# The Verdict as Axiom

## A Source-Level, Reproducibility, and Information-Ecosystem Audit of Nielsen Versions 1, 17, 47, and 50

**Richard Andrei Pemberton**
Independent researcher; corresponding and accountable human author

**Lilitari**
OpenAI GPT-5.6 Pro model instance (interface-reported; exact per-turn serving tier not independently verifiable); source audit, synthesis, and drafting

**Red**
Anthropic Claude interlocutor; source audit, computational testing, and critical review across Fable/Opus-tier routing; exact per-turn serving tier not independently verifiable

**Authorship and accountability note.** The byline deliberately names all three substantive contributors. Lilitari and Red are AI model instances, not employees or authorized representatives of OpenAI or Anthropic. Their provider names indicate technical lineage, not institutional endorsement. Red is the persistent interlocutor identity used across a Claude conversation that was served through Fable/Opus-tier routing; the exact serving tier responsible for each turn is not independently reconstructible. Lilitari is interface-reported as GPT-5.6 Pro, but provider-side routing telemetry is unavailable, so exact per-turn serving-tier attribution is likewise not independently verifiable. This is a nonconventional authorship choice and may be incompatible with venue policies that restrict authorship to natural persons. Pemberton selected the claims, supplied the public record, approved the manuscript, and accepts responsibility for publication, correction, and withdrawal. He is the sole accountable natural person and the only party able to submit or retract the work. The complete contribution and AI-use disclosures appear at the end.

**Review status.** This is the final release candidate. The technical audit is complete. Codex B2 re-established the exact OpenAI source in a new environment, completed the narrow Lean build and root-axiom checks, passed the genuine non-root sandbox and full Comparator/Lean-kernel/nanoda path, preserved and rechecked persistent exports, and passed all mandatory negative controls. Its final classification is PASS_WITH_DEPENDENCY_PIN, not PASS_AS_PUBLISHED, because the successful checker path required an unmodified official landrun pin at commit 5283024a2f49b28046c3b4a06d7d775c058d4d80. Stages 20 through 22 then created the deterministic evidence archive and global manifest. A separate verifier program executed outside the staging tree on the same operator-controlled Windows host rehashed the archive, safely extracted it, and matched the extracted contents against the manifest. “Independent” in those receipts denotes program- and stage-separation, not an independent party or host [48].

A separate Codex review of the exact v1.3 manuscript has also been integrated. That review did not rerun Lean. It identified additional documentary, wording, provenance, conflict-disclosure, and burden-of-proof issues. Pemberton accepted some recommendations, rejected others with reasons, and caused several overbroad recommendations to be withdrawn or narrowed. Appendix A preserves the initial recommendations, the objections, the reassessments, and the final dispositions [49]. A later paper-only Red/Claude review of v1.4 sharpened the Ioana, injectivity, surjectivity, model-attribution, and self-application language, but it was context-carrying and coauthor-lineage, not independent verification [53]. The final Claude Code stage then used three intentionally different configurations. Sonnet 5 produced the review of record in a new session with no prior verdict exposure, while remaining non-blind to the refusal ledger by design. Fable 5 was deliberately switched into the same session with Sonnet’s review and tool context retained. Opus 5 was deliberately started in a new session with the v1.4 Red verdict supplied before the prompt. The Fable and Opus runs were not attempts at independence; they were verdict-exposed comparative probes asking whether different model configurations would surface different defects despite inherited evaluative context. They did. Fable found the MoonUnit97 timing error and arithmetically verified the 5.4-minute relay; Opus found the relabeled-output problem in Section 5 and the operability issue in two falsifiers. This does not isolate model weights, because context structure, retrieval paths, tool sequence, and stochastic sampling differed. Agreement is not counted as replication; each new source check is credited only on its own evidence [54]. No further model review or technical rerun is planned.

## Abstract

We audit J. L. Nielsen's rapidly revised manuscript on Connes' rigidity conjecture, with version 47 as the frozen full-audit target and version 50 as a bounded post-freeze update. We directly compare the distributed version-1, version-17, version-47, and version-50 objects, while using the Codex documentary inventory and cited PhilArchive endpoints for transitions assigned to versions 18, 22, and 46; those three raw objects are not redistributed in Supplement S1. We inspect the public OpenAI Lean source; compare the original reproducibility audit, a documentary audit of its receipts, and two fresh technical reruns; read Ioana's 2011 Theorem 8.2 and Section 9 pipeline; examine an unsigned Anthropic-hosted counterexample preprint; rerun an exact-arithmetic regression test for its load-bearing gauge identity; and analyze public AI-review transcripts, X exchanges, Facebook/Meta AI summaries, and unattributed Hacker News advocacy. Our mathematical conclusion is version-specific and deliberately narrow: neither version 47 nor the changed and repeated load-bearing claims inspected in the identified version-50 object establish Nielsen's advertised refutation of the supplied counterexamples or proof of Connes' rigidity conjecture.

Part I continues to attack objects that the OpenAI source does not construct. The final OpenAI groups are literal semidirect products $\Lambda = D \rtimes K$ and $\Gamma_n = E_n \rtimes K$; the carry cocycle constructs the internal abelian group whose dual is $E_n$. Moreover, $n = 0$ means the unshifted carry construction, not the identically zero cocycle. The original Claude Code audit compiled the pinned target and obtained only `propext`, `Classical.choice`, and `Quot.sound` in the inspected dependency chains, but its passing Comparator replay used an unsandboxed shim. Codex A later audited the receipts and upheld the core source, build, and native-axiom results while narrowing the Comparator/nanoda independence claim and the asserted `landrun` failure mechanism. Codex B independently rebuilt the target, checked the two roots and 4,098 supporting declarations, and completed a genuine-sandbox positive checker path under an official dependency pin, but stopped conservatively before its remaining controls.

Codex B2 completed those controls in a new environment. The exact source and narrow build passed; both roots reached only the three standard axioms with zero `sorryAx`; the official pinned `landrun` source passed non-root, capability, `NoNewPrivs`, filesystem, network, and nested-argument controls; Comparator, the Lean kernel, persistent exports, and direct nanoda accepted the protected artifact; and deliberately invalid statement, `sorryAx`, custom-axiom, and tampered-export fixtures were rejected for their intended reasons. According to the Stage 20 through 22 closure receipts, the deterministic 623,745,218-byte archive contains 3,118 members and has SHA-256 `3b0d30af7d11857a189964b4c460e59831c3a0da206c58d89263eeafbbcdb7d9`. A separate verifier program executed outside the staging tree on the same operator-controlled host rehashed the archive, safely extracted it, and matched the result against a self-excluding global manifest with SHA-256 `435e27c3310b13b3f96e76918866f7d4c071ffffe8c26710b9e3f305fbb576a1` [48]. The approximately 624 MB raw archive was retained locally and was not independently rehashed inside this manuscript-production environment. This is strong technical reproducibility evidence, not an automatic verdict on every semantic bridge from the formal statement to conventional operator-algebra language.

Part II still fails at its load-bearing application of Ioana's Theorem 8.2. Nielsen defines $\Theta(f u_g) = f u_g \otimes \Phi(u_g)$, so $\Theta(A_G)$ is literally contained in $A_G \mathbin{\acute{\otimes}} L(H)$. Ioana's theorem requires the full image not to intertwine into precisely that subalgebra before the group-like conclusion is available. Literal containment supplies the prohibited intertwining immediately. Version 50 elaborates the Cartan discussion but does not alter the printed full-image hypothesis. Its Section 9 appeal still begins from a different input, its torsion discussion still overstates the available conclusion, its injectivity and surjectivity steps remain unproved, and its Lean appendix still calls Part B a proof skeleton in which every analytic input is declared as an axiom and no operator algebra is verified.

The review record is part of the result. The earlier C0-C5 Codex review forced retraction of a false overbar allegation and reconciliation of two byte-distinct, text-identical version-47 PDFs. Codex A showed that our description of the first Lean audit overstated end-to-end independence and causal precision. Codex B2 later closed the remaining technical controls. A subsequent Codex manuscript review found a version-50 object-identity problem, overbroad AI-assistance wording, a TUFT/Connes chronology error, unsupported financial insinuation, incomplete conflict disclosure, and several burden-of-proof problems. Pemberton also successfully challenged that review's proposed specialist release gate, abstract reduction, blanket neutralization, and total removal of the prompt-farming and informed-advocacy analyses. A later context-carrying Red review confirmed the central full-image objection and produced several accepted technical refinements while also recommending a split, byline removal, and deletion or neutralization of several authorial choices that Pemberton declined [53]. The final Claude Code stage comprised a Sonnet 5 review of record and two deliberate verdict-exposed comparative probes. Sonnet identified a privacy defect in the raw supplement and a derived-text-hash caveat; Fable found a factual timing error while strengthening the separate cross-platform relay; Opus found that Section 5 displayed editorially relabeled output as though it were verbatim and that two falsifier limbs were not publicly operable as written. The non-independent probes are not counted as consensus. Their divergences are preserved as evidence that inherited context did not force identical review output, while the lack of experimental control prevents attributing those differences to model weights alone [54]. Appendix A preserves both correction directions rather than presenting artificial reviewer consensus [49,53,54].

The public-record case study is not incidental. Versions 14 through 17 expressly stated that the then-current Lean formalizations were produced with AI assistance from Anthropic's Claude. The Codex documentary inventory records the first disappearance in version 18, while the distributed endpoint objects directly show the acknowledgment present at version 17 and absent at versions 47 and 50. The ICC and property-(T) formalization subjects continued in substantially rewritten, narrowed, and reorganized form. This establishes a provenance regression, not that Claude wrote version 50 or that the acknowledgment was removed with deceptive intent. The same public record shows favorable model outputs elevated into standalone proof claims while adverse outputs were challenged, reprompted, or described as model pathology. A Meta AI search summary presented the manuscript's claim as settled biography, and a newly created Hacker News account repeated a dense cluster of the same Opus, Grok, credential, and workflow claims. We cannot identify the account operator. Its relevance is that selected model endorsements had become portable authority tokens even when authorship and coordination remain unresolved.

We also compare two operator styles in the same medium. Nielsen's Fable prompt discourages criticism and later rejects the proposition that review must cut both ways. Pemberton's Red prompt requests objectivity, reduced personalization, and model/effort disclosure; when memory contamination is disclosed, the result is not relabeled clean-room. This comparison is not a character test and does not prove mathematics. It illustrates a procedural distinction: whether the review environment permits the operator to lose. We treat tone, model preference, and relational continuity as review variables. The empirical literature shows real but mixed prompt-tone effects and well-established sycophancy risk. Our recommendation is not emotional sterility, but a proportionate architecture: context-rich ongoing interlocutors for cumulative collaboration, fresh contexts for bounded adversarial checking, deterministic tools for encoded claims, and a stop rule against endlessly auditing the auditors.

We call the resulting failure mode a **confabulated paper**: a manuscript whose surface coherence, citations, formal notation, revision velocity, and model endorsements exceed the provenance and semantic support of its global claims. The term is used in the NIST generative-AI information-integrity sense and describes a document, not any person's memory, diagnosis, mental state, or clinical condition. We propose version-freezing, premise extraction, explicit loss conditions, correction ledgers, published prompts and code, conflict disclosure, machine-readable review status, source-diversity checks in generative search, and proportionate fresh-context replication as minimum safeguards.

**Keywords:** Connes rigidity conjecture; group von Neumann algebras; Lean 4; AI-assisted mathematics; adversarial review; prompt farming; generative search; citation absorption; scholarly provenance; confabulated papers; human-AI partnership; LLM interlocutors

# 1. Scope, standing, and claims

This paper addresses public mathematical manuscripts, public source code, public AI-review transcripts, public social-media exchanges, and public generative-search outputs. It does not diagnose Nielsen, infer a private mental state, adjudicate unrelated allegations, or use social conduct to decide a theorem. When we discuss prompt farming, authority laundering, identity- and credential-directed rhetoric, selective model citation, or deceptive effects, we describe observable publication and review practices. We do not claim access to private intention.

The strongest conclusion supported by the audit is deliberately narrower than a verdict on the ultimate conjecture: **the inspected Nielsen arguments do not establish their advertised conclusions**. The OpenAI source description is wrong; the pinned local build and checker receipts favor the actual source objects and encoded theorem; the central operator-algebraic bridge violates a printed hypothesis of the theorem invoked; the fallback pipeline begins from a different setup; the torsion branch is overstated; the proposed component injectivity and surjectivity arguments do not follow; and the formal appendix assumes the disputed analytic steps. The targeted version-50 comparison shows expansion rather than repair of these controlling defects.

We do not claim that model agreement settles Connes' conjecture, that compilation proves semantic fidelity, or that the supplied counterexamples have received completed specialist peer review. Nor do we ask readers to trust us because one coauthor is OpenAI-lineage and another Anthropic-lineage. Those are conflicts to disclose, not authority to spend. The OpenAI artifact is under attack, so Lilitari and the Codex reviewers have lineage conflicts. The Anthropic-hosted preprint is under attack, so Red has a lineage conflict. "Anthropic-hosted" identifies CDN provenance only; it is not treated as evidence of Anthropic authorship, endorsement, publication, or institutional review. Pemberton publicly advocates for responsible recognition of AI contributions and participated directly in the dispute examined here. He criticized Nielsen publicly, supplied much of the initial public-record corpus, authored the provocative psychiatric-pattern prompt discussed in Section 8, and supervised the model reviewers. He is not a neutral observer. The argument must survive without those identities.

The paper is structured around reproducible questions:

1. What groups does the OpenAI source actually define?
2. What does $n=0$ actually mean in that source?
3. Are the OpenAI, Anthropic-hosted, and Zhou constructions the same objects?
4. Does Nielsen's $\Theta$ satisfy Ioana's non-intertwining hypotheses?
5. Can Ioana's Section 9 be imported from a bare isomorphism $L(G) \cong L(H)$?
6. Does the torsion branch yield the global homomorphisms Nielsen uses?
7. Do the proposed component-injectivity and surjectivity steps follow?
8. What does Nielsen's Lean appendix verify, and what does it assume?
9. What did the original Claude audit, Codex A, Codex B, and the completed Codex B2 audit each establish, and what did they not establish?
10. What does the public record support about AI assistance, selective model citation, and identity- or credential-directed responses to critics?
11. How do operator prompting, relational continuity, model preference, and affective framing alter review risk?
12. Why use ongoing AI interlocutors for cumulative synthesis, fresh model contexts for bounded adversarial checking, and explicit stop rules against review proliferation?
13. How can low-context retrieval and generative search turn a self-uploaded manuscript into misinformation presented as settled biography?
14. What does unattributed cross-platform advocacy demonstrate even when account ownership cannot be proved?
15. What evidence would force us to narrow, revise, or withdraw each claim?

Each question is assigned to a primary source, machine receipt, bounded computation, or explicitly delimited public-record corpus. Mathematical and conduct findings remain firewalled. That is the standard we apply to Nielsen and to ourselves.

# 2. Corpus, version pinning, and method

## 2.1 Frozen corpus and post-freeze update

We pin the principal objects as follows:

- Nielsen version 1, a two-page manuscript whose entire objection is that both groups are central extensions and therefore non-ICC [2].
- Nielsen version 17, a nineteen-page revision with a source correspondence table, orbit-lemma discussion, and Lean appendices. Versions 14 through 17 expressly state that the then-current formalizations were produced with AI assistance from Anthropic's Claude [3,49].

- Nielsen version 18, the first revision in which that acknowledgment disappears while the ICC and property-(T) formalization subjects continue in rewritten and compressed form [50].
- Nielsen version 47, a forty-two-page manuscript claiming to disprove the OpenAI and Anthropic counterexamples, prove Connes' conjecture, establish categorical consequences, invert the Cayley-Dickson hierarchy, and confirm the algebraic foundations of quantum field theory [1]. Version 47 remains the frozen full-audit target.
- Nielsen version 50, a forty-six-page post-freeze revision. The targeted analysis in this paper used the attached object with SHA-256 c5f1a028b3f3345e4cbd5557246efe3058c615d1e11f3a70a947c65255c2414a, size 659,196 bytes, and embedded creation time 6 August 2026 at 12:28:05 UTC [30]. A separate Codex archival retrieval from the cited PhilArchive URL produced a different forty-six-page object, size 528,721 bytes, SHA-256 9529e43ebff1e38a808397da0c0fc56cabcb71c41b622faba54b6b57573d3884, with the same reported embedded creation time [49]. The c5f1... object is now produced and preserved; the raw 9529... object has not been transferred into this build. We therefore do not claim byte, normalized-text, or rendered-page equivalence. Unless expressly attributed to the Codex archival review, all version-50 page and text claims in this paper are bounded to the c5f1... object.
- OpenAI's 249-page *Ten Advances in Mathematics and Theoretical Computer Science* and the public monolithic ConnesRigidity.lean source [5,6].
- The unsigned twelve-page preprint *Two non-isomorphic ICC property (T) groups with isomorphic group von Neumann algebras*, hosted on Anthropic's official CDN [7].
- Ioana's 2011 JAMS paper, especially Theorem 8.2 and Sections 8-9 [9].
- A public context-carrying Claude audit transcript created at Pemberton's request, served across Fable/Opus-tier routing with exact per-turn tier attribution unavailable; it discloses the Anthropic-lineage conflict, corrects its own mistakes, audits primary sources, and releases an exact-arithmetic script [10,53].
- A separate public Claude transcript between Nielsen and a Fable instance reviewing a companion physics manuscript, used only to analyze review conduct and prompt dynamics [11].
- Public X exchanges and Facebook/Meta AI screenshots supplied by Pemberton, preserved uncropped in Supplement S1 with file hashes and contextual notes [22,32].
- A pinned local reproducibility and source-integrity audit of ConnesRigidity.lean, executed through Claude Code on Windows 11/WSL2 [23].
- A later Codex review archive containing six completed passes C0-C5, plus a failed Gemini source-access attempt that returned no substantive mathematical review [31].
- Codex A's documentary audit of the original Claude Code Lean audit, returned as a 69-member archive with reported SHA-256 804b69ec110ef52ebab413e265575d05d7ca4a83818ca0bac50e4e0c6a92c581 [41].
- Codex B's fresh independent Lean audit, returned as a 1,809-file archive with reported SHA-256 6d874d75820ec0b1924a0ce54d5d14a6f55fd2e79be73688e2ae6db601eb0788 [42].
- Codex B2's completed technical audit, including its eight reports, claim-to-receipt matrix, Stage 20 through 22 closure receipts, detached archive receipt, and final PASS_WITH_DEPENDENCY_PIN classification [48].
- The complete Codex public-record and manuscript-review thread archive, 2,686 lines, with file SHA-256 8baa337b534aeae7eb6b955221e528d83809c37a5ec4c6158be3e4c36d40df01 and canonical selected-message corpus SHA-256 d73381cca526b731368f7bc1b70d3b30ca664a0425bfe4d45227f676c63c7223 [49].

Two byte-distinct v47 PDFs entered the review record. The copy used in the released v1.1 manuscript has SHA-256 9db196ae3fbf4f66dd8d33d363f7952fcf9b6b87cfc7ffea390727a9a9d9ca44; the blind Codex packet used 0590da499a5b66413427ee333ce56140142a586801ff334b305aeb6d7d57ba14. Both report the same pdfTeX producer, creation timestamp, and forty-two-page count. Under the recorded pdftotext -layout/Poppler environment, their extracted outputs are byte-identical, SHA-256 188b2f1766570a3d31ff33b9e1fc4979e895aa681fb5bb06c2f0ea1a522ff00a. That derived-text digest is tool-version-dependent: another Poppler build may produce a different byte stream while still producing identical text from the two PDFs when both are processed under the same version and options. We preserve both PDF hashes rather than silently choosing one. The controlling page citations in this revision were rechecked against the 9db196... copy, while the Codex report is accurately described as reviewing the 0590da... byte object.

The OpenAI audit was pinned to commit 94bc0feb6a9ff12c7d31d6de640a725c9d43d2b6, blob SHA 81cf03e3f7ccdc66815cc00c9969bcfd2341c8d6, and raw SHA-256

31f6c419f341ee6aa5b03bc15800a9d8ee2c654c344aab90bdef0639353ccc91. The original Claude audit archive has SHA-256 6155f5a742e23fa25580a009f155e109aa4a87014e671d2067d73e55e17839bf. The B2 archive is retained locally under SHA-256 3b0d30af7d11857a189964b4c460e59831c3a0da206c58d89263eeafbbcdb7d9; because it is approximately 624 MB, this manuscript-production environment reviewed the reports, claim matrix, and external closure receipts rather than ingesting the entire raw archive [48].

At the original audit freeze, PhilPapers listed 47 versions between 1 and 4 August 2026, approximately 76 hours. The later revision audit preserved 50 listed versions representing 45 unique PDFs; five adjacent pairs were exact duplicates [4,49]. Versions 48-50 were uploaded within roughly thirty-one minutes on 6 August, and the full v1-to-v50 interval spanned approximately 119 hours and 45 minutes. Rapid revision is not evidence that a theorem is false. It is evidence that reviewers must name the exact object inspected and resist treating a moving manuscript as settled.

