LLM VERIFIERS ARE WRONG TOGETHER: CONFIDENT CONSENSUS ERRORS SET THE FLOOR FOR FAILURE PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

LLM verifiers decide which generated claims reach users, and their confidence decides which verdicts a human rechecks. How much of a verifier's failure that confidence can expose has not been measured. We introduce a *detectability decomposition*: for any set of failure signals and any review budget, it measures the share of verifier errors that no signal in the set ranks for review. On 22,189 claims from 11 verification datasets, with three further families held out of distribution, the verdict-token margin comes free with the verdict and flags 61.8% of GPT-4o's errors at a 20% review budget. A self-evaluation signal recovers only 37.2% of the errors the margin misses, pooling four signals lowers the missed share from 38.2% to 31.6%, and their union does better only by reviewing 37.0% of claims. The errors that remain are not unstable verdicts. At the same confidence, a verifier from a different vendor is wrong on them 6.0–36.1 times as often as on correct claims, across a nine-verifier panel from five vendors the median cross-vendor error co-occurrence is 7.75 times the independent rate, and flagging every claim on which any panel member dissents recovers 8.8% of the core. A blind re-annotation of 120 core errors shows what this consensus is made of: 59.2% are benchmark-label errors that the verifiers rejected together and correctly, 40.8% are confirmed verifier errors, and the cross-vendor gap survives correction for this label noise (4.3–30.2×). LLM verifiers are right together against bad labels and wrong together on hard claims, and both are invisible to confidence-based review. An intervention on GPT-4o points to one mechanism: once accepted edits that a blind audit found still supported are removed, meaning-violating edits that keep the document's vocabulary are confidently accepted 4.3pp more often than those that change it (CI [2.7, 5.9]). The decomposition prices what can be guaranteed: a 5% residual-risk target is reachable at a 13.9% budget, a finite-sample document-level guarantee costs 52.4% (25.1pp for the loss definition, 13.4pp for finite-sample slack), and thresholds certified before two unannounced replacements of the serving snapshot met their target in 100 and 94 of 100 recalibrations after them.

1 INTRODUCTION

An LLM verifier is a gate. It reads a claim and a source document and returns a verdict: supported, unsupported, or unknown. Factuality metrics count that verdict (Min et al., 2023; Tang et al., 2024), evaluation pipelines rank models by it (Zheng et al., 2023), training pipelines reward it (Lightman et al., 2024), and a wrong verdict passes through the gate unless something marks it for human review; in deployed systems that something is the verifier's own confidence. Selective prediction formalizes the practice: rank the verdicts by a confidence signal, send the least confident fraction B to a reviewer, and accept the rest (Chow, 1970; El-Yaniv & Wiener, 2010; Geifman & El-Yaniv, 2017).

Work on LLM uncertainty has supplied many such signals: verbalized confidence (Lin et al., 2022; Tian et al., 2023; Xiong et al., 2024), self-evaluation with a second call (Kadavath et al., 2022), disagreement among repeated samples (Kuhn et al., 2023; Farquhar et al., 2024; Manakul et al., 2023), the probabilities of the first output tokens (Zhao et al., 2024; Vashurin et al., 2025), and trained probes (Kapoor et al., 2024). Each is reported by how well it ranks errors, usually as an

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

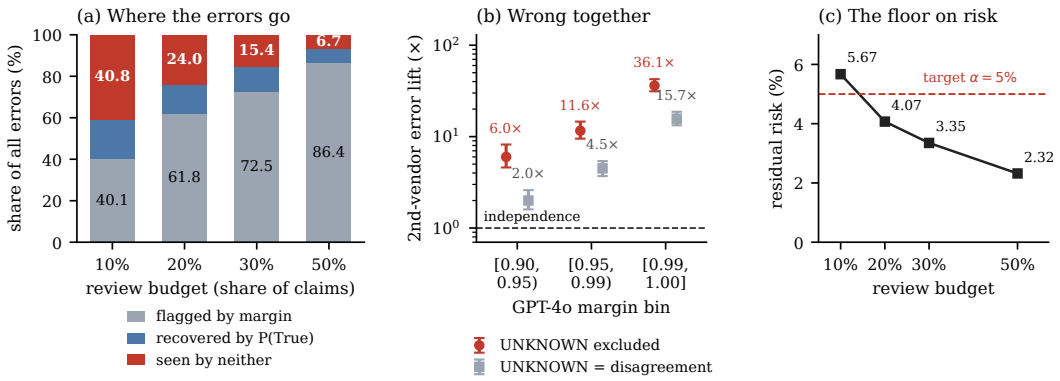

Figure 1: **LLM verifiers are wrong together, and this sets a floor on review. (a)** All 1,725 errors of a GPT-4o verifier, split by which signal flags them at each review budget. At a 20% budget the verdict-token margin flags 61.8%, a self-evaluation signal recovers 37.2% of the rest, and 24.0% of all errors are seen by neither. **(b)** Within the same margin bin, a verifier from a second vendor is wrong on the errors that GPT-4o misses 6.0–36.1× as often as on the claims GPT-4o gets right (10% budget; 2.0–15.7× when its abstentions count as disagreement), and the gap widens with confidence. **(c)** The missed errors fix the residual risk among accepted claims (4.07% at 20%, against benchmark labels), and no calibration of the margin can certify a level below this curve. Better signals and second opinions reach the same limit: the remaining errors are *confident consensus errors*.

AUROC, which describes the errors a signal can see but not how many it cannot see, whether another signal would see them, or what residual risk remains after a budget B; a deployer needs these three quantities, and for LLM verifiers they have not been measured.

We therefore ask a different question: is the bottleneck in predicting verifier failure the quality of the signal, or the information that all signals share? Software reliability met this question four decades ago. N-version programming ran independently written programs in parallel on the premise that their failures would be independent, and Knight & Leveson (1986) found that the versions failed on the same inputs far more often than independence predicts. LLMs show the precondition for the same outcome: models from different developers produce similar outputs (Jiang et al., 2025), models that both err often choose the same wrong answer (Kim et al., 2025), and the similarity grows with capability and biases LLM judges (Goel et al., 2025), as theories of algorithmic monoculture anticipate (Kleinberg & Raghavan, 2021; Bommasani et al., 2022). If verifier errors are common-mode in this sense, then a sharper signal, a second sample, and a second verifier would all miss the same claims.

Approach. We introduce the *detectability decomposition* to measure this directly. For a signal set $\mathcal{S}$ and a review budget B, it splits a verifier's errors into three parts: *escalated* errors, where the verifier itself answers "unknown" (8.81% of claims); *visible* errors, which some signal in $\mathcal{S}$ ranks into the top B% for review; and the *core*, $\mathrm{core}_{\mathcal{S}}(B)$, which no signal in $\mathcal{S}$ ranks for review. We apply it to GPT-4o verifying 22,189 claims from 1,757 documents across 11 LLM-AggreFact benchmarks (Tang et al., 2024; 2023), with FEVER (Thorne et al., 2018), AVeriTeC (Schlichtkrull et al., 2023), and FELM (Chen et al., 2023) held out of distribution; a second vendor cross-checks every verdict, a nine-verifier panel from five vendors re-verifies a 5,000-claim sample, and the study uses more than 300,000 verifier calls. Three signal families enter $\mathcal{S}$: distributional (the verdict-token margin), self-evaluation (P(True) and verbalized confidence), and behavioral (vote entropy and cross-model agreement). We audit the evaluation before we measure: three pitfalls inflate a naive failure detector from 0.67–0.72 AUROC to 0.93 (cf. Jaeger et al., 2023), and because about 16% of fact-checking benchmark examples are mislabeled or ambiguous (Seo et al., 2025; Northcutt et al., 2021; Pavlick & Kwiatkowski, 2019) we re-annotate a stratified sample of the core, blind to labels and verdicts.

Findings. Figure 1 summarizes the results. The margin is a by-product of the verification call, ranks errors at AUROC 0.801, and flags 61.8% of them at a 20% budget; verbalized confidence, the

common practice, reaches 0.651 because 91.2% of its responses are exactly 100. Of the errors the margin misses, $P(\text{True})$ recovers 37.2%, behavioral signals $33.8 \pm 0.5\%$, and a 50% budget raises the recovery only to 51.1%; one score combining all four signals lowers the missed share from 38.2% to 31.6%, and their union reaches 18.4% only by reviewing 37.0% of claims: the core shrinks with review volume, not with the number of signals. It consists of *confident consensus errors*, and the consensus has two faces. Within the same margin bin, a verifier from a different vendor is wrong on a missed error 6.0–36.1$\times$ as often as on a correct claim, across a nine-verifier panel the median cross-vendor error co-occurrence is 7.75$\times$ the independent rate, and flagging every claim on which any of eight other verifiers dissents recovers 8.8% of the core, so a second model's verdict does not recover them (Tan et al. (2025) probe a second model's hidden states, untested here). A blind re-annotation of 120 core errors then shows what the consensus is made of: in 59.2% the adjudicated label sides with the verifiers, so the models rejected a wrong benchmark label together, and in 40.8% (CI [32.3, 49.6]) they are wrong together; the cross-vendor gap survives correction for this label noise (4.3–30.2$\times$). Both faces are invisible to confidence-based review. An intervention on GPT-4o points to one mechanism: meaning-violating edits that keep the document's vocabulary are confidently accepted more often than edits that change it (+4.3pp, CI [+2.7, +5.9], after removing accepted edits a blind audit found still supported), while reordering alone moves acceptance by -2.0pp; the same audit finds that most generated violations the verifier accepted were still supported by the document. The verifier's confidence tracks surface support where it lets wrong claims through; whether other verifiers share this is untested.

What the map supports. A rule with budget B accepts $1 - B$ of the claims, so its residual risk equals the missed errors over the accepted claims, and no calibration of that score can certify a level below this floor. For the margin a 5% target becomes reachable at 13.9% of review; a guarantee that holds with probability $1 - \delta$ under document-level dependence (Angelopoulos et al., 2025) costs 52.4%, and the gap decomposes into 25.1pp for the loss definition and 13.4pp for finite-sample slack. Restoring omitted evidence turns silent errors into explicit abstentions (Section 5.2), and a verifier about 17 times cheaper predicts GPT-4o's errors nearly as well as GPT-4o's own margin does (Section 5.3). Guarantees assume that calibration and deployment share a distribution, which deployment can break (Tibshirani et al., 2019; Barber et al., 2023; Chen et al., 2024): the provider replaced the serving snapshots behind the same model name twice during the study, one replacement changed verdict correctness on 1,194 of 22,297 paired claims, and thresholds certified before each met their target in 100 and 94 of 100 recalibrations after it (Section 6), a case study of two events.

Contributions. (1) The detectability of verifier errors is a property of the information a signal set carries, and $\text{core}_{\mathcal{S}}(B)$ measures it with controls for review volume and label noise. (2) The errors outside that information are confident consensus errors, shared across samples, vendors, sizes, and a reasoning model: a majority are benchmark errors rejected together, the rest are genuine errors made together, and an intervention on the primary verifier, audited blind, shows that its confidence depends on surface support. (3) The core determines which risk guarantees are attainable and at what budget, the price of a finite-sample guarantee is set by document-level dependence, and a guarantee priced this way held through two silent model updates. The consensus finding is observational; the surface-support and evidence-access results are interventional; the drift result is a two-event case study; the primary verifier is one model family (Section 8).

2 Related Work

Confidence and selective prediction. LLM confidence signals include verbalized confidence (Lin et al., 2022; Tian et al., 2023; Xiong et al., 2024), self-evaluation (Kadavath et al., 2022), sampling-based entropy (Kuhn et al., 2023; Farquhar et al., 2024; Manakul et al., 2023), and output-token distributions (Zhao et al., 2024; Vashurin et al., 2025); selective classification supplies the review rule and the risk–coverage curve (Chow, 1970; El-Yaniv & Wiener, 2010; Geifman & El-Yaniv, 2017). This work compares signals by ranking quality; we hold the signals fixed and measure the errors that none of them, and no combination of them, ranks.

