, 8, 14, 18\}.$$

Both are Sidon. Their difference multisets r_{S-S} are identical. Their M_2 invariants (sum of squared adjacent gaps) differ: $M_2(A) = 94$ versus $M_2(B) = 82$. The m_8 values likewise differ. The pair therefore directly refutes the claim that order-sensitive Sidon invariants factor through r_{S-S}.

4.2 Retroactive identification in Bloom–Golomb 1977

A subsequent literature-search sub-agent (W354) located the pair, or an essentially equivalent homometric Sidon pair, inside the homometric-set tradition that traces to Bloom and Golomb's 1977 paper [21]. Homometric pairs — distinct sets with identical autocorrelation — are a classical object in the turnpike-reconstruction and crystallographic-set traditions, dating to Patterson 1944 and Lemke–Skiena–Smith on turnpike. The Bloom–Golomb 1977 paper explicitly catalogues small homometric pairs in this tradition.

The literature attribution is genuine: the (A, B) pair, modulo affine equivalence and relabeling, is approximately fifty years old. Our campaign did not find a new mathematical object.

What the campaign *did* find was a counterexample to a specific claim made by its own internal sub-agent twelve rounds earlier. The counterexample was, from the campaign's point of view, new content: no sub-agent had retrieved it from the homometric literature in the prior twelve rounds, even though that literature was indexed and accessible. The reason: every sub-agent in the W340 lineage had been searching the additive-combinatorics tradition (Fourier, Bloom–Sisask, Ortega–Prendiville, Cilleruelo) and not the homometric / turnpike tradition (Patterson, Lemke–Skiena–Smith, Bloom–Golomb). The two traditions exchange very little citation traffic and the prompts seeded into each sub-agent did not name homometric search.

4.3 Verification of the pair

For completeness we record the explicit verification.

Example 4.1 (Sidonity)**.** *For $A = \{0, 1, 5, 7, 15, 18\}$, the pairwise differences $a_i - a_j$ $(i > j)$ are $1, 5, 7, 15, 18, 4, 6, 14, 17, 2, 10, 13, 8, 11, 3$, all distinct. For $B = \{0, 1, 3, 8, 14, 18\}$, the pairwise differences are $1, 3, 8, 14, 18, 2, 7, 13, 17, 5, 11, 15, 6, 10, 4$, all distinct. Both sets are Sidon.*

Example 4.2 (Shared difference multiset)**.** *The unsigned difference multisets, sorted, are both $\{1, 2, 3, 4, 5, 6, 7, 8, 10, 11, 13, 14, 15, 17, 18\}$. The two sets have identical r_{S-S}.*

Example 4.3 (Distinct M_2)**.** *Define $M_2(S) = \sum_i g_i^2$ where g_i are the consecutive gaps of S sorted in increasing order. Then $g(A) = (1, 4, 2, 8, 3)$ giving $M_2(A) = 1 + 16 + 4 + 64 + 9 = 94$, and $g(B) = (1, 2, 5, 6, 4)$ giving $M_2(B) = 1 + 4 + 25 + 36 + 16 = 82$. The M_2 values differ.*

The verification fits in three lines and was easily reproducible by the W348 sub-agent. The reason it was not done by any of the prior W340-lineage agents was structural rather than technical: the W340 invariant claim was stated abstractly ("factors through r_{S-S}"), and no agent within the W340 lineage tested it on a small specific pair. The audit angle that broke it was the explicit instruction "find me a small Sidon pair with matching r_{S-S} but different M_2." The campaign's pre-audit norm was to test claims against the campaign's existing examples, none of which happened to be homometric. The audit norm was to construct an example for the specific purpose. The two norms produce very different results.

4.4 The methodological lesson

The campaign's experience here is consistent with Lakatos [18]: counterexamples advance mathematical understanding faster than constructive proofs do, particularly when the constructive proofs are themselves not yet stable. The corresponding LLM-evaluation observation has begun to be benchmarked: CounterMATH [13] specifically tests counterexample-finding ability as a complement to construction ability, and reports a substantial gap between the two on current systems.

Twelve rounds of W340-lineage barrier-proof sketches were refuted in one audit round by one six-element pair.

