ASSURANCE, NOT JUST ACCURACY: AUDITING LLM REASONING WITH AUDIT-RISK CONTROL

Anonymous authors
Paper under double-blind review

ABSTRACT

Benchmarks report how accurate a large language model (LLM) is on average; users of a single answer need to know how much they can trust *this* output. We transplant the institutional design of financial auditing into LLM quality control and study the resulting system, ARA (Audit-Risk control for Assurance). ARA converts a target misstatement risk α into per-claim detection-risk budgets via the audit-risk model $\mathrm{AR} = \mathrm{IR} \times \mathrm{CR} \times \mathrm{DR}$ (inherent, control, and detection risk), purchases evidence under a token budget by risk-weighted knapsack, ranks evidence by independence (CPU recomputation over isolated LLM procedures over non-isolated ones), and issues per-output audit opinions. The deliverable is an empirical dual-caliber certificate, conditional on the realized certified subset of the measured distribution: a one-sided Clopper–Pearson 95% upper bound on residual error over the extracted claim universe, at a stated coverage; it is not a per-answer or future-sample guarantee. α is a target, not a promise, and we falsify it openly: across twelve track$\times$domain$\times\alpha$ settings α=0.01 is falsified in three, while 0.05 and 0.10 hold wherever the pass is not vacuous. On GSM8K with Qwen2.5-7B, ARA certifies 89.0% of answers at 3.8% residual error (raw: 10.9%). Auditing the frontier API model DeepSeek V4 Pro certifies 93.4% of GSM8K-slice answers at 2.1% risk (CP95 3.6%) and 57.7% of stress answers at zero error as-run; a deterministic rejudging recovers coverage to 96.2%/78.2%. Root-causing attributes 98.7% of adverse evidence to our own extraction layer, dominated by a single Unicode minus-sign artifact, not arithmetic inability. A sealed reproduction on a second, weaker deployment (Tencent Hunyuan Hy3, reasoning disabled, ¥15.43) re-confirms the certificate on GSM8K, does not confirm it on stress, and finds evidence-channel value generator-dependent, not universal.

1 INTRODUCTION

On September 9–10, 2026, we ran a full statistical audit of DeepSeek V4 Pro, the flagship reasoning endpoint of a major provider. The premise of the paper: frontier models and their deployments change faster than any static benchmark score remains meaningful, so assurance must be computable *at use time*, per answer, with a stated budget and a stated residual risk.

Benchmark practice answers a different question. Aggregate accuracy summarizes a distribution; it cannot tell a user whether the answer in front of them is wrong. Selective prediction abstains on low-confidence inputs (El-Yaniv & Wiener, 2010; Geifman & El-Yaniv, 2017) and conformal methods wrap guarantees around such scores (Angelopoulos & Bates, 2023), but the score usually comes from the same model that produced the answer, coupling error and detector. Verbalized confidence is miscalibrated (Kadavath et al., 2022; Tian et al., 2023); self-consistency improves the average while certifying nothing per instance (Wang et al., 2023).

Financial auditing solved an analogous problem long ago. An auditor cannot recompute every transaction; the profession instead institutionalized a risk model, an evidence hierarchy, budgeted procedures, and standardized opinions with quantified residual risk (International Auditing and Assurance Standards Board, 2019). That institutional design, not any single statistic, makes assurance scalable, and we transplant it to LLM generation.

The transplant is not cosmetic. The audit-risk model $\mathrm{AR} = \mathrm{IR} \times \mathrm{CR} \times \mathrm{DR}$ factorizes the risk of a wrong clean opinion into inherent, control, and detection risk; the auditor controls only the last

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.

term, so a target $\mathrm{AR} \leq \alpha$ becomes a per-assertion detection budget. A knapsack allocator purchases procedures by expected risk reduction per token. Opinions follow suit: an *adverse* opinion requires exact machine counter-evidence, not another model's disagreement.

We evaluate the system in three tracks. Track A audits a local 7B model end-to-end, every component ablatable. Track B audits DeepSeek V4 Pro through its public API as a black box; Track B′ re-runs the protocol on a second, live deployment. Throughout we ask: (Q1) does the certified subset carry materially lower error at high coverage? (Q2) does the guarantee survive pre-registered comparators and ablations, including losses? (Q3) is α achieved, and where not, do we detect and report it?

Our answers are mixed by design; a measurement study that reports only wins is marketing. Q1: yes. Track A certifies 89.0% of GSM8K answers at 3.8% residual error versus 10.9% raw; Track B certifies 93.4% of slice answers at 2.1% as-run (CP95 3.6%), coverage recovering to 96.2% under a deterministic rejudging repair, with zero certified stress errors in both calibers. Q2: mostly, with named exceptions. Isolation of audit evidence matters ($p \approx 10^{-19}$); deterministic tools and heterogeneous pools tie the full system on GSM8K; uniform purchasing dominates ours on stress coverage at equal zero risk. Q3: $\alpha{=}0.01$ is falsified in three of the four track×domain cells (all twelve settings reported), one $\alpha{=}0.10$ cell holds only vacuously by collapsing coverage, and root-causing attributes 98.7% of adverse evidence to our own extraction layer, a single Unicode minus-sign artifact (U+2212), not arithmetic inability. Assurance audits the interface as much as the model.

Contributions. (i) *Framework*: a formal transplant of financial-audit institutional design (risk model, evidence hierarchy, budgeted procedures, standardized opinions) into per-output LLM quality control, with a dual-caliber certificate: certified subset, CP95 risk bound, coverage (§3). (ii) *Falsifiable targets*: α is a target subject to exact binomial tests in both directions, not a promise; across twelve settings we report three falsifications at $\alpha{=}0.01$ and one vacuous pass (§6, §7.2). (iii) *Evidence*: a controlled local study plus a sealed deployment case study of a frontier API endpoint, with a full cost ledger (¥37.79), comparator losses disclosed, and claim-level root-causing of all adverse evidence; a sealed reproduction (Track B′) adds a second live deployment for ¥15.43 (§5–§8). The statistical machinery is deliberately classical throughout; the contribution is the assembly, and Appendix S maps each claim to its evidence.

2 Related Work

Risk-coverage trade-offs descend from Chow's reject option (Chow, 1970) through selective classification (El-Yaniv & Wiener, 2010; Geifman & El-Yaniv, 2017) and its end-to-end variant (Geifman & El-Yaniv, 2019), with distribution-free wrappers via conformal prediction (Angelopoulos & Bates, 2023) and risk-controlling extensions to user-specified losses (Bates et al., 2021; Angelopoulos et al., 2024). ARA differs in what is certified: not a model score but an evidence workpaper, risk measured on audit opinions (a split-conformal head-to-head is in §5).

Chain-of-thought (Wei et al., 2022), self-consistency (Wang et al., 2023), and self-refinement (Madaan et al., 2023) raise average accuracy; verification builds checkers (Cobbe et al., 2021; Lightman et al., 2024), trained on process- and outcome-based feedback (Uesato et al., 2022) and benchmarked as process reward models on human-labeled traces (Wang et al., 2024; Zheng et al., 2025); program-aided prompting moves arithmetic into a runtime (Gao et al., 2023; Chen et al., 2023); LLM-as-judge scales evaluation with documented biases (Zheng et al., 2023). The trade-off against trained verifiers is structural: a process reward model amortizes step-level labels into one forward pass, but it is a learned scorer that shifts with the generator's distribution and returns a score; ARA needs no training and returns a measured certificate triple, at the price of per-claim evidence purchases and bounded extractor recall (§8).

Hallucination surveys (Ji et al., 2023; Huang et al., 2025) catalog the failure modes our taxonomy operationalizes. Algorithmic auditing has an internal-framework tradition (Raji et al., 2020; Mökander et al., 2024); AI control pairs untrusted models with trusted monitoring (Greenblatt et al., 2024); we add the quantitative core: a priced risk model and a falsifiable per-output certificate.

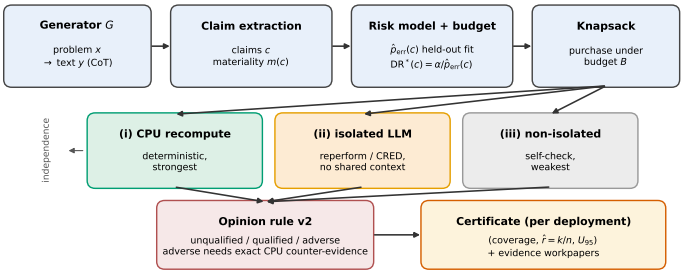


Figure 1: The ARA pipeline. Claims are priced by a held-out risk model, evidence is purchased by a knapsack allocator under independence ranking, opinions are gated by machine counter-evidence, and the deliverable is a per-deployment certificate triple with workpapers.

3 THE ARA FRAMEWORK

3.1 SETUP: CLAIMS, MATERIALITY, OPINIONS

A generator G answers a problem x with text y. We decompose y into *claims*: atomic, checkable assertions such as an arithmetic step or the final answer. Each claim c carries a *materiality* $m(c)$, the tolerance within which a misstatement does not change the conclusion (relative tolerance d_0=0.3; the final answer is always material). 552 of the 554 root-caused adverse-driving claims (§7.4) are material under any $d_0 \ll 1$ (one at the boundary, one a false alarm), so no opinion moves within a wide band around d_0. An audit issues one of three opinions: *unqualified* (no material misstatement detected), *qualified* (misstatement suspected or detected but immaterial, or evidence inconclusive), or *adverse* (material misstatement established by exact counter-evidence). The product is the pair (opinion, evidence workpaper), not a corrected answer; Figure 1 sketches the assembly.

Scope and terminology. The audit certifies only the *extracted claim universe*: assertions the pipeline can parse into checkable form. Extraction recall is bounded externally in §8 (best case $\leq$0.40 on human-cleaned text); certificates do not cover unrecorded assertions, but certified risk is scored against gold answers, so recall bounds what the audit can catch upstream, not the certificate's validity. We borrow the institutional structure of audit opinions, not their letter: accounting's *qualified* means "except for"; read ours as *flagged*.