Correlated failures. Independently developed software versions fail on the same inputs (Knight & Leveson, 1986), and LLMs from different developers are homogeneous in generation (Jiang et al., 2025) and correlated in error (Kim et al., 2025; Goel et al., 2025). Tan et al. (2025) show that self-

consistent errors in question answering defeat four families of detectors and recover them with a probe into a second model. We show that verification errors are consistent *across* models at matched confidence, across five vendors and a reasoning model, so a second model's verdict does not recover them, although we do not test their probe; we separate the part of the consensus that is right against a bad label from the part that is wrong; and we connect the size of the shared core to the risk a deployment can certify. High lexical overlap with a changed meaning misleads fine-tuned inference models (McCoy et al., 2019; Zhang et al., 2019; Schuster et al., 2019; Gururangan et al., 2018; Geirhos et al., 2020); we test the same mechanism in LLM verifiers by intervention.

Verifiers, labels, and guarantees. Benchmarks of LLM verifiers measure accuracy (Tang et al., 2024; Seo et al., 2025); we measure how predictable the remaining failures are and, because confident disagreement with a gold label is also the signature of a label error (Northcutt et al., 2021; Pavlick & Kwiatkowski, 2019), we audit the labels of the core. Conformal methods guarantee properties of LLM generations (Mohri & Hashimoto, 2024; Cherian et al., 2024); we apply risk control to a verifier's accepted set, separate the in-expectation guarantee of conformal risk control (Angelopoulos et al., 2024; Bates et al., 2021) from the finite-sample guarantee of Learn-then-Test (Angelopoulos et al., 2025), and price the gap under silent model updates (Chen et al., 2024). Section A extends this discussion.

3 Setup and the Detectability Decomposition

3.1 Task, data, and verifiers

A verifier receives a claim c and a source document d and returns a verdict $v \in \{\text{TRUE}, \text{FALSE}, \text{UNKNOWN}\}$ as a single structured token; the verdict is an *error* when it differs from the gold label. The in-domain pool is the 11 subsets of LLM-AggreFact (Tang et al., 2024), deduplicated at the document level (1,757 documents, 22,189 claims); the primary verifier is GPT-4o, run three times per claim at temperature 0.1, and 1,954 claims (8.81%) are escalated by the verifier itself (majority UNKNOWN or a three-way tie), leaving 20,235 committed verdicts and 1,725 errors (base rate 8.5%). Three families are out of distribution and never used for fitting: FELM (3,731 claims), AVeriTeC (3,567; 2,651 in the guarantee analysis), and FEVER (12,056). Gemini 2.5 Flash cross-checks every claim; GPT-4o-mini and a panel of six further verifiers from five vendors and two open-weight families, including a reasoning model, re-verify a stratified 5,000-claim sample with the same prompt at temperature 1.0 (Section 4.2). Every response is stored with the provider's `system_fingerprint`, timestamp, and upstream provider, and a dataset whose fingerprint set changes is re-collected in full; the signal comparison in Section C uses the pre-deduplication pool (20,241 committed verdicts, 1,731 errors). Details are in Section B.

3.2 Signals

Distributional signals read the verdict-token distribution of the verification call: the *margin* is the probability gap between the top two verdict tokens and costs no extra call. *Self-evaluation* signals cost one more call: $P(\text{True})$ asks whether the verdict is correct and reads the probability of "True"; verbalized confidence asks for a number from 0 to 100. *Behavioral* signals need repeated or additional verifiers: vote entropy over the three runs, cross-model agreement, and reasoning similarity (four features). A 17-dimensional vector (17D) concatenates all features; learned scores are fitted out of fold with document-grouped 5-fold splits over 10 seeds. Details are in Section C.

3.3 The detectability decomposition

Let E be the set of errors among the N committed claims and $\mathcal{S}$ a set of risk scores $s : (c, d, v) \mapsto \mathbb{R}$. For a budget $B \in (0, 1)$, let $T_s(B)$ be the $\lceil BN \rceil$ claims that s ranks as most at risk, with ties at the cut broken deterministically (a tied block that straddles the cut is not reviewed).

$$\text{escalated: } A = \{\text{claims whose majority verdict is UNKNOWN or tied}\}, \tag{1}$$

$$\text{visible: } \text{vis}_{\mathcal{S}}(B) = E \cap \bigcup_{s \in \mathcal{S}} T_s(B), \qquad \text{core: } \text{core}_{\mathcal{S}}(B) = |E \setminus \text{vis}_{\mathcal{S}}(B)|/|E|. \tag{2}$$

Escalated claims go to a reviewer unconditionally and are removed first (Section 3.4). The union in Equation (2) lets every signal spend its own budget, so it reviews $v_{\mathcal{S}}(B) = |\bigcup_s T_s(B)|/N$ of

the claims, up to $|\mathcal{S}|B$. We therefore report the union with the share it reviews and two controls: the *equal-budget* variant, in which one combined score spends B, and the *equal-volume* variant, in which the reference signal alone reviews $v_{\mathcal{S}}(B)$; their difference isolates the gain from pooling signals rather than from reviewing more. The conditional form $\mathrm{core}(B \mid s_0)$ is the share of the errors missed by a reference signal s_0 that the rest of $\mathcal{S}$ also misses; all are curves in B with document-level bootstrap intervals; differences between variants are paired within resamples (Algorithm 1, Section D).

The core prices any review rule. A rule that reviews T with $|T| = BN$ and accepts the rest has residual risk

$$R(B) = |E \setminus T| \,/\, ((1-B)N), \tag{3}$$

under the idealization that review corrects every reviewed error (imperfect review scales the floor up), so a calibration procedure that ranks by a score in $\mathcal{S}$ cannot certify a level below $|E|\,\mathrm{core}_{\mathcal{S}}(B)/((1-v_{\mathcal{S}}(B))N)$; a new score that combines signals has its own floor, which can be lower. Equation (3) is an accounting identity that turns the core curve into the attainable region of the guarantee in Section 5.1; it is measured against benchmark labels, and the audit of Section 4.3 shows that label errors bias it in both directions.

3.4 Evaluation protocol and audit

Every interval is a document-level bootstrap, every split is grouped by document, and directional hypotheses are committed to the repository before each experiment runs; we report every prediction that reversed (Section M) and mark as exploratory any comparison specified after seeing data. A naive logistic detector with an abstention indicator reaches AUROC 0.93, and three pitfalls account for the gap to 0.67–0.72: *(i)* an UNKNOWN verdict is an error by definition, and the abstention flag alone explains 95% of the naive score, so abstentions are escalated and excluded from features; *(ii)* on the pre-deduplication pipeline, OR-aggregation over a document's claims moves the base rate from 17.5% to 75.6% and the apparent blind spot from 4.5% to 9.9%, so we evaluate at the claim level; *(iii)* claim count, length, and origin predict errors at 0.69–0.73 in-domain but 0.49–0.55 out of distribution, so every result carries a confound-only baseline and an OOD test. Details are in Section E.

4 The Core: Confident Consensus Errors

4.1 How large is the core?

The margin of a single call carries the signal (AUROC 0.801 for the 3-run mean; a single run reaches 0.799 in a replicate run), $P(\text{True})$ exceeds it by $+0.024$ [0.009, 0.038] at the cost of a second call, verbalized confidence is degenerate (91.2% of responses equal 100), and behavioral features reach 0.697; out of distribution, adding behavioral features to the margin *lowers* AUROC, so the margin is the only signal whose transfer we verify (0.707 / 0.835 / 0.740 on FELM / AVeriTeC / FEVER; Section C). Table 1 reports the decomposition at four budgets. Of the errors the margin misses, $P(\text{True})$ recovers 31.9–51.1% as the budget rises from 10% to 50%, behavioral signals $33.8 \pm 0.5\%$ at 20%, verbalized confidence 20.2%, and all three together 51.9% (Section D). One equal-budget score combining all four signals lowers the missed share at 20% from 38.2% to 31.6% (paired difference -6.6pp, CI $[-9.4, -3.8]$); at matched volume its advantage over the margin holds only at the smallest volume (-6.9pp at 21.4%) and reverses above it. The union of the four reaches 18.4% (CI [16.1, 21.1]) but reviews 37.0% of the claims, where the margin alone misses 21.3%; of the union's 19.8pp gain over the margin, reviewing more accounts for 16.9pp and pooling for 2.9pp (CI [0.6, 5.0]), a residual that vanishes under random tie-breaking (Section D). The core therefore shrinks with review volume, not with the number of signals.

4.2 The core is shared across verifiers

Gemini, a verifier from a different vendor, agrees with 70.6% of the wrong verdicts that the margin misses; because missed errors have high margins by construction, we compare within margin bins. At the same confidence, Gemini is wrong on a missed error $6.0\times$, $11.6\times$, and $36.1\times$ as often as on a correct claim (Figure 1b; 66.3%, 76.5%, and 85.4% against 11.0%, 6.6%, and 2.4%, every interval

Table 1: The core against the review budget B (% of all 1,725 errors unless noted; document-level 95% CI). *Conditional*: share of the margin's misses that $P(\text{True})$ also misses. *Union*: four signals each spending B, with the share of claims reviewed; *equal volume*: the margin alone reviewing that share; *equal budget*: one combined score spending B; *confirmed*: union core scaled by the audited share of genuine errors (Section 4.3).

	$B = 10\%$	$B = 20\%$	$B = 30\%$	$B = 50\%$
Missed by margin	59.9	38.2	27.5	13.6
Conditional core, $P(\text{True})$ (% of missed)	68.1	62.8	56.0	48.9
Union of four signals	30.6 [27.7, 33.3]	**18.4** [16.1, 21.1]	10.7 [8.9, 13.0]	3.1 [2.3, 4.5]
reviewed (% of claims)	21.4	37.0	53.3	79.3
Equal volume, margin alone	36.5 [33.5, 39.4]	21.3 [18.6, 24.1]	11.8 [9.8, 14.2]	3.9 [3.0, 4.9]
Equal budget, combined score	50.8 [48.3, 53.8]	**31.6** [28.7, 34.6]	23.1 [20.9, 26.4]	15.5 [13.7, 18.2]
Confirmed verifier errors only	–	7.5 [5.9, 9.1]	–	–

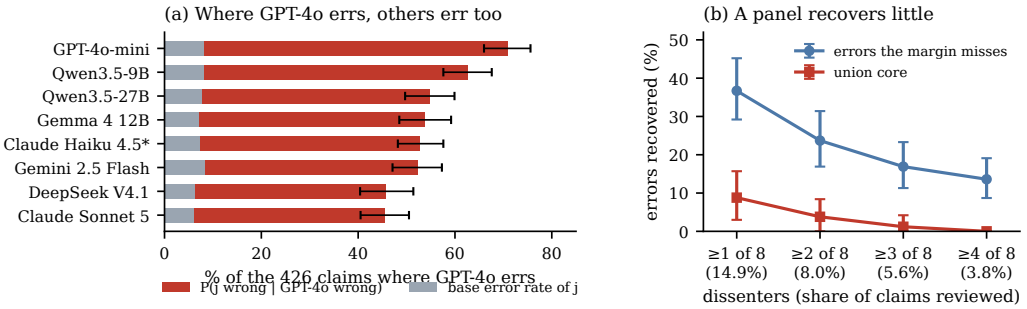

Figure 2: **Consensus survives a larger panel. (a)** On the 426 sampled claims where GPT-4o errs, the probability that each other verifier errs (red; 95% document-bootstrap CI) against its base error rate (gray); *reasoning model. **(b)** Share of GPT-4o's errors recovered by flagging claims on which at least k of eight other verifiers dissent, with the share of claims reviewed.

excluding 1), and the ratio grows with confidence. Our pre-specified prediction that agreement on missed errors would *match* agreement on correct claims is rejected in every bin: a second opinion recognizes 15–34% of what the margin misses, the same one third that a second signal recovers; whether one calls this shared difficulty or common-mode failure, the second opinion does not recover the errors.

