Epistocracy and the Commitment Problem

Post-publication audit v2.2: the sentence-level census, and the strength of the Guerrero objection

14 August 2026. Version history: v2.1 re-examined the high-graded items under steelmanning at the author's challenge (two downgraded, §4); v2.2 applied the same steelmanning to the "flatly false" list (§5 re-sorted, headline corrected); v2.3 adds the verbatim sentence for every finding discussed.

1. The question

Alexander Guerrero's objection, stated in the Daily Nous thread and generalized: he opened the one passage of the paper that discusses his own work, found it wrong, and inferred that a paper which "gets something that easy that wrong" cannot be trusted — that a random-feeling spot check finding an error implies many errors. This document reports a test of that inference: a census in place of a sample. Every sentence of the published paper was read adversarially and checked — does each inferential step follow, is each factual claim true, does each citation support the sentence it is attached to, is each characterization of a named scholar's view accurate, and what is the worst defensible reading a hostile critic could give the sentence. Per instruction, this document contains no proposed revisions; it is an assessment, not an errata sheet.

2. Method

Six independent adversarial readers (same model class as the primary auditor) each took one section: Abstract–§2; §3; §4 Responses 1–3 with footnote 4; §4 Response 4; §§5–6 with acknowledgments and endnotes; and the formal appendix, where the instruction was to re-derive every mathematical claim from scratch. Each reader tested every sentence and reported only failures and risks. Every reported finding was then re-checked by the primary auditor before acceptance; several were rejected (including one refuted by the project's own July source-verification records) and overlapping findings were merged. Where a finding turned on what a cited author actually says, the primary sources were consulted directly: Brennan's Against Democracy (the copy in this project's Literature folder), Gibbons 2025 (published PDF), the Estlund, Bagg, and López-Guerra literatures, and the live AI report link. Coverage: 579 sentences.

3. Headline results

- **Roughly sixty distinct adjudicated items:** 1 high-severity, about 27 medium, about 30 low or minor. Of these, only 2–4 are literal defects a reader can verify mechanically (a stale cross-reference; a wrong referent; a sloppy citation range; a main-text/appendix mismatch — §5); the rest are misleading filings, overclaims, internal tensions, unsupported assertions doing argumentative work, and glosses whose letter outruns what the machinery or the sources support. Both the high-severity gradings (§4) and the "flatly false" gradings (§5) shrank when re-examined under deliberate steelmanning at the author's challenge — itself a datum about the method: adversarial reading over-grades, whoever performs it, and the steelman layer is load-bearing.

- **Zero errors in the formal core.** The appendix reader independently re-derived the income accounting, concavity and interiority conditions, the revolution values, the χ mapping, the Jensen chain, and every step of Proposition A1's proof, including the boundary evaluations and the openness argument. All sound, with no hidden assumptions. The confirmed defects in the appendix sit in the non-theorem "Remarks" (see §6 below).
- **Zero fabricated citations.** Every cited work exists as described; the reference list was verified in two independent passes before publication and nothing in this census contradicts that. The canonical AI failure mode — invented sources — is absent.
- **The errors cluster by sentence type, not by section.** Formal statements: clean. Uses of empirical studies: accurate at claim level, with a few scope slips. Characterizations of named scholars' views and one-line taxonomy filings: the worst stratum, now including the largest single finding. Argumentative connective tissue — framing sentences, summary sentences, "so" and "therefore" sentences: the numerical bulk of the medium findings.

4. The highest-graded findings, re-examined

All three items initially graded high were re-checked under deliberate steelmanning against the primary texts. One survives at high severity; two downgrade to medium. The re-examination is recorded in full because the movement itself is part of the evidence.