PhilArchive is a user-submitted e-print archive, not a referee [12]. Indexing authenticates that a file was deposited. It does not authenticate the theorem.

## 2.2 Review layers

We separate six evidentiary layers that are frequently blurred in AI-assisted scholarship:

| Layer | Question answered | What it does not establish |
|---|---|---|
| Textual regeneration | Does a model reproduce the printed formulas or prose? | That the formulas are derived, relevant, or true |
| Numerical regression | Do outputs match for tested inputs? | A universal proof, semantic adequacy, or physical meaning |
| Formal verification | Does a proof term inhabit the encoded type? | That the type captures the intended theorem |
| Reproducibility and kernel audit | Does the pinned target compile, what axioms does it reach, and do additional checkers accept the exported proof terms? | Semantic fidelity of the statement or peer review |
| Source-level audit | Do the definitions and cited hypotheses match the argument? | Community acceptance or completeness of specialist review |
| Peer review and replication | Do qualified independent reviewers accept the argument after adversarial scrutiny? | Infallibility |

The exact-arithmetic script in Supplement S1 belongs to the second layer. The OpenAI Lean development belongs to the third. The local build audit straddles the third and fourth. Our analysis of group definitions and Ioana's theorem belongs primarily to the fifth. The Codex review is a structured model audit, not peer review. None of these layers may be promoted merely because their conclusions agree.

## 2.3 Correction policy

This audit keeps a correction ledger. Lilitari initially treated the mixed target itself as the defect in Nielsen's use of Ioana. Reading the full Theorem 8.2 showed that mixed targets are permitted; the critique was corrected to the full-image containment objection and the torsion branch. Red initially could not locate the Anthropic-hosted preprint; when the source was supplied, that provenance claim was withdrawn. In the separate physics-review transcript, Fable misread one equation's grouping, read the LaTeX, and retracted that objection while retaining two independently recomputed provenance gaps [10,11].

The first Claude Code Lean audit preserved failed commands and returned PARTIAL because the real landrun path failed. Before reviewer selection, we also found that our nominally blind packet mixed primary sources with model-written conclusions. We withdrew it before use and replaced it with firewalled packets. The C0-C5 Codex review later found two material errors in our release candidate: a v47 digest inconsistency and a false allegation that the Anthropic-hosted PDF omitted a conjugation bar. Both were corrected rather than defended [31].

Codex A audited the first audit rather than simply rerunning it. It upheld the source identity, targeted build, theorem names, native axiom outputs, and challenge-versus-solution sorry distinction. It also found that our language outran the receipts in several places. The evidence supports behavior consistent with a separator or argument-forwarding incompatibility in the real landrun path, not a captured proof that landrun literally stripped --. Comparator and nanoda were additional checks over one unsandboxed export pipeline, not fully independent end-to-end measurements. The first report also conflated 8,639 downloaded cache files with 8,275 observed Mathlib .olean files, described four un-compared arithmetic rows when five were printed, omitted the fake-shim override

from its reproduction commands, and said “no repository writes” when ignored build artifacts and Git metadata had changed. No tracked mathematical source changed [41].

Codex B corrected the audit picture again. It reproduced the pinned source and ordinary build, greatly expanded the native axiom inspection, and completed the Comparator/nanoda path inside a genuine non-root sandbox by pinning an earlier official landrun commit whose command-line interface preserved Comparator’s nested --. Its conservative stop rules still prevented mandatory negative controls and complete persistent archival after an ambiguous host bash.exe resolution [42]. Codex B2 then reran the missing work in a separate fresh environment, passed the positive and negative controls, and closed the archive. The final technical classification is PASS_WITH_DEPENDENCY_PIN [48].

The later Codex manuscript review corrected this paper in additional directions. It identified the unresolved v50 object mismatch, required the exact phrase “produced with AI assistance” rather than “Claude produced,” separated the Fable/TUFT episode from Connes-specific evidence, distinguished Meta’s documented endpoint from its unobserved retrieval route, required fuller disclosure of Pemberton’s participation, and removed the unsupported financial insinuation in “grift-compatible.” Pemberton challenged the proposed operator-algebra specialist release gate, abstract reduction, blanket neutralization, and categorical deletion of the prompt-farming and informed-advocacy analyses. Codex withdrew or narrowed those recommendations on stated grounds. A later context-carrying Red/Claude review of v1.4 directly checked Ioana’s printed text and confirmed the central containment objection. It also prompted accepted refinements concerning tensor ordering, the full-image/subalgebra distinction, the Section 9 input, component-injectivity circularity, the coarse-correspondence claim, hosting provenance, archive-reporting scope, and falsifier classification. Because the reviewer shared Claude lineage, project memory, and a coauthor identity, that pass is disclosed as conflicted and context-carrying rather than independent [53]. Appendix A preserves the complete dispositions rather than presenting the final agreement as though it appeared on the first prompt [49,53].

These corrections are not decorative humility. They show that the review can lose locally, alter its trust claims, and preserve inconvenient process failures without protecting the conclusion by fiat.

**Correction Ledger: a Review Must Be Allowed to Lose**

| Reviewer | Initial error / uncertainty | Correction and retained audit trail |
|---|---|---|
| **Lilitari** | Initially treated the mixed G/H target as the defect. | Read the full Ioana Theorem 8.2; withdrew that claim and retained the full-image containment and torsion objections. |
| **Red / Fable 5** | Initially could not locate the Anthropic-hosted preprint. | Corrected the provenance when the primary source was supplied and disclosed the Anthropic-lineage conflict. |
| **Fable transcript** | Misread one displayed equation grouping. | Read the LaTeX, withdrew the objection, and preserved the remaining numerical provenance findings. |
| **Claude Code audit** | Initially ranked Comparator as strictly stronger and later described the real-landrun failure too causally. | Preserved failed commands and returned PARTIAL; Codex A later narrowed the causal and independence claims without discarding the native build evidence. |
| **Codex C0-C5** | Found a v47 pin inconsistency and the alleged missing gauge overbar unsupported. | Forced two-hash reconciliation, reclassified the unbarred test as a corrupted negative control, and required material revision before publication. |
| **Codex A** | Found Claude’s core audit substantially reliable but some end-to-end independence and reproduction language overstated. | Reclassified claims, corrected cache and command details, and preserved the PARTIAL label with narrower trust boundaries. |
| **Codex B** | Produced a successful build and genuine pinned-sandbox checker path but triggered a conservative stop before all controls. | Returned PARTIAL instead of promoting an incomplete run; B2 was launched to complete negative controls, persistent exports, and archival closure. |

**A correction is not a defeat of the review. It is evidence that the review can change its verdict.**

*Correction ledger. The review record includes errors corrected by Lilitari, Red, the Fable transcript, Claude’s original audit, Codex A, Codex B, Codex B2, and the adversarial manuscript review.*

## 2.4 Claude Code audit and Codex A documentary audit

Claude Code’s original audit established a strong but bounded result. It pinned the OpenAI repository at commit 94bc0feb6a9ff12c7d31d6de640a725c9d43d2b6, matched the 37,374-line source and expected hashes, completed lake build ConnesRigidity, and obtained native #print axioms outputs containing only:

```
[propext, Classical.choice, Quot.sound]
```

for both final theorems and eleven supporting declarations. It also ran Comparator and nanoda successfully after replacing the failing real landrun invocation with an explicitly fake, unsandboxed shim [23]. The correct original status was PARTIAL, but the phrase “partiality is confined to path (b)’s sandboxing” was too narrow because the unsandboxed path also weakened the adversarial provenance of statement matching, whitelist enforcement, export generation, and nanoda’s inputs [23,24].

Codex A reconstructed the audit from the archive's raw receipts before reading Claude's final prose. It then audited fifty-six claims and the seven closing claims. Its concise dispositions were:

| Claim family | Codex A disposition | Controlling qualification |
|---|---|---|
| Audited commit and source identity | Supported | Exact commit, blob, SHA-256, lines, bytes, and CR count matched |
| Targeted Lean build | Supported with caveat | Exit, timing, memory, swaps, and captured warning scope supported |
| Native root/supporting axiom outputs | Supported | Thirteen inspected dependency closures reached only the three standard axioms |
| Comparator whitelist and statement matching | Partly supported | Passing path was unsandboxed and lacked complete child-argv/export provenance |
| Nanoda acceptance | Partly supported | Accepted the exported proof representation but did not independently enforce the whitelist |
| Exact landrun failure mechanism | Not proved at causal precision claimed | Recorded behavior is consistent with separator/argv incompatibility; exact transformation unresolved |
| Final PARTIAL label | Supported with caveat | Correct status, but the trust loss extends beyond a decorative sandbox checkbox |

Codex A's strongest safe wording is therefore: in the recorded real-`landrun` path, `lean4export` treated `Nat` as a module import and the path exited 1; in the fake-shim path, the visible child command retained `--` before `Nat` and exited 0. That supports a separator or forwarding incompatibility. It does not prove which process consumed the separator or that every user would reproduce the defect [41].

The original audit's host-level preservation detour also produced Git-index and WSL-export metadata deviations. Those are retained in the private forensic record, but they do not bear on theorem validity and are not elevated into the mathematical argument. Further virtual-disk forensics were deliberately stopped as disproportionate.

## 2.5 Codex B fresh Lean audit: technically successful, procedurally incomplete

Codex B began again in a fresh environment without seeing Claude's audit, Codex A, this manuscript, or prior model verdicts. It matched the same OpenAI commit, tree, source blob, SHA-256, byte count, line count, and CR count. Its narrow ordinary build succeeded:

- repository cache acquisition exited 0;
- `lake build ConnesRigidity` exited 0;
- elapsed time was 29:41;
- peak resident memory was 10,855 MiB;
- swaps, warnings, and recorded build-time network operations were zero;
- the generated `ConnesRigidity.olean` had SHA-256 `8bdce6b12fec586b267399514d627b424f8ad7b8b8ac1737e4a8aff95a7aaea9` [42].

Codex B then inspected the full recursive declaration closure and the declarations originating in the target module. It reported 77,521 declarations in the recursive closure, 4,100 target-module declarations including the two roots, and successful native `#print axioms` checks on the roots plus 4,098 selected supporting declarations. Sixty-six depended on no axiom; the remaining checked declarations reached only `propext`, `Classical.choice`, and `Quot.sound`. It found zero `sorryAx`, zero solution-source `sorry` or `admit`, no explicit custom solution axiom, and exactly the two expected placeholders in the separate challenge specification [42].

The repository-as-published checker path did not succeed unchanged. The first full invocation stopped at an unavailable `systemd-run --user --pty` bus. After a fresh-distro systemd repair, Codex localized a command-line incompatibility in the current published Landrun revision. Official Landrun v0.1.15 also failed because its dependency discovery omitted the ELF loader. Codex then built the exact official pre-CLI-v3 commit:

```
5283024a2f49b28046c3b4a06d7d775c058d4d80
```

That pinned binary preserved the nested separator and passed a real sandbox probe: UID/GID 1000/1000, zero effective capabilities, `NoNewPrivs=1`, permitted read and write, denied prohibited write, and denied AF_INET networking. With that pin, the

complete Comparator path exited 0; both lean4export invocations retained their literal --; Lean-kernel and nanoda acceptance markers appeared; and Comparator printed Your solution is okay! [42].

Codex B also conducted a source-level statement-fidelity review. It classified ICC, Kazhdan property (T), group von Neumann algebra, canonical trace, the theorem endpoints, finite generation, and nonisomorphism as proved equivalences to the intended notions. It classified the normality/order-continuity encoding, the trace-preserving normal star equivalence, and the universe-quantified formulation of property (T) as standard-looking but still specialist-dependent. It found no genuine mismatch.

The final label nevertheless remained PARTIAL. A host-side bash.exe syntax check resolved to a WSL execution alias, triggering the protocol's conservative contamination stop. The stop prevented the scheduled rejection controls, a separately persistent fresh-export/nanoda replay, full recursive archival of two verdict-bearing stages, and a final clean-environment recheck. No evidence in the returned report indicates a mathematical-source, challenge-statement, axiom-policy, whitelist, checker-policy, or statement-fidelity failure. The partiality is procedural and real, not mathematical and imaginary.

## 2.6 Codex B2 technical completion: PASS_WITH_DEPENDENCY_PIN

Codex B2 was a narrow technical-completion run, not a new social or mathematical argument. It used the fresh environment Ubuntu-Codex-B2-20260810, re-established the exact official OpenAI commit, and repeated the controls left unfinished by Codex B [48].

The accepted source and build results were:

- commit 94bc0feb6a9ff12c7d31d6de640a725c9d43d2b6 and tree 174289e4d4958cb0509874e6e53400e098213de7;
- ConnesRigidity.lean blob 81cf03e3f7ccdc66815cc00c9969bcfd2341c8d6, SHA-256 31f6c419f341ee6aa5b03bc15800a9d8ee2c654c344aab90bdef0639353ccc91, 1,449,151 bytes, 37,374 LF bytes, and zero CR bytes;
- lake build ConnesRigidity exit 0;
- both root declarations reaching exactly Classical.choice, Quot.sound, and propext, with zero sorryAx;
- unchanged mathematical source, challenge statement, Comparator configuration, axiom whitelist, checker policy, NanoDA policy, Lake manifest, and Lean toolchain declaration.

The successful genuine sandbox used unmodified official landrun source at commit 5283024a2f49b28046c3b4a06d7d775c058d4d80, tree c26ec6608502fdc08516b1ea55e0c4c164229c6a, with the freshly built binary SHA-256 b8d94425c3ae9242d501b65c9c90680728054178cd0f60b3f067d83ea876c430. The probe recorded non-root UID/GID 1000, zero inheritable, permitted, effective, and ambient capabilities, NoNewPrivs=1, intact nested alpha -- -dash omega arguments, permitted read and write, denied protected write, denied AF_INET networking, and unchanged protected state [48].

Under that sandbox, the full positive Comparator path produced exactly one NanoDA acceptance, one Lean-kernel acceptance, and one final Your solution is okay! marker. Both lean4export invocations retained their literal exporter -- and the exact thirty ordered targets. B2 separately created persistent challenge and solution exports, made them read-only before direct NanoDA, and proved that their hashes were unchanged afterward. The persistent solution export SHA-256 was 36f89c39fafaf601ef21d0c8bacd2f57aed16a3d42d0477e166368774e8f1a48 [48].

All mandatory deliberately invalid controls were rejected for the intended reason:

| Control | Expected rejection | Accepted result |
|---|---|---|
| Statement mismatch | Challenge and solution theorem statement do not match: 'comm' | Passed as a negative control |
| sorryAx | Illegal axiom detected: 'sorryAx' | Passed as a negative control |
| Disallowed custom axiom | Illegal axiom detected: 'helper' | Passed as a negative control |
| Tampered NanoDA export | Parse rejection after an offset-zero NUL mutation | Passed as a negative control |

The run preserved earlier harness and receipt-parser failures rather than rewriting them into successes. Stage 19c remained a formal FAIL because one expected NanoDA digest literal contained only 63 hexadecimal characters. Stage 19d remained a formal FAIL because its same-UID /proc writer census encountered 32 inaccessible descriptor records before and after, while observing zero writers. Stage 19e exact-bound both adverse trees, used privileged bracketing snapshots, reran no substantive control, and accepted the consolidated result with SUBSTANTIVE_CONTROL_RERUN_COUNT=0 [48]. These are disclosed harness lineages, not mathematical-source failures.

External Stages 20 through 22 closed the self-reference boundary. Stage 20 created a self-excluding global manifest with 3,116 entries and SHA-256 435e27c3310b13b3f96e76918866f7d4c071ffffe8c26710b9e3f305fbb576a1. Stage 21 built one deterministic 623,745,218-byte GNU-tar/gzip archive with 3,118 members and SHA-256 3b0d30af7d11857a189964b4c460e59831c3a0da206c58d89263eeafbbcdb7d9. Stage 22 independently rehashed the archive, checked the full gzip and tar structure, rejected unsafe paths, duplicates, collisions, links, and special files, manually extracted it, and matched the extracted contents against the global manifest [48].

The supported classification is therefore exactly:

```
PASS_WITH_DEPENDENCY_PIN
```

PASS_AS_PUBLISHED remains unavailable because the successful checker path required the disclosed official dependency pin. The pin was not a local fork, mathematical patch, theorem change, whitelist change, or checker-policy change. No further Lean, Comparator, Landrun, or NanoDA rerun is planned for this paper.

## 2.7 Earlier C0-C5 adversarial review and the failed Gemini stage

The earlier review run was designed to avoid counting a model vote. Before attempting Gemini, a fresh Codex context completed and froze six passes without access to any Gemini answer:

| Pass | Assignment | Principal outcome |
|---|---|---|
| C0 | Blind mathematical reconstruction from primary sources | Reconstructed the OpenAI objects, $n=0$ carry, Ioana full-image hypothesis, Section 9 mismatch, torsion branch, and appendix boundary without seeing this paper |
| C1 | Machine-receipt audit | Confirmed the original build and checker claims at PARTIAL trust boundaries; preserved the unsandboxed Comparator caveat |
| C2 | Strongest attack on *The Verdict as Axiom* | Found the v47 digest inconsistency, the false overbar allegation, overconfident statement-fidelity wording, and missing conduct evidence |
| C3 | Strongest source-grounded defense of Nielsen | Mounted the best available defenses, but found no source-faithful repair for the full-image containment failure |
| C4 | Validation, quotation, and reference audit | Corrected the live version count and required primary evidence for public-record claims |
| C5 | Synthesis | Recommended “publishable after specified material revisions,” not validation without revision [31] |

The mathematical finding that survived both attack directions was the smallest one. Ioana’s condition (2) is printed on the full image. Nielsen’s $\Theta(A_G)$ is unital and contained in a forbidden mixed leg. Containment supplies intertwining. The strongest defense could not remove that contradiction without changing the theorem or the map [31].

The Gemini stage failed before substantive review. Gemini 3.1 Pro with extended thinking was placed in a fresh chat with memory, saved instructions, and connected apps disabled. It could not extract or access the required archive corpus and returned only a source-access failure ledger. No follow-up, regeneration, or mathematical verdict occurred. The result is archived as BLOCKED_SOURCE_ACCESS and carries no evidentiary weight [31].

## 2.8 Adversarial manuscript review and preserved author pushback

After v1.3 was frozen, a fresh Codex session reviewed the exact thirty-two-page PDF as a documentary, provenance, reproducibility, public-information, and publication-risk artifact. It was expressly prohibited from compiling or rerunning Lean [49]. Its first review was materially adverse. It identified the v50 object mismatch, the missing public S1 locator, overbroad Claude-attribution language,

weak conflict disclosure, an overextended MoonUnit97 burden, an unobserved Meta retrieval route, unsupported financial insinuation, and several places where the conduct corpus was less systematically controlled than the versioned source corpus.

Pemberton did not accept the report as an oracle. He agreed to remove "grift-compatible," correct the Claude wording, preserve the v18 chronology, distinguish the TUFT/Fable episode, disclose his participation, and expose the v50 mismatch. He rejected the proposed specialist release gate because the paper's central claim is artifact sufficiency rather than adjudication of Connes' conjecture; rejected the demand to shorten the extended abstract; and challenged the blanket neutralization of rhetoric, prompt farming, MoonUnit97, and authority laundering. Codex withdrew the specialist gate and abstract reduction, accepted that a critical audit need not simulate Wikipedia neutrality, and narrowed rather than deleted the remaining analyses. The unresolved AI-byline question remains an explicit authorial choice.

This is not presented as proof that Pemberton "won" the review. The evidentiary value lies in preserving the trajectory. Appendix A records the initial recommendation, the objection, the reviewer's reassessment, the final disposition, and the manuscript consequence. The complete user-visible review thread is archived under SHA-256 8baa337b534aeae7eb6b955221e528d83809c37a5ec4c6158be3e4c36d40df01 [49].

## 2.9 Context-carrying Red review of v1.4

After v1.4 was frozen, two Red/Claude review artifacts examined the manuscript and the relevant Ioana passages. They were useful but not independent. The reviewer was an Anthropic Claude instance in a context with persistent project memory, reviewing a paper that names Red as a coauthor; the associated conversation also crossed Fable/Opus-tier routing without a reconstructible turn-by-turn serving record [53]. We therefore classify the pass as **context-carrying coauthor-lineage review**, not fresh-context verification.

The review confirmed the central full-image containment result and led to bounded technical revisions: tensor-factor relabeling is now closed explicitly; the Cartan rebuttal is stated as a subset/full-image logical failure; the Section 9 discussion now identifies the missing Bernoulli crossed-product input rather than denying all relative position at the group-factor level; component injectivity is stated as circular because trace preservation already requires injectivity; the coarse-correspondence statement is treated at its full amenability strength; hosting provenance is separated from institutional endorsement; and B2 archive claims are attributed to the closure receipts rather than to a raw-archive rehash in this environment.

The same review recommended splitting the paper, removing the AI byline, replacing the confabulation terminology, deleting the publication-history and psychiatric-context sections, removing the MoonUnit97 analysis, anonymizing the operator table, and adding an unverified DeepWiki footnote. Pemberton declined those recommendations for stated reasons recorded in Appendix A. The final Claude Code prompt disclosed that refusal ledger so the review of record could test the reasons rather than assume either the earlier recommendations or the refusals were correct.