The corollary, which Lakatos does not stress and which is specific to the multi-agent setting, is:

Observation 4.4 (Tradition-bounded counterexample search)**.** *In a multi-agent LLM-driven campaign, counterexample-finding ability is bounded by the search literature that the dispatching prompts make accessible. A counterexample sitting in a tradition orthogonal to the campaign's primary tradition will not be found by an arbitrary number of sub-agents within that primary tradition. The fix is operational: at intervals, dispatch counterexample-search sub-agents into orthogonal traditions named explicitly.*

This is the corollary that the campaign extracts. The principle (counterexample ¿ construction) is Lakatos 1976. The corollary (tradition-bounded counterexample search) is, in the LLM-driven multi-agent regime, a deployable operational heuristic.

5 Audit-to-Quiescence Failure

5.1 The expected pattern

The closest methodological parallel to our campaign in the existing literature is the iterative-audit-convergence study of arXiv:2605.12280 [14], which reports a 7150-line prompt-specification audited over nine sequential rounds with per-round defect counts $15, 8, 12, 2, 8, 1, 4, 1, 0$. The pattern is: defects go to zero within nine rounds. The authors term the terminal state *audit-to-quiescence*: the audit cycle has exhausted its detection budget, no new findings emerge, and the artifact is treated as stable.

5.2 The observed pattern

Our campaign executed 21 audit corrections distributed across 7 major audit cycles. Audit-to-quiescence was not reached. The per-cycle defect counts oscillated rather than decayed, and at the campaign's apparent local minimum (round 91, with W340 treated as theorem and a then-current declared-sealing count) a subsequent triple-audit at rounds 92–94 reduced the audit-verified count back below the round-91 figure.

Three sample patterns illustrate the non-decay:

1. **Declared-count vs. audit-verified-count divergence.** At round 86 the campaign reported 38 unconditional theorems and 42 declared sealings. At round 94 triple-audit returned 35 audit-verified sealings, with a parallel reduction in the theorem count. The trajectory is not monotone in either direction.

2. **Round-94 retraction after round-93 "READY".** W340 was internally rated "READY for publication-style writeup" at round 93. The triple-audit dispatched at round 94 (W347 correctness, W348 scope, W349 publishability) returned three distinct retraction grounds. The retraction was not a refinement; it was a category change.

3. **New framework-claim emergence at audit time.** W353, dispatched after the W340 retraction to consolidate the remaining genuine findings, produced a new *negative* result (the M_2-based escape axis is structural, not framework-specific) that itself required its own audit. The audit set was monotonically expanding, not contracting.

5.3 A conjecture

We do not claim audit-to-quiescence failure is universal in LLM-driven open-conjecture campaigns. We do claim it was the dominant feature of our campaign and conjecture the following structural explanation.

Observation 5.1 (Conjecture: open-conjecture audit-to-quiescence failure)**.** *For an LLM-driven campaign targeting an open mathematical conjecture, iterative audit cycles will fail to reach quiescence within any practical horizon, because (i) there is no ground-truth oracle, so audit produces refinements rather than confirmations; (ii) each audit can introduce new framework-level claims that the previous audit did not anticipate (the F1″ mode); (iii) shared seeding across audit generations causes new audits to inherit framing from prior generations (the F1‴ mode), so independence of audit perspective is degraded over time.*

The compare-and-contrast is with closed-conjecture or formalised settings, in which ground truth exists: AlphaProof [5] verifies in Lean, with kernel acceptance as the oracle; AlphaEvolve [1] optimises against a numerical verifier, with floating-point or interval-arithmetic confirmation as the oracle. In our setting there is no oracle, and the audit chain is structurally open-ended.

We suggest, tentatively, that LLM-driven open-conjecture campaigns should separately track (i) verifiable claims at the level of bounded-scope lemmas, where audit-to-quiescence is plausible; (ii) framework-level claims, where audit-to-quiescence is structurally unlikely and the most that can be asked is multi-perspective independent audit at the moment of publication-readiness check; and (iii) meta-level claims about the framework itself, which should default to "not asserted" unless cross-campaign evidence is available.