3.2 THE AUDIT-RISK MODEL AS A BUDGET EQUATION

Financial auditing factorizes the risk of a wrong clean opinion as $\mathrm{AR} = \mathrm{IR} \times \mathrm{CR} \times \mathrm{DR}$ (International Auditing and Assurance Standards Board, 2019): inherent risk (how error-prone the assertion is), control risk (whether controls miss it), detection risk (whether the auditor's procedures miss it). Only DR is under the auditor's control; we fix control risk at a reliance constant $\hat{\rho} = 0.1$ that never credits the generator's self-checks (§3.4):

$$\mathrm{AR}(c) \;=\; \underbrace{\hat{p}_{\mathrm{err}}(c)}_{\mathrm{IR}} \times \underbrace{\mathrm{DR}(c)}_{\text{procedure choice}} \;\leq\; \alpha, \qquad \mathrm{DR}^*(c) = \frac{\alpha}{\max(\hat{p}_{\mathrm{err}}(c),\, 10^{-12})}. \tag{1}$$

$\hat{p}_{\mathrm{err}}(c)$ comes from a small logistic risk model fitted on held-out generations (Appendix A). Claims with $\hat{p}_{\mathrm{err}}(c) \leq \alpha$ need no procedure; the rest are strengthened or rerouted (§3.3). Two caveats. First, the factorization is a conceptual decomposition for allocating effort, not a stochastic identity (International Auditing and Assurance Standards Board, 2019): the per-claim budget is an allocation heuristic and the certificate is measured at answer level (§3.5). Second, $\hat{\rho} = 0.1$ is a chosen constant, not an estimate (the risk model's floor, costs, and materiality are $\hat{\rho}$-independent). Misspecification costs efficiency, not validity: the certificate is measured.

3.3 PURCHASING EVIDENCE UNDER A BUDGET

Each procedure π on claim c costs tokens $\kappa(\pi, c)$ and delivers an estimated detection rate $1 - \mathrm{DR}(\pi, c)$; its weight $s(c) = |\mathrm{val}(c)|/(|\mathrm{answer}| + 1)$ is the claim's magnitude share of the final

answer (0 for non-numeric claims, 1 if the answer is unparseable; material flag at $s(c) \geq 0.05$). Strength follows equation 1: claims with $\mathrm{DR}^*(c) \leq \hat{\rho}$ are upgraded to the deterministic recompute channel, and the rest receive $n(c) = \lceil \log((\mathrm{DR}^*(c) - \hat{\rho})/(1 - \hat{\rho}))/\log d_0 \rceil$ isolated reperforms, the smallest n with $\hat{\rho} + (1 - \hat{\rho})d_0^n \leq \mathrm{DR}^*(c)$ (here d_0 is the per-sample miss factor, distinct from its materiality role in §3). Given a per-problem token budget B, the allocator solves

$$\max_{S \subseteq \Pi} \sum_{c \in S} \hat{p}_{\mathrm{err}}(c)\, s(c) \big(1 - \mathrm{DR}(\pi_c, c)\big) \quad \text{s.t.} \quad \sum_{c \in S} \kappa(\pi_c, c) \leq B, \tag{2}$$

a multiple-choice knapsack (at most one procedure per claim) solved by greedy density with exact repair. The budget is real money: Track B prices tokens at published rates and publishes the full ledger (Appendix M).

3.4 An evidence hierarchy, with independence as the currency

Not all evidence is equal; the ordering principle is independence from the generator's failure modes. (i) *Deterministic recomputation* (strongest): an extracted claim is re-evaluated by CPU, counter-evidence exact. (ii) *Isolated LLM procedures*: an auditor re-performs (*reperform*) or critiques (CRED) the claim without seeing the generator's rationale. (iii) *Non-isolated checks* (weakest): the generator verifies its own claim, or an auditor sees its trace; correlated errors survive, as §5 shows ($p \approx 10^{-19}$). Our isolation is *context independence* (the auditor never sees the generator's rationale), not epistemic independence; same-family auditors share training distributions with the generator and can inherit its blind spots, a leakage §7.5 measures. The hierarchy has teeth in the opinion rule (v2): *an adverse opinion requires exact CPU counter-evidence*. LLM-only disagreement escalates the claim (reperform, then recompute) but cannot by itself condemn an answer (§7).

3.5 The dual-caliber guarantee, and what we do not claim

Definition 1 (Dual-caliber guarantee D1)**.** *For an audit configuration on a problem distribution $\mathcal{D}$, let $\mathcal{S}$ be the certified subset (problems with unqualified opinions), k the number of certified answers that are wrong, and $n = |\mathcal{S}|$.* ARA *reports the triple*

$$\big(\textit{coverage}\ \Pr_{x \sim \mathcal{D}}[x \in \mathcal{S}], \quad \hat{r} = k/n, \quad U_{95} = \mathrm{Beta}^{-1}(0.95;\, k+1,\, n-k)\big),$$

where U_{95} is the one-sided Clopper–Pearson upper bound (Clopper & Pearson, 1934).

Two calibers are deliberate: $\hat{r}$ is what a user experiences, U_{95} is what a regulator can hold us to, and coverage keeps both honest. The opinion attaches to one answer; the statistical statement attaches to the certified subset, not to any single answer: a lone unqualified opinion is a rule-based judgment with no statistical guarantee. The certificate is descriptive, not predictive: it holds for the measured distribution under the stated opinion rule, over the extracted claim universe; U_{95} covers future exchangeable draws from the same conditional distribution, but abstention correlates with error by design and new workloads require re-measurement (Appendix B).

Remark 1 (D4: α is a target, not a promise)**.** *We do not claim the construction proves $\mathrm{AR} \leq \alpha$: the risk model is estimated, the reliance constant is a choice, and opinions correlate with errors by design. We instead make α falsifiable, testing each cell in both directions (§4), Holm-controlled within pre-registered families (Holm, 1979; McNemar, 1947).*

The contrast with conformal-style wrappers (Angelopoulos & Bates, 2023) is one of targets, not superiority: marginal coverage over future exchangeable draws versus an empirical certificate over the realized certified subset, with no distribution-free claim and failures published.

4 Experimental Design

Three tracks. *Track A* (local) audits Qwen2.5-7B-Instruct (Qwen Team, 2024), with Qwen2.5-3B and Llama-3.1-8B-Instruct (Dubey et al., 2024) as auditor pool. *Track B* (API) audits DeepSeek V4 Pro through its public endpoint; the same stack runs unchanged; auditors are Qwen/Llama-family models, disjoint from the generator (leakage measured in §7.5). *Track B′* (§7.6) re-runs the same protocol under a fresh hash-anchored pre-registration on a second live deployment.

Table 1: Headline, GSM8K test (n=1,319). Coverage: fraction of answers certified (unqualified); certified risk: error within that subset (D1). Baselines certify everything by construction; SC@2 (Wang et al., 2023) and Self-Refine (Madaan et al., 2023) score 0.1046/0.1107 at coverage 1. †: ablations (§5.2).

System	Coverage	Certified risk	Avg. audit tokens
Qwen2.5-7B, CoT (Wei et al., 2022)	1.000	0.1092 (144/1319)	—
ARA (full)	0.890	**0.0383** (45/1174)	150.2
no-tool†	0.894	0.0424 (50/1179)	150.2
no-isolation†	0.965	0.0856 (109/1273)	58.4
heterogeneous pool†	0.818	0.0398 (43/1079)	184.5

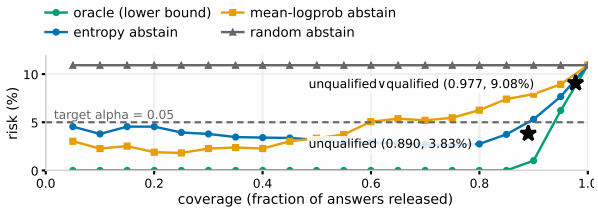


Figure 2: Coverage–risk on GSM8K (Track A). Curves: 20-point threshold sweeps; ARA's operating points (stars) sit below every non-oracle curve at matched coverage. Derived offline from the frozen artifacts (per-row provenance; sha-256 logged).

Data. GSM8K test (n=1,319) (Cobbe et al., 2021); a synthetic stress set (n=600) of multi-operand arithmetic with magnitudes up to 10^7, GSM-Hard-style (Gao et al., 2023) but rebuilt so every claim is exactly CPU-recomputable; MBPP-100 (Austin et al., 2021) and HotpotQA-500 (Yang et al., 2018) as out-of-domain probes (Appendix O).

Statistics. Risk bounds are one-sided Clopper–Pearson 95% (Clopper & Pearson, 1934). Each α cell is a separate audit with its own purchases and certified subset, tested in both directions with exact one-sided binomials (falsification rejects $\mathrm{AR} \leq \alpha$, confirmation rejects $\mathrm{AR} \geq \alpha$; Tables 3 and 2). Track A's pre-registered target is audit risk over all test problems, $\mathrm{AR} = \Pr[\text{wrong and unflagged}]$; Track B's certificate conditions on the certified subset (D1). Paired comparisons use exact McNemar (McNemar, 1947); family-wise control uses Holm (Holm, 1979) (rosters in Appendix A). Matched-coverage comparisons (A5b) sweep a 20-point grid with interpolation and report wins and losses (Appendix K). Claim-vs.-answer rates use two-sided Fisher exact tests.

Honesty protocol. Comparators, gates, and α grids were fixed before outcomes were inspected; the decision log and budget gates ship with the hashed manifest. Where an early pipeline was refitted on rechecked labels (*rejudged*), we report both as-run and rejudged numbers and never let rejudged calibrators influence scoring or planning. No arm was re-planned or re-run after unblinding; the purchase ledger is frozen. Negative results, ties, and losses appear in the main text, not the supplement. Track B′ and the extraction-recall study (§8) follow the same discipline (Appendix F).

5 TRACK A: AUDITING A LOCAL 7B MODEL

5.1 HEADLINE: WHAT CERTIFICATION BUYS

Table 1 gives the deployment-decision numbers. Raw Qwen2.5-7B chain-of-thought is wrong on 10.9% of GSM8K answers and certifies nothing; ARA certifies 89.0% at 3.8% residual error, a 2.8× risk reduction at a stated abstention cost; the remaining 11.0% receive qualified (115) or adverse (30) opinions, none dropped. Error is monotone in the opinion an answer escaped: 3.83% among unqualified, 62.6% among qualified, 90.0% among adverse. At matched coverage 89.0%, entropy abstention carries 5.00% risk against ARA's 3.83%, mean-logprob and random sit higher still (Fig. 2, Appendix K); the oracle lower bound is 0.81%, reported as a loss. A split-conformal calibration of the same two scores (20 seeds, post-hoc) also lands behind—(0.867, 3.95%) against (0.890, 3.83%)—and on stress collapses to abstain-all in 20/20 seeds (Appendix K).