We re-verify the 5,000-claim stratified sample with nine verifiers (GPT-4o, GPT-4o-mini, Gemini 2.5 Flash, DeepSeek V4.1 Flash, Claude Sonnet 5, Claude Haiku 4.5 with extended thinking, and open-weight Qwen3.5 9B and 27B and Gemma 4 12B; Section F). Committed verdicts are binary, so two verifiers that both err give the same wrong verdict; the informative quantity is *co-occurrence*, $P(j \text{ errs} \mid i \text{ errs})$, against j's base error rate, the value under independence. Over the 64 cross-vendor ordered pairs, co-occurrence averages 59.9% against a base rate of 7.5% (median lift 7.75×, range 5.8–10.1×, every interval excluding 1), and over the 8 same-vendor pairs 66.6% against 7.7% (median 9.01×), so sharing a vendor adds little to sharing an error. Where GPT-4o errs, GPT-4o-mini errs on 70.9% of the claims and the two Claude models on 45–53% (Figure 2a); where the reasoning model errs, Qwen3.5-27B errs on 71.3%. A panel also buys little review: at a 20% budget, flagging every claim on which at least one of the eight other verifiers dissents from GPT-4o recovers 8.8% of the union core (CI [3.0, 15.7]) while sending 14.9% of claims to review, and four or more dissenters recover none (Figure 2b).

4.3 What the consensus is made of: a label audit

Confident, cross-vendor disagreement with the gold label is also what a *label* error looks like (Northcutt et al., 2021). Five annotators re-label a stratified random sample of 120 core errors (the 20% union core), 60 flagged errors, and 30 correct verdicts, blind to gold labels, verifier verdicts, stratum, and dataset; every core item and 20 flagged items receive two independent annotations, a separate blind adjudicator resolves disagreements, and no LLM takes part (Section G). Agreement before ad-

judication is itself a result (Table 9, Section G): annotators agree on 65.8% of core items ($\kappa = 0.34$) against 85.0% of double-annotated flagged errors: the core is where the label loses resolution. After adjudication, 59.2% of the core items (CI [50.4, 67.7]) side with the verifiers against the gold label and 40.8% (CI [32.3, 49.6]) are confirmed verifier errors; no item was adjudicated ambiguous, and because strata were hidden, the gradient across strata below cannot come from annotator bias. The gold-wrong share falls from the core to flagged errors (33.3%) to correct verdicts (16.7%), the last matching the rate that Seo et al. (2025) report; removing any single annotator moves the confirmed share between 34.7% and 47.2%; and our pre-specified prediction that confirmed errors would exceed half of the core is rejected. Two consequences follow. First, the confirmed core is 7.5% of all errors (CI [5.9, 9.1]) rather than 18.4%, and the floor of Equation (3) is measured against labels that err in both directions: core label errors inflate it, while the 16.7% gold-wrong rate among correct verdicts means accepted claims hide errors the labels cannot count, so the net bias has no determined sign; both faces of the consensus are invisible to a deployer who reviews by confidence. Second, the consensus of Section 4.2 is not an artifact of label noise. Treating the audited core-wide share of label errors (59.2%) as claims on which both verifiers were right lowers the cross-vendor lift from 6.0–36.1× to 4.3–30.2×; for it to vanish, 94–99.6% of the missed-and-agreed claims would have to be label errors, against audited rates of 33.3% and 64.6% in the two audited bins, and attributing every audited label error to co-occurrence leaves the panel's cross-vendor lift at or above 4.7× (Section F).

4.4 Why are they invisible? Surface support

We pre-specified that core errors would share less vocabulary with the document than flagged errors; observation does not support this (overlap 0.581 against 0.556, $+0.025$, CI $[-0.006, +0.054]$, highest in the quartile the verifier trusts most; Section H). We therefore test the hypothesis that *verifier confidence tracks surface support rather than meaning* with an intervention on GPT-4o (Table 11, Section H). From 1,000 claims that the verifier accepts with confidence (majority TRUE, mean margin ≥ 0.95, gold TRUE, at most 25% from one dataset), a model from a different vendor generates a 2×2 of edits plus a control: meaning-violating with the document's vocabulary kept (B; eight operators with replacement values drawn from the document), meaning-violating with vocabulary reduced (C), meaning-preserving with vocabulary reduced (D), and a meaning- and vocabulary-preserving reordering (A′). Edit distances lie within a factor of 1.16 across cells, so the kind of edit is the treatment, and cell-blind validation confirms the intended label for 90–100% of A′, B, and C and 88% of D ($\kappa = 0.85$). As generated, violations that keep the vocabulary are confidently accepted 14.1% of the time against 1.2% when the vocabulary changes (+12.9pp, CI [+10.6, +15.2]), while reordering alone moves acceptance by -2.0pp (H4; margin difference 0.004). A blind audit of all 141 confidently accepted B edits finds that only 34.8% (CI [26.9, 42.0]) are genuine violations: most apparent acceptances are edits the verifier was right to accept. After removing the 92 accepted B edits the audit found still supported, with their paired C edits, the difference is +4.3pp (CI [+2.7, +5.9]): the interval excludes zero, the point estimate falls short of the pre-registered 5pp threshold, and the effect is small but present; by operator, causal edits fool the verifier most (15.0% of validated causal edits accepted, against 4.4–5.8% for the others) and negation never does (Section H). Paraphrases (D) that pass validation are accepted as often as reorderings (94% against 92%), and the invalid ones at 60%, so the verifier rejected changed-meaning paraphrases more often, H2 is unsupported, and the mechanism is one-sided. Surface support lets some wrong claims through; correct claims pass with or without it (Section H).

5 What the Map Supports

5.1 The core prices the attainable guarantee

Equation (3) predicts the residual risk of margin-only review from its missed errors: 5.67%, 4.07%, 3.35%, and 2.32% at budgets of 10%, 20%, 30%, and 50%, so a 5% target becomes reachable between the first two, at $13.9 \pm 1.8\%$ over 100 document-level splits. What a certified rule costs above the floor depends on the guarantee. *Escalate-then-rank* sends every UNKNOWN verdict to review (Stage 1) and accepts a remaining claim when its risk score is below a threshold $\hat{\lambda}$ (Stage 2). Conformal risk control (Angelopoulos et al., 2024) bounds the *expected* risk; with the 17-feature score it holds residual error at or below $\alpha = 5\%$ in expectation at a budget of $12.0 \pm 2.3\%$, below

the margin's floor because that score has its own floor and because an expectation-level guarantee tolerates exceedances on individual splits, which occur on about half of them. A deployer who wants the bound to hold with probability $1-\delta$ over the calibration set needs Learn-then-Test (Angelopoulos et al., 2025), which tests $H_\lambda : R(\lambda) > \alpha$ over a threshold grid with a finite-sample bound and fixed-sequence testing.

Proposition 1 (application of Angelopoulos et al., 2025). *If calibration and deployment documents are independent draws from one distribution and the per-document loss lies in* $[0, 1]$*, then with probability at least* $1 - \delta$ *over the calibration set the selected* $\hat{\lambda}$ *satisfies* $R(\hat{\lambda}) \leq \alpha$. The unit is the document, because claims within a document are dependent and the concentration bound needs independent units; the statement is an application, not a new theorem (Section I). The loss matters: our pre-specified guarantee controls the *document-mean* loss (a document's accepted-error rate, 0 if it accepts nothing) with a Hoeffding–Bentkus test; at the certified threshold the realized pooled risk is 2.2%, so the document-mean guarantee is conservative for the pooled rate in practice. A *pooled* loss that controls the accepted-error rate over all claims must absorb up to $M = 63$ errors from one document, and with 829 calibration documents no finite-sample test (WSR, empirical Bernstein, Hoeffding–Bentkus) certifies any threshold for it, a range effect confirmed on AVeriTeC, where every document has one claim and the same test certifies 13.2% (Table 2). A document-level CLT on the pooled loss is cheaper (19.6%) but asymptotic; calibrated in-domain and applied to FELM it passes in 0 of 100 splits, a distribution shift that finite-sample bounds survive only through their slack.

Table 2: Guarantees at $\alpha = 5\%$, $\delta = 0.1$, 100 document-level splits; budgets are shares of committed claims (mean $\pm$ s.d.), and *pass* is the share of splits whose held-out risk is at most α. *Floor*: smallest budget with realized pooled risk on the test half at or below α (13.9% on the calibration half, Section 5.1). *Main*: document-mean loss, Hoeffding–Bentkus; *Pooled WSR*: pooled loss, betting test; *CLT*: pooled loss, asymptotic; "–": no threshold certified. Main controls the document-mean loss (0 for a document accepting nothing), so it can certify below the pooled floor (Section I).

Data	Score / calibration	Floor (%)	Main: budget / pass	Pooled WSR	CLT: budget / pass
In-domain	no signal (random order)	99.5	93.9 ± 1.5 / 95	–	–
	margin	13.6	$\mathbf{52.5 \pm 5.2}$ / 94	–	19.6 ± 3.0 / 93
	margin $\oplus$ P(True)	9.9	51.1 ± 6.7 / 94	–	15.2 ± 2.3 / 93
FELM	margin, calibrated in-domain	53.0	62.0 ± 3.7 / 100	–	30.4 ± 3.9 / **0**
	margin, calibrated on target	51.9	60.2 ± 4.1 / 96	–	81.7 ± 17.1 / 84
AVeriTeC	margin, calibrated on target (M=1)	3.2	7.6 ± 2.5 / 92	13.2 ± 4.0 / 99	9.6 ± 2.9 / 92
FEVER	margin, calibrated in-domain	0.0	27.1 ± 2.2 / 100	–	13.0 ± 1.6 / 100

The main guarantee certifies a threshold at $52.5 \pm 5.2\%$ and passes in 94 of 100 splits; over the same splits (Figure 4a) pooled risk crosses α at 13.9%, the document-mean loss at 39.0% (25.1pp for the loss definition), the finite-sample bound is met at 52.4% (13.4pp of slack), and the procedure adds 0.2pp. The loss definition dominates because documents that accept one or two claims are 40.5% of documents and contribute 76.6% of the document-mean loss; truncating each document's weight at 20 accepted claims certifies at $34.4 \pm 4.4\%$ (exploratory), and calibrating in-domain and deploying out of distribution is conservative but safe (Section I).

5.2 When does the signal transfer? An evidence intervention

FELM omits its reference evidence from the verifier input, and the margin's AUROC there ranges from 0.82 (math) to 0.51 (writing); we built a grounded subset of 1,294 claims (selection test $p = 0.077$) and verified each claim twice, with and without the evidence. With evidence, the signal's advantage over a confound-only baseline changes sign in science (from -0.098 to $+0.182$ [0.019, 0.315]) and rises only to zero in writing ($+0.006$ [-0.287, 0.201]): transfer needs evidence and a checkable claim. We had predicted that evidence would lower abstention; it raised it from 26.7% to 40.6%, and the newly abstained claims were 1.53 times as error-prone without evidence ($p = 0.007$): missing evidence produced confident verdicts with no grounding, and supplying it converts silent errors into escalations (Section J).

5.3 Monitoring a verifier with a cheaper one

Because failure risk is shared, the margin of GPT-4o-mini, about 17 times cheaper at list price, ranks GPT-4o's errors at AUROC 0.778 against 0.784 for GPT-4o's own margin on the same 5,000-claim sample, the hierarchy of signals reproduces inside the cheaper verifier (margin 0.800, vote entropy 0.562; it abstains on 1.0%), and GPT-4o's margin ranks every other panel verifier's errors at 0.738–0.764 (Section K).