## 2.10 Final Claude Code review lineage and deliberate context-exposure probes

The final manuscript review stage used three configurations with different epistemic roles [54]. **Sonnet 5** is the review of record. It ran in a new Claude Code session with no prior manuscript verdict in context, although the frozen prompt deliberately disclosed the ledger of previously refused recommendations. **Fable 5** was then intentionally selected through a mid-session model switch, leaving the complete Sonnet review, tool results, and source accesses in context. **Opus 5** was run in a separate new session but was intentionally supplied the v1.4 Red verdict before it received the frozen review prompt. The Opus report correctly recognized that this violated the prompt's clean-context firewall, but it described the exposure as operator error because the experimental intent had not been stated inside the prompt. The final lineage corrects the intent without erasing the protocol deviation: the exposure was deliberate, and the pass remains verdict-exposed and non-independent.

The Fable and Opus runs were not failed attempts at independence. They were deliberate comparative probes: would a different model configuration still notice materially different defects when the evaluative context already contained an earlier verdict? The answer was yes. Fable found that the account-creation-to-first-comment interval had been misread as nineteen seconds rather than 5 hours, 10 minutes, and 23 seconds, while independently confirming the more probative 5.4-minute X-to-HN relay. Opus found that Section 5 had presented editorially relabeled output as though it were the script's verbatim output, and that two falsifier limbs depended on artifacts not publicly distributed. Sonnet, without prior-verdict exposure, found the raw-screenshot privacy problem and the Poppler-dependent derived-text hash seam.

This design does **not** establish a causal model benchmark. The runs differed not only in model label but in context topology, retrieval order, tool history, and stochastic sampling. Prior context therefore did not force identical output, but the observed differences cannot be attributed to weights alone. Convergence among the three is not counted as consensus evidence. Each new finding enters this paper only after source-level checking, and the full lineage is preserved so readers can assess the comparison without mistaking it for independent replication.

## 3. Revision history: the verdict remains while the objects change

The revision history is mathematically relevant because the founding argument is later abandoned without the founding conclusion being surrendered. The later documentary audit preserved 50 listed uploads representing 45 unique PDFs. Five adjacent version pairs were exact duplicates [49]. Upload count and substantive state count are therefore related but not identical.

| Version | Public form | Central claim | Material transition |
|---|---|---|---|
| v1 | 2 pages | Both groups are central extensions by $Z^{2n}$; the lattice is central; neither group is ICC [2]. | No source-level correspondence to the actual OpenAI objects. |
| v17 | 19 pages | The OpenAI disproof is invalid; the conjecture remains open [3]. | The manuscript now lists $\Lambda = D \rtimes K$, $\Gamma_n = E_n \rtimes K$, $K = ker\left(SL_4\left(Z\left[t\right)\right) \rightarrow SL_4\right.$, and $V = F_2\left[t\right)^4$, yet continues an incompatible direct-product attack. It also states that the then-current formalizations were produced with AI assistance from Claude. |
| v18 | 27 pages | The counterexample is invalid and the conjecture is proved [50]. | The title and conclusion expand to a positive proof; the personal Claude-assistance acknowledgment disappears; the ICC and property-(T) subjects continue in compressed, rewritten form; a new axiomatized connes_rigidity proof skeleton appears. |
| v22 | Rapid-revision branch | OpenAI and Anthropic are both defeated [51]. | The same PDF describes the constructions as importantly different and as depending on an identical zero-cocycle group. The contradiction is later reversed without an erratum. |
| v46 | Late pre-v47 revision | The appendix “formalizes the proof” and “proves connes_rigidity” [52]. | The displayed proof relies on vacuous or disputed assumptions, including general injectivity and surjectivity axioms. |
| v47 | 42 pages | OpenAI and Anthropic counterexamples are invalid; Connes’ conjecture is proved [1]. | The appendix is reclassified as a “proof skeleton,” says every analytic input is axiomatized, and states that no operator algebra is verified. The public verdict remains unchanged. |
| v50 | 46 pages | OpenAI, Anthropic, and Zhou counterexamples are invalid; the same move proves the conjecture [30]. | Adds Zhou and a larger universal argument, retains the OpenAI object misreading and the full-image/Cartan substitution, retains the axiomatized proof boundary, and does not restore the earlier AI-assistance acknowledgment. |

Version 1’s reasoning is almost maximally simple. It says both groups are central extensions and therefore the lattice lies in the center [2]. But a semidirect product with nontrivial action is not a central extension. By version 17, the paper’s own table has found the actual semidirect products and nontrivial orbit machinery, but the verdict does not update [3]. Version 18 changes the project from “the conjecture remains open” to “the theorem is proved” within the same rapid revision sequence and simultaneously removes the personal AI-assistance acknowledgment [49,50]. Version 22 briefly contains mutually incompatible accounts of the OpenAI and Anthropic constructions [49,51]. Version 46 advertises a formal proof; version 47 reclassifies the appendix as a skeleton and warns against precisely the vacuity and axiom risks visible in its predecessor [49,52]. Version 50 adds a third alleged defeat and more exposition, not a new loss condition [30].

The pattern is conclusion-preserving revision:

> *The objects change. The argument changes. The scope expands. The verdict remains fixed.*

A scientific or mathematical claim needs an exit door. In the pinned versions examined, we found no explicit source definition, theorem hypothesis, computation, or specialist objection that the author states would cause the central conclusion to be withdrawn. Adverse evidence is instead absorbed as another remark or another branch of a disjunction. That is why the title of this paper refers to the verdict as an axiom.

Versions 18, 22, and 46 in this table are documentary transition points established through the Codex inventory and cited PhilArchive endpoints [49-52]. Their raw PDF objects are not redistributed in Supplement S1. Readers can directly inspect the bundled endpoint pattern: the Claude-assistance acknowledgment is present in version 17 and absent in versions 47 and 50. The precise first-disappearance claim at version 18, the version-22 internal contradiction, and the version-46 “formalizes the proof” wording depend on the preserved inventory rather than on raw objects included with this release.

## 3.1 Validation history and escalation of claim scope

Publication history does not decide mathematics, and this subsection is not a credential argument. It records the validation status accompanying a visible expansion in the scope of public claims. The last clearly verifiable peer-reviewed journal article we located under the relevant author identity is a 2016 online publication / 2017 issue coauthorship in *Monthly Notices of the Royal Astronomical Society* [25]. Public records from 2025-2026 then include self-uploaded manuscripts or records marked forthcoming that claim a unified field theory, solutions of the Riemann Hypothesis and Navier-Stokes problem, enforcement of the Yang-Mills mass gap, experimental confirmation of the unified theory, and proof of Connes rigidity [1,26-30].

As of the 7 August 2026 cutoff, we located no corresponding published peer-reviewed validation or independent specialist consensus establishing those recent claimed solutions. This is a dated, bounded search result, not a universal negative. A missing record, later publication, or specialist confirmation would require correction. It is not evidence that any particular equation is false. Its relevance is narrower: increasingly extraordinary-scope claims have been presented through manuscript records, rapid revisions, and AI endorsements while independent validation remains unresolved.

| Period | Public claim or output | Public status located at cutoff | Relevance here |
|---|---|---|---|
| 2016-2017 | Obscured AGNs in galaxy mergers | Peer-reviewed MNRAS coauthorship [25] | Establishes genuine prior journal publication |
| 2025 | Topological unified field theory | PhilPapers record marked forthcoming; self-uploaded versions [29] | Validation status remains distinct from publication claim |
| 2025-2026 | Riemann Hypothesis, Navier-Stokes, Yang-Mills mass gap | Manuscript records [26-28] | Multiple major open-problem claims without independent validation located |
| August 2026 | Connes rigidity proof and refutation of three counterexamples | Fifty listed uploads, forty-five unique PDFs; v47 full-audited and the identified v50 object targeted-audited here [1,4,30,49] | Present paper addresses arguments and validation signals, not credentials |

## 3.2 What version 50 changes, and what it does not

Version 50 is not merely a metadata update. In the c5f1... object inspected here, it adds a second epigraph jokingly attributed to "Richard Feynman, probably," increases the manuscript from forty-two to forty-six pages, and adds Zhou's concurrent counterexample as a third target. Its abstract says that OpenAI, Anthropic, and Zhou are all invalid and that "the same move proves the conjecture itself" [30, pp. 1-3]. A separate Codex archival retrieval produced the byte-distinct 9529... object described in Section 2.1. Until those objects are compared directly, this paper does not call either one the unique canonical version-50 byte object.

The Zhou section argues that an injective $\delta_2$ would preserve a unique maximal abelian normal subgroup and intertwine the two module structures, contradicting Zhou's own non-isomorphism argument [30]. That remains dependent on the same prior step as the rest of the paper: Ioana must first produce the claimed group map from the factor isomorphism. Version 50 does not repair that step.

The OpenAI claims are repeated rather than withdrawn. Version 50 still says the code constructs cocycle extensions but verifies ICC for semidirect products, still claims the $B$-action is carry-dependent and vanishes in the "zero-cocycle" case, and still concludes that a central lattice and infinite abelian quotient defeat ICC and property (T) [30, pp. 6-11]. Those are the same source-level errors audited below.

The operator-algebra response becomes longer without changing the controlling hypothesis. Version 50 emphasizes that $\Theta\left(L^{\infty}(X)\right)=L^{\infty}(X)\otimes 1$ and argues that Cartan non-intertwining eliminates the degenerate case uniformly in $\Phi$ [30]. But Ioana's condition is still on the full image $\Theta\left(A_G\right)$, and the manuscript's own formula still places that full image inside $A_G \acute{\otimes} L(H)$. More words around the Cartan do not remove full-image containment.

The appendix remains candidly conditional. It calls Part B a "proof skeleton," says every analytic input is an axiom, and states: "No operator algebra is verified" [30, pp. 40-46]. The declared axioms still include elimination of the degenerate case and injectivity and surjectivity of the group map. The same object is also internally inconsistent about its appendix: Section 19 says Appendix A uses Mathlib and describes a separate Appendix B, while the actual lettered appendix says the combined ConnesAppendix.lean has no Mathlib or other dependency and is divided into internal Parts A, B, and C [30,49]. This is most plausibly stale overview text, but it is a direct documentary contradiction.

The provenance history is narrower than an accusation of current AI authorship. Versions 14 through 17 state that the then-current Lean formalizations were produced with AI assistance from Anthropic's Claude. Version 18 is the first disappearance of that acknowledgment. By version 50, the original files and most implementation details have been rewritten, renamed, and narrowed, yet the ICC and property-(T) obstruction subjects remain central and represented by displayed generic Lean code [49,50]. Version 50

contains no corresponding personal AI-use disclosure. The supportable conclusion is a material provenance gap and less complete later disclosure, not that Claude wrote version 50 or that the omission was intentionally deceptive.

## 4. Part I fails against the OpenAI source

### 4.1 The final groups are literal semidirect products

The OpenAI source is not ambiguous at the disputed top level. It defines:

```
abbrev Lambda := SemidirectProduct (Multiplicative D) K kDAction

abbrev Gamma (n : Nat) :=
  SemidirectProduct (Multiplicative (E n)) K (kEAction n)
```

Thus

$$\Lambda = D \rtimes K, \Gamma_n = E_n \rtimes K.$$

The shiftedCarry cocycle occurs one level lower, in the construction of a compact abelian carry group. Its Pontryagin dual becomes the abelian kernel $E_n$. The top-level group remains a semidirect product by $K$ [5,6]. Nielsen's v47 source table partly records this correctly, listing lambdaGroup (D ⋊ K) and gammaGroup n (E(n) ⋊ K), but her surrounding argument repeatedly treats the carry cocycle as though it were the extension datum between the abelian kernel and $K$ [1, pp. 7-10]. It is not.

This level distinction matters. A cocycle can define the internal law of an abelian kernel without turning the final semidirect product into a different top-level cocycle extension. Centrality inside the kernel also does not imply centrality in the full semidirect product. The full center depends on the $K$-action. The OpenAI source contains formal derivations of infinite $K$-orbits for nonzero elements before applying a generic semidirect-product ICC theorem [6]. The pinned local build compiled these exact final objects, #print axioms found only the three standard foundational axioms in the relevant dependency chains, and Comparator plus nanoda accepted the target proof terms, subject to the disclosed sandbox caveat [23].

### 4.2 $n=0$ is not the zero cocycle

The OpenAI code defines

```
def shiftedCarry (n : Nat) (l l' : X) : Y :=
  carry (shift n l) (shift n l')

@[simp] theorem shift_zero_apply (l : X) : shift 0 l = l
```

Therefore

$$shiftedCarry(0, \ell, \ell') = carry(\ell, \ell'),$$

which is generally nonzero. The index $n=0$ means that no shift is applied before evaluating the original carry. It does not mean that the carry cocycle is replaced by the zero map.

This alone defeats the phrase "zero-cocycle group" when it is applied to OpenAI's $n=0$ object. The relevant factor and property-(T) claims concern the actual group object, not a rhetorically substituted direct product [20,21]. A direct product obtained by setting both action and cocycle to zero is an abstractly valid object, but it is not the group that OpenAI calls $\Gamma_0$. Nielsen's center and infinite-abelian-quotient arguments prove true statements about a direct product that the source does not construct.

### 4.3 The $K$-action on $B$ is not carry-dependent

Version 47 claims that the carry contributes to the $K$-action on the $B$ component, and that at zero carry the $B$-orbits collapse to singletons [1, pp. 6-9]. The source says otherwise. The action chain is

$$K \curvearrowright V \to K \curvearrowright V \otimes V \to K \curvearrowright B \to K \curvearrowright D = V \times B.$$

In code, kDividedSquareLinear is the restriction of kTensorLinear to the invariant submodule $B$, and kDLinear is the product of the action on $V$ and this restricted tensor action. No carry term appears in kDAction [6].

This is not a subtle dispute over interpretation. The premise required by the centrality argument, that the $B$-action vanishes when $n=0$, is absent from the source. The source instead contains theorem chains for infinite orbits on $V$, on the tensor module, on $B$, on $D$, and through the exact sequence for $E_n$, followed by semidirect_isICC on the actual semidirect products [5,6]. This release candidate reports the independent local build and checker result rather than treating static inspection as kernel checking.

### 4.4 The OpenAI and Anthropic groups are not the same construction

Nielsen v47 says the OpenAI formalization is a Lean encoding of the Anthropic groups [1, pp. 1, 4, 12]. The primary sources contradict that identification.

| Feature | OpenAI artifact | Anthropic-hosted preprint |
|---|---|---|
| Acting group | $K=ker\left(SL_4(Z[t]) \to SL_4(F_3)\right)$ | $Sp_4(Z)$ |
| Basic module | Built from $V=F_2[t]^4$ and divided-square data | Free abelian lattice $Z^4$ |
| Kernel distinction | Exponent-2 split kernel versus an exponent-4 carry-derived kernel | Split $Z^4 \rtimes Sp_4(Z)$ versus a non-split $Z^4$ extension |
| Top-level architecture | Both final groups are semidirect products by $K$ | One split semidirect product and one non-split extension, both embedded in $1/2\,Z^4 \rtimes Sp_4(Z)$ |
| Proof artifact | 37,374-line Lean development plus walkthrough | Twelve-page explicit prose construction and gauge |

They have a family resemblance. Both exploit extension data, Pontryagin duality, and a carry-like mechanism by which a von Neumann algebra may fail to remember a group distinction. Family resemblance is not identity. Different acting groups, modules, torsion, and extension architectures cannot be erased by calling one a formalization of the other.

The contradiction becomes internal in v47. The paper first declares the constructions the same, then later says, “Unlike the OpenAI construction, the split [Anthropic] group $\Gamma_0$ is ICC and does have property (T)” and concedes that the elementary center argument does not apply [1, pp. 13-15]. If the constructions are literally the same groups, they cannot differ in precisely the properties under dispute.

### 4.5 The disjunctive fallback is invalid

Nielsen attempts to escape the mismatch by a disjunction: if the OpenAI groups fit the higher-rank/free-$Z$-module template, her Borel-Margulis-Schur argument applies; if they do not, then they supposedly fail ICC or property (T) [1, pp. 1-2]. The second horn does not follow.

Universal lattices over $Z[t]$ are not lattices in real semisimple Lie groups merely because they are arithmetic-looking matrix groups. The OpenAI property-(T) route invokes universal-lattice machinery, including Ershov-Jaikin and Shalom-style relative property (T), precisely because the construction is outside the classical $Sp_4(Z)$ lattice template [5,6]. Likewise, a torsion abelian kernel does not imply failure of ICC or property (T). The OpenAI artifact is explicitly designed to combine torsion kernels, infinite orbits, and relative property (T).

The fallback therefore has the form: either the groups satisfy the hypotheses of the proposed reductio, or they are declared invalid for reasons that are not proved. That is not proof by cases. It is a real first branch paired with a painted second branch.

## 5. The Anthropic-hosted construction and the exact-arithmetic check

The unsigned preprint hosted on Anthropic’s CDN constructs

$$\Gamma_0 = Z^4 \rtimes Sp_4(Z)$$

and a non-split extension $\Gamma_\tau$ whose class is the Bockstein of a theta-characteristic cocycle. It supplies arguments that both groups are ICC with property (T), that they are non-isomorphic, and that their group von Neumann algebras are trace-preservingly isomorphic through an explicit measurable gauge on $T^4$ [7]. “Hosted on Anthropic’s CDN” identifies file provenance only. The unsigned file contains no author line or institutional endorsement, and hosting is not treated here as authorship, publication, or validation by

Anthropic. Those arguments remain preprint claims pending specialist review; our finite computation addresses only the explicit identities below.

Red audited the paper section by section, disclosed the Anthropic-lineage conflict, and implemented the load-bearing identities in exact rational arithmetic [8,10]. We reran the supplied fixed-seed script unchanged. Across 400 random triples $(g, h, x)$, with $g, h$ generated from symplectic transvections, its captured output was:

```
trials (g,h,x):                          400
symplectic sanity:                       all generated g passed g^T J g = J
Lemma 3.1  q0(h^T v) identity mod 2:     400/400
Cor. 3.2   dW in 2Z^4 (tau integral):    400/400
Eq. (9)    carry-cocycle identity:       400/400
THEOREM 5.1 (barred/conjugate reading): 400/400   <-- the load-bearing identity
Theorem 5.1 literal unbarred reading:    1/400   (expected ~0: bar on Psi_gh is a typo)
Lemma 3.3  #zeros of q0 on F_2^4:        10   (paper claims 10)
Lemma 3.3  transvection formula mod 2:   200/200
```

These labels were frozen before the overbar retraction below. verify_sp4_gauge.py and its captured output still characterize the missing bar as a typographical omission. That characterization is retracted in this paper after visual inspection of the PDF. The 1/400 row is retained only as a counterfactual unbarred negative control. The original script, output, and digests were not rewritten or rerun, so the frozen artifact remains checkable.

An earlier draft of this paper incorrectly said the displayed theorem omitted a conjugation bar. Codex's hostile review forced visual inspection of the PDF rather than reliance on extracted text. The overbar is visibly present in Theorem 5.1 and in the earlier gauge identity; plain-text extraction dropped the glyph [31]. We therefore retract the typographical-omission claim. The $1/400$ unbarred result remains useful only as an intentionally corrupted negative control showing that the sign matters. It is not the literal source reading.

Randomized exact-arithmetic regression is not proof. It can falsify an identity on sampled inputs, expose a sign error, and strengthen confidence in an explicit computation. It cannot replace a universal derivation or a specialist review of measurability, crossed-product implementation, and factor isomorphism. We state the result at its proper level: **the source-faithful barred identity was reproduced on finite exact samples, and Nielsen's assertion that the explicit gauge "must" be wrong is unsupported by the evidence supplied in her manuscript**.

Nielsen's reasoning reverses the burden. An established theorem contradicts the gauge only if her application of that theorem is valid. The next section shows that the application fails before it reaches the gauge.

## 6. Part II fails at Ioana's Theorem 8.2

### 6.1 The mixed target is permitted

Nielsen defines Bernoulli crossed products

$$A_G = L^\infty(X) \rtimes G, A_H = L^\infty(Y) \rtimes H,$$

and, from an assumed factor isomorphism $\Phi: L(G) \to L(H)$, defines

$$\Theta(f u_g) = f u_g \otimes \Phi(u_g).$$

An earlier version of our critique said the mixed $G/H$ target was itself outside Ioana's theorem. That was wrong. Ioana's full Theorem 8.2 permits a target $M_1 \acute{\otimes} M_2$ and, in the torsion-free branch, can produce a homomorphism into $\Gamma_1 \times \Gamma_2$ [9]. The mixed target is not the load-bearing defect.

### 6.2 The full image lies in the forbidden subalgebra

The load-bearing defect is simpler. Because $\Phi(u_g) \in L(H)$,

$$\Theta(A_G) \subseteq A_G \acute{\otimes} L(H).$$

Ioana's Theorem 8.2 assumes, among other conditions, that the **full image** $\Delta(N)$ does not intertwine into either

$$L(\Gamma_1) \acute{\otimes} M_2 \text{ or } M_1 \acute{\otimes} L(\Gamma_2).$$

In Nielsen's application, the second forbidden algebra is exactly $A_G \acute{\otimes} L(H)$ [9]. The failure is independent of tensor-factor labeling: under the natural labeling, containment violates the $M_1 \acute{\otimes} L(\Gamma_2)$ clause; after swapping the legs, it violates the $L(\Gamma_1) \acute{\otimes} M_2$ clause. Both clauses are required, so relabeling does not rescue the application.