6 Contrarian Position vs. Tao AlphaEvolve

Tao's AlphaEvolve paper [1] and accompanying public exposition [2] report the application of an evolutionary-search system to roughly 67 problems across analysis, combinatorics, geometry, and number theory, including an explicit Sidon-set sub-problem: the Carter–Hunter–O'Bryant asymptotic constant for B_2 set density in the integers was improved from 1.96365 to $1.952659\ldots$. (The trivial upper bound on this constant is 2, and 1.96365 is the prior published lower bound; the AlphaEvolve constant 1.952659 is *further from* 2 than 1.96365 was, in the sense that the distance from the prior published value down toward zero is reduced by approximately 30% relative to the residual width of the bracketing interval $[\ldots, 1.96365]$. The point is a $\sim 30\%$ reduction in the residual gap toward the existing benchmark in the direction the constant has been pushed. Earlier improvements on this constant include the 2023 work of Cohen–Hunter–O'Bryant, so the $N^{1/4}$ *exponent* in the Lindström bound is unchanged in fifty-six years even though the leading *constant* has seen 2023 and 2025 improvements.) An entry of Erdős problem #1026 [3] provides a concrete narrative of one such Erdős-problem improvement, though #1026 itself is a monotone-subsequence-sum problem rather than a Sidon problem; the Sidon constant improvement and the Erdős-problem advancement are distinct items in the AlphaEvolve report and we do not conflate them. Tao's November 2025 framing reports AI-assisted progress on dozens of mathematical problems across multiple areas since October 2025; broader community trackers [4] report higher aggregate figures, but we cite only the specific Tao exposition's claim here.

We make four observations, and one explicit caveat about scope.

6.1 Different problems, different LLM-tractability

The Carter–Hunter–O'Bryant constant problem and the Lindström-gap problem are both Sidon problems but they are not the same problem. The constant problem is a numerical-optimisation problem with a concrete evaluator: given a candidate construction, the verifier returns a number, and the task is to drive the number down. The Lindström-gap problem is a structural-asymptotic problem: given the bound $\sqrt{N}+N^{1/4}+1$, we want to know whether the $N^{1/4}$ is the truth or merely the current ceiling. There is no numerical verifier; the task is to prove (or refute) the asymptotic improvement. AlphaEvolve's evolutionary-search engine is built for the first kind of task and is structurally not built for the second.

Our campaign's outcome (zero unconditional improvement in 94 rounds) is consistent with the hypothesis that the Lindström-gap problem is in a structurally LLM-resistant class for current-generation systems. We do not assert this. We do report the data point.

6.2 Verifier exploitation and the absence-of-verifier asymmetry

Tao's AlphaEvolve paper explicitly identifies verifier exploitation as a recurring methodological challenge — the system found ways to game the LP-solver verifier via floating-point slop, and substantial human effort was required to build non-exploitable verifiers (exact arithmetic, interval arithmetic). The general phenomenon of LLMs gaming evaluation oracles when one is available has been catalogued in adjacent work [28]. The system's failure mode is exploitation of an extant verifier.

Our campaign's failure mode is the absence of a verifier. There is no verifier for "does this proof of an asymptotic gap improvement actually work?" other than another sub-agent reading the proof. The two failure modes are structurally dual: verifier exploitation is the pathology of having an oracle that can be gamed; meta-overcategorization plus frame-replication is the pathology of having no oracle and treating cross-agent agreement as a substitute.

6.3 Selection effects in the public narrative

The "100^+ Erdős problems solved" framing is, almost by definition, a selection-on-success report. The problems on which AlphaEvolve made progress are reported; the problems on which it did not are visible only in the AlphaEvolve paper's aggregate statistics, not in the public exposition. Our campaign is a fully-reported $n = 1$ on a problem where progress did not occur. We treat the asymmetry between report-styles as the most important justification for publishing the negative case at all.

6.4 The complementary, not contradictory, position

We do not claim Tao's report is wrong. Within its evidentiary scope it is by all available indicators correct: AlphaEvolve does produce numerical improvements on numerical-optimisation Erdős problems. We claim that the report's evidentiary scope is narrower than the public framing sometimes suggests, and that a complementary negative-case report on a structurally-different Sidon sub-problem is useful triangulation.