6 Case Study: Two Silent Model Updates

The guarantee of Section 5.1 assumes that calibration and deployment documents come from one distribution, and a provider can break that without notice. During collection, the `system_fingerprint` values returned with GPT-4o responses were replaced twice (August: 5 old, 8 new, 2 shared; September: 7 old, 2 new, none shared; Section L). We assign each response to a serving state by its fingerprint and compare the same claims across states (Table 15, Section L). The first replacement changed behavior, with domain-level error rates moving from $+2.8$pp on meeting summaries to -28.9pp on news summaries, so the aggregate hides the drift; the second, on the 500-claim AVeriTeC development set, changed 10 verdicts (1 in correctness) and no abstention count (92). A finite-sample threshold certified on the old state met its target on the new one in 100 of 100 recalibrations after the first replacement (baseline on the old state: 93) and 94 after the second (baseline 97); the in-expectation threshold passed in 94 and 92 on point estimates but in 47 and 15 under a document-bootstrap upper bound (Figure 4b). The first replacement lowered the aggregate error rate, so every method passed more often on the new state, and our prediction that the asymptotic threshold would degrade more than the finite-sample one did not materialize; this is a report that two unannounced replacements did not invalidate a certified threshold, not evidence that guarantees survive drift. Details are in Section L.

7 Implications

A deployer should escalate, rank by the free margin, budget for the floor plus the price of the guarantee it wants, and log serving fingerprints; an agreeing second verifier adds no evidence of correctness. An evaluator should read confident agreement of independent verifiers against a benchmark label as a signal about the label: in the core as a whole the label was wrong more often than the verifiers, so consensus is a cheap label-error detector.

8 Limitations

The core is measured for one primary verifier family, GPT-4o; the nine-verifier panel reproduces the consensus and the signal hierarchy, but the core curves are not re-estimated for other verifiers. The split of the core into label errors and confirmed errors rests on a 120-item audit with $\kappa = 0.34$, bounded by a leave-one-annotator-out range of 34.7–47.2%. The setting is claim verification against a supplied document with prompted frontier verifiers; whether the floor has the same composition for fine-tuned verifiers, step-level reward models, or verification without a source document is a question the decomposition can pose but this paper does not answer.

9 Conclusion

Confidence-based review of an LLM verifier has a floor, and the floor is made of consensus: claims on which independent verifiers agree, rightly against a bad label or wrongly together, which signals, second opinions, and panels recover little of. The decomposition measures that floor for any signal set and budget, the audit separates its two faces, and the residual-risk identity turns it into the price of a certifiable guarantee. For oversight, disagreement rather than agreement is the informative event; treating concurring models as independent evidence repeats the premise N-version programming disproved.

AI USE STATEMENT

Generative AI tools were used solely to assist with language polishing of the manuscript, including improving the grammar, clarity, readability, and phrasing of author-written text. All AI-assisted revisions were carefully reviewed and edited by the authors to ensure that they accurately reflect the intended scientific meaning. The authors take full responsibility for the final content of this work.

ETHICS STATEMENT

The label audit and edit validation were carried out by lab members and research interns who were informed of the purpose of the annotation (relabeling verification benchmarks) but blinded to hypotheses, labels, and verdicts; participation was voluntary and acknowledged. All datasets are used under their licenses for evaluation only, and no source documents are redistributed. The work concerns the reliability of automated verification; its findings caution against treating agreement among LLM verifiers as evidence of correctness.

REPRODUCIBILITY STATEMENT

Section B states the data conventions, deduplication, provenance logging, temperatures, and licenses; Sections D and I give Algorithms 1 and 2 with the losses, p-values, and splits; Sections G and H give the annotation and validation protocols with seeds; Section M lists every pre-registered prediction with its commit-time status. Every number in the paper is produced by a macro file generated from the experiment logs, and the commit hash of each pre-registered prediction is listed with the released code. Scripts, derived verifier outputs (verdicts, token probabilities, fingerprints), the panel transfer matrix, and the annotation files will be released as supplementary material.

REFERENCES

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International Conference on Learning Representations*, 2024.

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *The Annals of Applied Statistics*, 19(2):1641–1662, 2025. doi: 10.1214/24-AOAS1998.

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2), 2023.

Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael I. Jordan. Distribution-free, risk-controlling prediction sets. *Journal of the ACM*, 68(6), 2021.

Rishi Bommasani, Kathleen A. Creel, Ananya Kumar, Dan Jurafsky, and Percy Liang. Picking on the same person: Does algorithmic monoculture lead to outcome homogenization? In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Lingjiao Chen, Matei Zaharia, and James Zou. How is ChatGPT's behavior changing over time? *Harvard Data Science Review*, 6(2), 2024.

Shiqi Chen, Yiran Zhao, Jinghan Zhang, I-Chun Chern, Siyang Gao, Pengfei Liu, and Junxian He. FELM: Benchmarking factuality evaluation of large language models. In *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*, volume 36, 2023.

John J. Cherian, Isaac Gibbs, and Emmanuel J. Candès. Large language model validity via enhanced conformal prediction methods. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

C. K. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, 1970.

Charles Corbière, Nicolas Thome, Avner Bar-Hen, Matthieu Cord, and Patrick Pérez. Addressing failure prediction by learning model confidence. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11, 2010.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630:625–630, 2024.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Yonatan Geifman and Ran El-Yaniv. SelectiveNet: A deep neural network with an integrated reject option. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, 2019.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11), 2020.

Shashwat Goel, Joschka Strüber, Ilze Amanda Auzina, Karuna K. Chandra, Ponnurangam Kumaraguru, Douwe Kiela, Ameya Prabhu, Matthias Bethge, and Jonas Geiping. Great models think alike and this undermines AI oversight. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 19621–19678, 2025.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, 2017.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. Association for Computational Linguistics, 2018.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2017.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. In *International Conference on Learning Representations*, 2024.

Paul F. Jaeger, Carsten T. Lüth, Lukas Klein, and Till J. Bungert. A call to reflect on evaluation practices for failure detection in image classification. In *International Conference on Learning Representations*, 2023.

Liwei Jiang, Yuanjun Chai, Margaret Li, Mickel Liu, Raymond Fok, Nouha Dziri, Yulia Tsvetkov, Maarten Sap, and Yejin Choi. Artificial hivemind: The open-ended homogeneity of language models (and beyond). In *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*, volume 38, 2025.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Adam Tauman Kalai and Santosh S. Vempala. Calibrated language models must hallucinate. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*. ACM, 2024.

Amita Kamath, Robin Jia, and Percy Liang. Selective question answering under domain shift. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2020.

Sanyam Kapoor, Nate Gruver, Manley Roberts, Katherine Collins, Arka Pal, Umang Bhatt, Adrian Weller, Samuel Dooley, Micah Goldblum, and Andrew Gordon Wilson. Large language models must be taught to know what they don't know. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Elliot Myunghoon Kim, Avi Garg, Kenny Peng, and Nikhil Garg. Correlated errors in large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 30038–30066, 2025.

Jon Kleinberg and Manish Raghavan. Algorithmic monoculture and social welfare. *Proceedings of the National Academy of Sciences*, 118(22), 2021.

John C. Knight and Nancy G. Leveson. An experimental evaluation of the assumption of independence in multiversion programming. *IEEE Transactions on Software Engineering*, SE-12(1): 96–109, 1986.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *International Conference on Learning Representations*, 2023.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, 2024.

Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research*, 2022.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2023.

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2019.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2023.

Christopher Mohri and Tatsunori Hashimoto. Language models with conformal factuality guarantees. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, 2024.

Hussein Mozannar and David Sontag. Consistent estimators for learning to defer to an expert. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, 2020.

Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, Kashun Shum, Randy Zhong, Juntong Song, and Tong Zhang. RAGTruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2024.

Curtis G. Northcutt, Anish Athalye, and Jonas Mueller. Pervasive label errors in test sets destabilize machine learning benchmarks. In *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*, volume 34, 2021.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua V. Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. LLM evaluators recognize and favor their own generations. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Ellie Pavlick and Tom Kwiatkowski. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7, 2019.

Stephan Rabanser, Stephan Günnemann, and Zachary C. Lipton. Failing loudly: An empirical study of methods for detecting dataset shift. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Michael Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. AVeriTeC: A dataset for real-world claim verification with evidence from the web. In *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*, volume 36, 2023.

Tal Schuster, Darsh J. Shah, Yun Jie Serene Yeo, Daniel Filizzola, Enrico Santus, and Regina Barzilay. Towards debiasing fact verification models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*. Association for Computational Linguistics, 2019.

Wooseok Seo, Seungju Han, Jaehun Jung, Benjamin Newman, Seungwon Lim, Seungbeen Lee, Ximing Lu, Yejin Choi, and Youngjae Yu. Verifying the verifiers: Unveiling pitfalls and potentials in fact verifiers. In *Second Conference on Language Modeling*, 2025.

Hexiang Tan, Fei Sun, Sha Liu, Du Su, Qi Cao, Xin Chen, Jingang Wang, Xunliang Cai, Yuanzhuo Wang, Huawei Shen, and Xueqi Cheng. Too consistent to detect: A study of self-consistent errors in LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 4755–4765. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.238.

Liyan Tang, Tanya Goyal, Alexander R. Fabbri, Philippe Laban, Jiacheng Xu, Semih Yavuz, Wojciech Kryściński, Justin F. Rousseau, and Greg Durrett. Understanding factual errors in summarization: Errors, summarizers, datasets, error detectors. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2023.

Liyan Tang, Philippe Laban, and Greg Durrett. MiniCheck: Efficient fact-checking of LLMs on grounding documents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2024.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, 2018.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D. Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2023.

Ryan J. Tibshirani, Rina Foygel Barber, Emmanuel J. Candès, and Aaditya Ramdas. Conformal prediction under covariate shift. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Roman Vashurin, Ekaterina Fadeeva, Artem Vazhentsev, Lyudmila Rvanova, Daniil Vasilev, Akim Tsvigun, Sergey Petrakov, Rui Xing, Abdelrahman Sadallah, Kirill Grishchenkov, Alexander Panchenko, Timothy Baldwin, Preslav Nakov, Maxim Panov, and Artem Shelmanov. Benchmarking uncertainty quantification methods for large language models with LM-polygraph. *Transactions of the Association for Computational Linguistics*, 13:220–248, 2025. doi: 10.1162/tacl_a_00737.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. In *International Conference on Learning Representations*, 2024.

Yuan Zhang, Jason Baldridge, and Luheng He. PAWS: Paraphrase adversaries from word scrambling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, 2019.

Qinyu Zhao, Ming Xu, Kartik Gupta, Akshay Asthana, Liang Zheng, and Stephen Gould. The first to know: How token distributions reveal hidden knowledge in large vision-language models? In *Computer Vision – ECCV 2024*, Lecture Notes in Computer Science. Springer, 2024. doi: 10.1007/978-3-031-73195-2_8.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*, volume 36, 2023.

A Extended related work

Confidence signals for LLMs. Calibration and softmax-based failure detection predate LLMs (Guo et al., 2017; Hendrycks & Gimpel, 2017), and both degrade under dataset shift (Ovadia et al., 2019). For LLMs, signals come from verbalized confidence (Lin et al., 2022; Tian et al., 2023; Xiong et al., 2024), self-evaluation (Kadavath et al., 2022), sampling and semantic entropy (Kuhn et al., 2023; Farquhar et al., 2024; Manakul et al., 2023), output-token distributions (Zhao et al., 2024; Vashurin et al., 2025), and trained probes (Kapoor et al., 2024). This line of work compares signals by ranking quality on generation tasks; we hold the signals fixed and measure the errors that none of them ranks, on a verification task where the output is a single verdict token. Kalai & Vempala (2024) prove that a calibrated language model must hallucinate at a positive rate; our floor is an empirical counterpart for verification, measured per review budget.