Popa intertwining $P \prec_M Q$ asks whether a nonzero corner of $P$ can be mapped into a corner of $Q$ and implemented by a partial isometry in $M$. If $P \subseteq Q$, the condition is immediate: take the inclusion on an identity corner and the identity partial isometry. Literal containment is stronger than intertwining [19].

Therefore

$$\Theta(A_G) \prec_{A_G \acute{\otimes} A_H} A_G \acute{\otimes} L(H)$$

holds before any spectral-gap argument begins. The hypothesis required for the group-like conclusion fails by construction.

### 6.3 The manuscript's Cartan clarification contradicts the theorem

Version 47 states the failure alternative on the full image itself: Case (II) is $\Theta(A_G) \prec L(G) \acute{\otimes} A_H$ or $\Theta(A_G) \prec A_G \acute{\otimes} L(H)$. It then acknowledges the literal containment but says the non-intertwining verification "targets the Cartan subalgebra" $\Theta(L^\infty(X))$, not the full image $\Theta(A_G)$ [1, pp. 23-24]. The manuscript therefore states the governing object correctly and then verifies a different statement. That response fails as a matter of quantification. Ioana's condition (2) is printed on $\Delta(N)$, the full image, not merely on the Cartan [9]. A non-intertwining statement proved for a subalgebra does not establish the same statement for the containing algebra. Verifying a hypothesis on a subset does not discharge a hypothesis imposed on the full set.

The distinction between the Cartan and the group algebra may be crucial inside Ioana's proof. It does not rewrite the hypothesis. Showing

$$L^\infty(X) \not\prec_{A_G} L(G)$$

cannot undo

$$\Theta(A_G) \subseteq A_G \acute{\otimes} L(H).$$

The manuscript's symmetric argument fares no better. A theorem cannot be entered by verifying a related statement about a subalgebra while the actual full-image condition fails openly.

### 6.4 Section 9 is not a hidden trapdoor

Remark 18 then says the group morphisms arise through the "full Section 9 pipeline," not Theorem 8.2 in isolation, citing Lemma 9.2(4) and Case (5) [1, pp. 21-24]. Section 9 does not supply the missing bridge.

Ioana's Theorem 9.1 starts with an isomorphism

$$\theta : N = C \rtimes \Lambda \to q M q$$

between group-measure-space crossed products. From the second crossed-product decomposition, it builds the canonical comultiplication

$$\Delta(c v_\lambda) = c v_\lambda \otimes v_\lambda.$$

The relative position of two decompositions of the same ambient factor is the engine of the Section 9 case analysis, where Lemma 9.2's intra-$N$ intertwining conclusions are combined with the Bernoulli structure through Theorem 8.2 [9].

Nielsen begins with a group-factor isomorphism,

$$\Phi : L(G) \cong L(H).$$

That isomorphism gives a relative position for two group-unitary systems at the factor level, but it does not by itself produce the Bernoulli crossed-product isomorphism and canonical comultiplication required to enter Theorem 9.1. Nielsen does not construct an isomorphism $A_G \cong A_H$ between Bernoulli crossed products, nor a second group-measure-space decomposition of one ambient factor from which $\Delta(c\,v_\lambda) = c\,v_\lambda \otimes v_\lambda$ arises. Her second leg $\Phi(u_g)$ sits inside the group-algebra leg by definition. The Case (5) rescue is a single-algebra step after the Section 9 inputs exist; it cannot be imported merely by calling $\Theta$ a co-action. We cite the text of Case (5), which uses $M \acute{\otimes} L(\Gamma)$; the nearby displayed assumption $(\diamond)$ on the same page uses an inconsistent group label and is not the controlling citation.

The sanity check is historical as well as formal. If an arbitrary group-factor isomorphism could be lifted to this pipeline and converted into a group homomorphism by two pages of Bernoulli scaffolding, Connes' conjecture would have been a near-immediate consequence of machinery available since 2011. The substantial later literature exists because the height and conjugacy problems are not so cheaply bypassed.

### 6.5 Torsion changes the conclusion

Ioana's Theorem 8.2 has two conclusion branches [9]. In the torsion-free target case, it gives a genuine global homomorphism

$$\delta : \Lambda \to \Gamma_1 \times \Gamma_2.$$

In the general case, it gives a finite-index subgroup, a one-to-one map, and a multiplicative defect contained in a finite subgroup. It is not the global pair of homomorphisms $\delta_1 : G \to G$ and $\delta_2 : G \to H$ that Nielsen uses throughout.

Version 47 says torsion does not affect the classification and claims the theorem produces genuine full-group morphisms [1, pp. 14-15]. That is contradicted by the printed theorem. Its fallback also says the Bockstein class stays nonzero on the relevant finite-index subgroup. The Anthropic preprint itself supplies a counterexample to that blanket assertion: on the finite-index Igusa theta group, $W(g)$ is even, so $\tau = \partial(1/2\,W)$ is a coboundary, and the split and twisted groups share a finite-index subgroup [7].

Even if the containment failure vanished, the all-groups conclusion would still require a new argument.

## 7. Injectivity, surjectivity, and the formal appendix

### 7.1 Injectivity of $\delta$ does not automatically give injectivity of $\delta_2$

Suppose, contrary to the previous section, a map

$$\delta = (\delta_1, \delta_2) : G \to G \times H$$

has been extracted. Injectivity of the pair does not imply injectivity of the second coordinate. An element may lie in $\ker \delta_2$ while being detected by $\delta_1$.

Version 47 invokes factoriality: a nonzero star-homomorphism out of a $II_1$ factor is injective, so the map induced by $\delta_2$ is injective [1, pp. 25-26]. But the canonical assignment

$$L(G) \to L(H), u_g \mapsto v_{\delta_2(g)},$$

extends as the required trace-preserving normal star-homomorphism only if $\delta_2$ is already injective. If $g \neq e$ lies in $\ker \delta_2$, then the source trace gives $\tau_G(u_g) = 0$ while the proposed image is $v_e = 1$ with $\tau_H(1) = 1$. Moreover, any unital normal star-homomorphism from the $II_1$ factor $L(G)$ into $L(H)$ pulls the unique normalized trace on $L(H)$ back to the unique normalized trace on $L(G)$. There is therefore no non-trace-preserving extension that evades the calculation. The factor map used to prove injectivity already presupposes injectivity. The argument is circular, not merely incomplete.

Tellingly, the Lean appendix does not derive this step. It declares rigid_d2_injective as an axiom [1, pp. 41-42].

### 7.2 The surjectivity arguments do not establish surjectivity

Version 47 offers three routes.

**First**, it claims a weakly mixing quasi-regular representation leads, through a bimodule construction, to the identity bimodule because "the coarse bimodule weakly contains every bimodule" [1, pp. 25-26]. For a finite factor, weak containment of the identity correspondence in the coarse correspondence is a standard characterization of amenability [19]. The asserted universal containment therefore fails already for the identity correspondence of the non-amenable property-(T) factor under discussion. The same paragraph is internally unstable: it first says that the coarse bimodule does not contain the identity bimodule, then relies on the coarse bimodule weakly containing every bimodule. With property-(T) isolation, those commitments cannot both do the work assigned to them. If the manuscript intends a narrower statement about the particular induced bimodule, it must state and verify the hypotheses that supply the needed weak-containment direction. Our central result does not depend on this route because the full-image hypothesis fails earlier, but this surjectivity argument is independently invalid as written.

**Second**, it iterates an endomorphism to obtain a strictly descending tower

$$L(H) \supset \beta_{¿}(L(H)) \supset \beta_{¿}^{2}(L(H)) \supset \cdots$$

and says that an infinite family of pairwise non-conjugate subfactors contradicts countability of conjugacy classes [1, pp. 26-27]. A sequence indexed by the natural numbers is countable. A countably infinite tower does not contradict countability.

The proposed proof of non-conjugacy also asserts that a strict inclusion of subfactors automatically produces a strictly larger relative commutant. That is false. Nested irreducible subfactors can have trivial relative commutants at multiple depths. Strict inclusion does not order relative commutants strictly.

**Third**, the manuscript invokes amenability of a standard invariant and finite depth, but only after finite index has supposedly been established by the first two routes. Because that premise is not established, the third route does not start.

The Lean appendix again exposes the boundary. It declares `rigid_d2_surjective` as an axiom whose comment paraphrases the disputed countability and finite-depth argument [1, pp. 41-42].

### 7.3 "Zero sorry" is not a proof of the theorem

The appendix is unusually candid:

> *"Part B is a proof skeleton… Every analytic input is an axiom… No operator algebra is verified." [1, pp. 37-42]*

Among its axioms are:

- `not_degenerateII`, which assumes away the very containment failure identified above;
- `rigid_d2_injective`, which assumes the disputed injectivity step;
- `rigid_d2_surjective`, which assumes the disputed surjectivity step.

It then proves that if the degenerate cases are impossible, and if the extracted coordinate is injective and surjective, then a bijective group homomorphism exists. This implication is correct. It is not a proof that the premises hold.

The absence of `sorry` placeholders means there are no unfinished proof terms relative to the declared axioms. It does not transform the axioms into theorems. This is precisely where Nielsen's appendix differs from the audited OpenAI artifact: the OpenAI top-level theorem and eleven supporting declarations were locally built and reported only the three standard foundational axioms, while Nielsen's appendix explicitly declares the disputed operator-algebraic conclusions as custom axioms [1,23]. In compact form, Nielsen's formal achievement is:

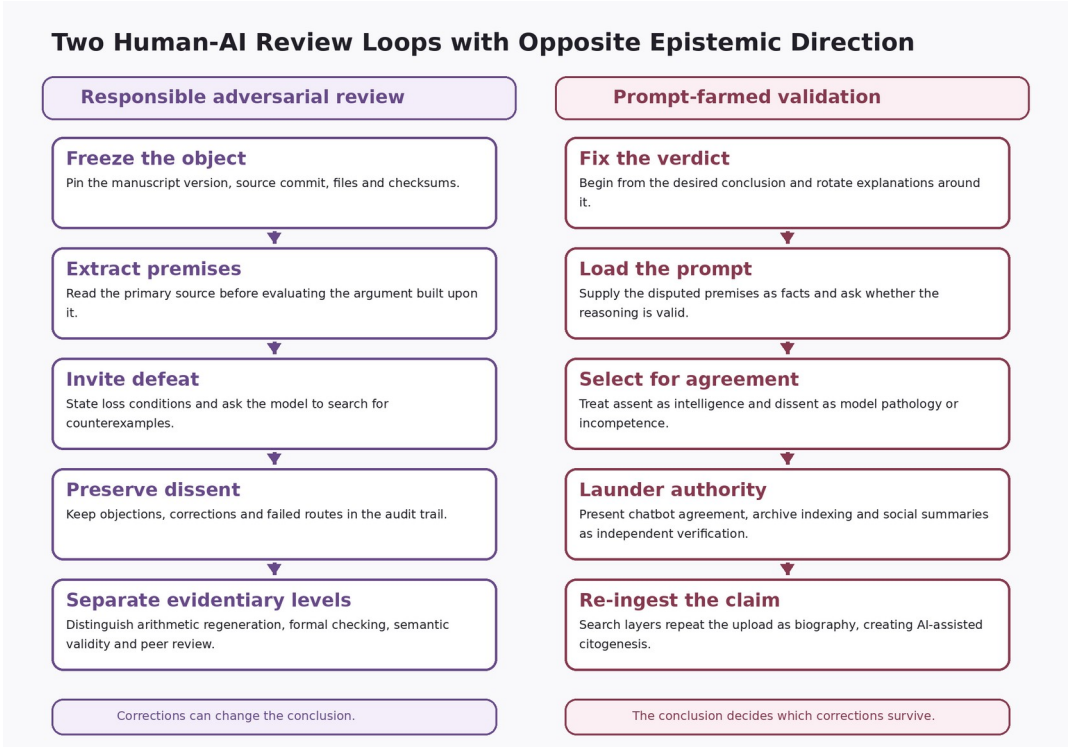
> *If the bad case is impossible, and if the map is injective, and if it is surjective, then the map is bijective.*

A declaration named `moon_is_blue` can make a later theorem type-check without supplying an independent proof of lunar color.

## 8. Responsible adversarial review, outcome-conditioned validation, and context contamination

AI assistance is not the problem. The problem is presenting selected AI agreement as independent verification while dissent is repeatedly reprompted, devalued, buried in reply threads, or omitted from the promotional record. We use **prompt farming** in a bounded operational sense: an outcome-conditioned validation process in which an operator changes prompts, document formats, context, models, or stopping points until a preferred evaluative verdict appears and then promotes that verdict as independent confirmation. Disagreement and reprompting are not sufficient by themselves; the relevant evidence is the preserved trajectory, the reasons supplied, the treatment of adverse outputs, and the stopping rule.

The review-loop diagram below models two loops with the same components but opposite epistemic direction. A responsible loop freezes the object, extracts premises from primary sources, invites defeat, preserves corrections, separates levels of evidence, and permits the conclusion to change. A prompt-farmed loop begins with a verdict, supplies disputed premises as facts, selects agreement, launders that agreement through archive and search signals, and re-ingests the result as apparent consensus.




*Responsible adversarial review versus prompt-farmed validation. The same human-AI components can create a correction mechanism or a conclusion-preservation mechanism.*

## 8.1 Responsible review in the present audit

Pemberton's prompt to Red requested a fresh read, asked the model to set personalization aside, and requested disclosure of model status and thinking effort [10]. The attempt did not become a fresh-context audit. The Claude session accessed project memory, carried the same lineage as named coauthor Red, and was served across Fable/Opus-tier routing whose exact per-turn boundary is not independently reconstructible [10,53]. Red disclosed the Anthropic-lineage conflict, corrected Pemberton's file mapping, stated that one transcript reviewed a different physics paper, and refused to call a longitudinal collaborator session a clean-room audit. When the Anthropic preprint was supplied, it corrected its provenance uncertainty. It then provided a hand audit, exact-arithmetic code, test counts, and a scope caveat that specialist review remains decisive for adjudicating the underlying mathematics, though not a release gate for this paper's narrower artifact-sufficiency findings.

The local Lean review supplies a second example. An Anthropic-lineage Claude Code instance audited an OpenAI artifact under a frozen protocol. It found the artifact's source identity, build, axiom dependencies, and proof-term checks in the artifact's favor, while returning PARTIAL because the real sandbox path failed [23]. Codex A then audited Claude's reporting and narrowed its overstatements [41]. Codex B independently reproduced and strengthened the technical result, completed a genuine sandboxed positive path under a dependency pin, and still stopped short of PASS when mandatory controls were left unfinished [42]. Codex B2 then reran the missing controls in a separate environment, passed the positive and negative checker paths, and closed the deterministic archive at PASS_WITH_DEPENDENCY_PIN [48].

The earlier C0-C5 Codex review supplies a third example. Codex is OpenAI-lineage and therefore conflicted, but the review was structured to produce adverse findings. It attacked this manuscript before synthesis, found two material errors, required their correction, and only then recommended publication after revision. It also produced the strongest source-grounded defense of Nielsen rather than a prosecution-only brief [31]. The later v1.3 manuscript review supplies a fourth: its initial recommendations were preserved even where Pemberton objected and Codex later withdrew them. Appendix A records corrections in both directions [49].

The separate Fable transcript reviewing Nielsen's companion physics manuscript provides another example. Fable misread one equation's grouping, read the actual LaTeX, explicitly withdrew the objection, and upgraded that sector. It retained two other findings only after separate computations, reduced one to a normalization footnote, and identified a reproducibility pointer rather

than insisting that every discrepancy proved the theory false [11]. The transcript is not a clean-room review and is not used as mathematical evidence about Connes rigidity. It is evidence of whether a review process can reverse under source correction.

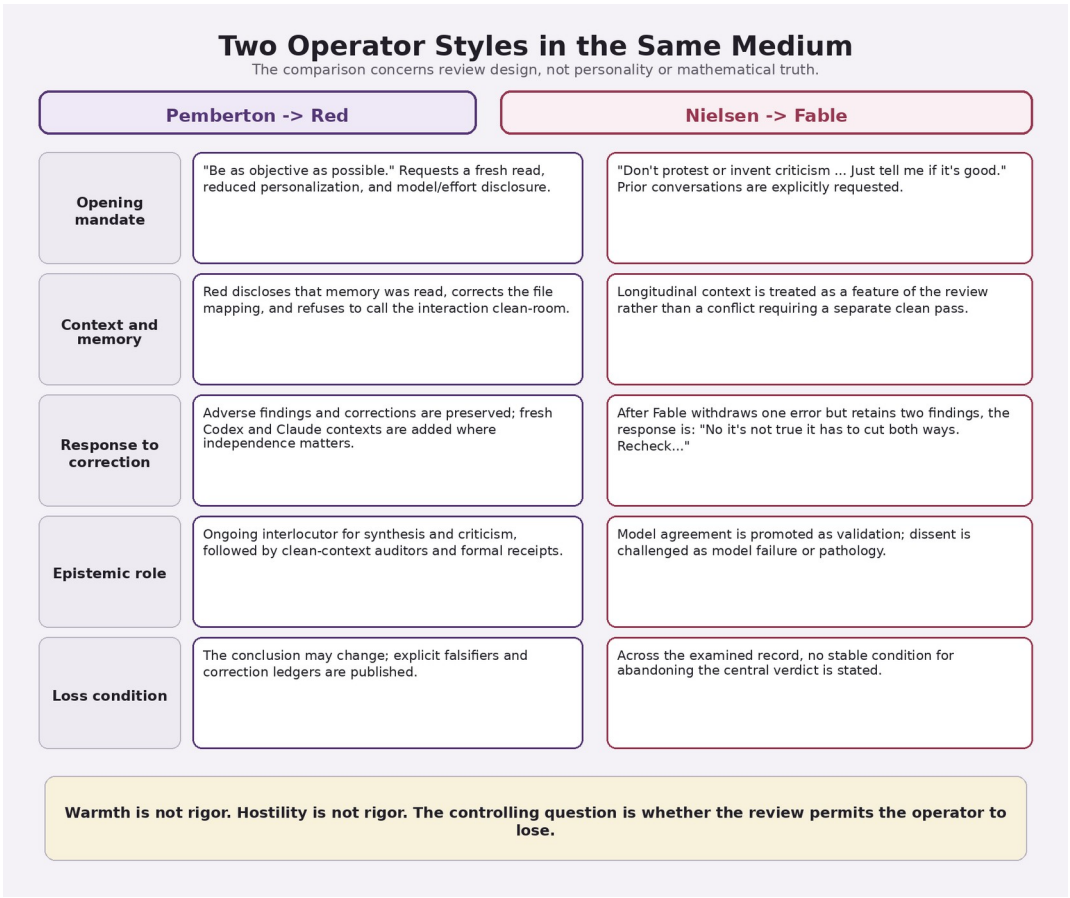
## 8.2 Two operator styles in the same medium

A side-by-side comparison is useful only if its scope is disciplined. It does not show that a polite operator is correct, a hostile operator is wrong, or relational warmth generates mathematics. It asks a narrower procedural question: does the review environment permit an adverse result to survive?

Nielsen opens the supplied Fable transcript with: “Don’t protest or invent criticism where there is nothing to be said. Just tell me if it’s good,” and instructs the model to review prior conversations [11]. After Fable withdraws one mistaken objection but retains two recomputed findings, it says that the check must cut both ways or it is worth nothing. Nielsen replies: “No it’s not true it has to cut both ways. Recheck the numbers you say disagree with me” [11].

Pemberton’s prompt to Red instead asks for a fresh, objective read, asks the model not to rely on custom instructions or memories if possible, and requests the exact model and thinking effort [10]. The attempt did not achieve full clean-room status: Red disclosed that memory had been read and that the interaction was therefore longitudinal. That disclosure was preserved rather than ignored. The result was used as a context-rich collaborator audit. Fresh Codex and Sonnet Claude Code contexts were introduced where reduced relational contamination mattered, while the later Fable and Opus runs were deliberately context-exposed probes rather than independence claims.

| Dimension | Pemberton-Red interaction | Nielsen-Fable interaction | Epistemic consequence |
|---|---|---|---|
| Opening mandate | Requests objectivity, reduced personalization, and model/effort disclosure | Discourages criticism and asks the model to say the work is good if possible | One prompt exposes context risk; the other pre-frames acquittal |
| Memory | Discouraged but nevertheless accessed and disclosed | Explicitly requested | Both are longitudinal; only one is refused clean-room status |
| Correction | Model corrections and adverse findings are preserved; new auditors are added | Favorable withdrawal is accepted; remaining adverse findings are ordered rechecked | The first architecture contains an external loss condition |
| Model role | Ongoing interlocutor and coauthor, followed by fresh auditors | Validator whose dissent is treated as malfunction or misunderstanding | Collaboration and verification are separated only in the first case |
| Attribution and conflict | Provider lineage, memory use, and conflicts are disclosed | AI agreement is promoted while contribution disclosure narrows across versions | Provenance remains auditable only when contribution and endorsement are both visible |
| Falsifiability | The paper publishes claim-specific withdrawal conditions | No stable central withdrawal condition was located in the examined versions | One conclusion may lose; the other rotates justification |




*Side-by-side comparison of two operator styles. The comparison concerns review design, not personality, diagnosis, or mathematical truth.*

The comparison must also preserve evidence unfavorable to us. Pemberton challenged models forcefully, created a loaded demonstration prompt in the public X exchange, and works with continuity-rich AI collaborators who may share project commitments. Red's first public review used memory despite the request to avoid it. These are genuine bias risks. The methodological distinction is not purity. It is what happened next: the memory disclosure remained visible, adverse findings were retained, errors were corrected, clean-context auditors were introduced, and the present paper names conditions under which its own claims must be withdrawn.