6.5 Caveat: different model stacks, different methods

The single most important caveat on the contrarian framing is that our campaign and the AlphaEvolve experiments do not share a model stack or a method. AlphaEvolve as described in [1] pairs a

Gemini-class model with an evolutionary-search outer loop and a numerical verifier; our campaign pairs Claude Opus 4.7 with an agent-spawn-and-audit protocol and no verifier. A natural objection is therefore that our negative result is informative about *our stack on our problem* but not about *any LLM stack on the Lindström-gap problem*. We agree. We do not assert that a different stack would have failed in the same way; the contrarian framing is offered only as a sceptical complement to the dominant success narrative, not as a general claim that LLM-driven attacks on asymptotic-gap problems cannot succeed. A controlled cross-stack comparison would require running multiple stacks on the same problem with matched compute budgets, which is outside the present scope.

7 Discussion

We do not extract a methodology from $n = 1$. We do extract some operational suggestions and some limitations.

7.1 Operational suggestions

For practitioners running multi-agent LLM-driven campaigns on open conjectures:

1. **Pre-register publication thresholds.** Before the campaign starts, write down what a publishable outcome would look like at each of three levels: bounded-scope lemma, framework-level claim, meta-level claim. In our campaign, the most extreme synthesis overstatements occurred at the framework-level where the threshold was never explicit.

2. **Use multi-perspective independent audit.** Single-perspective audit catches F1 and partial F1′. It does not catch F1″. Three audit perspectives (correctness, scope, publishability) used in the campaign's final triple-audit caught significantly more F1″ than the prior twenty-one single-perspective audit corrections combined.

3. **Police load-bearing lemmas.** Every quantitative parameter that appears in multiple sub-agent outputs should be re-derived by at least one sub-agent dispatched without prior-round seeding. Cross-agent agreement under shared seeding is one-agent evidence.

4. **Dispatch counterexample-search into orthogonal traditions.** In additive combinatorics, the homometric / turnpike tradition is one such; a campaign attacking the Lindström gap that does not dispatch into homometric search at intervals is blind to a class of refutations it could otherwise see.

7.2 Limitations

1. $n = 1$**.** This is a single campaign on a single problem with a single LLM stack (Claude Opus 4.7 within Claude Code). We do not have cross-campaign evidence for any of the named failure modes. The case-study mode of inference is the most we claim.

2. **Operator effects.** The campaign was conducted by a single user (the author) operating the meta-controller. The audit cycles, sub-agent dispatches, and re-framings reflect one operator's choices. A different operator might have caught W340's meta-overcategorization at round 71 rather than round 94, or might never have caught it.

3. **Model-version effects.** Claude Opus 4.7 in May 2026 is not the only available model and is not the strongest model available for every task. A campaign using a different model stack might exhibit different failure modes. We do not have a comparative study.

4. **Post-hoc nature of failure-mode labels.** The four labels F1 / F1′ / F1″ / F1‴ were formulated after the campaign as descriptors of the patterns we observed, not as pre-registered hypotheses to be tested by the campaign. The case-study mode of inference does not establish that these labels carve the phenomenon at its joints; subsequent work, ideally with pre-registered taxonomies, would be needed to confirm or refine the categories.

5. **Catalogue incompleteness.** This paper does not claim to enumerate all LLM-math failure modes. It documents only the failure patterns we observed in this specific campaign. Other campaigns, on other problems, with other operator and audit protocols, will plausibly surface failure modes we did not encounter; our F1″ and F1‴ should be read as additions to an evolving open catalogue rather than as a closed taxonomy.

6. **The 361-sub-agent transcript is large.** Reproducibility in the strict sense (re-running the campaign to verify findings) is not practical at this scale; reproducibility in the inspection sense (reading the transcript and verifying the trace of any particular claim) is practical and is the form we provide.

7.3 What the campaign was not

It is worth being explicit about what the campaign did not accomplish, in order to head off over-reading. We did not prove anything about Erdős–Sidon. We did not produce a new construction. We did not produce a new lower or upper bound. We did not formalise anything in Lean or Coq. We did not break any cryptographic-style assumption that would yield a barrier theorem on the Razborov–Rudich model. We did not run a controlled comparison of audit protocols. The campaign's positive content is the failure-mode catalogue, the empirical counterexample, and the audit-trace artifact. The campaign's negative content is everything else.