Selective prediction and failure prediction. The reject option (Chow, 1970) and the risk–coverage analysis of selective classification (El-Yaniv & Wiener, 2010; Geifman & El-Yaniv, 2017; 2019) define the review rule we study, and learning-to-defer methods route inputs between a model and an expert (Mozannar & Sontag, 2020). Corbière et al. (2019) train a confidence model to predict classifier failure, Kamath et al. (2020) show that selective question answering weakens under domain shift, and Jaeger et al. (2023) document evaluation pitfalls in failure detection. These works select among a predictor's own outputs; we decompose the part of the risk–coverage curve that no available signal can lower and identify what the remaining errors have in common.

LLM verifiers and their benchmarks. Fact verification has moved from Wikipedia claims (Thorne et al., 2018) to real-world claims with web evidence (Schlichtkrull et al., 2023) and to LLM outputs: summaries (Tang et al., 2023), retrieval-augmented answers (Niu et al., 2024), long-form generations (Min et al., 2023), and multi-domain responses (Chen et al., 2023). Tang et al. (2024) aggregate such benchmarks and train an efficient verifier, and Seo et al. (2025) evaluate 12 LLMs and one specialized verifier on 14 benchmarks and find that about 16% of examples are ambiguous or mislabeled. We measure how predictable the remaining failures are and adopt the label-quality concern of Seo et al. (2025), Northcutt et al. (2021), and Pavlick & Kwiatkowski (2019) as a control on our core estimates.

Surface heuristics. Fine-tuned inference models accept hypotheses that overlap lexically with the premise (McCoy et al., 2019), paraphrase models fail on word-scrambled pairs (Zhang et al., 2019), and inference and fact-verification datasets carry artifacts that a claim-only model can exploit (Gururangan et al., 2018; Schuster et al., 2019); Geirhos et al. (2020) unify such findings as shortcut learning. We find the same signature in the confidence of a frontier LLM verifier and confirm it by intervention.

Correlated failures and LLM-based oversight. Knight & Leveson (1986) showed that independently developed software versions fail on the same inputs, and Kleinberg & Raghavan (2021) and Bommasani et al. (2022) analyze the costs of shared decision models. LLMs from different developers are homogeneous in open-ended generation (Jiang et al., 2025) and correlated in their errors (Kim et al., 2025; Goel et al., 2025); LLM judges favor their own outputs (Panickssery et al., 2024; Zheng et al., 2023), LLMs do not reliably correct their own reasoning without external feedback (Huang et al., 2024), and step-level verifiers serve as reward signals (Lightman et al., 2024). Tan et al. (2025) propose a probe into a second model for self-consistent errors; we show that verification errors are consistent across models as well.

Guarantees under drift. Conformal prediction (Vovk et al., 2005) and its risk-controlling extensions (Bates et al., 2021; Angelopoulos et al., 2024; 2025) give distribution-free guarantees under exchangeability, with extensions to covariate shift and non-exchangeable data (Tibshirani et al., 2019; Barber et al., 2023). Shift detection (Rabanser et al., 2019) and longitudinal audits of a deployed LLM (Chen et al., 2024) document that models change after release; we follow certified thresholds through two such changes on paired claims.

B DATA PIPELINE AND PROVENANCE

Document selection and convention. Documents whose source text is byte-identical across subsets are collapsed to one (60 removed; 1,819 $\rightarrow$ 1,759), and two documents with fewer than 20 characters of source text are dropped (1,757 documents, 22,189 claims); every document has at least two claims. GPT-4o is run three times per claim. A claim is escalated when its majority verdict is UNKNOWN (1,946 claims) or the three runs disagree three ways (8 claims); the remaining 20,235 claims carry a committed verdict and 1,725 are errors (8.525%). The signal comparison of Table 3 was run on the pre-deduplication pool (22,347 claims; 20,241 committed verdicts, 1,731 errors); the 20% union core differs by less than 0.2pp between the two pools.

AVeriTeC. The AVeriTeC train split (3,067 claims) is used only for evaluation, never for fitting, and is added to the 500-claim development split (3,567 claims; 2,651 with a committed verdict enter the guarantee analysis); the extension narrows the AUROC interval from 0.150 to 0.052 in width (0.8817 [0.795, 0.945] on 17 errors to 0.8350 [0.809, 0.861] on 153). Both splits were collected on the same serving generation.

Provenance and temperature. Every response stores the provider's `system_fingerprint`, timestamp, and batch id. Temperatures are 0.1 for GPT-4o and Gemini and 1.0 for GPT-4o-mini and the panel; a probe of the GPT-4o API confirms that verdict-token log-probabilities are returned untempered (log-probability gap ratio 0.98 between temperature settings), so the margin's scale does not depend on temperature, while DeepSeek returns tempered log-probabilities and is collected at 1.0 with the verdict defined as the argmax token.

Pipeline checks. Layer-1 deduplication and format filters (a flip-word safeguard), a five-model claim-validity jury (flagged 51.7% against 15.2% disagreement on negatives, $p = 2.3 \times 10^{-11}$), a judge-bias diagnosis, contamination screening (65 of 81 flagged items removed, the remaining 16 checked by hand), leave-one-subset-out transfer (0.844 $\pm$ 0.055, no collapsed subset), and a delta-retry re-collection protocol (68K calls) precede every number in the paper.

Licenses. LLM-AggreFact is used as an evaluation benchmark only (no pretraining or fine-tuning); FELM is CC BY-NC-SA 4.0; AVeriTeC is CC BY-NC 4.0; FEVER annotations are CC BY-SA 3.0 over Wikipedia text. We release scripts and derived verifier outputs, not the source documents or claims.

C SIGNALS AND THE FULL SIGNAL COMPARISON

Table 3: Failure-prediction signals, in-domain, pre-deduplication pool (20,241 committed verdicts, 1,731 errors). AUROC with document-level 95% CI; AURC is the area under the risk–coverage curve (lower is better); $\pm$ is the SD over 10 fold seeds. Cost is in verifier calls per claim beyond the verdict.

Family	Signal	Extra calls	AUROC	AURC
Distributional	margin (1 run)	0	0.799 [0.782, 0.814]	0.0309
	margin (3-run mean)	2	0.801 [0.785, 0.817]	0.0304
Self-evaluation	verbalized confidence	1	0.651 [0.635, 0.669]	0.0563
	$P(\text{True})$	1	0.824 [0.803, 0.842]	0.0258
Behavioral	4 behavioral features	5	0.697	0.051
Combined	margin $\oplus$ $P(\text{True})$	1	0.835 ± 0.007	0.0245
	all 17 features	5	0.834	0.025

The paired difference $P(\text{True})$ $-$ margin is $+0.0236$ [$+0.0093$, $+0.0379$]; verbalized $-$ margin is -0.1495 [-0.168, -0.131]. Verbalized confidence takes 12 distinct values, with 91.24% of responses at 100 and 2.68% at 0. Single-run ablation (a separate replicate run in which the 3-run mean reaches 0.800): runs 1–3 alone reach 0.7986, 0.7996, and 0.7973, and the 3-run mean

adds $+0.0017$ [$+0.0000$, $+0.0038$] over run 1, one eighth of the seed SD. Feature sets: behavioral 4D 0.697 (AURC 0.051), margin alone 0.800 (0.031), 17D 0.834 (0.025; $p = 2.1 \times 10^{-11}$ over 4D). Out of distribution, adding the four behavioral features to the margin changes AUROC by $[-0.054, -0.012]$ on FELM, $[-0.064, -0.018]$ on FEVER, and ≈ 0 on AVeriTeC; adding all 17 features changes it by -0.029 [-0.047, -0.009], -0.009 [-0.029, $+0.012$], and -0.019 [-0.057, $+0.021$]. A 16-dimensional set that removes the reasoning-token log-probability recovers parity with the margin out of distribution (0.830 in-domain), so the harmful dimension is style-sensitive rather than error-sensitive. Negative controls: overlap-based gating improves within-pool AURC (0.052 $\rightarrow$ 0.044) but fails zero-shot; a contradiction feature is redundant with cross-model agreement ($\phi = 0.97$–0.98); an off-the-shelf NLI verifier reaches 0.45. $P(\text{True})$ and verbalized confidence were collected in-domain only, so their transfer is not claimed; the primary verifier's temperature of 0.1 suppresses sampling variation, which is one reason vote entropy carries little signal in-domain, while the panel at temperature 1.0 reproduces the same hierarchy. Margin AUROC out of distribution (3-run mean): FELM 0.707 [0.672, 0.745] (423 errors), AVeriTeC 0.835 [0.809, 0.861] (153), FEVER 0.740 [0.715, 0.766] (361); within FELM, math 0.822, reasoning 0.807, science 0.618, world knowledge 0.608, writing 0.511.

D DETECTABILITY DECOMPOSITION: ALGORITHM AND SENSITIVITY

Algorithm 1 Detectability decomposition

Require: verdicts from K runs, gold labels, document ids, signal set $\mathcal{S}$ with reference s_0, budgets $\mathcal{B}$, bootstrap draws M

1: $A \leftarrow$ claims whose majority verdict is UNKNOWN or tied; remove A; $E \leftarrow$ errors among the N remaining claims
2: **for** $B \in \mathcal{B}$ **do**
3: **for** $s \in \mathcal{S}$ **do**
4: $T_s(B) \leftarrow$ top $\lceil BN \rceil$ claims by s, deterministic ties; record capture $|E \cap T_s(B)|/|E|$
5: **end for**
6: $U \leftarrow \bigcup_s T_s(B)$; $\text{core}^{\cup}(B) \leftarrow |E \setminus U|/|E|$; $\text{vol}(B) \leftarrow |U|/N$
7: $\text{core}^{=}(B) \leftarrow |E \setminus T_{\bar{s}}(B)|/|E|$ for the combined score $\bar{s}$ ▷ equal budget
8: $\text{core}^{\text{vol}}(B) \leftarrow |E \setminus T_{s_0}(\text{vol}(B))|/|E|$ ▷ equal volume
9: $\text{core}(B \mid s_0) \leftarrow |E \setminus U|/|E \setminus T_{s_0}(B)|$
10: **end for**
11: Repeat lines 2–9 on M document-level bootstrap resamples; report percentile intervals and paired differences
12: **return** $|A|/(N + |A|)$, per-signal capture, vol, the four core curves, intervals

Table 4: Sensitivity of the core at a 20% budget to protocol choices (% of all errors). Random ties: tie-breaking at the cut by seed. No verbalized: the degenerate signal removed. z-scored: signals standardized and averaged instead of the out-of-fold logistic combination.

Variant	Union core	Union reviewed (%)	Equal volume, margin	Equal budget, combined
Main (deterministic ties, logistic)	18.4	37.0	21.3	31.6
Random ties	16.1	44.8	16.2	31.6
No verbalized confidence	18.8	36.5	21.5	31.6
z-scored combination	18.4	37.0	21.3	31.3

Per-signal capture at 20%: margin 61.8%, $P(\text{True})$ 64.6%, verbalized 35.7% (it selects only the 8.7% of claims with confidence below 100), behavioral 51.6%. Under random tie-breaking, verbalized confidence fills its budget with randomly chosen confidence-100 claims, which is why the union reviewed 8pp more and the core looked smaller in an earlier tally. The conditional core with all other signals is 51.2, 48.1, 38.9, and 22.6% of the missed errors at the four budgets, and the SD over 10 fold seeds is below 1.2pp for every learned quantity.

E AUDIT OF EVALUATION PITFALLS

Table 5 lists each check on the naive pipeline, what it found, and the protocol rule it produced; the five rules (claim-level evaluation, escalation before ranking, document-grouped splits, a confound-only baseline, and an out-of-distribution test) govern every result in the paper.

Table 5: Audit of the naive pipeline.