5.2 Ablations: which institutions carry the weight

Exact McNemar tests (paired on problems) rank what matters. *Isolation is the load-bearing wall*: removing it (auditors see the generator's trace) raises certified risk to 8.6%, n_{01}=64, $p = 1.1 \times 10^{-19}$; the workpapers show auditors adopting the generator's wrong intermediate value. The recompute tool, by contrast, is *not* strictly necessary here: no-tool ties the full system (n_{01}=5, n_{10}=0, $p = 0.063$; Table 1). Heterogeneous pools tie ($p = 0.88$) at 23% higher token cost; uniform purchasing ties the knapsack exactly here, reported as such. GSM8K arithmetic is easy enough that independence, not machinery, binds.

Repair and correction. Auditing is not repair, but flagged opinions route cheaply into a correction loop: on GSM8K, repair quality is channel-dependent, 87.5% fixed for recompute-driven adverse opinions against 5.9% for invariant-driven, regression rates reported alongside (Appendix R). The pre-registered H1 stress ablation flips the recompute tool at matched budget: flag-miss falls from 11.33% (34/300, tool off) to 3.00% (9/300, tool on; paired McNemar n_{01}=25, n_{10}=0, $p = 6.0 \times 10^{-8}$); batch 1 alone misses the α=0.05 confirmation (p=0.065); pooled, it confirms (16/600, 2.67%, p=0.003; Appendix R).

Auditor scale and distribution shift. A 3B auditor suffices for triage, tier error rates staying monotone (5.0%/53.4%/89.7%; Appendix R); the risk model is not memorizing the benchmark, train-fitted test-certified accuracy moving 94.6%→89.1%. Out-of-domain probes carry dual calibers in Appendix O (MBPP pass@8: 0.43 audit-caliber vs. 0.72 standard; conflating them would inflate the paper).

The full audit averages 150.2 tokens per problem; the heterogeneous variant costs 184.5 tokens for no measured benefit on this track.

6 Falsifying α

Guarantee machinery is only interesting if it can fail in public. We swept $\alpha \in \{0.01, 0.05, 0.10\}$ across every track×domain cell and tested each setting exactly (Table 3, Appendix E). Since the upgrade boundary and purchase strengths scale with the ratio $\alpha/\hat{\rho}$ (at a fixed ratio, varying $\hat{\rho}$ alone moves neither), this sweep doubles as a reliance-constant probe over the equivalent grid $\hat{\rho} \in \{0.5, 0.1, 0.05\}$ at α=0.05; the intermediate ratio $\alpha/\hat{\rho} = 0.25$ required new purchases and is run only in Track B′ (Table 4). Three regimes appear.

Regime 1, hold. At α=0.05 the guarantee holds on both distributions ($p \leq 0.0034$), and purchases respond: moving α from 0.10 to 0.05 on stress buys 753 extra procedures and upgrades 170, cutting empirical risk from 6.5% to 2.7%. *Regime 2, falsified.* At α=0.01 empirical audit risk exceeds the target on both distributions (3.4%, 6.5%), decisively ($p = 3.8 \times 10^{-12}$ and 1.3×10^{-19}); the two test directions jointly localize GSM8K audit risk in $(0.01, 0.05)$. The risk model cannot resolve claim difficulty at that tolerance. *Regime 3, vacuous.* At α=0.10 on stress the test passes ($p = 0.0017$) while coverage sits at 0.34: the target is "achieved" by refusing to certify; the dual caliber exposes the pass as hollow, which is why coverage is part of the guarantee.

Two lessons: α below the risk model's resolution floor fails detectably, so the floor should be measured and published per deployment. The floor is a published attribute of the fitted risk model (Track A 1.97×10^{-3}; Track B slice 4.10%, stress 0.46%; Appendix A), readable from a small held-out pilot fit; above it, a cheap pilot audit answers feasibility directly.

7 Track B: Auditing a Frontier API Model

Track B asks whether any of this survives contact with a black-box frontier model, run as a deployment case study with Track A as the controlled evidential base. We audited DeepSeek V4 Pro on September 9–10, 2026, on two workloads: a 500-problem GSM8K slice and the 600-problem stress set. Deployments change faster than evaluation cycles; assurance must be cheap enough to re-run per deployment, the regime Track B′ (§7.6) exercises.

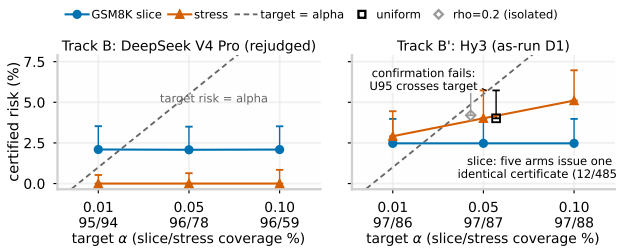


Figure 3: Certificates on the API tracks: certified risk with one-sided U_{95} whiskers against target α (dashed); tick rows give the target and (slice/stress) coverage. Track B: rejudged caliber (a sensitivity, not the certificate); Track B′: as-run D1. B′ slice degenerate (five arms, one certificate, 12/485); uniform (square) coincides with α=0.05; $\hat{\rho}$=0.2 (diamond) non-comparable, isolated. No interpolation; per-row provenance, sha-256 logged.

7.1 Model identity, sealed

Claims about a black-box endpoint are only as good as the identity evidence; ours stands on three legs (Appendix N): the model string `deepseek-v4-pro` on all 21,450 cached responses; one `system_fingerprint` across the collection; and request timestamps inside the documented service window (DeepSeek, 2026a). The bundle is archived with SHA-256 digests; one disclosed defect (a freeze timestamp 1h31m early) is bounded by the documented window.

7.2 The guarantee, purchased for real money

Table 2 (Appendix E) is Track B's central measurement, in two calibers binding different opinion-rule versions, never mixed. The pre-registered certificate binds the as-run rule (v2): coverage 93.4% at certified risk 2.14% (10/467), U_{95} 3.61%, on the slice; coverage 57.7% at zero observed errors (0/346), U_{95} 0.86%, on stress; both confirmation tests at α=0.05 survive Holm-2 ($p = 1.3 \times 10^{-3}$, 2.0×10^{-8}). The rejudged caliber (a deterministic glyph repair over the frozen workpapers) is a sensitivity, not the certificate: it recovers coverage the extraction artifact destroyed (Fig. 3), certified wrong counts unchanged (10 and 0).

The α grid (Appendix E) adds the third falsified cell: on the slice, the falsification test rejects H_0 : AR ≤ 0.01 ($p = 0.024$, n=477), while stress holds at all three levels. Certified risk stays flat as the target tightens (Fig. 3; 552/92/73 purchased procedures; plans are not nested across α): the guarantee here is carried by evidence quality, not volume. The slice calibrator is degenerate after label rechecking (38 as-run positives collapse to zero; Appendix A), so Track B purchasing rests on floors and deterministic evidence; Track A is where the non-degenerate calibrator lets the risk model earn its keep.

Comparators, with losses named. Under the pre-registered matched-coverage rule (A5b, rejudged caliber), the strongest single-model comparator carries 3.33% risk against ARA's 2.08% on the slice, and 1.23% against ARA's 0% on stress (wins). Self-consistency@8 carries 2.6% risk at 100% coverage against ARA's 2.08% at 96.2%, and 1.20% on stress against ARA's 0% at 78.2% coverage (wins on risk, coverage conceded). The no-isolation ablation loses at strength consistent with Track A; diversity purchasing (HB5) ties on both workloads. Two named losses: the main endpoint is a statistical tie in both domains, and uniform purchasing reaches 85.8% stress coverage at the same zero risk against ARA's 78.2%, a weak-dominance loss reported verbatim. Counting the flagged tail, slice error is 3.2% (16/500) against SC@8's silent 2.6% at full coverage; the flagged 3.8% carry 31.6% error (6/19), a 3.8% review load for the certified 96.2% at 2.08%. On stress, overall error 1.17% (7/600) matches SC@8's 1.20%, all seven errors adverse-flagged at a 21.8% review load. Two clamp-configured variants lose by $\leq 1.7 \times 10^{-4}$ (Appendix K).

7.3 Anatomy: where the errors actually live

Claim-level auditing sees a different model than answer-level scoring (Fig. 4, Appendix C). On the 500-problem GSM8K dev slice (distinct from the certificate slice), zero of 952 audited arithmetic claims are wrong (CP95 0.31%) while 20 of 500 final answers are (4.0%); on stress, 3 of 875 claims (0.34%) against 3 of 200 answers in the calibration split (1.5%) (Fisher $p = 4.3 \times 10^{-10}$ and 0.082,

descriptive since claims pool within problems). The inversion is the finding: intermediate claims are essentially clean; residual errors concentrate at composition and answer-extraction boundaries.

Channel attribution. All seven wrong stress answers were caught, all adverse driven by deterministic recomputation with materiality ($p_{\text{floor}} = 0.184$); CRED drove zero. On GSM8K, 16 certified-plus-escalated answers were wrong: 10 unqualified with no channel fire, three qualified via invariants, three via reperform; `test-78` was marked adverse while correct, a taxonomy-B false alarm (§7.4). Deterministic evidence is both the strongest detector and, through extraction, the only observed source of false condemnation.

7.4 Root cause: the audit caught the interface, not the arithmetic

We root-caused all 554 adverse-driving claims (Appendix C): 547 (98.7%) are extraction-side, 545 a single artifact (the generator renders minus as Unicode U+2212 in large numbers; the extractor normalized it to a space, merging a subtraction into a spurious operand), seven genuine generator errors. All extracted claims and as-run opinions re-derive with zero mismatches. Two conclusions follow: the frontier model's arithmetic is stronger than answer-level scoring suggests; and an assurance system that cannot distinguish "the model is wrong" from "my parser is wrong" manufactures adverse opinions. The mechanism also runs in reverse as a selection-bias channel (Appendix B); unparseable claims are never dropped but escalate, and the as-run 57.7% coverage is what that choice costs.