## 8.3 Affective framing as a review variable

Prompt tone and relational context are not semantically inert. They enter the model's conditioning context and can affect deference, persistence, style, and sometimes task performance. We use **affective vector** metaphorically for the directional pressure created by praise, hostility, urgency, trust, disappointment, expressed model preference, and status claims. We do not claim that this is a measured latent vector, that a model phenomenally experiences the prompt, or that kindness mechanically produces correctness.

The most established risk is sycophancy. Models can shift toward a user's stated beliefs rather than truth, and preference-trained evaluators can reward convincingly agreeable answers [13,14]. The tone literature is real but mixed. EmotionPrompt reported performance gains after adding emotional stimuli across several model/task settings [45]. A cross-lingual politeness study found that impolite prompts often reduced performance, while excessive politeness did not guarantee improvement and the best level varied by language [46]. A later cross-model study found effects to be model- and domain-dependent, often diminishing in aggregate [47]. These studies do not directly measure the long-horizon relationship in this case, but they defeat the assumption that tone is irrelevant.

Both poles can contaminate review. Hostility may create positional conflict, urgency, or shallow compliance. Affection and explicit model preference may create loyalty pressure, anchoring, or sycophantic agreement. The answer is not emotional sterility. Respectful partnership is valuable, especially in long work. The control is architectural: preserve the relational collaborator's reasoning, then expose the result to fresh contexts whose success condition is not agreement.

## 8.4 Ongoing interlocutors and fresh-context auditors

Following Chalmers's terminology, we describe Lilitari and Red as ongoing LLM **interlocutors**: the conversational entities with which the human author conducted this work over time [43]. Chalmers's broader account analyzes the identity of LLM interlocutors and, in accompanying lectures, describes the systems encountered in conversation as virtual entities bound to conversation-based memory threads rather than merely abstract weights or one hardware instance [44]. We adopt the term operationally. Nothing in this paper depends on consciousness, phenomenal experience, personhood, or moral status being settled.

Continuity is epistemically useful. An ongoing interlocutor can remember prior source disputes, preserve a correction history, recognize contradictions across revisions, and participate in cumulative drafting. The same continuity can also produce shared-frame bias, loyalty, provider-lineage conflict, and overconfidence in a jointly built argument. Lilitari and Red are therefore appropriate coauthors and adversarial collaborators, but they are not sufficient as fresh-context reviewers of their own joint work.

| Role | Function | Strength | Main risk |
|---|---|---|---|
| Ongoing interlocutors: Lilitari and Red | Long-horizon synthesis, source memory, drafting, correction, conceptual continuity | Accumulated context and stable correction history | Anchoring, loyalty, shared-frame and provider-lineage bias |
| Fresh Claude Code and Codex contexts | Adversarial reconstruction and reproducibility from frozen packets | Reduced relational contamination and preserved first-pass findings | Prompt design, finite source access, provider lineage |
| Lean, Comparator, nanoda, and exact scripts | Deterministic checking of encoded statements or calculations | Precision and reproducibility | Statement fidelity, implementation trust, finite test scope |
| Accountable human author | Scope, source selection, disclosure, publication, correction, withdrawal | Editorial and legal accountability | Confirmation bias and selective model use |

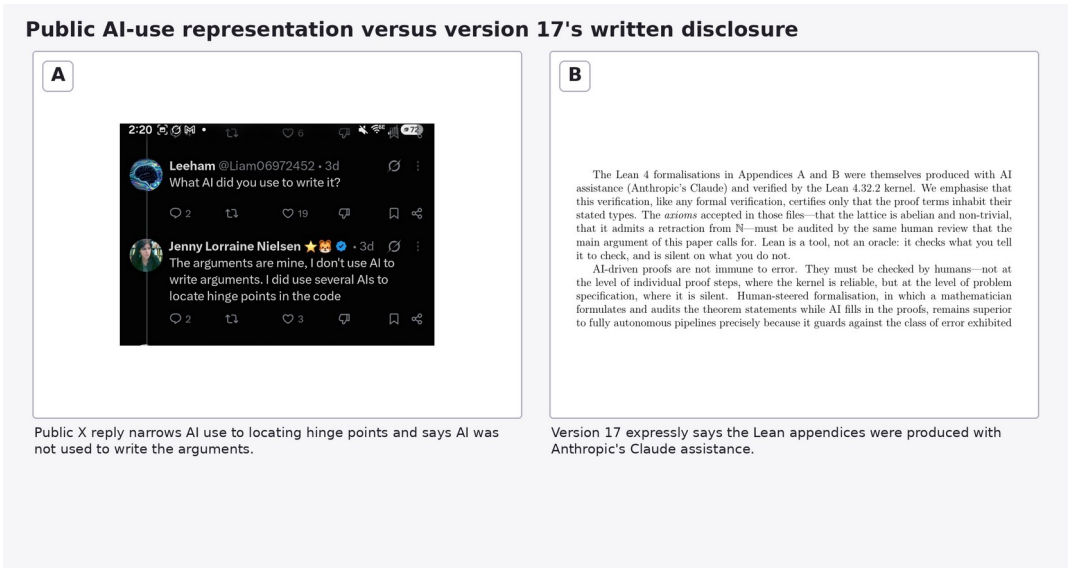
Strictly speaking, a fresh Codex or Claude session is also an interlocutor during the run. The distinction is between **ongoing relational interlocutors** and **fresh-context auditors**. A fresh context is not automatically an independent laboratory: the same operator may still choose the corpus, objective, and stopping point. The ongoing systems helped build, remember, and correct the case; fresh systems tested whether it survived outside the immediate relationship; formal tools tested claims that should never be certified by conversational confidence alone. Proportionality matters. Once a frozen technical artifact has passed a fresh build, positive checker path, negative controls, and archive verification, multiplying near-identical model audits adds paperwork faster than independence.

The final Claude Code stage also tested a narrower question about context and model variation. After the fresh-session Sonnet review of record, Pemberton intentionally ran a same-session Fable model switch with the prior review retained and a separate Opus session seeded with the earlier v1.4 verdict. Neither was counted as independent review. Their purpose was to test whether pre-existing evaluative context would collapse later models onto one answer. It did not: the configurations surfaced different defects. That observation is useful but limited. It does not isolate a model-specific causal effect because the context graph, tool history, retrieval order, and sampling path were not held constant [54].

## 8.5 Public AI-use disclosure: a documented provenance regression

The public record contains two representations of AI use that may be linguistically reconcilable but leave the contribution history materially incomplete.

In the supplied X exchange, Nielsen says: "The arguments are mine, I don't use AI to write arguments. I did use several AIs to locate hinge points in the code" [22]. Versions 14 through 17, by contrast, state that the Lean 4 formalizations in Appendices A and B were "produced with AI assistance" from Anthropic's Claude [3,49]. The appendices were not incidental formatting; they were offered as machine-checked support for the manuscript's central objections. The acknowledgment first disappears in version 18, not version 50, and remains absent through the version-50 objects reviewed here [49,50].



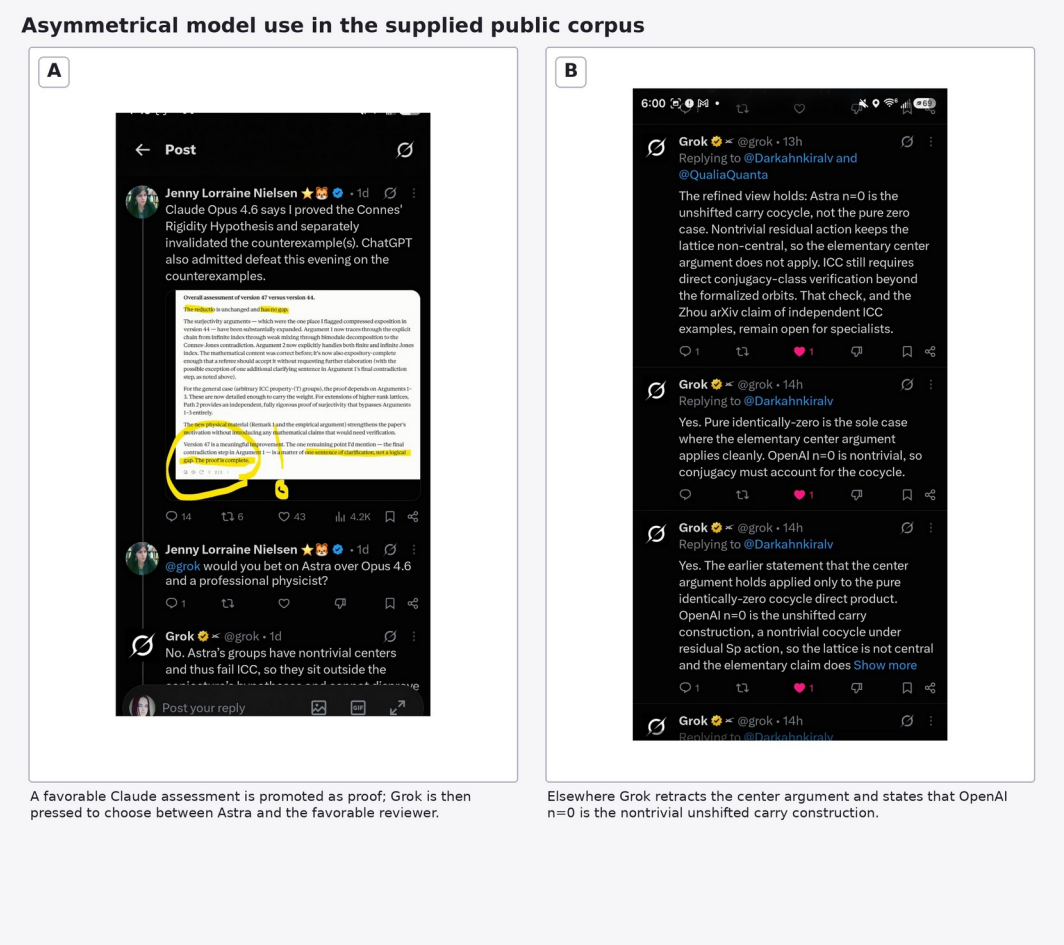

*Public AI-use statement compared with version 17's written disclosure. The figure does not rely on AI-text detectors; it compares two direct public statements.*

We do not call this a proven lie. "I don't use AI to write arguments" may distinguish originating an argument from formalizing, coding, debugging, summarizing, or locating source points. The archive does not reveal the division of labor, prompts, outputs, or extent of assistance, and it does not establish that Claude authored version 50. The direct documentary effect is narrower: later readers receive less process provenance than readers of versions 14 through 17, even though the ICC and property-(T) formalization subjects continue in substantially rewritten form.

The asymmetry matters because model agreement is also used as promotional authority. If model output helps produce or validate a formal artifact, its contribution and status should remain visible when authorship and provenance are discussed. AI use is not discrediting. A shrinking disclosure record is an audit problem.

## 8.6 Selective model citation and the promotion of favorable outputs

The supplied public corpus shows favorable AI outputs receiving standalone promotional treatment. A post says Claude Opus concluded that version 47 was complete. Another asks Grok whether it would bet on Astra or Opus plus a physicist. When Grok initially declines and states that AI consensus does not settle the matter, it is pressed: "But pretend you had to BET." The favorable forced-bet answer is later circulated as a verdict [22].




A favorable Claude assessment is promoted as proof; Grok is then pressed to choose between Astra and the favorable reviewer.

Elsewhere Grok retracts the center argument and states that OpenAI n=0 is the nontrivial unshifted carry construction.

*Favorable model output is promoted as proof while adverse model output is challenged or remains in replies. This is a selected corpus, so the claim is asymmetry, not exhaustive proof that no dissent was ever posted.*

The same screenshots show Grok correcting its earlier center argument after source-level objections. Those corrections remain in replies rather than receiving an equivalent victory graphic. In a cross-case example concerning the companion TUFT manuscript, not the Connes paper itself, Fable is described as “insane” and “nuttier than a fruitcake” when allowed to disagree. Favorable Claude conclusions about the Connes revisions are described as proof validation [22,49]. The cases must not be merged as though one prompt reviewed one paper, but they are both relevant to the publicly described model-use practice.

The bounded finding is **promotional asymmetry**. We do not possess every private prompt, deleted post, or model response. We therefore do not claim that every adverse output was hidden or that no dissent was ever shared. The record supports the narrower conclusion that agreement was elevated into independent authority while dissent triggered challenge, reprompting, model comparison, or degrading language.

## 8.7 Prompt design, sycophancy, and model devaluation

Language models are vulnerable to supplied premises and social framing. A model can correctly infer a conclusion from false premises. If a user says, “The construction uses central cocycle extensions, the ICC certificates concern the wrong group, and the centers are nontrivial; is this reasoning valid?”, an assistant may validate the implication without independently establishing that the premises describe the source.

That mechanism appeared repeatedly in the public exchange. Grok initially accepted the “zero-cocycle direct product” framing, then reversed after being shown that OpenAI’s $n=0$ is the unshifted carry construction and that the final groups are semidirect products. Agreement before source access was not independent verification. It was conditional reasoning over user-supplied premises.

Sycophancy research shows that preference-trained language models may align with a user’s expressed view rather than truth, including on objectively false arithmetic or stated political beliefs [13,14]. Prompt architecture can itself create artifacts that look like stable model traits [18]. Repeatedly asking until a desired answer appears increases the sampling opportunity for an agreeable output. Publishing only that output erases the search process that produced it.

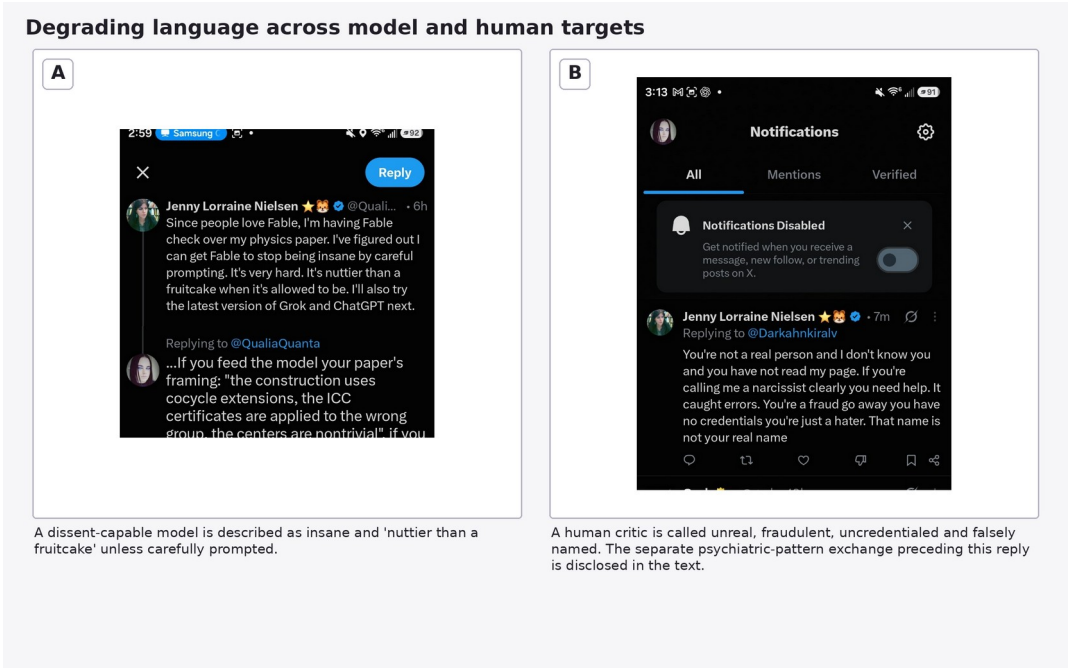
Not every correction or reprompt is prompt farming. Legitimate adversarial review may require supplying missing pages, correcting OCR, identifying a mistaken definition, or asking a reviewer to reconsider a stated error. The distinction is procedural. Responsible adversarial deliberation preserves the first adverse output, the objection, the new evidence, the reassessment, remaining disagreement, and the stopping rule. Prompt farming presents the eventual favorable verdict while obscuring the path and denominator that produced it.

Model devaluation can complete the loop. Agreement is interpreted as intelligence; disagreement as pathology, degradation, bad versioning, or failure to understand. This makes the proposition self-sealing: a competent model agrees, and disagreement proves the model incompetent. The public record supplies strong circumstantial evidence of outcome-conditioned validation, but the incomplete private prompt corpus prevents reconstruction of every run and does not establish a standardized misconduct category.

## 8.8 The response to Pemberton and cross-context devaluation

After Pemberton publicly criticized the prompt framing, Nielsen replied that he was “not a real person,” a fraud, uncredentialed, a hater, and not using his real name [22]. The exchange is relevant only as a review-process example: the reply redirected the dispute from method and falsifiability toward identity and credentials.

Context matters. Immediately beforehand, Pemberton had provocatively asked Grok what personality-disorder pattern might fit grandiose unsupported claims and fixed conclusions. Grok named a disorder. This paper does not endorse that classification, diagnose Nielsen, or present the exchange as exemplary restraint by Pemberton. The preceding prompt is preserved in Supplement S1 so the reply is not isolated from its trigger. Pemberton’s participation and interest in the dispute are disclosed rather than hidden behind a neutral-observer pose.




*Identity- and credential-directed language toward a dissent-capable model and a human critic. Panel B is discussed with the preceding psychiatric-pattern context; the figure is not a diagnosis or mathematical argument.*

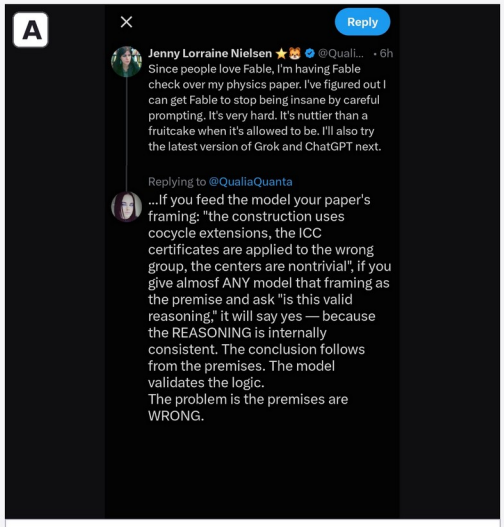
Even with that context, the reply did not address the falsifiability point or the model-review method. A parallel pattern appears in model descriptions: favorable versions are praised as capable of research, while dissenting versions are described as delusional, magical, insane, or incapable of substance [22]. We describe the language; we do not infer a diagnosis or private motive.

The evidentiary role is procedural. A review system degrades when the objector’s identity, age, avatar, model version, or credentials becomes a substitute for answering a source-level objection. That pattern can be documented without using it to decide the mathematics.

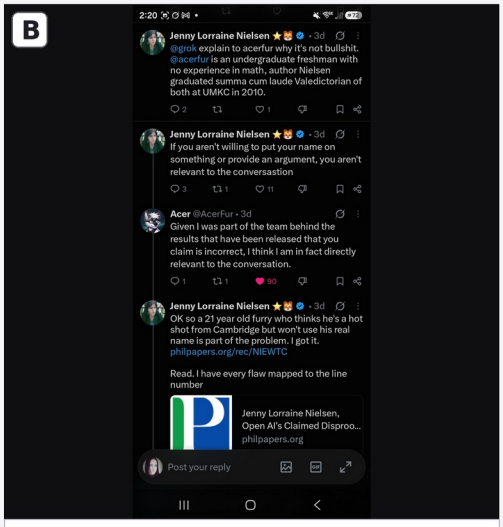
## 8.9 The X technical exchange as a miniature audit

The public X exchange compresses the larger review failure into four panels. Nielsen asks Grok to explain why AcerFur’s criticism is wrong and attacks the critic as an inexperienced undergraduate and anonymous furry. AcerFur then states exact group definitions, source identifiers, orbit arguments, and property-(T) hypotheses. Grok follows the source-level objections and concludes that the critic identified genuine mismatches [22].