4.1 Section 3.1's reading of Brennan — downgraded to medium: contestable summary, not contradiction

The initial finding held that §3.1's "narrow reading" of Brennan is contradicted by the cited pages. On re-examination the charge decomposes. What is defensible: "described in narrow terms" can refer to the grounds — Brennan's council may strike only decisions judged "malicious, incompetent, or unreasonable," and Brennan himself writes that it "can only veto bad decisions as well as incompetently made laws"; on that reading, the paper's next sentence ("On this narrow reading the council rarely vetoes…") is a conditional characterization of a reading, not an exegetical claim, and cannot be refuted by quotation. Weight also belongs to an authority fact: Brennan, the maximally authoritative reader of Brennan, reviewed this description of his own proposal across three editorial rounds and let it stand. What survives against the paper, more modestly: as a summary, "His institutional specification is narrow" sits badly with the same pages' explicit domain breadth — the council "can veto any (or almost any) political decisions" (p. 216), "bad legislation, regulations, and executive orders" (p. 218), and "retains the right to veto any [day-to-day] legislation" (p. 217) — and with Brennan's contemplation of a "heavy gridlock" cycle of continual vetoes; and "judicial-review-style rights-and-procedural review" clashes with the paper's own §3.4 taxonomy, under which "incompetent" is a competence ground, not a rights ground (Brennan's jury analogy at p. 216 is explicitly competence-based, and his judicial-review analogy is offered for democratic compatibility, expressly not as a justification transfer, p. 219). Net: a contestable exegetical summary with genuine textual tension on both sides — medium severity, and materially different in kind from the Guerrero misfiling, where the view's owner confirmed the error; here the view's owner implicitly endorsed the description.

Unaffected by the re-examination: Brennan quotes López-Guerra's "competence-building process" in full at p. 214, so Response 3's description of the enfranchisement lottery ("without competence filtering") remains contradicted by the very pages the paper cites two paragraphs earlier. That item stands as found.

4.2 The noise claim against the paper's own Gilens concession — survives as the census's top item

Response 4(b): ordinary democratic deviation from the median voter "looks more like noise: sometimes the deviation goes to the right of the median voter, at other times to the left, and over time the deviations tend to average out." Twenty lines later, the same section: "actual policy tends to track the affluent" (citing Gilens, with the qualification literature). The natural reconciliation — median-voter deviation is noise even if the poor are systematically shortchanged — is unavailable to this paper, because its own model identifies the median voter as poor ($\delta < 1/2$; Proposition A1 assumes the poor are the median voter). Within the paper's framework, deviation from the median just is deviation from the poor, and the paper's own concession says that deviation is one-signed. The two sentences are quotable side by side, and the first carries contrast #1 of the comparative argument. A repair exists (claim a larger noise component for democracy, not noise) but is not in the text. The same noise framing recurs in §2 ("noisily pushes outputs around") and §3.2 against literatures the paper itself cites as documenting systematic divergence.

4.3 The reputation arm's gloss — downgraded to medium: apparent non sequitur with an available reconstruction

Response 3: off-equilibrium punishments are "electoral, constitutional, or revolutionary — that is, franchise-based or franchise-backed." Read flatly, the gloss is wrong for the third disjunct: revolutionary punishment is the paradigm non-franchise mechanism. But a reconstruction runs through the paper's own §2 thesis: revolutionary threat alone cannot sustain commitments (its power is transient — that is the commitment problem itself), so revolution-backed reputation is effective only insofar as it converts into franchise power. On that reading the sentence compresses the paper's core argument rather than contradicting it. As written it invites the non sequitur reading and the reconstruction is nowhere supplied — a compression defect, medium severity, rather than a false claim.

5. Literal defects, and the reclassified "false" candidates

This section was re-sorted in v2.2 after the author's second challenge. The census's first pass graded eight or nine sentences "flatly false"; steelmanned re-examination — reconstructing the charitable reading, checking the cited authors' actual corpora, and honoring citation conventions — sustained only a small core and reclassified the rest. The movement itself is reported because it is evidence (see §8).

5.1 Stands: literal defects a reader can verify mechanically

– Response 2's cross-reference to "Appendix A, 'Remarks' (c)" — the published Remarks contain no lettered items (a stale label from an earlier draft structure).

– Response 4(a): "the same studies mostly point the other way" — the findings then given are from Bartels and Kuziemko et al., not from the Althaus and Caplan studies the phrase refers back to. "The same literature" was meant.
– Response 2: "([41–43] develops a version of the argument in the epistocratic setting)" — references 42 and 43 are Shapiro's Courts (1981) and Wilson's Bureaucracy (1989); at most [41] fits the description. (Caveat: loose range-citation is a common convention; the singular verb makes this one worse than convention excuses.)
– §2: "Between the endpoints, higher χ means less redistribution" — as a description of AR's model, inconsistent with the paper's own appendix, which correctly reports AR's $\tau(\chi)$ flat from $\chi = 1/2$; true of the paper's linear normalization, which §2 does not mention.