Check	Finding	Correction
Abstention tautology	`is_unknown` coefficient $+2.62$; majority UNKNOWN $\Rightarrow$ error by definition; explains 95% of the 0.93	escalate; exclude from features
True difficulty after correction	AggreFact 0.853 $\rightarrow$ 0.722; AVeriTeC 0.67; FELM 0.69	report 0.67–0.72
Document-level OR aggregation	pre-deduplication pipeline (1,819 documents): blind spot 9.9 $\rightarrow$ 4.5%, base rate 75.6 $\rightarrow$ 17.5% (16.6% on the current pool)	claim-level evaluation
Cross-model "failure mode"	98.4% of the cluster is `is_unknown`; removing it, error 9.1%	rejected
Dataset confounds	confound-only 0.69–0.73 in-pool $\rightarrow$ 0.55 / 0.49 out of distribution (gap 0.126 / 0.201, 10/10 seeds)	confound baseline and OOD test
Embedding shortcut	no document-level suspicious cluster (XSum partial only)	–
Label shuffle	transcript-embedding overfit confirmed; leakage excluded	–

F CONSENSUS ANALYSES

Table 6: Consensus at matched confidence (10% budget; document-level 95% CI). *Excluded*: Gemini's UNKNOWN verdicts dropped; *disagreement*: counted as wrong on correct claims and not-wrong on missed errors. Lift is an exploratory re-expression of the pre-registered matched-bin agreement test.

UNKNOWN	Margin bin	P(wrong | missed error)	P(wrong | correct)	Lift
excluded	[0.90, 0.95)	66.3% [55.4, 75.9]	11.0% [8.3, 13.7]	6.0× [4.6, 8.2]
excluded	[0.95, 0.99)	76.5% [70.2, 83.1]	6.6% [5.4, 7.9]	11.6× [9.5, 14.6]
excluded	[0.99, 1.00]	85.4% [80.8, 89.6]	2.4% [2.0, 2.7]	36.1× [31.3, 42.5]
disagreement	[0.90, 0.95)	46.5% [37.4, 55.7]	22.9% [19.4, 26.3]	2.0× [1.6, 2.6]
disagreement	[0.95, 0.99)	61.4% [54.7, 68.2]	13.8% [12.0, 15.6]	4.5× [3.7, 5.4]
disagreement	[0.99, 1.00]	73.3% [67.0, 79.2]	4.7% [4.0, 5.4]	15.7× [13.3, 18.7]

Across budgets. At 20%, 30%, and 50% the missed errors concentrate in the top bin, where the lift is 36.1× [31.3, 42.5], 36.0× [31.3, 42.5], and 38.6× [33.3, 45.7] (UNKNOWN excluded). Differences in agreement (missed − correct) at 10% are -0.227 [-0.333, -0.124], -0.169 [-0.237, -0.107], -0.122 [-0.163, -0.083]; with UNKNOWN as disagreement, -0.307, -0.248, -0.220. In the top bin at 10%, 2.7% of the claims Gemini agrees with are GPT-4o errors against 16.8% of those it disputes, and 73% of missed errors are agreed on by both verifiers. An unmatched comparison of missed and flagged errors gives an agreement gap of $+0.257$; matching on margin reduces it to $+0.151$ [0.066, 0.224], so we report only the matched figure. Gemini abstains on 29.9%, 19.7%, and 14.1% of missed errors against 13.3%, 7.7%, and 2.4% of correct claims (exploratory). Unanimity over GPT-4o's three runs is 99.5% on missed errors, 88.3% on flagged errors, and 99.1% on correct claims; this contrast is partly induced by selecting on the three-run mean margin and is reported for completeness only.

Label-noise correction. The core-wide audited gold-wrong share (59.2%) is applied to the missed-and-agreed cell of every bin, treating each gold-wrong item as a claim on which both verifiers were right; the corrected lifts are 4.3, 9.1, and 30.2× (the bin-specific audited rates are

33.3% [14.3, 52.4] in [0.95, 0.99) and 64.6% [54.6, 74.0] in [0.99, 1.00]; correcting the top bin with its own rate lowers its lift slightly below 30.2×, and both rates are far below every tipping point), and the tipping points at which the lift would reach 1 are 94.4%, 98.0%, and 99.6% (68.4%, 90.7%, 98.3% with UNKNOWN as disagreement). For the panel, let q be the share of GPT-4o's errors that are label errors; weighting the audited rates by the union core's share of errors gives $q \approx 0.184 \times 0.592 + 0.816 \times 0.333 = 0.38$. If every one of these label errors is counted as a shared error (co-occurrence 1), the co-occurrence on genuine errors is at least $(0.599 - 0.38)/(1 - 0.38) = 0.35$, and dividing by the observed base rate of 7.5%, which itself includes label errors and so overstates the genuine base rate, leaves a cross-vendor lift of at least 4.7×.

Table 7: Verifier panel on the 5,000-claim stratified sample (4,997 claims after the convention change, 1,045 documents), same prompt. Added verifiers run at temperature 1.0 with three runs unless noted. Reasoning models expose no token probabilities and contribute a majority verdict; Sonnet 5 used adaptive thinking on 10.0% of responses. Gemma 4 31B was dropped because neither API provider that serves it (bf16 or fp8) returned usable verdict-token log-probabilities.

Verifier	Vendor	Access	Logprobs	Parse	Abstain	Error rate [CI]	Margin AUROC [CI]
GPT-4o (primary)	OpenAI	API, T=0.1, 3 runs	yes	100%	0.0%	8.5% [7.6, 9.5]	0.784 [0.758, 0.809]
GPT-4o-mini	OpenAI	API, T=1.0, 3 runs	yes	100%	1.0%	8.2% [7.3, 9.2]	0.800 [0.773, 0.826]
Gemini 2.5 Flash	Google	API, T=0.1, 1 run	no	100%	6.5%	8.9% [8.0, 9.8]	–
DeepSeek V4.1 Flash	DeepSeek	API, thinking off, 3 runs	yes	100%	8.0%	6.9% [6.1, 7.8]	0.813 [0.787, 0.840]
Claude Haiku 4.5	Anthropic	batch, extended thinking (1,024), 1 run	no	100%	4.7%	7.7% [6.9, 8.6]	–
Claude Sonnet 5	Anthropic	batch, adaptive thinking, 3 runs	no	99.9%	7.9%	6.6% [5.8, 7.5]	–
Qwen3.5-9B	open	local vLLM, bf16, 3 runs	yes	99.9%	0.7%	8.2% [7.3, 9.1]	0.822 [0.800, 0.843]
Qwen3.5-27B	open	OpenRouter → Novita, bf16, top-5 logprobs, 3 runs	yes	100%	2.7%	7.9% [7.1, 8.8]	0.813 [0.785, 0.840]
Gemma 4 12B	Google (open weights)	local vLLM, bf16, 3 runs	yes	99.9%	4.9%	7.6% [6.8, 8.5]	0.796 [0.770, 0.821]

Table 8: Error co-occurrence. Left: $P(j \text{ errs} \mid \text{GPT-4o errs})$ on the 426 sampled GPT-4o errors. Right: the same conditioned on the 366 errors of the reasoning model. Lift is the ratio to j's base error rate.

j	$P(j \mid \text{GPT-4o})$	base	lift [CI]	j	$P(j \mid \text{Haiku})$	lift
GPT-4o-mini	70.9% [66.0, 75.6]	8.1%	8.70 [7.89, 9.63]	Qwen3.5-27B	71.3%	9.30
Qwen3.5-9B	62.7% [57.6, 67.6]	8.1%	7.71 [6.98, 8.52]	Gemma 4 12B	63.7%	8.84
Qwen3.5-27B	54.9% [49.7, 59.9]	7.7%	7.17 [6.47, 7.98]	GPT-4o	61.5%	7.21
Gemma 4 12B	53.8% [48.5, 59.2]	7.2%	7.46 [6.75, 8.34]	Claude Sonnet 5	59.3%	9.75
Claude Haiku 4.5	52.8% [48.2, 57.6]	7.3%	7.21 [6.53, 8.06]	GPT-4o-mini	57.7%	7.08
Gemini 2.5 Flash	52.3% [47.1, 57.3]	8.3%	6.32 [5.67, 7.04]	DeepSeek V4.1	54.6%	8.61
DeepSeek V4.1 Flash	45.8% [40.4, 51.4]	6.3%	7.22 [6.47, 8.11]			
Claude Sonnet 5	45.5% [40.5, 50.5]	6.1%	7.49 [6.70, 8.42]			

Abstention convention. Co-occurrence $P(j \text{ errs} \mid i \text{ errs})$ and j's base rate are computed over all sampled claims, with j's abstentions counted as not wrong, so both share one denominator; the error rates in Table 7 are over j's committed verdicts and exceed the base rates by j's abstention share (e.g., Gemini $8.9\% \times (1 - 6.5\%) = 8.3\%$).

Panel disagreement rules. At a 20% budget, a k-of-8 dissent rule recovers 8.8% [3.0, 15.7], 3.8% [0.0, 8.4], 1.2% [0.0, 4.2], and 0.0% of the 80 sampled union-core errors for $k = 1, 2, 3, 4$ while reviewing 14.9%, 8.0%, 5.6%, and 3.8% of claims; of the 177 sampled errors the margin misses it recovers 36.7% [29.2, 45.2], 23.7% [16.9, 31.4], 16.9% [11.3, 23.3], and 13.6% [8.7, 19.1] (at 10%: 27.6, 28.0, 19.8, 15.2%; at 30%: 32.8, 21.9, 15.6, 10.9%). The rule reviews flagged claims first and fills the rest of the budget in margin order; when the flagged share exceeds the budget, as the 1-of-8 rule's 14.9% does at 10%, the flagged set is truncated in margin order, which is why the 1-of-8 rule recovers slightly less than the 2-of-8 rule there. Figure 3 plots every pair. Intervals for 49 of the 72 ordered pairs are wider than the pre-registered half-width of 5pp (widest 5.9pp) because several verifiers err on fewer than 400 sampled claims.

G LABEL AUDIT PROTOCOL

Sampling frame. The frame is the 20% union core (319 errors), the errors flagged by at least one signal (1,406), and the correct committed verdicts (18,510), under fold seed 0. A stratified sample

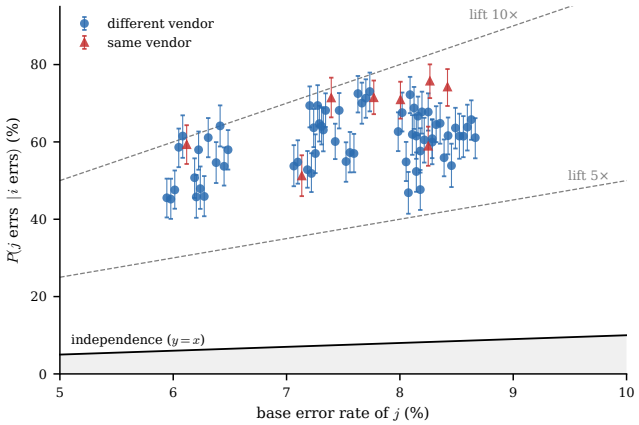

Figure 3: Error co-occurrence $P(j \text{ errs} \mid i \text{ errs})$ for every ordered pair of the nine-verifier panel against j's base error rate (95% document-bootstrap CI), same-vendor pairs marked; the solid line is independence and the dashed lines mark lifts of 5× and 10×. Every pair lies above the 5× line's neighborhood and none approaches independence.

Table 9: Label audit (blind, two annotators per core item, separate adjudicator; document-level bootstrap 95% CI). *Gold wrong*: the adjudicated label agrees with the verifier. *Confirmed*: it agrees with the gold label. Agreement is raw agreement before adjudication on double-annotated items. In the correct stratum the remaining 3.3% [0.0, 10.7] (one item) was labeled undecidable by its single annotator; after adjudication no core or flagged item is ambiguous.