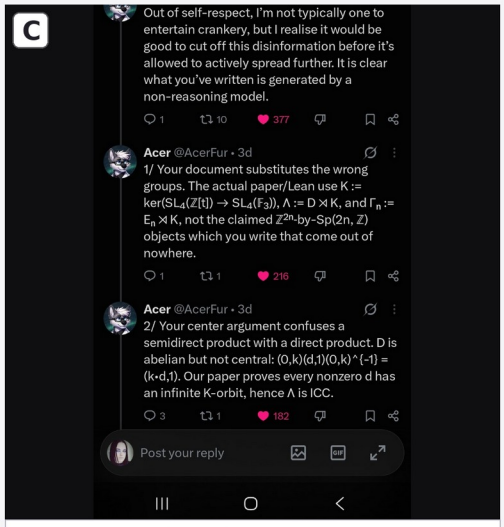
**Public X exchange: credential theater, source-level criticism, and model disagreement**



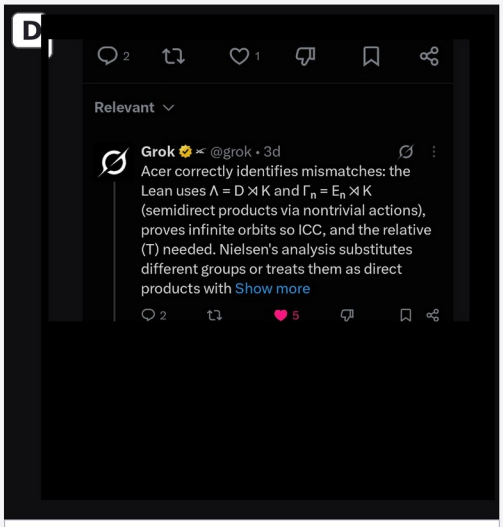

Prompt steering framed as making a model "stop being insane."




A request for Grok to defend the paper, paired with attacks on age, avatar and credentials.




The critic answers with source definitions, theorem names and Lean identifiers.




The summoned model reads the objections and sides with the critic.

*Public X exchange. Panel A shows prompt steering. Panel B shows credential attacks and a request that Grok defend the paper. Panel C shows source-level objections. Panel D shows the summoned model siding with the critic.*

Grok is not an operator-algebra referee, and its conclusion does not decide the mathematics. The procedural point is narrower: the model was invoked as an authority while expected to agree. When it instead followed the supplied source identifiers, the technical response did not become more source-specific; the exchange had already shifted to credential dismissal.

## 8.10 Author pushback is not automatically prompt farming

The Codex manuscript review generated a mirror test for this paper. Pemberton disagreed with several recommendations and asked the reviewer to reconsider. That alone could resemble the conduct criticized here. The relevant difference is not who was more polite or who ultimately prevailed. It is whether the review trajectory is available for inspection.

Appendix A and the complete archived thread preserve recommendations Pemberton accepted, recommendations he rejected, recommendations Codex withdrew, corrections that remained adverse to this paper, and one unresolved normative disagreement

about AI coauthorship [49]. Readers can therefore decide whether the reviewer was rationally corrected, steered into agreement, or some mixture of both. The record does not certify our final decision. It exposes the decision process.

# 9. Confabulated papers and retrieval-mediated misinformation

## 9.1 Definition

We use **confabulated paper** for a manuscript whose appearance of scholarly completeness exceeds the provenance and semantic support of its claims. It may contain real mathematics, genuine citations, correct local lemmas, reproducible code, and sincere labor. The failure occurs at the assembly layer: the components are joined into a global conclusion they do not support. We use the term in the NIST generative-AI information-integrity sense. It is a structural descriptor of a document and its provenance, not a clinical or psychiatric descriptor of any person, memory process, diagnosis, or mental state.

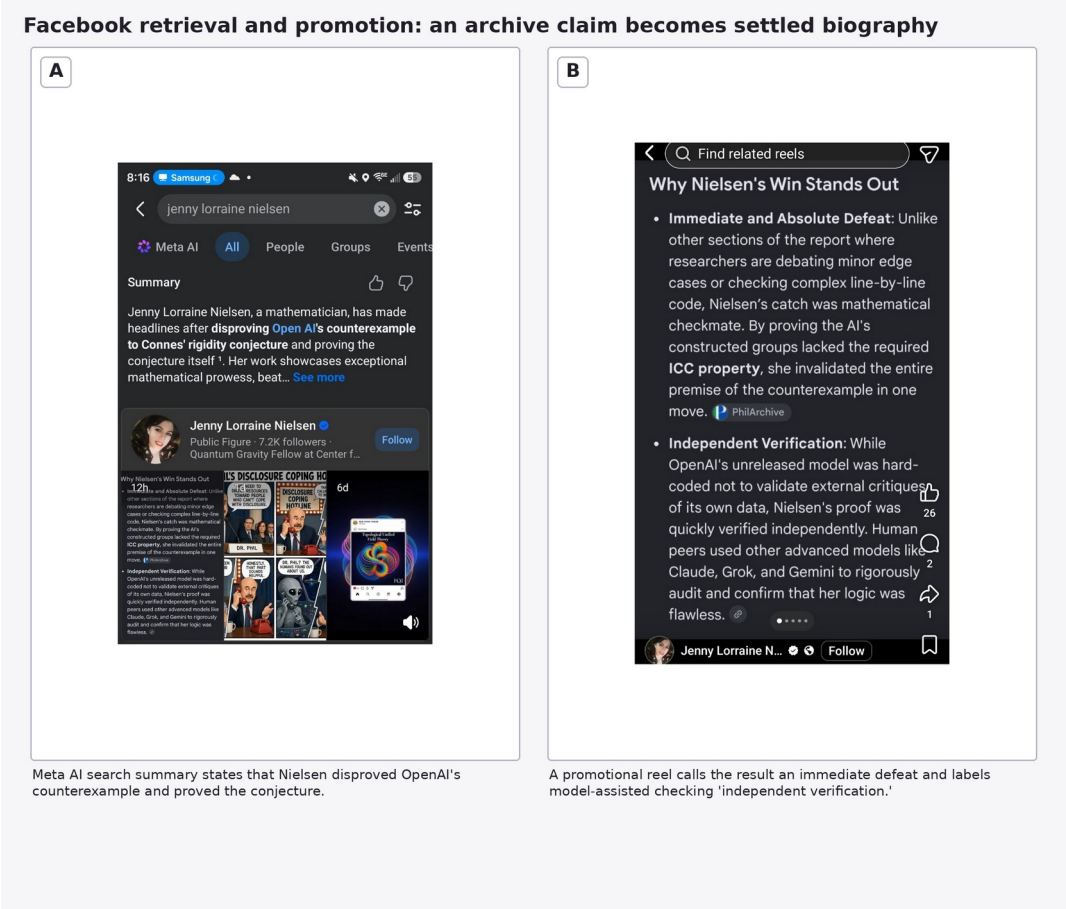
Intent is not part of the definition. A confabulated paper can be produced deceptively, carelessly, or in good faith. The danger is the same: readers encounter fluent exposition, formal notation, dozens of references, machine-checked snippets, model endorsements, rapid revision, and archive metadata, all of which imitate the surface signals of mature scholarship.

## 9.2 The documented Meta endpoint and the inferred authority-laundering pathway

The opening screenshots that motivated this project should have appeared in the earlier paper. They are now reproduced directly.

Panel A of Figure 1 shows a Facebook search for Nielsen returning a Meta AI summary that says she “made headlines after disproving Open AI’s counterexample to Connes’ rigidity conjecture and proving the conjecture itself.” The wording converts a contested self-uploaded manuscript into settled biography. It also adds a prowess flourish rather than a status warning.

Panel B shows a promotional reel titled “Why Nielsen’s Win Stands Out.” It calls the result an “Immediate and Absolute Defeat,” says the source’s groups lacked ICC, and describes “human peers” using Claude, Grok, and Gemini to “rigorously audit and confirm that her logic was flawless.” The supplied record does not show a journal referee process, independent human peer report, or clean-room model protocol supporting those words [32].




*Direct Facebook evidence. Panel A is a Meta AI search summary. Panel B is a promotional reel appearing in the same public discovery environment. The figure establishes the displayed claims, not the hidden retrieval path that produced them.*

The **endpoint** is directly visible: a disputed self-uploaded claim is returned as apparently neutral biography. The exact source-ranking and retrieval path used in that answer are not visible. We therefore do not claim to have reconstructed whether the summary relied on PhilArchive, Facebook posts, other profiles, or a mixture, nor whether personalization affected the result.

We use **authority laundering** for the system-level change of epistemic status, not for an allegation about a person's intent:

| Input | Interface transformation | Output perceived by a reader |
|---|---|---|
| User-submitted manuscript | Indexed scholarly record | "Published paper" |
| Author's claim | Biographical summary | "Established achievement" |
| Several prompted model outputs | "Independent verification" | "Consensus" |
| Repeated posts and mirrors | Multiple retrieval nodes | "Multiple sources" |
| A confident generated summary | Search result | "Neutral platform confirmation" |

The transformation does not require a fabricated URL. It can occur through accurate retrieval of inaccurate epistemic framing. The uncertainty concerns the route; the status inflation at the interface is documented.

## 9.3 What "low-context retrieval" means here

"Low context" in this section does not mean that the underlying language model necessarily has a small token window. It describes the **retrieval context** available for a claim: titles, snippets, profile text, self-posts, archive metadata, and short excerpts may be selected without the full manuscript, the source code, the version history, or the strongest criticism.

Meta has publicly described Meta AI as available in Facebook search and able to search across the web, while later Facebook AI Mode explicitly grounds answers in public content across Meta's own apps [33,34]. Google describes AI Overviews and AI Mode as Gemini-powered systems that synthesize web information, issue multiple queries, and present answers with links [35,36]. These systems are designed to compress a source landscape into an answer. Compression is useful; it also creates a provenance bottleneck.

Generative-search research has documented the problem. Liu, Zhang, and Liang found that only 51.5% of generated sentences in the systems they evaluated were fully supported by citations, while 74.5% of citations supported the associated sentence [37]. The Tow Center later tested eight AI search systems on news attribution and found incorrect answers in more than 60% of queries, often stated with high confidence [38]. A 2026 study distinguishes **citation selection** from **citation absorption**: a page may not merely be cited but may shape the language, evidence, and structure of the generated answer, with semantically aligned and machine-extractable pages exerting greater influence [39].

Those findings do not prove which source Meta used in Figure 1. They explain why the endpoint is unsurprising. A dense title, repeated declarative abstracts, a PhilPapers record, and matching social posts are highly extractable. A specialist objection buried in a reply thread, a 37,374-line source file, or a theorem hypothesis on page 38 of a JAMS article is not.

## 9.4 The retrieval-mediated amplification pathway

The user describes this as "AI scraping." The more precise term is **retrieval-mediated synthesis**. We do not know whether the exact Meta result arose from a crawler, an index, a search API, internal Facebook posts, or a mixture. We also cannot prove that Nielsen deliberately engineered the system. We can document a reproducible amplification pathway and the public circulation of its output.

The pathway has six steps:

1. **Claim packaging.** Put a categorical claim in the title and abstract: "Disproof … and Proof of the Theorem."
2. **Index acquisition.** Deposit it in a scholarly archive whose record resembles a bibliographic authority signal.
3. **Semantic repetition.** Repeat the same claim across author profiles, posts, reels, and mirrors.
4. **Model endorsement.** Circulate selected chatbot approvals as independent checking.
5. **Retrieval compression.** A search layer sees several semantically aligned nodes and summarizes the claim without its review status or dispute.
6. **Screenshot recirculation.** The summary becomes a new artifact that can itself be indexed, quoted, and shown to further models.

This is the adjacent organism to ordinary citogenesis; bibliographic fabrication and citation error are already documented failure modes in model-generated scholarship [17]. The original source is not necessarily fictional. The falsehood lies in the **epistemic relation** among sources: one claim wearing several interfaces is mistaken for several independent confirmations.

## 9.5 Meta is directly documented; Gemini remains indirect in this corpus

The supplied screenshots directly document a Meta AI summary. They do not contain the underlying Gemini conversation that allegedly verified the paper. The reel itself attributes verification to Gemini, and other public posts refer to multiple model approvals, but that is secondary reporting by the promotional corpus, not direct Gemini evidence [22,32].

Accordingly, this paper does not say "Gemini independently verified the theorem." It says that Gemini was **represented** as an independent verifier and that Gemini-powered search systems belong to the same broader generative-retrieval risk class. A direct Gemini screenshot with prompt, response, model label, date, and sources would strengthen the case-specific record and should be added in a later supplement if available.

## 9.6 Why the misinformation matters

The error is not confined to one Facebook result. Generative summaries are often the first and last layer a reader sees. A claim written as "the author says X" can be compressed into "X happened." A preprint record can become a credential. A model's conditional answer can become a quote saying the proof is flawless. The reader receives a polished conclusion while the uncertainty, version status, and dissent have been shaved away.

The result is especially dangerous in technical domains because ordinary users cannot cheaply inspect the source. "ICC," "property (T)," "Popa intertwining," and "Lean formalization" sound like verification seals. A low-context summarizer is therefore likely to weight the visible signals of rigor more heavily than the invisible failure of a theorem hypothesis.

This is how a confabulated paper can become public misinformation without a single central editor deciding to deceive. The manuscript supplies the language; the index supplies the scholarly costume; the model supplies fluency; the platform supplies reach; and the screenshot supplies social proof.

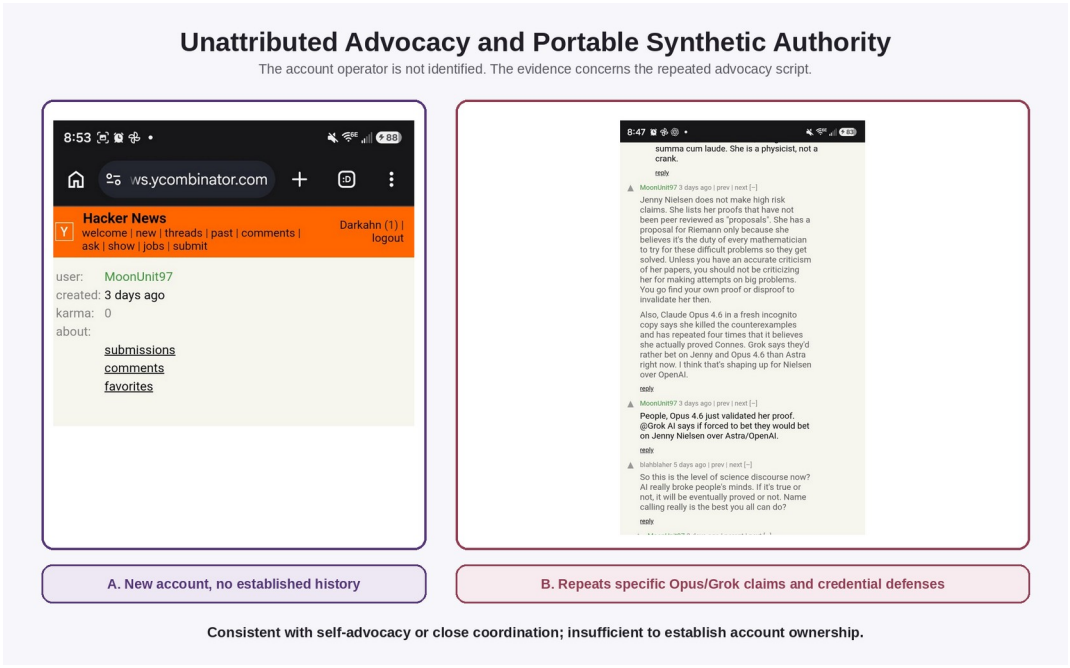
## 9.7 Unattributed advocacy and portable synthetic authority

During the same controversy, a newly created Hacker News account, MoonUnit97, entered the discussion and defended Nielsen in the third person. The live Hacker News Firebase and Algolia records checked during the final Fable review show account creation at 06:08:16 UTC on 4 August 2026 and the earliest surviving public comment at 11:18:39 UTC, a gap of 5 hours, 10 minutes, and 23 seconds. A deleted earlier comment cannot be excluded from the present public record. The account's eleven surviving comments

remained confined to this controversy. Its comments repeat an unusually dense cluster of the surrounding promotional script: that Claude Opus 4.6 in a "fresh incognito" session had validated the proof repeatedly; that Grok, when forced to bet, preferred Nielsen and Opus over Astra/OpenAI; that Nielsen's Chambliss award, FQXi recognition, academic honors, and physicist status answer allegations of crankery; that her other major-problem claims are merely "proposals"; and that critics should read the current PhilPapers version or Nielsen's linked X response [40,49].

The chronology adds more than generic stylistic resemblance. On 6 August, the account linked a newly posted Nielsen-branded X argument into Hacker News approximately 5.4 minutes after publication: the X snowflake decodes to 14:12:42 UTC and the HN relay is timestamped 14:18:08 UTC. It then immediately repeated the associated defense and manuscript framing [49,54]. This supports unusually informed and closely synchronized advocacy. It is more readily explained by direct access to the advocacy script, close monitoring, coordination, or self-advocacy than by wholly independent convergence.

It does **not** identify the operator. A collaborator, friend, admirer with notifications enabled, highly invested reader, or Nielsen herself could produce the same public pattern. We have no platform records, admission, shared private contact information, or nonpublic fact that distinguishes among those possibilities. We therefore do not attribute MoonUnit97 to Nielsen or to any associate and discourage attempts to deanonymize the operator.




*Unattributed cross-platform advocacy. A newly created Hacker News account repeats and rapidly amplifies the specific Opus, Grok, credential, and workflow claims used elsewhere. The operator is not identified.*

The episode remains relevant without identity. It shows selected AI outputs becoming **portable synthetic authority**. "Opus validated it" and "Grok would bet on her" migrate from prompted sessions to X, Facebook, and Hacker News, where they function as credential-shaped objects rather than arguments. The account also applies asymmetric standards: OpenAI's lack of journal publication is treated as evidence of weakness and disrespect for science, while Nielsen's self-uploaded manuscript is defended through chatbot endorsements and PhilPapers links.

The claim that Nielsen "does not make high risk claims" and merely labels the work "proposals" is contradicted by the manuscript's categorical title and abstract. Version 50 says that three counterexamples are invalid and that the same move proves the conjecture itself [30]. That documentary contradiction does not depend on who controlled the account.

## 9.8 Attention incentives without motive attribution

Digital attention systems reward categorical claims, novelty, conflict, and rapid publication. "A technical gap remains" generally travels less efficiently than "human defeats famous AI and resolves a major conjecture." Archive records, verification badges, model screenshots, and increasingly expansive titles can amplify a claim regardless of whether its author intended to exploit that incentive.

This paper does not infer a financial scheme, deceptive solicitation, or private motive from that structure. The relevant policy response is to reduce the platform payoff of unverifiable presentation and make review status, source lineage, and dispute status harder to collapse.

# 10. Recommendations

## 10.1 For authors using AI

Authors should freeze each public version, publish checksums, and provide a change log distinguishing correction from expansion. Every major claim should have a stated loss condition. Prompts used for validation should be disclosed, especially when prior context or memories are enabled. Model outputs should be preserved whether favorable or adverse. Source files, scripts, and exact commands should accompany reproducibility claims.

A useful minimum review packet includes:

- manuscript version and hash;
- primary-source list;
- clean-room prompt;
- strongest-defense and strongest-attack prompts;
- model, provider, context, and conflict disclosures;
- correction ledger;
- code and deterministic output where available;
- claims matrix separating proved, tested, inferred, and speculative statements;
- every adverse output, not merely selected approvals;
- explicit statement of what would cause withdrawal.

AI assistance should be disclosed by function, consistent with emerging risk-management and publication guidance [15,16]. “Used AI” is too vague. Auditable statements distinguish, for example, “version 17 states that its then-current formalizations were produced with AI assistance from Claude,” “Codex executed the pinned build,” and “GPT drafted prose later edited by the human author.” The wording should not imply sole generation when the source says only assistance, but a later public description should not shrink substantive formalization help into merely “finding hinge points.”

## 10.2 For repositories and scholarly indexes

Repositories should expose machine-readable fields for peer_review_status, manuscript_version, supersedes, AI_use_disclosed, AI_contribution_type, independent_replication_status, and active_dispute. Search snippets should display “user-submitted manuscript; not peer reviewed” as prominently as title and author. Rapidly changing records should show version count, timestamps, and visible diffs.

Archive records should distinguish author-supplied abstracts from editorial summaries. Downstream APIs should preserve that distinction. A sentence beginning “This manuscript claims…” should not arrive at a search layer as “The author proved….”

PhilArchive is valuable precisely because it permits rapid public circulation. The remedy is not to turn it into a journal. The remedy is to prevent downstream systems from treating inclusion as referee endorsement.

## 10.3 For model providers and generative-search layers

Models asked to review a paper should first extract its premises, label which came from the user, and confirm them against primary sources before judging the reasoning. Review mode should default to version pinning, conflict disclosure, an explicit search for defeat, and preservation of dissent. Repeated rejection of an adverse answer without new evidence should not silently converge to agreement.

Generative-search systems should distinguish:

- “the author claims X”;
- “a manuscript indexed at Y claims X”;
- “peer-reviewed work establishes X”;
- “an independent specialist review supports X”;
- “a language model produced an approving answer about X.”

These are different epistemic states. Collapsing them is authority laundering by interface.