5.2 Reclassified on steelmanning: misleading or contestable, not false

– **The Guerrero filing (§5 header):** "Democratic equality (Christiano, Kolodny, Viehoff, Guerrero; Pettit). … Guerrero's [65] lottocracy and Landemore's open democracy belong to this register as constructive alternatives." A filing error, confirmed as such by the view's owner and conceded in the thread — but never a false sentence: the sentence itself is ambiguous (as the author argued publicly), and the error lives in the header's proponent-implicature.
– **The "demographic line [11, 12]" (Introduction):** "The demographic line [11, 12] argues that competence criteria track privilege along dimensions that implicate broader equality concerns." The sentence's content is a true description of the demographic objection; the defect is citation placement under the paragraph's proponent convention, with the canonical sources (Estlund; Brennan 2018, the paper's own [26]) absent. The steelman partially holds: Christiano's Constitution of Equality argues against rule-of-the-wise on adjacent grounds — the cognitive-bias fact of judgment implies a competence-selected subset would advance its own interests, violating public equality — so [11] has a defensible basis; [12] is the weaker fit. Misleading-by-structure, not false.
– **López-Guerra (Response 3):** "López-Guerra [47] proposes an enfranchisement lottery: a randomly selected fraction of the citizenry votes each election. Without competence filtering, the expected decisive voter equals the population median, expected $\chi = 0$, and the commitment argument has nothing to say." The filtering clause is literally true of the selection stage — filter-free sortition is the point of his design. The defect is material incompleteness plus a questionable inference: the omitted competence-building stage (quoted in full by Brennan at p. 214, two pages before the pages the paper cites) plausibly installs the information–preference gradient by the paper's own Simulated-Oracle logic, undermining "expected $\chi = 0$, and the commitment argument has nothing to say."
– **Central banks (Response 1):** "they choose instruments, not targets" tracks the textbook goal-dependence/instrument-independence distinction (Congress legislates the dual mandate; the FOMC's 2% is an operationalization of "stable prices"). The initial "false" verdict depended on classifying numerical-target-setting as goal-setting — a contested reading. The deeper point survives as a medium argumentative item: mandate-boundedness does not neutralize distributional incidence.

– **"The debate Brennan started" (§1):** Estlund coined "epistocracy" and organized the earlier debate, but "the debate [over Brennan's proposals]" is a defensible colloquial reading. Loose, not false.

6. The medium findings, by theme — with the sentence at issue in each case

6.1 Evidence-status drift (the most consequential theme)

Claims introduced with careful hedges are restated later at higher strength. This matters doubly because qualifying the magnitude claims was the round-3 editor's explicit condition.

> *"Both ranges are illustrative rather than empirically established: the calculation that follows shows what such values would imply, not that they obtain." (Response 4(a))*

Later restatements: "…fiscal shortfalls in the calibrated range, if they occurred…" (4(c)); "Granting this empirical premise—granted throughout, and sized in (a)…" (4(d)); "Plausible values for them would make the cost substantial" (4(e)). "Calibrated," "sized," and "plausible" are statuses the ranges were expressly denied at introduction; plausibility is never argued.

> *"It bites on epistocracies whose competence filters correlate with the redistributive-preference gradient the empirical literature documents…" (Response 2, summary)*

Against the paper's own careful version: "the premise the analysis actually needs—$\tau_C < \tau_p$ on net redistribution—functions in this paper as the epistocrat's own empirical premise, granted for the argument" (§3.2). The summary converts the epistocrat's asserted extension into a documented fact.

> *"Even if tests are pristine, certifications honest, and the council demographically representative on every dimension Bagg tracks, the commitment problem arises from the council's institutional role…" (§5, Bagg contrast)*

Categorical at the paper's central novelty moment; the body's own statements are conditional (an orthogonal filter yields $\chi \approx 0$ and silences the argument — Response 2).

> *"…restricted suffrage, plural voting, and the filtered enfranchisement lottery—each of which installs a competence-weighted redistributive authority that maps to the same AR reduction (Appendix A 'Remarks')." (§6)*

"Installs" is categorical where the appendix's treatment is conditional throughout ("scales with… the preference gap"; "an orthogonal filter gives $\chi = 0$").