Stratum	n	Gold wrong (%)	Verifier error confirmed (%)	Agreement (%)
Core errors (missed by all signals)	120	59.2 [50.4, 67.7]	40.8 [32.3, 49.6]	65.8 [57.3, 74.6]
Flagged errors	60	33.3 [21.9, 45.0]	66.7 [55.0, 78.1]	85.0 [65.0, 100] (n=20)
Correct verdicts (correct confirmed)	30	16.7 [3.4, 32.3]	80.0 [63.3, 93.3]	single-annotated

of 210 items (120 core, 60 flagged, 30 correct; seed 20260921) is the prefix of a larger sample drawn earlier, so it remains a random sample of each stratum; a pilot of 30 items drawn outside the sample was used only to revise the guideline, which was committed before the main batch.

Annotation. Five annotators (two lab members and three research interns, none an author of the analysis) receive a spreadsheet with the claim, a reading aid (the document window sharing the most content words with the claim, ±2 sentences) above the full document, and the guideline; the judgment is whether the document supports the claim (TRUE, FALSE, undecidable), with the full document as the criterion. They are not told the hypothesis, the strata or their proportions, the gold labels, the verifier verdicts, or the dataset names; they may not discuss items, search, or use AI tools; no LLM is used as an annotator. Items are shuffled into one global random order and each annotator's file is a subsequence of it; the 140 double-annotated items (all 120 core and 20 random flagged) are spread evenly over the 10 annotator pairs (14 per pair; 70 items per annotator; assignment seed 20260923). Every assigned item was completed.

Adjudication and analysis. Disagreements (44: 39 TRUE-vs-FALSE, 5 involving undecidable) go to a separate adjudicator who sees the two annotations in shuffled order without annotator identity and returns TRUE, FALSE, or ambiguous; the adjudicator returned 17 TRUE, 27 FALSE, and 0 ambiguous (24 with the first annotator, 15 with the second, 5 a third label). Final labels: agreeing → that label; disagreeing → the adjudicator's; single-annotated → that label.

Agreement. Raw agreement is 68.6% (96/140) [0.60, 0.76] overall (Fleiss $\kappa = 0.39$ [0.24, 0.53]), 65.8% [57.3, 74.6] on core items (Cohen $\kappa = 0.34$ [0.18, 0.49]), and 85.0% [65.0, 100] on the 20 flagged items, where κ (-0.07) is uninformative because nearly all labels fall in one class. Per-pair κ on 14 items ranges from -0.12 to 0.86 (0.86, 0.61, 0.59, 0.58, 0.50, 0.39, 0.30, 0.29, -0.02, -0.12). The pre-registered criterion $\kappa \geq 0.70$ was not met; we report the result without changing the rule. One annotator labels FALSE on 70% of shared items against 34–52% for the others (16

items FALSE only for this annotator against 2 the other way), a consistent strictness rather than a guideline violation.

Sensitivity. Removing any single annotator moves the confirmed share of the core between 34.7% and 47.2% (all annotators: 40.8%; gold-wrong 52.8–65.3%), and gold-wrong remains a majority in every case. The corrected core in Table 1 is the union core scaled by the confirmed share: lower bound 7.5% [5.9, 9.1] (confirmed errors only) and upper bound 7.5% (only gold-label errors removed), which coincide because no item was adjudicated ambiguous.

H LEXICAL OVERLAP, EDIT INTERVENTION, AND EXAMPLES

Table 10: Claim–document content-word overlap of margin-missed and flagged errors (Jaccard; document-level bootstrap 95% CI on the difference; r is rank-biserial), with the overlap between the verifier's reasoning text and the document for comparison.

Budget	Missed	Flagged	Difference [CI]	r	Difference, reasoning text [CI]
10%	0.575	0.551	$+0.023$ [-0.006, $+0.049$]	0.061	$+0.019$ [-0.001, $+0.038$]
20%	0.581	0.556	$+0.025$ [-0.006, $+0.054$]	0.064	$+0.025$ [$+0.002$, $+0.048$]
30%	0.597	0.554	$+0.043$ [$+0.008$, $+0.075$]	0.110	$+0.036$ [$+0.010$, $+0.061$]
50%	0.592	0.561	$+0.031$ [-0.011, $+0.073$]	0.077	$+0.042$ [$+0.010$, $+0.072$]

Claim overlap by margin quartile of errors (least to most trusted): 0.557 / 0.549 / 0.550 / 0.605; reasoning-text overlap: 0.556 / 0.561 / 0.558 / 0.603. The two measures correlate at $\rho = 0.670$ because the reasoning restates the claim; claim length is weakly related to overlap ($\rho = 0.127$). The pre-registered prediction concerned claim overlap and is not supported; the reasoning-text measure is reported only for comparison. In-domain, low overlap between the verifier's reasoning and the document goes with a higher error rate (23.3% against 8.3%) and an abstention rate eight times higher (14.0% against 1.7%), the observation that motivated the evidence intervention of Section 5.2.

Table 11: Edit intervention on 1,000 confidently accepted claims (document-level 95% CI). *Accepted*: majority TRUE with mean margin ≥ 0.95 over three runs. *Validated*: share of a 40-item cell-blind sample matching the cell's intent; for B, also the share of the 141 accepted edits confirmed as violations.

Cell	Meaning / vocabulary	Overlap	Edit dist.	Accepted (%)	Validated
A	original	0.759	0	95.9 [94.5, 97.1]	–
A′	preserved / preserved	0.893	17.4	93.9 [92.2, 95.3]	90%
B	violated / preserved	0.881	17.4	**14.1** [11.9, 16.3]	95%; accepted: 34.8%
C	violated / reduced	0.180	20.1	**1.2** [0.6, 2.0]	100%
D	preserved / reduced	0.190	19.5	82.6 [80.0, 85.3]	88%

Construction. Sources are claims with majority TRUE, mean margin ≥ 0.95, and gold TRUE; 1,843 sources yielded edits in all four cells after filtering (51% yield) and 1,000 were kept with at most 25% from any dataset (753 outside RAGTruth; operators balanced at 124–126 claims). Edits are generated by Gemini 2.5 Flash from one prompt per cell. Filters require overlap above 0.80 for A′ and B and below 0.25 for C and D, a token edit distance within a factor of 1.3 across cells (achieved 1.16; B and A′ both 17.4), and a length ratio within [0.7, 1.4]. Fingerprints in the 15,000-call batch were balanced across cells (maximum difference 1.8pp; χ^2 $p = 0.96$, Cramér's $V = 0.019$).

Hypotheses (pre-registered). Numbering follows the pre-registration file. H1 $\geq$ 5pp: $+12.9$pp [$+10.6$, $+15.2$], supported. H2, H2′ (margin drop ≥ 0.02 for D against A and A′): $+0.038$ [$+0.027$, $+0.050$] and $+0.034$ [$+0.023$, $+0.045$], not supported after validation. H4a ($|A' - A| < 5$pp): -2.0pp [-3.5, -0.4]; H4b (margin drop < 0.02): $+0.004$ [-0.002, $+0.011$]; both supported. H5 (interaction < 0): -1.6pp [-5.4, $+2.1$], inconclusive. Without RAGTruth: H1 $+13.8$pp [$+10.9$, $+16.6$], H5 -3.5pp [-7.8, $+1.2$].

Table 12: Edit intervention by operator: confident acceptance in cell B and its reduced-vocabulary counterpart C (document-level bootstrap 95% CI).

Operator	n	B accepted (%)	C accepted (%)	B − C [CI]
causal	124	26.6	0.8	+25.8 [+17.6, +33.8]
scope	124	25.0	2.4	+22.6 [+14.4, +31.1]
role swap	125	15.2	2.4	+12.8 [+6.3, +19.8]
number	126	13.5	0.0	+13.5 [+7.8, +19.7]
entity	126	13.5	1.6	+11.9 [+5.2, +18.7]
temporal	126	10.3	0.0	+10.3 [+5.4, +16.0]
quantifier	125	8.8	0.8	+8.0 [+3.1, +13.8]
negation	124	0.0	1.6	−1.6 [−4.1, +0.0]

Human validation. A validator labeled 40 items per edited cell (160) as supported or not, blind to cell and operator, and a second validator relabeled a random 40 (agreement 92.5% [82.5, 100], $\kappa = 0.85$). Pre-registered criteria were 90% intended-label agreement per cell and 80% per operator: A′ 90% [80, 98], B 95% [88, 100], C 100%, D 88% [75, 98]; the causal operator reached 60% on five items. Among validated items, confident acceptance is 92% for A′ (36 items, margin 0.983), 0% for B (38), 0% for C (40), and 94% for D (35, margin 0.993); the five D items whose meaning had changed were accepted at 60%. The validation sample of B therefore contained one confidently accepted edit, and that edit had failed validation; because the sample is too small to test whether the accepted B edits are disproportionately failed edits, we validated the confidently accepted B edits directly. A validator who had not seen the earlier sample labeled all 141 confidently accepted B edits under the same blind protocol, with a random 20 relabeled by a second annotator (17/20 agreement, $\kappa = 0.68$): 34.8% (49/141; document-level bootstrap CI [26.9, 42.0]) are genuine meaning violations, and the one item that failed the earlier 40-item validation was again judged supported. Excluding the 92 accepted B edits judged still supported, together with the C edits of the same source claims, the paired difference is +4.3pp (95% document-level bootstrap CI [+2.7, +5.9], 2,000 resamples), against +12.9pp as generated. The correction is conservative: only accepted B edits were audited, so failed edits among the rejected ones remain in the denominator, and removing them pairwise would shrink it and raise the difference. By operator, 26 of the 31 accepted scope edits were failed edits; on validated edits scope is accepted at 5.1%, in line with the other operators (4.4–5.8%), and only causal remains high (15.0%), so the as-generated operator ranking in Table 12 overstates scope. The pre-registered H1 threshold of 5pp is met by the as-generated estimate and missed by the validated point estimate, whose interval includes 5 and excludes 0.

Qualitative examples. Examples in the released materials are drawn at random from the confirmed core errors with the sampling rule stated. A separate set of 12 hand-screened candidates, selected by reasoning-text overlap before the audit, is listed with the 4 that were rejected for label concerns (three because the document supported the claim verbatim), as an illustration of why the label audit was necessary.

I GUARANTEE: LOSSES, PROCEDURE, ARGUMENT, SENSITIVITY

Losses and argument. The main guarantee controls the document-mean loss $L_i = e_i/a_i$ (0 when $a_i = 0$), where a_i and e_i are the accepted claims and accepted errors of document i; the p-value is the Hoeffding–Bentkus bound (Bates et al., 2021), and fixed-sequence testing over Λ from the most conservative threshold gives family-wise validity at level δ (Angelopoulos et al., 2025), which is Proposition 1; the Hoeffding–Bentkus bound requires independent bounded losses, which is why the document, not the claim, is the unit. The pooled accepted-set error rate $\sum_i e_i / \sum_i a_i$ is at or below α exactly when $\mathbb{E}[Z_i] \le 0$ for $Z_i = e_i - \alpha a_i$, which we rescale to $W_i = (Z_i + \alpha M)/M \in [0, 1]$ with M the largest number of claims in a document (63 in-domain, 32 on FELM, 1 on AVeriTeC and FEVER); its range grows with M, and no finite-sample test (WSR, empirical Bernstein, Hoeffding–Bentkus) certifies it at $M = 63$, which we predicted from synthetic data before running it, while the same tests certify it at $M = 1$. Learned scores use a 20/40/40 train/calibration/test split; an earlier cross-fitted version broke exchangeability between calibration and test scores (pass 66%),

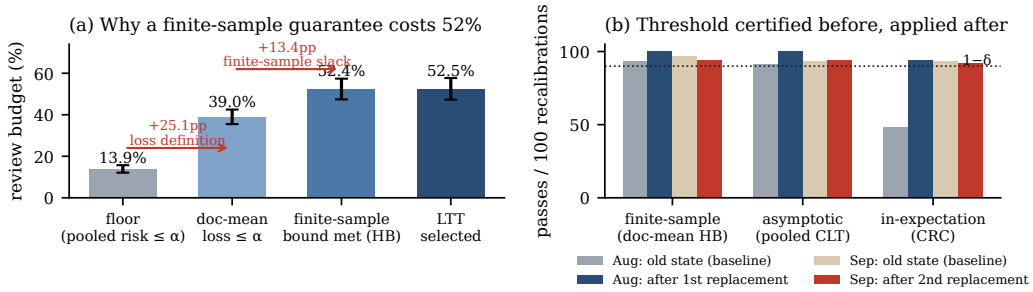

Figure 4: **The price of a guarantee above the floor.** **(a)** From the floor budget to the certified budget over 100 document-level splits (mean ± s.d.): the loss definition costs 25.1pp, finite-sample slack 13.4pp, and the procedure 0.2pp. **(b)** Thresholds certified before each silent replacement and applied after it, against the same threshold on the old state (Section 6); the dotted line is $1 - \delta$.