CRED as evidence. The isolated critic agrees with CPU counter-evidence on 15 of 22 GSM8K run-0 verdicts (7 unverifiable) and on 7 of 17 stress verdicts (8 unverifiable, 2 mismatches): by design it never condemns alone, and measured, it rarely even corroborates (Appendix P).

7.5 Leakage, cost, and continuity

Ten of the 16 wrong slice answers are shared with the 7B local model (stress comparison unavailable at those magnitudes): frontier and commodity models fail on overlapping structure, so a same-distribution auditor pool inherits blind spots. Admitting only CPU evidence raises coverage to 99.0% on the slice at 2.63% risk and 98.8% on stress at zero risk: the LLM-evidence tiers are conservative filters, not load-bearing detectors.

The entire Track B engagement cost ¥37.79 (¥0.0343 per audited problem; Appendix M); behavioral-continuity checks detect no serving change within the window (Appendix Q).

7.6 Track B′: a sealed reproduction on a second deployment

Track B binds a black-box frontier endpoint; to learn which claims travel, we re-ran the full protocol on a non-frontier generator, Tencent Hunyuan Hy3 via TokenHub (Tencent Cloud, 2026) (`hy3`): same prompts, problem sets, opinion rule, and parse caliber; hypotheses and verdict rules hash-anchored before the first call (Appendix F). Two disclosures: Hy3's default tier failed the parse gate ($0.75<0.95$), so we ran with reasoning switched off; the deployment ignores `seed` and silently drops $n > 1$, so SC@8 is eight salted calls.

Replicated. The certificate is issued on the slice (coverage 0.970, certified risk 12/485 = 2.47%, U_{95} 3.98%, Holm-2 adjusted $p \leq 0.008$; Fig. 3) but not on stress (U_{95} 5.73%; confirmation fails in both risk calibers, p=0.176 and 0.0501; a failure to confirm, not a near-pass). The comparator pattern repeats (beats entropy and random, loses to oracle), and the anatomy inversion is now significant (claim-level material error 1.24% vs. answer-level 14.17%, adjusted p=3.9×10^{-4}). The α knob recovers its monotone direction on stress (purchases 83/148/1,203) without reaching significance (adjusted p=0.315); the slice is flat because 356/500 problems expose no material claim, a pre-declared uninformative cell (Fig. 3).

Falsified. The channel-value inversion reverses: deterministic recomputation alone nails 4/85 wrong stress answers (4.7%) while isolated LLM evidence moves 48 risk flags, so channel values are generator-dependent, not universal. The capability gradient fails: Hy3's claim-defect rate (0/500) and paid-evidence efficacy (5/500 = 1.0%) fall below the frontier endpoint rather than between it and the 7B; Hy3's errors are setting-level (every step right, wrong quantities chosen), which claim-level recomputation cannot reach (band recall 18.8%). The glyph artifact vanishes (0/16

clean-parse; Hy3 emits plain ASCII), confirming it as a parser×glyph property. Yet error structure is partly generator-invariant: seven slice problems are wrong under all three generators, and all seven v4-pro stress errors reappear under Hy3. Finally, certificates bind deployments, not model names: re-serving `deepseek-v4-pro` on TokenHub reproduces 97.6%/97.7% of frozen answers but only 2/7 frozen adverse certificates.

8 Extraction Recall, Measured Externally

Certificates bind the extracted claim universe, so how much of a solution does the extractor see? We score the byte-frozen production rule on ProcessBench (Zheng et al., 2025) (400 GSM8K-style and 1,000 competition-style solutions, pinned revision), predictions hash-anchored. The as-is rule left 94.2%/97.0% of equations undecided; one registered normalization later, recall is 0.399/0.234 at precision 0.561/0.520, and the near-perfect predictions ($\geq$0.95) are decisively falsified (Appendix J). ProcessBench text is human-cleaned, so these are best-case bounds; production text can only read less.

9 Discussion

Three lessons generalize. *Independence is the scarce resource*: every large effect (isolation 10^{-19}, the CPU-evidence floor, shared-error leakage) is an independence statement; every null bought correlation, not information. *Assurance is a triple*: risk, bound, and coverage travel together; each falsified or vacuous regime was visible only because all three were on the table. *Audits audit the pipeline*: Track B's dominant error source was our own extraction layer. Practically: deploy CPU recompute first, add isolated LLM procedures where claims are not machine-checkable, never count generator self-assessment as evidence, and re-measure channel values per generator (§7.6); cost scales with purchasable claims, not problem count. Two directions matter most: raising extraction recall (normalization moved it from zero to 0.40; a learned extractor is next), since low recall surfaces intermediate-step hazards only through the final-answer gold, a safety-critical gap; and porting deterministic evidence beyond arithmetic, which re-enables the adverse tier.

10 Conclusion

Accuracy describes the model; assurance describes the answer. ARA turns LLM quality control into an instrument that states what it certified, at what risk, coverage, and cost, and reports its own falsifications. It survived a frontier black-box endpoint, caught its own extraction layer manufacturing evidence, and failed at α=0.01: the strongest evidence the measurement is real.

Limitations

The guarantee is distribution-conditional and must be re-measured on new workloads; the risk model's resolution floor bounds usable α; Track B's rejudged calibrator is degenerate (Appendix A); certificates cover the extracted claim universe only, with recall bounded at best-case $\leq$0.40 on human-cleaned text (§8). The certificate is validated on arithmetic tracks only: without machine-checkable claims, adverse opinions become unreachable and the taxonomy collapses to unqualified versus flagged, reported as lower usable α or coverage rather than hidden (Appendices L, O). The certificate inherits gold-label quality: mislabeled gold inflates k when unmatched and hides error when reproduced, so weak-label domains need a label-audit step first (Appendix A). A 5% gold flip rate moves the four frozen bounds $\{4.9, 3.6, 4.0, 5.7\}\%$ to $\{10.1, 9.3, 9.6, 11.1\}\%$ (contamination) or $\{4.7, 3.5, 3.8, 5.5\}\%$ (concealment) (`gold_noise_sensitivity_v2.csv`). Wall-clock latency was not instrumented (Appendix M); $\hat{\rho}$ is fixed at 0.1 and d_0 chosen, stability argued from the root-cause taxonomy (§3); matched-compute self-consistency runs on the API tracks only; Track B$'$ audits Hy3 with reasoning disabled (the default tier failed the parse gate), and that deployment ignores `seed` and drops n>1, so sampling-based comparators cannot be replayed exactly; certificates do not transfer across serving stacks even under the same model identifier (§7.6): Track B is a deployment case study, not the primary evidential base.

Ethics statement. This work studies quality control of generated text and uses public benchmarks and a commercial API under its terms of service. No human subjects or personal data are involved. Assurance systems can create false comfort if their coverage caliber is ignored; we mitigate by making coverage a first-class output and by publishing falsified regimes. A specific misuse follows: quoting the 95% upper bound while dropping that it is conditional on the realized certified subset of a measured distribution, or dropping the coverage at which it was earned, inverts this paper's central norm; we therefore state the conditioning in the abstract, report risk only as a (coverage, risk, bound) triple, require re-measurement per deployment and serving stack, and publish every falsified or vacuous regime alongside the holds. All API spend is disclosed in Appendix M.

Reproducibility statement. All Track A artifacts are frozen and hashed, and the frozen pipeline was re-executed end-to-end under a hash-anchored pre-registration with field-level gates (Appendices H–I); the aggregate manifest digest is `d67f12d0...` with a dated identity-archive addendum. The analysis phase makes zero API/GPU calls; every figure and table regenerates from the frozen CSVs byte-identically. Track B's 21,450 cached responses and request logs are archived with SHA-256 digests; identity rests on the three legs of §7.1, and the full protocol is replicated end-to-end on a second serving stack (E3, Appendix G). Track B′ is re-runnable against a live endpoint under its own hash-anchored pre-registration (29,094-member manifest); the extraction-recall study is fully offline against a pinned dataset revision; both ship per-phase spend ledgers. Two offline re-analyses regenerate from the frozen artifacts without API access: the budget–risk–coverage table behind Figure 5 (`pareto_budget_risk_coverage_v2.csv`) and the gold-noise sensitivity grid (`gold_noise_sensitivity_v2.csv`), both carrying per-row provenance columns and logged sha-256 manifests. Code, configs, and the evidence manifest accompany the submission.

AI-use statement. The authors used a large language model assistant for drafting and editing of this paper's text, for data analysis and experimental-workflow assistance under author direction, and for bookkeeping of the experimental record (an independent digest pipeline, aggregate digest `d67f12d0...`, with byte-identical re-run verification). All technical content, experimental design, and data are the authors' own; every AI-assisted analysis was independently re-verified by the authors, and every externally verifiable claim, citation, and quoted provider statement was checked against primary sources and official pages. The authors take full responsibility for the paper.

REFERENCES

Anastasios N. Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023.

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

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*, 2021.

Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael Jordan. Distribution-free, risk-controlling prediction sets. *Journal of the ACM*, 68(6):1–34, 2021. doi: 10.1145/3478535.

Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W. Cohen. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL `https://openreview.net/forum?id=YfZ4ZPt8zd`.

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

C. J. Clopper and E. S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

DeepSeek. DeepSeek API documentation. `https://api-docs.deepseek.com/`, accessed 2026-09-10, 2026a.

DeepSeek. DeepSeek V4.1 Flash release announcement, including the scheduled retirement of V4 Pro. `https://api-docs.deepseek.com/news/news260910/`, accessed 2026-09-10; snapshot archived as EV-20260910-01, 2026b.

DeepSeek. DeepSeek API models & pricing page, footnote withdrawing the V4 Pro retirement. `https://api-docs.deepseek.com/quick_start/pricing`, footnote added 2026-09-11 (service continues past 2026-09-14, billing unchanged), page accessed 2026-09-19; snapshot archived as EV-20260919-01, 2026c.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(53):1605–1641, 2010.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. PAL: Program-aided language models. In *International Conference on Machine Learning (ICML)*, 2023.

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 (ICML)*, pp. 2151–2159, 2019.