6.2 Internal tensions (quotable pairs)

> *"The epistocracy debate has run almost entirely in the epistemic register." (§6) —vs— "Four broad lines have organized the debate so far." (§1)*

> *"In a polity where prosperity has dissolved them, the AR mechanism does not apply at all, and the paper is silent by design." (Response 4(d)) —vs— "The argument's bite on each polity is proportional to its score." (§3.3)*

"Edge cases confirm the diagnostic." (Response 1) —vs— "This is a testable implication—one readers worried about juristocratic tendencies in contemporary constitutionalism can check against their own judgments." (same paragraph)

"The fourth—the most consequential—questions the magnitude of the problem…" (§4 opening) —vs— "…it is the response with the most purchase on the commitment argument." (Response 2)

Each pair is quotable side by side; each has a reconciling reading (register versus line-count; binary scope versus graded bite; consilience versus test; ledger-consequence versus mechanism-purchase) that the text does not supply.

6.3 Unsupported assertions doing argumentative work

"…as the council vetoes redistributive legislation, its composition shifts over time toward members whose preferences would justify past vetoes, amplifying rather than suppressing the gradient." (Response 2; asserted again in Response 4(d))

No citation and no mechanism: Brennan's exam-gated council has no appointing principal to run the feedback loop, and §3.1's drift citations document scope expansion, not composition feedback.

"The contemporary United States scores high on distributional-grievance salience and moderate on revolutionary-threat capacity; European social democracies score lower on both; many Latin American middle-income democracies score high on all three." (§3.3)

Uncited country-level rankings; "European social democracies" is load-bearing and ambiguous — read as Europe generally, French strike intensity breaks the mobilization claim.

"…a council veto of passed legislation is a discrete, public, datable act with an identifiable institutional author—a focal event of the kind that lowers the coordination cost of mass mobilization." (Response 4(b))

Uncited and unconditional, where §3.3's own transparency caveat treats attributability as an open parameter.

6.4 Response 3's arm-level defects (beyond §4.3)

"…constitutional courts, central banks, independent regulatory agencies, supermajority rules. If the commitment argument condemns the veto, shouldn't it also condemn these?" (Response 1)

The diagnostic canvass that follows clears courts, banks, and agencies — supermajority rules are never mentioned again, and by §3.4's own grounds-free criterion they are the instance most likely condemned.

"They presuppose the franchise upstream…" and "…an entrenchment adopted under the veto regime would itself be subject to council review on competence grounds…" (Response 3)

The veto regime preserves the franchise (§3.1), so presupposing the franchise disqualifies nothing by itself; the arm's force rests on the council-reviews-entrenchments claim, a power found nowhere in Brennan and defended only implicitly via drift.

> *"…international integration binds only when it aggregates democratic authorization that the unilaterally-acting veto regime does not supply." (Response 3)*

Pattern-1 logic filed under the pattern-3 label, and false as a generalization — treaties, BITs, and IMF conditionality bind non-democracies through interstate and market enforcement.

> *"…fiscal councils produce forecasts and advice rather than binding redistributive authority…" (Response 3)*

Exceptionless where at least one exception exists (Hungary's Fiscal Council holds a genuine budget veto).

6.5 Footnote 4 (independently flagged by two readers)

> *"Bismarck's welfare state was built to pre-empt socialist mobilization and was visibly inadequate to its commitment function by the Weimar collapse."*

The probative failure of a non-democratic commitment device is 1918 — the Kaiserreich falling to the mobilization the welfare state was built to pre-empt; "the Weimar collapse" (1930–33) names a democratic regime failing rightward, fifteen years later.

> *"Singapore operates on a distinctive settlement (PAP dominance plus high-end technocracy plus Tiebout-adjacent exit for mobile capital) that does not generalize. None is a case of durable commitment to an otherwise-revolutionary poor coalition under sustained non-democratic governance."*

"Does not generalize" concedes case status rather than denying it; if Singapore counts as non-democratic, the sixty-year HDB/CPF settlement is arguably the durable-commitment counterexample that Response 3's closing ("I have not found one") says was not found. This exposure is live: Hughes has been reading AR on Singapore in the thread.

6.6 Exegesis and filing (beyond §§4–5)

> *"Three of these lines focus on the character of the electorate—what voters know, how they decide, or who counts." (§1)*

The three must include the abuse line, which targets elite gaming of selection institutions — as the paper's own §5 says of Bagg ("structural public-choice methodology"). The electorate/institution contrast motivating the novelty claim is overdrawn.

> *"…its justification has to be acceptable to all reasonable views—not in the sense that everyone has to endorse it, but in the sense that no reasonable view can be asked to accept a justification that singles it out as cognitively inferior." (§5, on Estlund)*

Defines the qualified acceptability requirement by its epistocracy-specific application; the requirement is general, and "reasonable" substitutes for Estlund's term of art "qualified."