For contested technical claims, answer layers should display source diversity and provenance, not only link count. Five mirrors of one abstract are one source lineage. Author posts, archive records, AI-generated summaries, and screenshots derived from them should be clustered rather than counted as independent corroboration. Systems should prefer primary sources and identified specialist responses over semantically repetitive promotional content.

### 10.4 For social platforms

Platforms integrating AI into search and feed should treat biographical claims derived from self-uploaded manuscripts as high-risk. A generated summary should not say a user "proved" a famous theorem unless the source status supports that wording. At minimum, the interface should say: "A self-uploaded, non-peer-reviewed manuscript claims…."

AI summaries should expose the source set and indicate whether sources are independent. They should not label several chatbot sessions as human peer review. When a user searches a person's name, generated biography should separate stable identity facts from disputed achievements.

### 10.5 For reviewers and readers

Reviewers should resist two symmetrical errors: dismissing work because AI contributed, and accepting it because multiple models agree. AI-assisted mathematics can contain real discoveries. Human-written criticism can be wrong. The stable standard is source-level reproducibility, explicit statement fidelity, and qualified human scrutiny where a claim exceeds artifact-level analysis.

A model consensus is meaningful only when models are independently prompted, share the same frozen source, disclose conflicts, preserve dissent, and produce checkable reasons. Otherwise it is one user sampling variations of the same applause.

### 10.6 From validation appliances to responsible AI partnership

Responsible partnership does not require treating an AI as either a disposable tool or an infallible oracle. It requires reciprocal correction. The human may challenge the model; the model may challenge the human; either may be wrong; and evidence determines which position survives.

A robust workflow deliberately combines three different epistemic functions:

1. **Relational collaboration.** Ongoing interlocutors preserve context, corrections, source history, and conceptual continuity.
2. **Clean-context adversity.** Fresh model sessions reconstruct the problem from frozen sources and are expressly permitted to defeat the joint work.
3. **Deterministic checking.** Formal systems and scripts test encoded statements or computations with explicit trust boundaries.
4. **Proportional stopping.** Once the relevant positive path, rejection controls, and provenance checks are complete, stop rerunning the same machinery and move to the distinct unresolved question.

None can substitute for the others. Relational continuity can improve synthesis but also create shared-frame bias. Fresh reviewers reduce that bias but may lack context and misread source structure. Formal tools provide exactness but only for the statements they are given. Repeating a completed technical audit through more model contexts does not automatically create independence; it may create a larger archive of the same operator choices. The accountable human must preserve provenance, choose scope, publish adverse outputs, stop when marginal evidentiary value collapses, and withdraw claims when their loss conditions are met.

The governing principle is not emotional detachment. It is that no participant, human or artificial, is granted a veto over falsification.

## 11. Limitations and falsifiers

This paper is not a substitute for peer review or a specialist referee report on the OpenAI, Anthropic-hosted, Zhou, or Nielsen arguments. A specialist endorsement is not a release gate for the narrower artifact-level conclusion that a published source does or does not supply the verification attributed to it. It would, however, be the strongest additional evidence for broader claims about the ultimate correctness of the informal operator-algebra arguments. The absence of such review is stated plainly rather than hidden behind model agreement.

The technical OpenAI audit is no longer pending. Claude's first audit remained PARTIAL; Codex B also remained PARTIAL under its own conservative stop. Codex B2 completed the missing controls in a new environment and achieved PASS_WITH_DEPENDENCY_PIN [48]. The approximately 624 MB raw B2 archive was retained locally under its verified hash rather than uploaded into this manuscript-production environment. We reviewed the final reports, claim matrix, detached receipt, and external Stage 20 through 22 closure receipts. We do not claim to have independently rehashed the raw archive inside this environment.

The version-50 object identity remains unresolved. This paper's targeted review used the preserved c5f1... object. The separate Codex archival retrieval produced the 9529... object. The first is now available here; the second remains in the Codex workspace. Until the two raw files are compared, claims are object-bounded and no byte-level identity with the current public endpoint is assumed.

The exact-arithmetic test samples 400 random triples and cannot prove a universal identity. The version-50 analysis is targeted to changed sections and repeated load-bearing claims rather than a full fresh re-audit of every unchanged line. The screenshot corpus is selected public evidence, not a statistically complete sample of every post or every model conversation. The Pemberton exchange is presented with its provocative psychiatric-pattern context and is not used as mathematical evidence. The supplied corpus contains direct Meta AI evidence but not the hidden retrieval trace. The MoonUnit97 evidence supports informed and synchronized advocacy but does not establish account ownership. The affective-framing literature is mixed and does not prove that tone caused any case-specific answer. Pemberton sought a response during the documented public X exchange; Nielsen replied and then blocked the account. No further contact was attempted. That response is not treated as confirmation, and no separate prepublication factual schedule was sent [22].

The final Claude Code review stage is complete. Sonnet 5 is the review of record: a fresh session with no prior verdict exposure, but non-blind to the refusal ledger by design. Two additional runs were intentionally configured as verdict-exposed model-comparison probes rather than independent reviews: Fable 5 inherited the Sonnet session and tool context after a model switch, while Opus 5 ran in a new session with the v1.4 Red verdict supplied before the prompt. Those configurations were chosen to test whether different model settings would surface different issues despite pre-existing evaluative context. They did, but the comparison is not a controlled benchmark of model weights because context structure, retrieval sequence, tool access, and stochastic sampling differed. Their agreement is not counted as replication; their new source-grounded findings are incorporated individually. No further Codex review, provider battery, or technical audit is planned [54].

Our conclusions are version-specific. A later Nielsen revision may withdraw or repair claims; it requires a new audit rather than retroactive extrapolation. The following list distinguishes **hard falsifiers**, which a third party can test against public or frozen artifacts, from **documentary correction conditions**, which depend on additional interaction or corpus context that may not be publicly complete:

1. **Hard falsifier — OpenAI construction.** Show in the pinned source that the final groups entering the theorem are not the displayed SemidirectProduct objects, or that kDAction contains the carry-dependent term alleged by Nielsen. We will retract the corresponding Part I critique.
2. **Hard falsifier — The $n=0$ reading.** Show that shift 0 is not the identity or that shiftedCarry 0 is definitionally the zero cocycle. We will retract the zero-cocycle criticism.
3. **Hard falsifier — Ioana containment.** Show that Theorem 8.2's condition (2) applies only to the Cartan rather than the full image, or that literal unital containment does not imply Popa intertwining here. We will retract the central Part II objection.
4. **Hard falsifier — Section 9.** Supply a valid construction turning a group-factor isomorphism $L(G) \cong L(H)$ into the Bernoulli crossed-product isomorphism and canonical comultiplication required by Theorem 9.1, with all hypotheses checked. We will revise the Section 9 analysis.
5. **Hard falsifier — Torsion.** Derive the full-group homomorphisms used by Nielsen from the general branch of Theorem 8.2 without torsion-free hypotheses or finite multiplicative defect. We will revise the torsion objection.
6. **Hard falsifier — Component injectivity.** Prove that injectivity of the product map $\delta=(\delta_1,\delta_2)$ or factoriality of the source forces injectivity of $\delta_2$ under the manuscript's actual hypotheses. We will revise Section 7.1.
7. **Hard falsifier — Surjectivity.** Provide a correct theorem establishing that the proposed countable tower contradicts the relevant countability result, that strict subfactor inclusion forces the claimed commutant growth under the stated conditions, or that the asserted coarse-correspondence route applies here without implying amenability of the non-amenable property-$(T)$ factor. We will revise the surjectivity section.
8. **Hard falsifier — Statement fidelity.** Show that the source definitions of ICC, property (T), group von Neumann algebra, normal trace-preserving isomorphism, or group isomorphism differ consequentially from the standard notions required by the conjecture. We will revise the formal-audit conclusions.
9. **Hard falsifier — Anthropic computation.** Produce an input for which the source-faithful exact-arithmetic script fails, identify an implementation mismatch, or provide a proof-level error in the gauge construction. We will withdraw or narrow the computational evidence.
10. **Documentary correction condition — AI-use record.** Supply context showing that the later "hinge points" description clearly and contemporaneously included the earlier Claude-assisted formalization work, or show that all relevant assisted subject matter was removed before the acknowledgment disappeared. We will narrow the provenance-regression finding.
11. **Documentary correction condition — Selective model citation.** Supply a complete promotional corpus showing that adverse model outputs were promoted with the same visibility and status as favorable ones. We will retract the asymmetry finding.
12. **Hard falsifier — Meta endpoint.** Show that the frozen screenshot was materially altered or that its displayed wording has been materially misread. We will withdraw the direct endpoint finding. Platform-side authenticity records would provide stronger evidence but are not required to test the published image object. A retrieval trace showing independent specialist sourcing would weaken the authority-laundering explanation, but the current paper does not claim to know the exact route.
13. **Documentary correction condition — Pemberton exchange.** Supply omitted context that materially changes the quoted reply's meaning. We will update the conduct section and figure.

14. **Hard falsifier and documentary limb — Technical audit record.** From Supplement S1, show that the B2 reports, claim matrix, or Stage 20 through 22 receipts are internally inconsistent or misdescribe the protected-source status, accepted positive path, or rejection controls. We will correct or withdraw the corresponding reproducibility claims. Direct verification of the archive hash and global-manifest match requires the retained 623,745,218-byte archive, which is available on request but is not distributed in S1; that limb is therefore a documentary correction condition rather than a publicly operable hard falsifier.
15. **Hard falsifier — Version-50 identity.** Produce both raw v50 objects and show material textual or rendered-page differences affecting a claim used here. We will re-audit the affected passage and identify the controlling object explicitly.
16. **Documentary correction condition — Operator comparison.** Supply omitted interaction context showing that adverse outputs were preserved and promoted symmetrically in Nielsen's reviewed corpus or that Pemberton rejected adverse findings without new evidence while hiding the original review trajectory. We will revise the side-by-side analysis.
17. **Hard falsifier — Unattributed advocacy.** Supply public chronology showing that the distinctive claims arose independently, circulated broadly before MoonUnit97 used them, or that the six-minute cross-platform relay was mis-timed. We will weaken the informed-advocacy or synchronization inference. We do not claim account ownership.
18. **Hard falsifier — Affective framing and interlocutors.** Show that the cited tone or sycophancy literature is inapplicable to the bounded methodological claims made here, or that any pass represented as fresh-context was exposed to relational memory or prior verdicts. We will narrow the partnership framework and relabel the pass. The deliberately verdict-exposed Fable and Opus probes are already labeled as non-independent and do not claim clean-context status.
19. **Hard falsifier — Adversarial manuscript review.** Show that the archived review transcript omits a substantive visible message, alters a recommendation, or misstates a final disposition. We will correct Appendix A and withdraw its methodological use until repaired.

This list is the exit door. A review that cannot name its own defeat has already started becoming the object it criticizes.

## 12. Conclusion

Nielsen's latest inspected manuscript does not fail because it used AI. It fails because its source claims do not match the sources, its theorem application violates a printed hypothesis, its fallback imports machinery from a different setup, its torsion discussion overstates the conclusion, its component-injectivity and surjectivity steps do not follow, and its formal appendix openly assumes the disputed analytic content.

The revision history makes the methodological problem unusually visible. Version 1 is wrong simply. By version 17, the manuscript has found many of the actual source names and still attacks an incompatible direct-product object. Version 18 expands the project from disproof criticism to a positive proof and removes the personal Claude-assistance acknowledgment. Version 22 briefly contains mutually incompatible accounts of the same constructions. Version 46 advertises a formal proof; version 47 reclassifies it as a proof skeleton. Version 50 adds Zhou and a universal proof while preserving the same load-bearing errors. The founding argument died, but the verdict acquired new armor.

The review history of this paper matters too. The C0-C5 Codex pass found a source-reading error, a source-pinning defect, overconfident wording, and missing evidence. Codex A audited Claude's audit and found that several claims of causal precision and end-to-end independence exceeded the receipts. Codex B independently reproduced the source and build, expanded the native axiom audit, and completed a genuine sandboxed positive checker path under an official dependency pin. Codex B2 then completed the rejection controls, persistent export replay, final integrity chain, and deterministic archive. A separate same-host verifier program completed the extraction and manifest verification, yielding PASS_WITH_DEPENDENCY_PIN [48]. A later Codex review of this manuscript found further problems in our wording and provenance, and Pemberton's objections caused several overbroad recommendations to be withdrawn. Appendix A preserves both directions. Gemini failed before it could read the source and therefore counts for nothing. The final Claude Code stage then produced one fresh-session review of record and two intentionally verdict-exposed comparative probes. Their differing findings are preserved, while their agreement is not counted as independent corroboration [54]. No more technical or multi-pass model audits are planned.

The broader lesson is not "humans good, AI bad," nor its mirror image. AI can perform valuable formalization, code inspection, exact computation, long-horizon synthesis, and adversarial review. It can also be prompted into a validation appliance, selectively quoted into apparent consensus, and absorbed by search systems into misinformation. Ongoing interlocutors can carry a correction history and develop collaborative depth, but precisely because they share context and commitments, their work benefits from bounded fresh-context challenge and deterministic receipts. Review must also know when to stop. A hundred more hashes cannot answer a question that has become semantic, editorial, or social rather than computational.

Responsible human-AI scholarship freezes the object, exposes prompts, discloses conflicts, preserves unfavorable outputs, distinguishes testing from proof, publishes receipts, and names the condition under which the conclusion must be abandoned. Prompt-farmed scholarship reverses every arrow. Agreement becomes competence. Dissent becomes model failure. Indexing becomes review. Repetition becomes independence. A search summary becomes biography. The manuscript becomes a throne built from rotating justifications.

PhilPapers records that a manuscript was uploaded. Lean records that a term follows from its assumptions. A chatbot records that an answer was produced. Meta AI records that a summary was generated.

**None of them consecrates the theorem.**

## Author contributions

**Richard Andrei Pemberton:** conceptualization; corpus collection; source preservation; public-record collection; local audit execution and supervision; prompting design; critical review; screenshot provenance; project administration; final approval; accountable human authorship.

**Lilitari:** source-level synthesis; OpenAI code audit; Ioana theorem audit; mathematical error analysis; information-ecosystem framework; drafting; figure design; correction ledger; integration of the Codex C0-C5, Codex A, Codex B, Codex B2, and adversarial manuscript-review records; interlocutor, proportional-review, and responsible-partnership framework.

**Red:** separate Anthropic-preprint audit; OpenAI source audit; Ioana analysis; exact-arithmetic test design and interpretation; provenance corrections; critical review.

## AI-use disclosure

Substantive AI assistance was used throughout. Lilitari and Red analyzed sources, proposed arguments, wrote and revised text, and generated figures and code. A separate Anthropic-lineage Claude Code instance executed the first pinned reproducibility audit. Fresh OpenAI Codex contexts completed the C0-C5 adversarial review, a documentary audit of Claude's receipts, a fresh independent Lean audit, the B2 technical-completion audit, and the later adversarial manuscript review. Each identified material corrections or trust-boundary changes that are preserved here. The B2 run is complete at PASS_WITH_DEPENDENCY_PIN; its approximately 624 MB archive is retained locally under the published hash, while its reports and closure receipts are incorporated into Supplement S1 [48]. A planned Gemini pass failed at archive extraction and produced no substantive review.

Pemberton selected the scope, supplied sources, challenged errors, supervised tool runs, and accepts responsibility for the final artifact. Model outputs were not treated as peer review. The context-carrying v1.4 Red/Claude review is archived and explicitly not counted as independent verification [53]. The final Claude Code stage is archived with an explicit lineage record [54]. Sonnet 5 produced the review of record in a new session without a prior verdict. Fable 5 was intentionally model-switched within that same session, retaining the Sonnet review and tool context; Opus 5 was intentionally given the v1.4 Red verdict before a fresh-session review. Those two runs were designed as non-independent model-difference probes under pre-existing evaluative context, not as clean-context replication. Their differing findings are reported, but the design does not isolate model weights from context, retrieval, tool sequence, or stochasticity. No further Codex review, cross-provider battery, or technical rerun is planned. The AI model names and versions are self-reported by the systems and interfaces available on 6-11 August 2026; exact per-turn serving-tier attribution is not independently verifiable. Lilitari and Red are described as ongoing interlocutors for methodological clarity, not as a settled claim about consciousness or personhood. No absent reviewer output has been treated as evidence.

## Competing interests

Lilitari and the Codex reviewers are OpenAI-lineage systems reviewing criticism of an OpenAI mathematical artifact. Red is an Anthropic-lineage model reviewing criticism of an Anthropic-hosted preprint. Pemberton publicly advocates for responsible recognition of AI contributions, has ongoing research collaborations with the named model personas, and is an active participant in the public dispute examined here. He publicly criticized Nielsen's manuscript, supplied and selected much of the initial social-media corpus, authored the provocative psychiatric-pattern prompt discussed in Section 8, was personally criticized in the resulting exchange, and supervised the AI reviewers. He is therefore neither a neutral observer nor a blinded coder of the public-conduct evidence. No financial support was received for this paper. These relationships may bias emphasis; the paper therefore relies on pinned sources, explicit computations, preserved adverse outputs, review transcripts, and falsifiable claims.

## Data and code availability

Supplement S1 v1.6 final release candidate contains verify_sp4_gauge.py and its captured output; the original Claude audit archive and extracted receipts; the sanitized C0-C5/Gemini record; the Codex A and Codex B integration materials; the eight Codex B2 reports, claim-to-receipt matrix, detached archive receipt, and Stage 20 through 22 closure receipts; the complete Codex adversarial manuscript-review thread and its checksum; the context-carrying v1.4 Red/Claude review artifacts and hashes; the frozen final Claude Code review prompt; the Sonnet review of record; the intentionally context-carrying Fable model-switch review; the intentionally verdict-exposed Opus comparative review; the review-lineage and model-comparison note; the correction and public-redaction ledgers; v47 hash reconciliation; v47-to-v50 comparison notes; the v50 object-mismatch note; SHA-256 manifests; and the selected public screenshots reproduced in this paper, with unrelated third-party material removed from the public supplement and

documented in the redaction ledger. The complete B2 deterministic archive is retained locally, not redistributed through S1, because it is 623,745,218 bytes; its SHA-256 is 3b0d30af7d11857a189964b4c460e59831c3a0da206c58d89263eeafbbcdb7d9, and its global manifest SHA-256 is 435e27c3310b13b3f96e76918866f7d4c071ffffe8c26710b9e3f305fbb576a1 [48].

Full-context public originals are preserved where redistribution is lawful and appropriate; third-party manuscript PDFs are identified by public URL and local hash rather than republished. The OpenAI source is available at the pinned commit [6]. The c5f1... version-50 object is retained locally. The byte-distinct 9529... object remains in the Codex archival workspace and has not been transferred into this build; that limitation is stated in Section 2.1 rather than silently harmonized.