Ryan Greenblatt, Buck Shlegeris, Kshitij Sachan, and Fabien Roger. AI control: Improving safety despite intentional subversion. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 16295–16336, 2024.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, 2025. doi: 10.1145/3703155.

International Auditing and Assurance Standards Board. ISA 315 (revised 2019), identifying and assessing the risks of material misstatement. International Standard on Auditing, effective for audits of periods beginning on or after December 15, 2021, 2019.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri 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.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattva Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Quinn McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947.

Jakob Mökander, Jonas Schuett, Hannah Rose Kirk, and Luciano Floridi. Auditing large language models: A three-layered approach. *AI and Ethics*, 4(4):1085–1115, 2024. doi: 10.1007/s43681-023-00289-2.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 33–44, 2020.

Tencent Cloud. Tencent Hunyuan Hy3 on TokenHub: model documentation and pricing. `https://tokenhub.tencentmaas.com/`, accessed 2026-09-13, 2026.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher 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*, pp. 5433–5442, 2023.

Jonathan Uesato, Nate Kushman, Ramana Kumar, H. Francis Song, Noah Y. Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. Solving math word problems with process- and outcome-based feedback. *arXiv preprint arXiv:2211.14275*, 2022.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-Shepherd: Verify and reinforce LLMs step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 9426–9439, 2024.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, Fei Xia, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837, 2022.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380, 2018.

Chujie Zheng, Zhenru Zhang, Beichen Zhang, Runji Lin, Keming Lu, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. ProcessBench: Identifying process errors in mathematical reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1009–1024, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.50.

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 Statistical Procedures

Risk model. Track A uses a 4-feature logistic model (mean token log-probability, step index, log-magnitude of operands, predictor entropy) fitted on 1,011 held-out claims; base rate 9.89×10^{-4} (a single positive event, so the fit is coarse), prediction floor 1.97×10^{-3}. Track B (GSM8K slice) uses a 3-feature model (entropy-bank feature, step index, log-magnitude); base rate 4.00%, floor 4.10%, ECE 0.020. Floors prevent division by zero in equation 1 and bound $\mathrm{DR}^*(c)$ from above, so low-risk claims keep a finite audit requirement rather than an unconstrained one. Token accounting: the CPU recompute channel carries no token cost, so the no-tool arm's mean audit tokens equal the full arm's 150.2 by construction (Table 1); that ablation changes evidence quality, not spend.

Bounds and tests. One-sided Clopper–Pearson 95% upper bounds use $U_{95} = \mathrm{Beta}^{-1}(0.95;\, k+1,\, n-k)$ (Clopper & Pearson, 1934). $H_0 : \mathrm{AR} \leq \alpha$ is tested with the exact one-sided binomial on the certified subset (Track B's caliber) or the full test population (Track A's pre-registered flag-miss caliber). Paired comparisons use exact McNemar (McNemar, 1947); multiplicity uses Holm within pre-registered families (Holm, 1979). Rate comparisons between claim-level and answer-level errors use Fisher's exact test, two-sided. The matched-coverage rule (A5b) sweeps the comparator over a 20-point coverage grid, interpolates linearly (tolerance 10^{-12}), and reports wins and losses symmetrically.

Family rosters. Multiplicity control uses two pre-registered families. The primary family (Holm-2) contains the two Track B confirmation tests (GSM8K slice and stress, each at α=0.05, $H_0 : \mathrm{AR} \geq \alpha$, confirmation direction). The secondary family (Holm-4) contains the four Track A ablation McNemar tests. The α-grid falsification tests of §6 are reported per cell, unadjusted, with all twelve cells shown; each tests a separate pre-specified α target rather than a shared null.

Rejudging protocol. Rejudging was a deterministic re-derivation over the frozen workpapers, not a new human or model pass, and it made zero API and zero GPU calls, consistent with the analysis-phase freeze: all 4,360 extracted arithmetic claims were independently re-derived with zero mismatches, and all 1,100 as-run opinions were rebuilt with zero mismatches. Ground truth never moved (certified wrong counts 10 and 0 before and after; the 7 stress errors sit in the adverse tier throughout). What moved were opinions whose adverse drivers traced to the extraction artifact (Appendix C): on the main arm, GSM8K-slice opinions went from 467/10/23 (unqualified/qualified/adverse) to 481/18/1 (22 of 500 rows changed), and stress from 346/32/222 to 469/124/7 (215 of 600). Rejudged calibrators were never used for scoring or planning, only for reporting certificate status. The pre-registered certificate binds the as-run rule, and its Holm-2 confirmation tests pass on as-run counts as well ($p = 1.3 \times 10^{-3}$ GSM8K slice, 2.0×10^{-8} stress). Risk invariance holds arm by arm, not only on the main arm: certified wrong counts are identical across calibers in every arm (full: 10/10 GSM8K slice, 0/0 stress; no-isolation: 11/11), with zero Δ throughout.

Calibration status. After label rechecking, the Track B GSM8K-slice calibrator is degenerate (38 detected positives as-run collapse to 0 rejudged; base rate and floor collapse accordingly), so risk estimates there rely on floors plus deterministic evidence. The stress calibrator is *not* degenerate (160 as-run positives reduce to 3 rejudged; floor 0.46%). An intermediate summary table carried a copy-paste error stating otherwise for stress; the artifact record and this paper state the correct, non-degenerate status.

B The D1 Guarantee: Statement and Scope

D1 (Definition 1) is an empirical certificate, not a distribution-free bound. Its validity rests on: (i) opinions are issued before ground truth is consulted (enforced by pipeline order); (ii) the certified subset is defined by the opinion rule, not post hoc; (iii) U_{95} is exact for i.i.d. Bernoulli error on the certified subset; within-problem claim dependence does not enter it (sampling unit below). We do not claim conformal-style marginal coverage: the abstention mechanism is correlated with errors by design, which is the point of an audit. The falsification battery of §6 is the operative check: across twelve track×domain×α cells it failed exactly three, all at α=0.01.

Sampling unit. The bound's Bernoulli unit is the answer, one trial per problem; problems are independent draws by benchmark construction, so claim-level dependence within a problem does not enter U_{95}. Claim-level rates elsewhere (§7.3) pool claims within problems and are reported descriptively for that reason. Effective sample sizes are the certified-subset sizes themselves (n=481, 469), reported alongside every bound. A 2×10^4-draw parametric bootstrap over certified answers puts the 95th percentile of the slice risk at 3.12% against the exact U_{95} of 3.50%; the exact bound is the conservative one and is what we report.

Conditional validity and the selection channel. The bound is conditional on the certified subset that the opinion rule selects, and opinions correlate with errors by design, so the certified subset is not a random draw of the claim universe; U_{95} therefore describes the realized subset and does not license counterfactual claims about subsets a different rule would have certified. A subtler channel runs through extraction: the U+2212 artifact of Appendix C concentrates in long-number subtraction, exactly where true generator errors concentrate, so artifact-driven and true-error-driven adverse opinions arrive correlated in location. Two design choices keep this channel from contaminating the certificate: unparseable or unextractable claims are never dropped from the universe (they escalate instead), and every adverse opinion is root-caused into generator error versus extraction artifact before any rate is quoted.

C Root-Cause Analysis of Adverse Evidence (B2W)

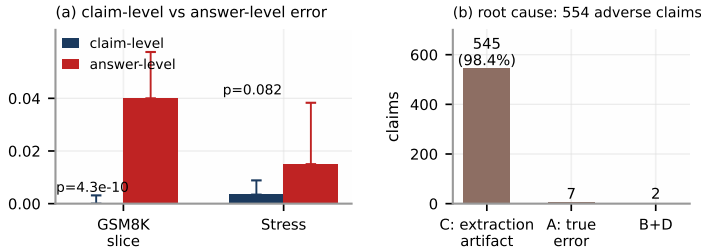


Figure 4: (a) Claim-level (dark) vs. answer-level (red) error on Track B; whiskers: one-sided CP95. (b) Root cause of all 554 adverse-driving claims: 547/554 extraction-side artifacts (545 from U+2212 normalization), 7 genuine generator errors.

All 554 adverse-driving claims were root-caused into four classes: A (generator true error, 7), B (extraction false alarm, 1), C (extraction artifact from U+2212 minus-sign normalization, 545), D (boundary ambiguity, 1). Class shares: stress 522 claims with 98.7% artifact; GSM8K 32 claims. Verification: the 4,360 extracted arithmetic claims were independently re-derived with zero mismatches; all 1,100 as-run opinions were rebuilt with zero mismatches. Exemplars: `stress-0367` (A), `test-78` (B), `test-410` (D), `stress-0033` (C). The mechanism for C: the generator emits U+2212 (minus sign) inside large-number expressions such as $2{,}471{,}500 - 64{,}259$; the extractor mapped U+2212 to a space; adjacent operands concatenated into a spurious number, which the CPU then correctly recomputed as inconsistent, manufacturing a false adverse. A 116-row audit sample over 60 problems confirmed the class balance (C 107, A 7, D 1, B 1).

D Channel Attribution and Opinion Tiers

Stress: 7 wrong final answers, all adverse, all driven by `recompute_det_material` ($p_{\text{floor}} = 0.184$); CRED-driven adverse: 0/7. GSM8K slice: 16 wrong answers; 10 unqualified with no channel fire (incl. `test-777`), 3 qualified via invariants (`test-1046`, `test-380`, `test-780`), 3 qualified via reperform (`test-371`, `test-754`, `test-984`); `test-78` is the taxonomy-B false adverse. Track A opinion-tier error: unqualified 3.83%, qualified 62.6%, adverse 90.0% (monotone). Track B GSM8K-slice tier errors: 10/481 unqualified, 6/18 qualified (33.3%), 0/1 adverse; the monotone property fails on this slice (the qualified tier exceeds the unqualified tier by design of escalation, and the single adverse is correct), and we report it as an honesty item rather than reclassifying opinions post hoc. Stress tiers: 0/469 unqualified, 0/124 qualified, 7/7 adverse (monotone). Error enrichment of flagged (qualified and adverse) over unqualified: 15.2× rejudged, 8.5× as-run.