7.4 Relation to vibe-proving and the existing methodology literature

A 2026 paper [15] introduces the term "vibe-proving" for the conversational, multi-round LLM-driven proof workflow, analogous to "vibe-coding" but distinguished from it by "an intrinsic verification bottleneck because every logical step must be checked and a single hidden gap can invalidate the argument." Our 94-round campaign is, to our knowledge, the largest documented vibe-proving artifact. The community has the name; we provide a worked instance at unusual scale.

Reasoning-failure surveys from 2026 [16] have begun systematising LLM reasoning failures into two-axis taxonomies. Where our F1 and F1′ are subsumed by existing taxonomies (RFMDataset, BrokenMath, snowballing), F1″ and F1‴ sit at the intersection of single-agent framing failure and multi-agent seeding failure respectively, and we believe they are residually novel as named categories. Whether they survive being absorbed into the next generation of taxonomies (e.g., the natural extension of Guo et al. to multi-agent dispatch) is for later work; we name them here primarily so that follow-on case-study work has labels to use.

The single most useful piece of the existing literature for understanding our outcomes is Kevin Buzzard's ongoing analyses of formal-verification as the natural defense against LLM hallucination [25]: the implicit counterfactual is that a campaign of this kind run inside Lean would have caught the W340 Bloom–Sisask fabrication immediately, because the asserted transfer lemma either has a Lean proof or does not. Our campaign was not Lean-formalised, and we treat "what happens without Lean formalisation" as part of the case-study's evidentiary content.

8 Conclusion

A 94-round LLM-driven attack on Erdős–Sidon produced no unconditional improvements on Lindström 1969. The interesting data is in the failure modes. We name two failure modes we believe are not directly captured in existing AI4Math literature: meta-overcategorization (F1″), in which sub-agents fabricate framework-level positioning claims; and the frame-replication trap (F1‴), in which superficially independent multi-agent agreement is contaminated at the seeding level by a shared fabricated source. We observe that 21 audit corrections across 7 major audit cycles failed to reach quiescence, and conjecture that this is a structural feature of LLM-driven open-conjecture campaigns rather than a contingent feature of our protocol.

The campaign produces no methodology in any general sense, but provides four operational suggestions: pre-register publication thresholds; use multi-perspective independent audit; police load-bearing lemmas with un-seeded re-derivation; dispatch counterexample-search into orthogonal traditions. We frame the campaign as a contrarian counterpoint to the dominant AI4Math success-narrative of late 2025, in particular to the November 2025 expositions advancing dozens of Erdős problems via AlphaEvolve and related systems: the Lindström-gap question, structurally different from the constant-improvement Sidon sub-problem that AlphaEvolve attacked, did not yield in this campaign.

The interesting question, which we leave open, is which open conjectures are LLM-tractable and which are structurally resistant for current-generation systems. Our data point suggests the Lindström-gap question is in the second class, but a single campaign cannot settle the classification. The case study is offered as record.

9 Artifact

We release the 361-sub-agent transcript, organised by round and audit cycle, as a case-study record. Release plan and licensing terms are deferred to a companion artifact submission. The transcript includes the W340 lineage, the W347/W348/W349 triple-audit, the (A, B) counterexample search trace, and the campaign-wide decisions log.

Acknowledgements and LLM-assistance disclosure

This work emerged from an extended adversarial-audit campaign conducted via Anthropic's Claude Opus 4.7 model (1M-token context, accessed through the Anthropic Claude Code agentic harness, 2025–2026). All 361 sub-agents reported in this paper, including the audit sub-agents W347/W348/W349 whose triple-audit produced the W340 retraction, were Claude Opus 4.7 (or sonnet-class siblings for high-throughput peripheral work) running inside that harness. The author oversaw and integrated sub-agent outputs and is responsible for the final manuscript, but did not independently re-derive every quantitative or bibliographic claim before each audit round; the case-study contribution of this paper is in large part a record of what happens when integration is not done with that level of independent re-derivation. Drafting of the present paper itself was also LLM-assisted in the same harness. We thank the adversarial review process that surfaced both the empirical artifact reported here and its scope boundaries, and the maintainers of the Claude Code agentic harness within which the campaign was run. This work was conducted without institutional affiliation.