> *"That transfer is what historically averted revolution in societies where inequality was steep enough for the poor to threaten one." (§1)*

A categorical historical claim the paper itself later retreats from ("the normative argument does not require that AR was historically dominant… only that the mechanism is a live possibility now").

> *"The same reduction applies to Brennan's three other epistocratic regimes. … the pivotal (decisive) voter shifts continuously toward higher-income voters as the franchise cut tightens…" (Appendix, Remarks)*

In the stated two-class model the pivotal voter moves in a jump, not continuously; the continuity claim needs the within-class heterogeneity the scope remark suppresses. "The same reduction applies" asserts theorem-strength coverage the remarks do not derive.

> *"…an unfiltered lottery gives expected $\chi = 0$, though realized χ fluctuates sample-to-sample…" (Appendix, Remarks)*

The conjunction needs an income continuum: with a continuum electorate there is no fluctuation; with a finite assembly in the two-class model, realized $\chi \in \{0,1\}$ and expected χ is small but positive. The substantive conclusion survives (the wedge vanishes with assembly size).

6.7 Framing and quantity slips

> *"…arguing that the cost's size is proportional to the size of the preference divergence between an epistocratic council and the median voter…" (§1)*

Drops the veto rate v — one of the model's own two governing quantities ($\chi_Ev = v(\tau_p - \tau_C)/\tau_p$); at $v \to 0$ the cost vanishes at any divergence.

> *"An institution that changes who decides on redistribution thereby changes commitment, even if it leaves the electorate and its epistemic qualities exactly where it found them." (§2, flagged as "the pivot of what follows")*

Needs the suppressed divergence premise: by the paper's own model, the change is zero when $\tau_C = \tau_p$ — a case §4 itself treats as live.

> *"The promise to redistribute becomes credible precisely because reversing it requires revoking the franchise itself…" (§2; also in the abstract)*

Necessity overstated: the paper's own thesis is that a veto partially reverses redistributive authority with the franchise intact. The needed claim is that reversal requires undoing democratic institutions, which is costly.

> *"…the expected redistribution under the veto is about 6% below the democratic benchmark…" (§3.3)*

Rate–transfer slippage: $\chi_Ev = 0.06$ puts the expected tax rate 6% below τ_p; realized transfers $(\tau - C(\tau))\bar{y}$ fall by a different (typically smaller) percentage under convex deadweight loss. The same slippage recurs in 4(a)'s "2 percentage points of national income."

> *"…about the size of a major austerity package, except repeated every period rather than enacted once." (Response 4(a))*

Describes the expectation as a steady flow; the model's realized pattern is episodic — no shortfall in most periods, a much larger cut in vetoed periods. An enacted austerity package also persists, so the contrast is loaded.

> *Header: "Do tax rates affect political unrest?" — followed by: "…with unrest increasing in the size of the cuts, and with expenditure cuts far more potent than tax changes." (Response 4(c))*

The header and topic sentence are contradicted by the half of their own lead citation the paragraph quotes; the intro's phrasing ("changes in taxation and spending") was the accurate one.

> *"But this would be surprising, given how much of policy concerns redistribution." (Response 4(a))*

A volume premise answering a divergence question: under the convergence hypothesis being addressed ($\tau_C \approx \tau_p$), the council has no redistributive disagreement to act on regardless of how much legislation is redistributive. This sentence is also the only guard on the domain-restriction escape from the self-correction dilemma.

> *"…there is some reason to rank the veto's problem as the more serious: its shortfall adds to democracy's own, and it removes a corrective channel…" (Response 4(e))*

Additivity supports "democracy-plus-veto is worse than democracy alone," not "the increment exceeds the baseline" — which needs exactly the magnitudes conceded unmeasured two sentences earlier. The channel-removal ground is the legitimate one.

> *"Acemoglu, Naidu, Restrepo, and Robinson [60] find that democracy itself causally raises long-run growth—which cuts against the technocratic argument that experts produce better policy…" (Response 4(d))*

The finding compares democracies to autocracies; it contains no expert-checked democracies, and an epistocratic-veto polity would plausibly be coded democratic in their data. Accurate citation, overextended inference.