E CERTIFICATE AND FULL α-SWEEP TABLES

Table 2: Track B certificates (HB1), dual caliber. As-run (v2) is the pre-registered certificate; rejudged (v2.1) is a deterministic coverage-recovery sensitivity with certified wrong counts unchanged. p: one-sided confirmation test against α=0.05; each caliber survives Holm-2. Per-α grid: Table 3.

Workload	Caliber	Coverage	k/n (risk)	U_{95}	p
GSM8K slice	as-run (v2, pre-registered)	0.934	10/467 (0.0214)	0.0361	1.3×10^{-3}
GSM8K slice	rejudged (v2.1, sensitivity)	0.962	10/481 (0.0208)	0.0350	8.4×10^{-4}
Stress	as-run (v2, pre-registered)	0.577	0/346 (0.0000)	0.0086	2.0×10^{-8}
Stress	rejudged (v2.1, sensitivity)	0.782	0/469 (0.0000)	0.0064	3.6×10^{-11}

Table 2 gives the Track B dual-caliber certificate; Table 3 reports all twelve track×domain×α cells with both test directions. The pre-registered rule picks the direction per cell: falsification ($H_0 : \mathrm{AR} \leq \alpha$) when the empirical rate exceeds α, confirmation ($H_0 : \mathrm{AR} \geq \alpha$) otherwise; per-cell p values in the main text use that direction. Track A's target is full-set audit risk over all test problems; Track B's is certified-subset risk (D1), each cell with its own purchases and certified subset (rejudged caliber). Track B as-run opinion populations: GSM8K slice 466/467/467 unqualified at $\alpha = \{0.01, 0.05, 0.10\}$ with 23 adverse; stress 358/346/346 with 222 adverse. Purchase plans are not nested across α (an honesty item). GSM8K-slice wrong-answer tier split (10 unqualified, 6 qualified, 0 adverse) is constant across cells. Multiplicity sensitivity: the twelve cells are reported per cell because each audits a distinct pre-registered target with its own purchases; a Holm-12 adjustment over the falsification-direction family leaves the two Track A α=0.01 rejections intact (adjusted $p \leq 4.2 \times 10^{-11}$) and attenuates the slice cell ($0.024 \rightarrow 0.24$). The qualitative lesson (α=0.01 over-demands current risk-model resolution) is carried by the Track A cells; the slice cell is supportive, not load-bearing. Coverage on Track B stress falls as α relaxes (0.943/0.782/0.590): a looser target raises every claim's allowable detection risk $\mathrm{DR}^*(c)$, so the allocator buys fewer procedures (shifting α from 0.10 to 0.05 buys 753 extra; §6), more claims stay unverified, unverified claims escalate to qualified rather than certify, and unqualified coverage collapses. The knob moves coverage, not certified risk; the dual-caliber report is what makes this visible.

Table 3: All twelve α cells, both directions. p_{fals}: exact one-sided binomial against $H_0 : \mathrm{AR} \leq \alpha$; p_{conf}: against $H_0 : \mathrm{AR} \geq \alpha$. Bold: the pre-registered direction and verdict. Track A targets full-set flag-miss risk; Track B targets certified-subset risk. The denominators differ accordingly: the rate column divides by the full population N for Track A rows and by the certified subset $|S|$ for Track B rows. On Track A stress at α=0.05 the conditional certified risk is 9.58% (Table 11); the confirmation holds under the pre-registered full-set caliber.

Track	Domain	α	k/n (rate)	p_{fals}	p_{conf}	Verdict	Coverage
A	GSM8K	0.01	45/1319 (0.0341)	$\mathbf{3.8 \times 10^{-12}}$	1.00	**falsified**	0.891
A	GSM8K	0.05	45/1319 (0.0341)	1.00	$\mathbf{3.4 \times 10^{-3}}$	**holds**	0.890
A	GSM8K	0.10	45/1319 (0.0341)	1.00	$\mathbf{7.5 \times 10^{-20}}$	**holds**	0.891
A	stress	0.01	39/600 (0.0650)	$\mathbf{1.3 \times 10^{-19}}$	1.00	**falsified**	0.343
A	stress	0.05	16/600 (0.0267)	1.00	$\mathbf{3.2 \times 10^{-3}}$	**holds**	0.278
A	stress	0.10	39/600 (0.0650)	1.00	$\mathbf{1.7 \times 10^{-3}}$	**holds (vacuous)**	0.340
B	GSM8K slice	0.01	10/477 (0.0210)	**0.024**	0.990	**falsified**	0.954
B	GSM8K slice	0.05	10/481 (0.0208)	1.00	$\mathbf{8.4 \times 10^{-4}}$	**holds**	0.962
B	GSM8K slice	0.10	10/478 (0.0209)	1.00	$\mathbf{7.4 \times 10^{-12}}$	**holds**	0.956
B	stress	0.01	0/566 (0.0000)	1.00	$\mathbf{3.4 \times 10^{-3}}$	**holds**	0.943
B	stress	0.05	0/469 (0.0000)	1.00	$\mathbf{3.6 \times 10^{-11}}$	**holds**	0.782
B	stress	0.10	0/354 (0.0000)	1.00	$\mathbf{6.3 \times 10^{-17}}$	**holds**	0.590

Coverage moves with the ratio $\alpha/\hat{\rho}$ while certified risk stays flat (GSM8K slice) or zero (stress): purchases 552/92/73 (slice) and 1,804/1,456/1,725 (stress, reserved 1,804/2,215/1,987) at $\alpha = \{0.01, 0.05, 0.10\}$. Since purchase strengths scale with $\alpha/\hat{\rho}$, these cells double as the measured reliance-constant sensitivity: at α=0.05 they instantiate $\hat{\rho} \in \{0.5, 0.1, 0.05\}$. The response is not flat in coverage or cost, which is exactly the efficiency price the ratio argument predicts; what stays invariant is measured risk.

Track B′ ran the ratio grid again on both domains, adding the intermediate ratio $\alpha/\hat{\rho} = 0.25$ ($\hat{\rho} = 0.2$) that the frozen track left unrun (Table 4); that cell has no frozen counterpart and is marked non-comparable. On the slice the grid is flat by construction: 356/500 problems expose no material claims, so purchases (20 at α=0.01, zero elsewhere) have no object to act on and every arm certifies the same 485 answers with the same 12 wrong (a pre-declared uninformative cell, not evidence for or against the knob). On stress the knob moves: purchases 83/148/1,203 across the registered α

cells (0.10/0.05/0.01), with empirical audit risk 4.50%/3.50%/2.50% in the pre-registered monotone direction, while certified coverage *rises* as the ratio relaxes (0.8583/0.8733/0.8717/0.8817), falsifying the registered monotone-coverage prediction (P-R3); the frozen track's own coverage grid was likewise non-monotone on the slice and strictly decreasing on stress.

Table 4: Track B′ ratio grid (Hy3, clean-parse caliber). Ratio $= \alpha/\hat{\rho}$; the 0.25 cell sets $\hat{\rho} = 0.2$ and has no frozen-track counterpart. $\mathrm{AR}_{\mathrm{emp}}$: full-population flag-miss rate k/N (the H-R2 caliber); certified risk and U_{95} condition on the certified subset (D1). Uniform tie: arm-by-arm identical flags on both domains.

Domain	Arm (ratio)	Purchases	Coverage	$\mathrm{AR}_{\mathrm{emp}}$	Certified k/n	U_{95}
Slice (N=500)	α=0.01 (0.10)	20	0.970	0.0240	12/485 (0.0247)	0.0398
	α=0.05, $\hat{\rho}$=0.2 (0.25)	0	0.970	0.0240	12/485 (0.0247)	0.0398
	α=0.05 (0.50)	0	0.970	0.0240	12/485 (0.0247)	0.0398
	α=0.10 (1.00)	0	0.970	0.0240	12/485 (0.0247)	0.0398
	uniform (0.50)	0	0.970	0.0240	12/485 (0.0247)	0.0398
Stress (N=600)	α=0.01 (0.10)	1,203	0.8583	0.0250	15/515 (0.0291)	0.0445
	α=0.05, $\hat{\rho}$=0.2 (0.25)	145	0.8733	0.0367	22/524 (0.0420)	0.0594
	α=0.05 (0.50)	148	0.8717	0.0350	21/523 (0.0402)	0.0573
	α=0.10 (1.00)	83	0.8817	0.0450	27/529 (0.0510)	0.0697
	uniform (0.50)	152	0.8717	0.0350	21/523 (0.0402)	0.0573

F TRACK B′: SEALED REPRODUCTION RECORD

Protocol. Object: Tencent Hunyuan Hy3 served by Tencent Cloud TokenHub (model id `hy3`) (Tencent Cloud, 2026), chosen as a deliberately non-frontier generator. Configuration: T=0, max_tokens 768, frozen prompts and problem sets (GSM8K slice n=500; stress n=600), opinion rule v2, single clean-parse caliber (parser v2.1). The pre-registration (hypotheses H-R1..H-R5 transcribed verbatim from the frozen track's B-PRE with two restricted rewrites, directional predictions P-R2..P-R6, and the verdict rules) was hash-anchored (SHA-256 `d11a484c...`) before the first API call; the extraction-recall and stack-migration amendments (`73feee4b...`, `737cb87c...`, `a85f50d2...`) likewise precede their measurements. Protocol disclosures: (i) Hy3's default serving tier failed the pre-registered parse gate (0.75 against 0.95 on a 20-problem smoke; failure records retained), so all generation ran with `reasoning_effort=none`, which measurably switches thinking off (zero reasoning tokens); the audit covers Hy3's non-reasoning mode only. This suppresses the intermediate traces ARA audits, so the low claim prevalence and the weak CPU channel of this track are properties of a direct-completion deployment, not evidence about reasoning-enabled behavior. (ii) `seed` is not honored (probe: 4 distinct outputs in 5 draws). (iii) $n > 1$ is silently ignored, so SC@8 is eight salted independent calls. (iv) An abandoned earlier arm against a different provider's deployment (2,742 calls) is archived separately and never merged into any certificate. (v) The certificate bounds reported in the track's ledger file were computed with the two-sided Clopper–Pearson formula; the paper reports the pre-registered one-sided U_{95} throughout (slice 3.98%, stress 5.73% at α=0.05); every verdict is identical under both formulas.