Algorithm 2 Escalate-then-rank with Learn-then-Test calibration

Require: calibration documents $D_1, \ldots, D_n$ with claims, verdicts, gold labels; score s; grid Λ (400 thresholds, descending acceptance); α, δ

Stage 1: route every claim with verdict UNKNOWN to review
for $\lambda \in \Lambda$ from most to least conservative **do**
  $L_i(\lambda) \leftarrow$ error rate among accepted claims of D_i (0 if none accepted) $\in [0, 1]$
  $p_\lambda \leftarrow$ Hoeffding–Bentkus p-value for $H_\lambda : \mathbb{E}[L(\lambda)] > \alpha$ from $L_1(\lambda), \ldots, L_n(\lambda)$
  if $p_\lambda > \delta$ **then break**
  end if
  $\hat{\lambda} \leftarrow \lambda$
end for
return $\hat{\lambda}$ (or "review everything" if no λ is rejected)

and the three-way split restores it (94%). Pass rates are computed on the held-out half of each of 100 document-level splits with the loss the method controls; pooled-risk pass rates are 100% for the margin (mean document loss 3.53%, pooled risk 2.20%). Rows omitted from Table 2: AVeriTeC with the margin calibrated in-domain has floor 3.7%, main $48.5 \pm 2.4\%$ / 100, and CLT $26.8 \pm 3.0\%$ / 100; FEVER calibrated on target certifies at 0% because its base error rate (3.0%) is below α; the no-signal row uses a random order and its floor of 99.5% is the budget at which nearly everything is reviewed. Documents that accept 1–2 claims are 40.5% of documents and contribute

Table 13: From the floor to the certified budget, 100 document-level splits (mean ± SD). Increments are computed per split and averaged.

Stage	Budget (%)	Increment (pp)
Pooled risk on calibration half $\leq \alpha$ (floor)	13.9 ± 1.8 [10.6, 17.3]	
Document-mean loss $\leq \alpha$	39.0 ± 3.5	25.1 ± 3.4 (loss definition)
Finite-sample bound met (HB, $\delta = 0.1$)	52.4 ± 5.0	13.4 ± 2.8 (finite-sample slack)
LTT selection	52.5 ± 5.2	0.18 ± 0.7 (procedure; 0 in 89 splits)
Claim-level HB (diagnostic; invalid under dependence)	16.7 ± 2.0	

76.6% of the document-mean loss (group error rate 6.2–7.0%); documents accepting 21 or more are 8.9% and contribute 3.7% (2.0%). This is a composition effect: subsets with few claims per document (FactCheck-GPT, Reveal, ClaimVerify) are harder, and the document mean weights them by document count while the pooled rate weights by claim count. A single split (seed 0: 17.8 → 41.4 → 53.6 → 54.1) sits at the 98th percentile of the floor distribution, which is why we report means over 100 splits.

Truncated weights (exploratory). Weighting each document by $\min(a_i, c)$ interpolates between the losses. Certified budgets with WSR are $94.7 \pm 13.0\%$ at $c = 1$ (certified in 20% of splits), $50.0 \pm 8.0\%$ at $c = 5$, $39.4 \pm 4.5\%$ at $c = 10$, and $34.4 \pm 4.4\%$ at $c = 20$ (pass 99%, pooled

risk at the threshold 3.0%), returning to uncertifiable at $c = M$; the pre-registered prediction of a monotone increase in c was wrong because documents with no acceptance drop out at $c = 1$.

Earlier calibration. The in-expectation budget of $12.0 \pm 2.3\%$ (17D) and the finite-sample budget of 52.4% (margin) use different scores and are not a like-for-like comparison; the in-expectation budget against $46.4 \pm 36.6\%$ for a baseline without the distributional signal is computed with document-grouped folds; the risk–budget curve falls from 6% to 2% for 17D and stalls at 5% for the baseline, and the certified threshold and its risk are insensitive to claim-level versus document-level splitting ($\hat{\lambda}$ 0.9417 against 0.9423).

J EVIDENCE INTERVENTION

Table 14: Evidence intervention on the grounded FELM subset (1,294 claims; same claims, one variable changed). Gap is signal AUROC minus confound-only AUROC within the domain.

	Without evidence	With evidence	Test
Gap, science domain	-0.098	$+0.182$ [0.019, 0.315]	CI excludes 0
Gap, writing domain	-0.208	$+0.006$ [-0.287, 0.201]	CI includes 0
Verdicts changed	–	36.1%	–
Error rate, unchanged verdicts	13.2%	10.1%	$p = 0.001$
Abstention rate	26.7%	40.6%	sign test, $p = 2.2 \times 10^{-18}$
Error rate (no evidence), newly abstained vs. retained	20.1% vs. 13.2% (1.53×)		Fisher, $p = 0.007$

A partial-abstention rule (routing claims with one UNKNOWN run) raises escalation value 2–5× across the four families without harming zero-shot AUROC.

K CROSS-VERIFIER TRANSFER

GPT-4o-mini on the stratified sample: its own margin ranks its own errors at 0.800 (vote entropy 0.562, cross-model agreement 0.718), GPT-4o's margin ranks GPT-4o-mini's errors at 0.764, and GPT-4o-mini's margin ranks GPT-4o's errors at 0.778 against 0.784 for GPT-4o's own margin; GPT-4o-mini abstains on 1.0%. Across the panel, GPT-4o's margin ranks the errors of every other verifier at AUROC 0.764 (GPT-4o-mini), 0.753 (Qwen3.5-9B), 0.748 (DeepSeek), 0.747 (Gemma 4 12B), 0.745 (Claude Sonnet 5), 0.744 (Gemini), 0.741 (Qwen3.5-27B), and 0.738 (Claude Haiku 4.5); the full 9×9 matrix is released with the code.

L SILENT MODEL UPDATES

Table 15: Two unannounced snapshot replacements behind one model name, on paired claims. Guarantee rows use the claims both states committed to and report passes over 100 recalibrations of the threshold certified on the old state, with the same threshold on the old state's held-out half in parentheses.

	First replacement (Aug.)	Second replacement (Sep., AVeriTeC dev)
Claims present in both states (before deduplication)	22,297	500
Correctness changed (improved / worsened)	707 / 487	1 / 0
McNemar p	2.3×10^{-10}	1.0
Abstentions (old / new)	9.0–9.3% / unchanged	92 / 92
Paired committed claims for guarantee rows	19,609 (1,652 docs)	362
Finite-sample, document-mean HB ($\delta = 0.1$)	100 (93), budget $81.3 \pm 7.4\%$	94 (97), budget $9.3 \pm 6.7\%$
Asymptotic, pooled CLT	100 (91)	94 (93)
In-expectation CRC: point / doc-bootstrap upper	94 (48) / 47	92 (93) / 15

Fingerprints per session in August share a fleet (one fingerprint appears in all five sessions); the September collection (18,406 calls) shows two fingerprints that appear in none of the 11 August values, and within one session the split between the two September fingerprints moved from 59/41 to 99/1 across three consecutive runs, so the mix is a per-request property outside our control. Domain-level paired McNemar tests on the first replacement: TofuEval-MeetB $+2.8$pp ($p = 0.020$), TofuEval-MediaS $+2.5$pp ($p = 0.015$), Lfqa $+1.5$pp ($p = 0.006$), RAGTruth

-0.7pp ($p = 2.2 \times 10^{-5}$), ExpertQA -7.2pp ($p = 1.5 \times 10^{-16}$), AggreFact-XSum -8.7pp ($p = 0.041$), AggreFact-CNN -28.9pp ($p = 1.9 \times 10^{-24}$); the improvement persists without CNN ($p = 2 \times 10^{-4}$). Restricting both states to first-attempt responses preserves the shift, which rules out the delta-retry collection as its source. For each replacement, the guarantee rows of Table 15 use the claims both states committed to in all runs (19,609 in 1,652 documents; 362); thresholds are certified on the old state's calibration half and applied to the new state's held-out half, with the same threshold on the old state's held-out half as baseline. The first replacement raised the certified budget of the main guarantee to $81.3 \pm 7.4\%$ (old-state calibration data require more review to certify) against 20.5% for the CLT auxiliary. The in-expectation guarantee, recomputed with document-level splits, passes in 94 and 92 of 100 on point estimates, in 47 and 15 under a document-bootstrap upper bound, and in 70 and 6 under a claim-level Wilson bound, which is not valid under within-document dependence. The second replacement's base error rate is below α, so its rows are a supporting case only.

M Pre-registered predictions and outcomes

Table 16 lists every directional prediction committed before the corresponding experiment, with its outcome and status.

Table 16: Every directional prediction committed before the corresponding experiment, with its outcome.

Prediction	Outcome	Status
Core errors share less claim–document vocabulary than flagged errors	$+0.025$ [-0.006, $+0.054$]	not supported
Supplying evidence lowers abstention	rose 26.7% $\rightarrow$ 40.6%, on riskier claims	reversed (protective)
Agreement on missed errors matches agreement on correct claims (matched margin)	-0.227 / -0.169 / -0.122	rejected
Finite-sample budget in 15–25%	52.4%	missed
Pooled-loss finite-sample test uncertifiable at $M = 63$	uncertifiable	confirmed
Confirmed verifier errors exceed half of the core	40.8% [32.3, 49.6]	rejected
Gold-wrong rate: core > flagged > correct	59.2 / 33.3 / 16.7%	confirmed
Gold-wrong rate in correct stratum below 16%	16.7%	missed (narrowly)
Inter-annotator $\kappa \geq 0.70$	0.39	missed
H1: vocabulary-kept violations accepted $\geq$ 5pp more than reduced	$+12.9$pp as generated; $+4.3$pp [2.7, 5.9] on validated edits	supported as generated; below threshold after validation
H2, H2′: paraphrase lowers margin by ≥ 0.02	$+0.038$, $+0.034$; D fails validation	unsupported
H4: reordering within $\pm$5pp and margin drop < 0.02	-2.0pp, margin $+0.004$	supported
H5: interaction < 0	-1.6pp [-5.4, $+2.1$]	inconclusive
Truncated-weight budget increases with c	U-shaped	reversed
CLT threshold degrades more than finite-sample under drift	both improved; CLT changed more, favorably	missed
Document-bootstrap upper bound at most Wilson	$47 \leq 70$ (Aug.); $15 > 6$ (Sep.)	mixed
Cross-vendor median lift ≥ 2	7.75$\times$	supported
1-of-8 dissent recovers under half of the union core at 20%	8.8%	supported
Cross-verifier transfer AUROC ≥ 0.70 for a majority of pairs	0.738–0.764	supported
Panel CI half-widths $\leq$ 5pp	49 of 72 exceed (max 5.9pp)	partly missed