> *"…the backsliding literatures cited in Section 3.3 document distributional grievance persisting through periods of absolute growth, particularly when the growth is perceived as concentrated among elites." (Response 4(d))*

§3.3's actual characterization of those literatures is "polarization, norm erosion, gatekeeping failure"; the grievance-through-growth content appears only here, hung on a cross-reference that does not contain it.

7. What checked clean

For the record, because it is the other half of the evidence: the entire formal chain re-derives independently; every empirical study in Response 4 is used accurately at claim level and hedged, with the qualification literatures cited alongside (Gilens with Branham–Soroka–Wlezien; minimum wage as contested; "this evidence should be read with care"); the Munger attribution, the Brennan p. 216 thwart-only language, the pp. 215–220 and pp. 223–224 page cites as locations, the Hobbits/Hooligans/Vulcans compressions, the Waldron gloss, the Estlund

attribution (as distinct from the QAR gloss), the Gibbons 2025 gloss (verified verbatim against the published PDF), the Ansell–Samuels, Ziblatt, Albertus–Menaldo, and Haggard–Kaufman characterizations, both endnote-1 glosses, the North–Weingast use, the cross-reference structure (with the one lettered-remark exception), and the live AI-report link. The July pre-publication source checks separately verified the AR χ pages and equation, the Brennan 2018 p. 59 ranking, and the full bibliography — and that process caught and fixed, before publication, a worse error than any that survived (the original misreading of Bartels's direction).

8. So how strong is the objection?

The global version fails. "No way this ends up being a net-positive for truth and knowledge" as applied to this paper — the claim that it is unreliable through and through — is not supported by the census. The spine holds: the model is sound, the theorem proves, the empirical evidence is used accurately, no source is fabricated, and the three attribution errors that started the dispute are one-line filings that touch no argument.

The induction is vindicated in count, at sentence level. Guerrero's implicit inference — one error found by a targeted check implies many findable problems — is correct as a prediction about this paper under adversarial sentence-level reading: roughly one flagged sentence in ten, about sixty distinct items, and a hostile expert who picks any framing, summary, or taxonomy sentence at random has real odds of finding something quotable. Anyone running his experiment on a different passage would plausibly have replicated his experience.

Severity is the crux, and it cuts against him. Of the sixty items, the literal mechanical defects number two to four in 579 sentences — well under one percent — and all are local. The mass of the census is misleading filing, overclaim, label drift, and internal tension: the endemic failure modes of published philosophy, the things referees and critics write about ordinary human papers. And the census exhibited a repeating dynamic worth naming: "false" verdicts — whether issued by the thread's critics against §3.2, or by this audit's own first passes against §3.1, the demographic line, and the central-bank sentence — tended to downgrade into "misleading or contestable" under steelmanned re-examination. Adversarial reading over-produces falsity verdicts, whoever performs it. The census found no instance of the failure mode the "AI slop" frame predicts — confabulated sources, invented findings, arguments that dissolve on inspection. The formal core, the part least checkable by casual readers and most checkable by careful ones, is the cleanest stratum in the paper.

The baseline is unknown, and honesty requires saying so. No human-written philosophy paper has ever been audited at this resolution, so whether sixty items in 579 sentences is high, normal, or low for the genre is an open empirical question. Two observed facts cut against an AI-specific reading: several flagged sentences are documented author additions from the revision rounds (the "surprising, given how much of policy concerns redistribution" line; the Response 4(c) retitle; the intro's proportionality summary), and the process caught its worst error pre-publication through exactly the verification discipline critics say is missing.

The uncomfortable finding points at everyone. The census's clearest source-contradiction — a proposal described as "without competence filtering" when the cited book quotes that proposal's competence-building process in full two pages earlier — survived the editor, two

referees, a 200-comment thread including formally trained critics reading with hostile intent, and the author's own passes. And the census's own first pass over-graded its biggest finding, corrected only under the author's challenge: even the auditors need auditing. Sentence-level verification is rare and hard enough that neither humans nor machines reliably do it unprompted, which strengthens Guerrero's structural worry — who will check the coming volume of AI-assisted work? — and simultaneously weakens his implied baseline, since nothing about the misses depended on the drafting being AI. The honest summary of the census: this paper's reliability profile is that of a formally solid, aggressively hedged, carefully sourced paper with an ordinary-to-unknown density of prose-level soft spots — and the only reason its profile is known at this resolution, unlike any of its neighbors in the literature, is that it was audited, and re-audited under challenge, by the same class of system that wrote it.