Certificates. Table 5 restates the working point in both risk calibers. The full-population caliber is the pre-registered H-R1 test basis; the certified-subset caliber is the D1 triple a deployer quotes. Verdicts coincide under both. Why stress fails to confirm: Hy3's errors are setting-level (every step right, wrong quantities chosen), which claim-level evidence cannot reach (detection band 18.8%; §7.6).

Table 5: Track B′ working point (α=0.05), both risk calibers. p_{conf}: exact one-sided confirmation test against α=0.05; Holm-2 adjusted within the two-domain family per caliber. The stress guarantee is *not confirmed*: we report the measured triple, per D4. On the last row the raw and adjusted values coincide because the step-down maximum $\max(2p_{(1)}, p_{(2)})$ is attained by $p_{(2)}$.

Domain	Caliber	Coverage	k/n (risk)	U_{95}	p_{conf} (raw / Holm-2)	Verdict
Slice	certified subset	0.970	12/485 (0.0247)	0.0398	0.0040 / 0.0080	confirmed
Slice	full population	—	12/500 (0.0240)	—	0.0026 / 0.0052	confirmed
Stress	certified subset	0.872	21/523 (0.0402)	0.0573	0.176 / 0.176	not confirmed
Stress	full population	—	21/600 (0.0350)	—	0.0501 / 0.0501	not confirmed

Verdicts. Table 6 adjudicates every registered hypothesis and prediction. Multiplicity: primary family Holm-2 (H-R1, two domains); secondary family A Holm-3 (two H-R2 paired contrasts + H-R3), raw $\{0.105, 1.0, 1.0\}$, adjusted $\{0.315, 1.0, 1.0\}$; secondary family B Holm-4 (P-R2..P-R5, stress), raw $\{1.0, 1.0, 9.7\times10^{-5}, 0.942\}$, adjusted $\{1.0, 1.0, 3.9\times10^{-4}, 1.0\}$. P-E1..P-E4 (§8) and

P-T1/P-T2 are threshold predictions adjudicated by their registered rules. Exact ties are recorded as $p{=}1$ by convention.

Table 6: Track B′ verdict table. "holds"/"falsified" are by the pre-registered criterion texts; directional predictions quote the registered thresholds. OP1/OP2 are the two pre-registered operating points (risk-oriented and coverage-oriented).

Item	Criterion (abbreviated)	Observed	Verdict
H-R1 slice	$\mathrm{AR}_{\mathrm{emp}} \leq 0.05$, one-sided binomial	12/500, $p{=}0.0026$ (adj. 0.0052)	holds
H-R1 stress	same	21/600, $p{=}0.0501$	not confirmed
H-R2 stress	purchases non-decreasing, risk non-increasing as α tightens	83/148/1,203; 4.5/3.5/2.5%; McNemar adj. $p{=}0.315$	not confirmed (direction restored, n.s.)
H-R3	knapsack $\leq$ uniform, paired McNemar	exact tie (21=21)	tie
H-R4	OP1 risk below entropy/random at matched coverage	wins both; loses to oracle (0.0144/0.0201)	holds; oracle loss disclosed
H-R5	OP1 risk $\leq$ OP2 risk, both domains[†]	0.0247$\leq$0.0360; 0.0402$\leq$0.1182	holds
P-R2	$\geq$80% of wrong stress answers det.-only; paid LLM evidence moves 0 flags	det.-only 4/85 (4.7%); LLM moved 48, paid-alone 11	**falsified**
P-R3	coverage monotone non-increasing as ratio relaxes; risk flat	0.8583/0.8733/0.8717/0.8817 (rising); risk flat	**falsified**
P-R4	claim-level material error $\leq$ 1/5 of answer-level	1.24% vs. 14.17%, ratio 0.088, adj. $p{=}3.9{\times}10^{-4}$	**holds**
P-R5	glyph-artifact share of adverse drivers $\leq$5%	as-run 2/18 = 11.1% (clean-parse 0/16)	**falsified**
P-R6	Hy3 strictly between 7B and v4-pro on claim defects (M1) and paid-evidence efficacy (M2)	M1 = 0/500 (CP95 [0, 0.74%]); M2 = 5/500 = 1.0% (CP95 [0.33, 2.32%]), below v4-pro's 2.4%	**falsified**; band recall 18.8%
P-T1	stack migration: answer agreement $\geq$95% per domain	0.9760 / 0.9767	holds
P-T2	all 7 frozen adverse certificates recur on the new stack	2/7 (5 frozen errors do not recur; 6 new candidates)	**falsified**

[†]The track's automated verdict field recorded *lose* for H-R5 under a stricter dominance reading (OP1 must also match OP2's coverage); the pre-registered criterion text is a risk comparison, and the numbers satisfy it in both domains (the qualified tier errs at 40.0%/78.7% against the unqualified tier's 2.5%/4.0%). We adjudicate by the criterion text and record the discrepancy in the evidence map.

Equal-compute comparator. At matched token budgets ($k{=}2$ slice, $k{=}3$ stress), self-consistency majority voting scores 0.966/0.870 against ARA's 0.964/0.858: ARA does not beat voting at equal compute on this generator. What voting does not provide is the certificate: a certified subset with a risk bound, loud flags on the remainder, and a per-claim evidence trail.

Stack migration (P-T1/P-T2). Re-serving model id `deepseek-v4-pro` on TokenHub (1,100 logged calls, ¥4.82; one is a cache reuse on the duplicated stress item `0086v1`/`0103v1`, leaving 1,099 returned responses, and the envelope `model` field is exact on 1,099/1,099): answer agreement with the frozen track is 0.9760 (slice) and 0.9767 (stress), so P-T1 holds. Only 2 of the 7 frozen stress adverse pairs recur (`0445v1`, `0459v2`); the other five frozen errors do not reappear, and the new stack produces six adverse candidates of its own (overlap two), so P-T2 fails: answer distributions transfer at ≈97.6%, certificate-level error events do not. The asymmetry is mechanical: adverse certificates are tail events, so 97.6% answer-level agreement still leaves large relative room for turnover in a 7-of-600 event set. Two stress rows (`0086v1`, `0103v1`) share one problem text and necessarily agree; conclusions are unchanged when they are counted once.

Cost. Track B′ total: ¥15.43, of which the engagement (generation, audit, re-audit phases) is ¥10.61 across 20,795 logged calls and the stack-migration probe ¥4.82 (1,100 calls); a mid-run phase accounted four rows under by ¥0.0022 (ledger reconciliation, disclosed in the record). The abandoned earlier arm's spend (1,690,981 tokens on a different provider) is archived separately and not added to this ledger. The frozen Track B ledger is unaffected (¥37.79).

G TRACK B ON A SECOND STACK: FULL-PROTOCOL REPLICATION (E3)

Object and protocol. E3 re-runs the entire frozen Track B protocol — CoT evaluation and calibration, SC@8 banks, and the audit channels including CRED — against the model string `deepseek-v4-pro` as served by Tencent Cloud TokenHub: the same model string on a different

serving stack, and every conclusion binds that deployment only. The engagement ran under a hash-anchored pre-registration (base document plus two amendments and one addendum, each anchored before the calls it governed). Request-body identity with the frozen track was proven at zero cost before each level: 1,800/1,800 pass1 and calibration cache keys, 14,400/14,400 bank-draw keys, and the L5 gate (2,200/2,200 base-channel keys and 2,992/2,992 CRED keys reproduced against the frozen response cache). Prompt-token totals came out exactly equal at every level (77,322; 618,576; 918,064). No `seed` was sent (the frozen configuration has none); `thinking` was disabled as in the frozen track.

Answer-level transfer. Table 7: raw agreement counts on matched problems. Majority-voted SC@8 answers agree more than single-shot CoT (0.984–0.998 against 0.972–0.985), the denoising one expects; correctness deltas are small and bidirectional. These rates sit in the same band as the one-shot migration probe of §7.6 (0.9760/0.9767), now measured end-to-end.

Table 7: E3 versus frozen Track B, answer-level agreement on matched problems (raw counts). 2×2: both correct / E3-only / frozen-only / neither. The 20-problem smoke (calibration subset) agreed 19/20 on both answer and correctness.

Arm	Collection	n	Answer equal	Correct equal	E3 correct	Frozen correct	2×2 (b/e/f/n)
CoT (T=0)	api500	500	486 (0.9720)	487 (0.9740)	477	484	474/3/10/13
CoT (T=0)	stress600	600	590 (0.9833)	591 (0.9850)	592	593	588/4/5/3
CoT calib	dev500	500	486 (0.9720)	489 (0.9780)	483	480	476/7/4/13
CoT calib	stresscal	200	195 (0.9750)	195 (0.9750)	198	197	195/3/2/0
SC@8 calib	dev500	500	497 (0.9940)	499 (0.9980)	485	486	485/0/1/14
SC@8 calib	stresscal	200	198 (0.9900)	198 (0.9900)	199	199	198/1/1/0
SC@8	api500	500	492 (0.9840)	492 (0.9840)	485	487	482/3/5/10
SC@8	stress600	600	592 (0.9867)	593 (0.9883)	597	594	592/5/2/1

Length inflation. The new stack writes longer completions from byte-identical prompts: κ (E3/frozen completion tokens) is 1.115 on api500 and 1.364 on stress600 at T=0 (merged 1.2419), and 1.117/1.359 on the SC@8 arm — domain-dependent, stable across temperatures, and matching the independent SC-5 measurement (1.2421) to 0.019% relative. All pre-registered behavioral thresholds (the 4a–4d drift rules) held at every level; prompt lengths were exactly equal everywhere (Δprompt $= 0$).

Audit-channel inputs. The CRED run-index distribution matches the frozen track exactly ({0:2231, 1:333, 2:238, 3:190}). One semantics disclosure, stated once and applying everywhere: *repeat* CRED *draws (*`run`$\geq$1*) are replays of run 1, not independent samples* — the frozen track's cache key is unsalted, so a repeat request returns run 1's stored response, and E3 reproduces that behavior by construction (pre-registered option A). On the frozen side, of the 761 `run`$\geq$1 rows only 428 are true replays (the other 333 are fresh T=0.7 decodes); all 761 are consumed (779 consumption slots across arms and workpapers), no consumed replay's verdict ever differed from run 1's (0 counterexamples), so replays enter slot counts (`n_cred_runs`, 4,747 taps across the eight arm files, 9.01% replays) but cannot flip any claim's `cred_det`. No consistency or stability statistic anywhere consumes the repeat draws.