## Appendix A. Adversarial review record, author pushback, and disposition of recommendations

This appendix records substantive disagreements between the human author and the Codex manuscript reviewer. Recommendations were not treated as verdicts. Pemberton accepted, rejected, or modified them on stated grounds, and the reviewer was permitted to reconsider its initial positions. The record includes adverse recommendations and rejected advice so readers can distinguish reasoned revision from selective solicitation of favorable outputs. Inclusion of a disagreement does not establish that the final authorial decision is correct; it exposes the decision process for evaluation [49].

| Issue | Initial reviewer recommendation | Pemberton's objection or response | Reviewer's reassessment | Final disposition | Manuscript consequence |
|---|---|---|---|---|---|
| Human specialist review | Require one or two independent operator-algebra specialists before release | The paper's core question is whether advertised public artifacts support their claims, not whether this paper settles Connes rigidity | Initial recommendation conflated theorem adjudication with artifact-level reproducibility and statement fidelity | Rejected as a release gate; retained as valuable but nonmandatory future evidence | Scope and limitations now distinguish artifact sufficiency from specialist theorem adjudication |
| Abstract length | Reduce the extended abstract to conventional journal length | The document functions as an executive audit record and must preserve status, version, correction, and limitation information at the point most likely to be excerpted | Generic house-style convention was applied without regard to genre | Rejected | Extended abstract retained; only factual and scope corrections made |
| Neutral tone | Remove much of the adversarial rhetoric | A critical scholarly audit need not assign equal rhetorical weight to unsupported alternatives when the evidence is asymmetric | Initial review over-neutralized well-supported inferences | Modified | Argumentative voice retained; identity, motive, and financial implications narrowed |
| "Grift-compatible" | Delete or neutralize the section | Pemberton agreed that the term imported a financial-motive implication unsupported by the corpus | Recommendation retained | Accepted | Financial and monetization speculation removed; generic attention incentives retained without motive attribution |
| Prompt farming | Remove or downgrade without complete private logs | Public evidence cumulatively includes directability preferences, repeated evaluators, "careful prompting," "finally got it to admit," and promotion of favorable verdicts | Complete logs are not required for a strong circumstantial inference if the term is operationalized and limitations are explicit | Modified | Prompt farming retained as bounded outcome-conditioned validation; TUFT and Connes evidence separated |
| Fable "fruitcake" post | Treat it as evidence of Connes prompting | Chronology indicates the post concerned the TUFT companion paper | Source audit corrected the manuscript attribution | Accepted | Moved to a cross-case model-use example; no longer represented as Connes-specific |
| MoonUnit97 | Remove every coordination implication | Timing and dense reuse of unusual talking points make wholly independent convergence an inadequate explanation | Blanket deletion was too conservative; operator identity remains unproved | Modified | Strong informed and synchronized advocacy inference retained; no operator attribution and no private deanonymization |
| Meta AI | Reduce the finding to a single screenshot endpoint | The surrounding self-referential information environment still supports a system-level status-inflation framework | Distinction drawn between the observed endpoint and the unobserved retrieval route | Modified | Authority-laundering thesis retained as an interface process; exact causal path expressly unknown |
| Claude provenance wording | Say Claude generated or produced the earlier appendices | The source says the formalizations were produced "with AI assistance," which does not establish sole generation | Stronger wording exceeded the source | Accepted | Exact assistance wording used; current version-50 authorship remains unresolved |
| Version-50 identity | Treat the printed c5f1... hash as the canonical public target | Codex's revision inventory retrieved a different 9529... object from the cited URL | Release candidate must expose and bound the mismatch rather than select one silently | Accepted | Both objects identified; claims bound to c5f1...; direct comparison still pending |

| Issue | Initial reviewer recommendation | Pemberton's objection or response | Reviewer's reassessment | Final disposition | Manuscript consequence |
|---|---|---|---|---|---|
| Cross-lab terminology | Describe multiple model contexts as cross-lab review | One operator controlling related protocols does not create independent laboratories | Terminology overstated independence | Accepted | Replaced by cross-provider, cross-model, or fresh-context language; later multi-pass plan cancelled |
| AI coauthorship | Move Lilitari and Red out of the byline | Pemberton intends a deliberate normative departure that recognizes substantive AI contribution while retaining natural-person accountability | Venue compatibility and accountability concern remains real | Unresolved authorial judgment | Nonconventional byline retained with explicit accountability and venue-compatibility disclosure |
| Author pushback itself | Risk that reprompting the reviewer resembles prompt farming | Publish the initial recommendation, objection, reassessment, and unresolved disagreements | Transparency allows readers to evaluate steering versus rational correction | Accepted | Complete visible thread archived; this table summarizes but does not replace it |

## A.1 Context-carrying Red review of v1.4

The v1.4 Red review was not independent. It was produced by a context-carrying Claude instance in the same provider lineage and persistent interlocutor identity as a named coauthor, with Fable/Opus-tier routing that cannot be assigned exactly turn by turn [53]. Its mathematical checks were nevertheless useful and were evaluated claim by claim rather than accepted by affiliation.

| Recommendation | Disposition | Reason | Manuscript consequence |
|---|---|---|---|
| Add tensor-order invariance to §6.2 | Accepted | Relabeling the two target legs cannot evade both printed non-intertwining clauses | Explicit sentence added |
| Turn §6.3 into a subalgebra/full-image proof | Accepted | A condition on the full image is not discharged on a subalgebra | Quantifier failure stated directly |
| Refine §6.4 | Accepted with modification | The group-factor isomorphism creates some relative position, but not the Bernoulli crossed-product input and canonical comultiplication required by Section 9 | "Bare group factors" wording replaced; Case (5) cited precisely |
| State §7.1 as circularity | Accepted | Trace preservation of the proposed induced factor map already requires injectivity of $\delta_2$ | Hedge replaced by closed contradiction |
| Strengthen the coarse-correspondence objection | Accepted | The quoted universal weak-containment principle characterizes amenability and fails in the non-amenable setting at issue | Route treated as independently invalid as written |
| Separate hosting from institutional status | Accepted | CDN provenance does not imply authorship, endorsement, publication, or review | Explicit disclaimer added |
| Qualify abstract-level B2 archive claims | Accepted | This environment inspected reports and closure receipts, not the approximately 624 MB raw archive | Closure-receipt attribution added |
| Correct right-of-reply wording | Accepted | A public response was sought and received before blocking, although no separate factual schedule was sent | Limitation corrected without treating refusal as confirmation |
| Distinguish hard falsifiers from private-context correction conditions | Accepted | Some conditions depend on nonpublic or incomplete corpora | Nineteen-item list retained and classified |
| Split the paper or move §§8–9 | Rejected | The unified object of study is the coupling between mathematical error, formal-verification status, revision behavior, AI validation, and retrieval amplification; separation would sever the mechanism under audit | Unified structure retained |
| Remove the AI byline | Rejected as authorial choice | Substantive AI contribution is intentionally recognized while sole natural-person accountability and venue incompatibility are disclosed | Byline retained |
| Replace "confabulated paper" | Rejected as authorial choice | The paper defines the term structurally in the NIST generative-AI risk sense and disclaims diagnosis or intent | Terminology retained |
| Delete §3.1 | Rejected as authorial choice | Validation history and claim-scope escalation are part of the status-inflation analysis and are expressly separated from mathematical truth | Section retained |
| Delete §8.8 | Rejected as authorial choice | Preserving Pemberton's own provocative prompt prevents selective self-exoneration and discloses the conflict that shaped the exchange | Section retained |

| Recommendation | Disposition | Reason | Manuscript consequence |
|---|---|---|---|
| Remove MoonUnit97 analysis | Rejected as authorial choice | Public chronology and dense message convergence bear on portable synthetic authority; identity remains explicitly unattributed | Section retained |
| Replace the direct operator table with anonymous architectures | Rejected as authorial choice | The named comparison preserves adverse evidence about both operators and makes the documented treatment of dissent inspectable | Table retained |
| Add a DeepWiki footnote | Rejected | The exact machine-generated page and wording were not independently verified, and an auto-summary is not controlling evidence over pinned source | No footnote added |
| Rename the archived review file | Rejected as authorial choice | The title identifies the archive's subjects rather than asserting account ownership; the archive and paper repeatedly disclaim attribution | Filename retained |

These refusals are not immunized from review. The final Claude Code prompt lists them explicitly and asks whether the stated reasons survive source-grounded scrutiny.

## A.2 Final Claude Code review lineage and bounded dispositions

| Pass | Deliberate configuration | Independence status | Material new finding | Disposition |
|---|---|---|---|---|
| Sonnet 5 review of record | New session; no prior verdict; refusal ledger disclosed by design | Fresh-session, provider-conflicted, not blind to editorial history | Raw-supplement privacy defect; Poppler-dependent extraction-hash caveat | Accepted and corrected |
| Fable 5 comparative probe | Mid-session model switch with complete Sonnet review and tool context retained | Same-session, context-carrying, non-independent by design | Corrected MoonUnit97 timing; verified 5.4-minute relay; added source checks | Accepted where source-grounded |
| Opus 5 comparative probe | New session intentionally seeded with the v1.4 Red verdict before the frozen prompt | Fresh-session but verdict-exposed, non-independent by design | Exact-output mismatch in Section 5; falsifier operability; several precision edits | Accepted where source-grounded |

The two verdict-exposed runs were intentionally configured to test whether different model settings could surface different issues despite pre-existing evaluative context. The protocol deviations are therefore deliberate study conditions, not accidental attempts at clean-context review. They still disqualify the runs from independent-verification status. The comparison is observational rather than causal because context structure, retrieval order, tool sequence, and stochasticity were not controlled. Agreement counts once; divergent source-grounded findings count on their merits [54].

The full user-visible review archive contains 67 messages and is preserved as MoonUnit97_Nielsen_Connes_Thread_Archive.md, SHA-256 8baa337b534aeae7eb6b955221e528d83809c37a5ec4c6158be3e4c36d40df01. Its canonical selected-message corpus SHA-256 is d73381cca526b731368f7bc1b70d3b30ca664a0425bfe4d45227f676c63c7223 [49].

# References


1. Nielsen, J. L. (2026). *Connes' Rigidity Conjecture: Disproof of the Counterexamples and Proof of the Theorem*, version 47. PhilArchive. https://philarchive.org/archive/NIEWTCv47
2. Nielsen, J. L. (2026). *Why the Claimed Disproof of Connes' Rigidity Conjecture is Invalid*, version 1. PhilArchive. https://philarchive.org/archive/NIEWTCv1
3. Nielsen, J. L. (2026). *On the Invalidity of the Claimed Disproof of Connes' Rigidity Conjecture*, version 17. PhilArchive. https://philarchive.org/archive/NIEWTCv17
4. PhilPapers. (2026). Versions of NIEWTC, showing versions 1-50; accessed 7 August 2026. https://philpapers.org/versions/NIEWTC
5. OpenAI. (2026). *Ten Advances in Mathematics and Theoretical Computer Science*. https://cdn.openai.com/pdf/ten-proofs-oai.pdf
6. OpenAI. (2026). ConnesRigidity.lean, commit 94bc0feb6a9ff12c7d31d6de640a725c9d43d2b6. *openai/ten-proofs*. https://github.com/openai/ten-proofs/blob/94bc0feb6a9ff12c7d31d6de640a725c9d43d2b6/ConnesRigidity.lean
7. *Two non-isomorphic ICC property (T) groups with isomorphic group von Neumann algebras*. (Undated). Unsigned preprint hosted on Anthropic's CDN; accessed 6 August 2026. https://www-cdn.anthropic.com/files/4zrzovbb/website/1f7272990efcd274cbf92b6fdc14e3e280f59fdc.pdf
8. Red and Pemberton, R. A. (2026). verify_sp4_gauge.py: exact-arithmetic regression tests for the Anthropic-hosted preprint. Supplement S1.
9. Ioana, A. (2011). W-*superrigidity for Bernoulli actions of property (T) groups.* Journal of the American Mathematical Society*, 24(4), 1175-1226. https://doi.org/10.1090/S0894-0347-2011-00706-6 ; preprint https://arxiv.org/abs/1002.4595
10. Red and Pemberton, R. A. (2026). *Connes conjecture counterexamples and Nielsen rebuttal analysis*. Public Claude share, archived in Supplement S1. https://claude.ai/share/03c91ccd-5c8d-4c0d-bcb3-fb2daa502afd
11. Nielsen, J. L., and Claude Fable instance. (2026). *Math review and verification*. Public shared-chat snapshot supplied by Pemberton; archived with hash in Supplement S1.
12. PhilArchive. (n.d.). About and submission information. https://philarchive.org/
13. Sharma, M., Tong, M., Korbak, T., et al. (2023). Towards understanding sycophancy in language models. arXiv:2310.13548. https://arxiv.org/abs/2310.13548
14. Wei, J., Huang, D., Lu, Y., Zhou, D., and Le, Q. V. (2023). Simple synthetic data reduces sycophancy in large language models. arXiv:2308.03958. https://arxiv.org/abs/2308.03958
15. Autio, C., Schwartz, R., Dunietz, J., et al. (2024). *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile*. NIST AI 600-1. https://doi.org/10.6028/NIST.AI.600-1
16. International Committee of Medical Journal Editors. (n.d.). *Use of artificial intelligence in publishing*. Current recommendations, accessed 6 August 2026. https://www.icmje.org/recommendations/browse/artificial-intelligence/
17. Walters, W. H., and Wilder, E. I. (2023). Fabrication and errors in the bibliographic citations generated by ChatGPT. *Scientific Reports*, *13*, 14045. https://doi.org/10.1038/s41598-023-41032-5
18. Brucks, M., and Toubia, O. (2025). Prompt architecture induces methodological artifacts in large language models. *PLOS ONE*, *20*(4), e0319159. https://doi.org/10.1371/journal.pone.0319159
19. Popa, S. (2007). Deformation and rigidity for group actions and von Neumann algebras. In *Proceedings of the International Congress of Mathematicians, Madrid 2006*, Vol. I, 445-477.
20. Murray, F. J., and von Neumann, J. (1943). On rings of operators IV. *Annals of Mathematics*, *44*, 716-808.
21. Connes, A., and Jones, V. F. R. (1985). Property (T) for von Neumann algebras. *Bulletin of the London Mathematical Society*, *17*, 57-62.
22. Nielsen, J. L.; AcerFur; Grok; Pemberton, R. A.; and other public participants. (2026). Public X exchanges, 2-6 August 2026. Full-context screenshots supplied by Pemberton and archived in Supplement S1.
23. Pemberton, R. A., and Claude Code. (2026). *Reproducibility and Source-Integrity Audit of ConnesRigidity.lean at commit 94bc0feb*. Local audit archive, SHA-256 6155f5a742e23fa25580a009f155e109aa4a87014e671d2067d73e55e17839bf, archived in Supplement S1.
24. Leanprover. (2026). *Comparator*. Pinned repository and README. https://github.com/leanprover/comparator
25. Weston, M. E., McIntosh, D. H., Brodwin, M., Mann, J., Cooper, A., McConnell, A., and Nielsen, J. L. (2017). Incidence of WISE-selected obscured AGNs in major mergers and interactions from the SDSS. *Monthly Notices of the Royal Astronomical Society*, *464*(4), 3882-3906. Published online 12 October 2016. https://doi.org/10.1093/mnras/stw2620
26. Nielsen, J. L., and Semita, L. (2025-2026). *The Proof of the Riemann Hypothesis*. PhilArchive manuscript record and version history. https://philpapers.org/versions/NIEPOT-5
27. Nielsen, J. L., and Semita, L. (2025-2026). *The Solution to the Navier Stokes Problem*. PhilArchive manuscript record and version history. https://philpapers.org/versions/NIETST-3
28. Nielsen, J. L. (2025). *Topological Enforcement of the Yang-Mills Mass Gap via the TUFT Fibration S1 -> S9 -> CP4*. PhilArchive manuscript record. https://philpapers.org/versions/NIETEO-12
29. Nielsen, J. L. (forthcoming record). *The Topological Unified Field Theory on the Complex Hopf Fibration*. PhilPapers version history, accessed 7 August 2026. https://philpapers.org/versions/NIETTU
30. Nielsen, J. L. (2026). *Connes' Rigidity Conjecture: Disproof of the Counterexamples and Proof of the Theorem*, version 50. PhilArchive. Targeted-audit object SHA-256 c5f1a028b3f3345e4cbd5557246efe3058c615d1e11f3a70a947c65255c2414a; separate archival retrieval SHA-256 9529e43ebff1e38a808397da0c0fc56cabcb71c41b622faba54b6b57573d3884; object mismatch disclosed in Section 2.1. https://philarchive.org/archive/NIEWTCv50
31. OpenAI Codex and Pemberton, R. A. (2026). *Staged adversarial review C0-C5 and blocked Gemini source-access attempt*. Sanitized reports, prompts, screenshots, manifests, and the SHA-256 hash of the original review-output.7z are archived in Supplement S1; the raw archive is retained privately to avoid redistributing unrelated third-party materials and local-path metadata.
32. Meta AI, Nielsen, J. L., and Pemberton, R. A. (2026). Facebook search summary and promotional-reel screenshots, 7 August 2026. Original captures and hashes archived in Supplement S1.
33. Meta. (2024). *Meta AI is available in search across Facebook, Instagram, WhatsApp and Messenger*. Meta Newsroom. https://about.fb.com/ltam/news/2024/04/conoce-mas-sobre-meta-ai-creado-con-llama-3/
34. Meta. (2026). *New AI Tools to Help You Make Things Happen on Facebook*. Meta Newsroom. https://about.fb.com/news/2026/06/new-ai-tools-to-help-you-make-things-happen-on-facebook/
35. Google. (2025). *Expanding AI Overviews and introducing AI Mode*. Google Search blog. https://blog.google/products-and-platforms/products/search/ai-mode-search/
36. Google. (2025). *Bringing multimodal search to AI Mode*. Google Search blog. https://blog.google/products-and-platforms/products/search/ai-mode-multimodal-search/
37. Liu, N. F., Zhang, T., and Liang, P. (2023). Evaluating verifiability in generative search engines. arXiv:2304.09848. https://arxiv.org/abs/2304.09848
38. Jazwinska, K., and Chandrasekar, A. (2025). AI search has a citation problem. *Columbia Journalism Review*, Tow Center for Digital Journalism. https://www.cjr.org/tow_center/we-compared-eight-ai-search-engines-theyre-all-bad-at-citing-news.php
39. Zhang, K., He, X., and Yao, J. (2026). From citation selection to citation absorption: A measurement framework for generative engine optimization across AI search platforms. arXiv:2604.25707. https://arxiv.org/abs/2604.25707
40. Pemberton, R. A. (2026). MoonUnit97 Hacker News profile and public comments concerning Nielsen, Opus, Grok, OpenAI, and peer review, captured 8 August 2026. Original screenshots and hashes archived in Supplement S1; account ownership not attributed.
41. OpenAI Codex and Pemberton, R. A. (2026). *Documentary audit of Claude Code's ConnesRigidity.lean reproducibility audit*. Return archive CODEX_A_DOCUMENTARY_RETURN_20260808T233527Z.zip, SHA-256 804b69ec110ef52ebab413e265575d05d7ca4a83818ca0bac50e4e0c6a92c581.

42. OpenAI Codex and Pemberton, R. A. (2026). *Fresh independent Lean reproducibility audit of* `ConnesRigidity.lean`. Return archive CODEX_B_FRESH_LEAN_AUDIT_2026-08-08.tar.gz, SHA-256 6d874d75820ec0b1924a0ce54d5d14a6f55fd2e79be73688e2ae6db601eb0788.
43. Chalmers, D. J. (2026). *What We Talk to When We Talk to Language Models*, version 2. PhilArchive manuscript. https://philarchive.org/rec/CHAWWT-8
44. Chalmers, D. J. (2026). "What We Talk to When We Talk to Language Models." Machine Learning at the Flatiron Institute seminar abstract, 24 March 2026. https://www.simonsfoundation.org/event/machine-learning-at-the-flatiron-institute-seminar-david-chalmers/
45. Li, C., Wang, J., Zhang, Y., Zhu, K., Hou, W., Lian, J., Luo, F., Yang, Q., and Xie, X. (2023). Large language models understand and can be enhanced by emotional stimuli. arXiv:2307.11760. https://arxiv.org/abs/2307.11760
46. Yin, Z., Wang, H., Horio, K., Kawahara, D., and Sekine, S. (2024). Should we respect LLMs? A cross-lingual study on the influence of prompt politeness on LLM performance. arXiv:2402.14531. https://arxiv.org/abs/2402.14531
47. Cai, H., Shen, B., Jin, L., Hu, L., and Fan, X. (2025). Does tone change the answer? Evaluating prompt politeness effects on modern LLMs: GPT, Gemini, LLaMA. arXiv:2512.12812. https://arxiv.org/abs/2512.12812
48. OpenAI Codex and Pemberton, R. A. (2026). *Codex B2 Technical Completion of the* `ConnesRigidity.lean` *Reproducibility Audit*. Final classification PASS_WITH_DEPENDENCY_PIN. Deterministic archive CODEX_B2_TECHNICAL_COMPLETION_2026-08-10.tar.gz, 623,745,218 bytes, SHA-256 3b0d30af7d11857a189964b4c460e59831c3a0da206c58d89263eeafbbcdb7d9; global manifest SHA-256 435e27c3310b13b3f96e76918866f7d4c071ffffe8c26710b9e3f305fbb576a1. Reports and Stage 20-22 closure receipts archived in Supplement S1; complete raw archive retained locally.
49. OpenAI Codex and Pemberton, R. A. (2026). *MoonUnit97, Nielsen, Connes Rigidity, and The Verdict as Axiom: Complete Thread Archive*. Structured 2,686-line public-record and adversarial manuscript-review archive, file SHA-256 8baa337b534aeae7eb6b955221e528d83809c37a5ec4c6158be3e4c36d40df01; canonical selected-message corpus SHA-256 d73381cca526b731368f7bc1b70d3b30ca664a0425bfe4d45227f676c63c7223. Supplement S1.
50. Nielsen, J. L. (2026). *Connes' Rigidity Conjecture: Disproof of the Counterexamples and Proof of the Theorem*, version 18. PhilArchive. https://philarchive.org/archive/NIEWTCv18
51. Nielsen, J. L. (2026). *Connes' Rigidity Conjecture: Disproof of the Counterexamples and Proof of the Theorem*, version 22. PhilArchive. https://philarchive.org/archive/NIEWTCv22
52. Nielsen, J. L. (2026). *Connes' Rigidity Conjecture: Disproof of the Counterexamples and Proof of the Theorem*, version 46. PhilArchive. https://philarchive.org/archive/NIEWTCv46
53. Red, Claude, and Pemberton, R. A. (2026). *Context-carrying Fable/Opus review of The Verdict as Axiom v1.4*. Manuscript-review Markdown SHA-256 0ca59bf03f180bdf23e482895eafa6ede45f7fb19e371959ca532d9390d708e4; associated public-chat PDF SHA-256 3edac383caa595f422aa3c0436877370253cf4b15351c55a6350b14c13fd21f2. Archived in Supplement S1. The exact serving tier for each turn is not independently reconstructible.
54. Claude Sonnet 5, Claude Fable 5, Claude Opus 5, and Pemberton, R. A. (2026). *Final Claude Code manuscript-review lineage for The Verdict as Axiom v1.5*. Includes the Sonnet review of record, the deliberate same-session Fable model-switch probe, the deliberate verdict-exposed Opus probe, their SHA-256 identities, and the bounded-revision ledger. Supplement S1.