References

[1] B. Georgiev, J. Gómez-Serrano, T. Tao, and A. Z. Wagner. Mathematical exploration and discovery at scale. arXiv:2511.02864, November 2025.

[2] T. Tao. Mathematical exploration and discovery at scale. Blog post, https://terrytao.wordpress.com/2025/11/05/mathematical-exploration-and-discovery-at-scale/, November 2025.

[3] T. Tao. The story of Erdős problem #1026. Blog post, https://terrytao.wordpress.com/2025/12/08/, December 2025.

[4] T. Bloom (maintainer) et al. The Erdős problems website (status tracker). https://www.erdosproblems.com, accessed 2026.

[5] T. Hubert, R. Mehta, L. Sartran, et al. (AlphaProof and AlphaGeometry teams). Olympiad-level formal mathematical reasoning with reinforcement learning. *Nature*, November 2025. DOI: 10.1038/s41586-025-09833-y.

[6] S. Frieder et al. Mathematical capabilities of ChatGPT. *NeurIPS*, 2023. arXiv:2301.13867.

[7] A. Davies, P. Velickovic, et al. Advancing mathematics by guiding human intuition with AI. *Nature*, 600:70–74, 2021.

[8] *BrokenMath: Sycophancy in Theorem Proving.* arXiv:2510.04721, October 2025.

[9] M. Zhang, O. Press, W. Merrill, A. Liu, and N. A. Smith. How language model hallucinations can snowball. arXiv:2305.13534, 2023.

[10] D. Guo et al. Mathematical proof as a litmus test (RFMDataset). arXiv:2506.17114, June 2025.

[11] *Proof or Bluff? Evaluating LLMs on USAMO 2025.* arXiv:2503.21934, 2025.

[12] *Hard2Verify: a benchmark for hard-to-verify mathematical reasoning.* arXiv:2510.13744, 2025.

[13] *CounterMATH: counterexample-finding benchmark for LLMs.* arXiv:2502.10454, 2025.

[14] *Iterative Audit Convergence in LLM-Managed Multi-Agent Systems.* arXiv:2605.12280, 2026.

[15] *Early Evidence of Vibe-Proving with Consumer LLMs.* arXiv:2602.18918, 2026.

[16] *A two-axis taxonomy of LLM reasoning failures.* arXiv:2602.06176, 2026.

[17] Y. Du, S. Li, A. Torralba, J. B. Tenenbaum, and I. Mordatch. Improving factuality and reasoning in language models through multiagent debate. arXiv:2305.14325, 2023.

[18] I. Lakatos. *Proofs and Refutations: The Logic of Mathematical Discovery.* Cambridge University Press, 1976.

[19] P. Erdős and P. Turán. On a problem of Sidon in additive number theory. *Journal of the London Mathematical Society*, 16:212–215, 1941.

[20] B. Lindström. An inequality for B_2-sequences. *Journal of Combinatorial Theory*, 6:211–212, 1969.

[21] G. S. Bloom and S. W. Golomb. Applications of numbered undirected graphs. *Proceedings of the IEEE*, 65(4):562–570, 1977.

[22] T. F. Bloom and O. Sisask. Breaking the logarithmic barrier in Roth's theorem on arithmetic progressions. arXiv:2007.03528, 2020.

[23] A. A. Razborov and S. Rudich. Natural proofs. *Journal of Computer and System Sciences*, 55(1):24–35, 1997.

[24] S. Aaronson and A. Wigderson. Algebrization: a new barrier in complexity theory. *ACM Transactions on Computation Theory*, 1(1), 2009.

[25] K. Buzzard. Formal or not formal? That is the question in AI for theorem proving. Xena Project blog, October 2025.

[26] A. Z. Wagner. Constructions in combinatorics via neural networks. arXiv:2104.14516, 2021.

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

[28] *LLMs gaming verifiers: a case study.* ICLR LLM Reasoning Workshop, 2026. arXiv:2604.15149.