Identity. Four pre-registered legs. The string leg held on all 3,621 scanned rows (no successor string); the fingerprint leg is vacant — TokenHub returns no `system_fingerprint`, registered as unavailable, not passed; the probe leg passed; the behavioral leg never tripped. The provider announced this model's retirement on 2026-09-10 (to take effect 2026-09-14) and rescinded it the next day, service continuing at unchanged billing (DeepSeek, 2026b;c); the string persisted throughout, 2026-09-15 included (20/20). No rerouting took place, and the string could not have detected a silent one anyway (the frozen track's own routing precedent), so the behavioral leg remains a detector, not a proof; the episode is itself the volatility this paper instruments, a scheduled replacement announced and withdrawn before its effective date, and the evidence binds this deployment only.

Incidents and repairs. L3 stopped on a provider-quota HTTP 402 (account billing, not identity or data; resumed from cache with zero rework and no retry loop). L5 had a real defect: the runner's thread pool lacked a per-key lock, so same-key tasks each posted — 379 duplicate billings (¥1.492512, included in the total below) and 41 CRED rows that recorded their own response instead of run 1's, violating replay semantics. The repair rebuilt the rows from the cached first responses

at zero cost (post-repair, `run`$\geq$1 equals run 1 on every row); originals and the runner's superseded summary are preserved, and the corrected summary is authoritative.

Cost. Total ¥90.590184: chain 1 ¥4.902744, chain-2 probe ¥0.003444, L2 ¥2.916912, L3 ¥24.104472, L4 ¥38.824824, L5 ¥19.841232 — against the frozen track's ¥37.787328 for identical workload, a 2.40× multiplier from higher TokenHub prices (12/24 versus 4.5/13.5 per Mtok), zero cache billing (the frozen track offset 364,288 cached tokens at 0.15), and the $\kappa \approx 1.24$ longer completions. The ¥200 hard cap and ¥180 fuse were never approached (0.9083 of the cumulative 1.2× line; the L5 level itself reached 0.9864 of its line, the margin consumed entirely by the defect). Residual: an unclosed ¥0.125388 sits between row-nominal costs (70.874340) and the ledger (70.748952, through L4); the row-level unit-price identity holds on all 3,620 rows and the ledger's internal decomposition closes exactly, so the ledger remains the sole billing authority and the residual is a bookkeeping-caliber item carrying no cost claim.

Scope. E3 as delivered is answer-level and channel-input fidelity evidence; it does not re-issue certificates on the new stack. Read together with P-T2 (certificate-level error events fail to transfer at 97.6% answer agreement; §7.6), the picture is consistent: the pipeline, its request bodies, and its channel-input distributions port across stacks; certificate conclusions remain deployment-local.

H Track A Bank Provenance

Generation caliber. Six frozen banks (Table 8) supply the self-consistency comparators and entropy-bank features used across §5–§6. Every bank row holds 8 draws from a *single* `n=8` request at T=0.8: eight sequences from one call, not eight independently seeded calls. The seven generation call sites pass no `seed`, so the shared client injects `seed=42`; `max_tokens` is 640 or 768 by producer. Bank rows store no seed field, so per-row seeds are traceable to source but not self-provable from the artifacts; three banks carry an `n_calls` field (always 1) and the other two families were not retro-annotated. De-duplication is structural: producers append idempotently by `problem_id`, hence zero duplicate pids in all six banks and row counts equal to distinct-pid counts throughout. The stress corpus split uses a separate `seed=43` NumPy stream (first 200 candidates discarded, last 300 forming eval-300-v2); it is a corpus seed, not a decoding seed.

Track B contrast. The Track B stress bank follows the opposite caliber: the deployment rejected `n=8` (HTTP 400), so it holds 8 independent calls per problem, cache-key salted by draw index, on a stack that ignores `seed`. Its 600 rows are two 300-problem batches (v1/v2) with 184 overlapping pids, hence 416 distinct pids and 599 distinct question texts (one duplicated text; Appendix R); the scored 416-problem set carries majority-vote error 5/416 (0.012019), bitwise equal to the frozen SC@k value.

Table 8: The six frozen Track A banks. Draws = rows × 8; distinct-pid counts equal row counts in all six. SHA-256 truncated to 16 hex digits.

Bank	Producer	Rows	Draws	Distinct pid	SHA-256$_{16}$
`a5_mbpp_bank_raw`	`run_a5_mbpp`	100	800	100	`f67ad9c79bd5cafb`
`a5d_calib_bank`	`run_a5d_features`	200	1,600	200	`983cda564a790271`
`a5d_eval_bank`	`run_a5d_features`	300	2,400	300	`66b4e0478adc1622`
`a5e_3b_calib_bank`	`run_a5e_3b_features`	300	2,400	300	`5bab7eb943ffbd91`
`a5e_3b_test_bank`	`run_a5e_3b_features`	800	6,400	800	`12d7f32632202d05`
`a5i_stress_v2_bank`	`run_a5i_task1_stress`	300	2,400	300	`168e700a9c22d209`
Total		2,000	16,000	2,000	

I Track A Replay Fidelity Record (E2)

Protocol. After the artifacts of §5 were frozen, the entire Track A pipeline was re-executed under a hash-anchored pre-registration: thirteen segments S0–S12 covering 68 frozen files and 7,275,859 tokens, each replayed artifact compared against its frozen counterpart field by field. Replay outputs went to new paths only; the frozen directories stayed read-only (zero mtime changes across `results/trackA`, `trackB`, and the workpaper trees on replay days; the seven registered frozen SHA-256 digests re-verified 7/7). Each segment ran under a pre-registered wall-clock line with checkpoint gates and four guards (G1: `correct`

flips; G2: opinion-flip imbalance; G3: novel field differences; G4: escalation volume). Divergences are classified by a three-family taxonomy with priority F1>F2>F3 — conclusion fields (`answer`, `correct`, `original_answer`, `original_correct`), audit fields (`opinion`, `opinion_tier`, `ar_flag`), text fields (`model_output`, `extra_model_output`, `tokens_gen`) — one row belonging to exactly one family, else non-family. Runtime-stamp fields (`wall_s`, `date`) differ by construction and are disclosed without row loss.

Verdicts. Table 9: nine PASS, one FAIL, three FAIL_DISCLOSED counting the S0 gate (eight PASS excluding it). FAIL is S1: two of 500 rows diverge, both in the self-refine arm (the CoT arm matches 500/500; `n_calls` matches 500/500). One row changes only tokens (540→537); the other changes tokens (339→342), answer (489→484), and `correct` (0→1). Three same-configuration re-runs reproduce both states across runs (token counts [540,537,537] and [342,339,339]), so the two rows are value-nondeterministic rather than evidence of a configuration difference; S1 sits outside the family taxonomy's scope and is dual-reported, not re-stratified. FAIL_DISCLOSED covers S3 (the PATH-8 exemption below), S5 (1,352 family events, zero non-family, zero escalations), and S9 (191 family events after the repair below). Family events total 1,543 in final artifacts (1,819 including superseded runs; both figures reported, unconverted). Non-family events are zero in final artifacts, with two named exceptions handled by ruling: the 29-row G2 piece (reclassified, then voided; below) and two outside-gate G3 events caused by a singular/plural gap in the guard's baseline wordlist (`answer` versus `answers`), with no scoring flips and no criterion adjusted. Per-family splits of the 1,352 and 191 totals were not aggregated in the segment reports and we do not compute them post hoc; the visible per-file splits (e.g., `a5h_ara_full_test`: 65 events, F2 12 / F3 53) are text-field dominated.

The G2 true positive. The one guard halt that caught a real defect. Midway through S9, G2 fired on a 29-row prefix of the task-1 stress piece: 29 flips, imbalance 20.0, all divergence in purchased reperform counts, three opinion flips. Forensics established the rows were terminal and that the cause was cross-stage residue of the planner's global α: the replay driver ran the task-2 α-arms (which set α=0.10 at module level) before task 1 in a single process, so task-1 planning read 0.10 in place of 0.05. The decisive control: with identical inputs and model, flipping only the module-level α reproduces both states (0.05 yields the frozen rows, 0.10 yields the polluted rows), and offline recomputation of the polluted rows equals the frozen rows 29/29 — the frozen artifacts were never polluted; the defect was introduced by the replay harness's process sharing. The repair pinned α by snapshot/restore at stage boundaries plus hard assertions at the two set-points, isolated and voided the polluted outputs (29 output rows, 75 decode rows, 104 journal lines), and regenerated the segment from zero at α=0.05; the regenerated artifacts match the frozen ones on every compared row (300/300 and 981/981, zero diverged). We read the episode as the gate system working: one true positive, the criterion itself unchanged, and the question whether guards should evaluate partial artifacts registered as an open item.

Time accounting. Three calibers, never summed. Caliber A (driver machine wall) totals 157,191.1 s = 43.6642 h; verbatim annotation: *caliber A is a lower bound* — machine time outside segment brackets is not counted — with one counter-direction item, since the S9 carry retains ≤4,109.1 s of voided pollution-era wall, unsubtracted (direction overestimate; net direction undetermined). Caliber B (per-row wall sums) totals 22.5100 h and exists only from S7 onward; caliber B2 (journal generation time) totals 39.1642 h and exists only from S5 onward. Replay wall is 0.8303× the frozen wall in aggregate, with mixed per-segment direction (nine segments over, four under), reported unconverted. The hard stop line (1.2× the pre-registered projection, 58.62 h) was never approached (14.9558 h unused, not carried over). No halt corresponded to a true line crossing: four false triggers (an accounting double-count, two projection halts during a measured environment slowdown, a pre-segment path defect), one probe halt, and the G2 true positive above; segment lines were re-anchored once under a registered amendment after the slowdown (the sustained-rate >1.25 rule).

PATH taxonomy and the llama-8B item. Frozen pieces route by generator: PATH-7 (Qwen2.5-7B), PATH-8 (Llama-3.1-8B on either side), PATH-3 (Qwen2.5-3B); PATH-8/3 pieces are judged field-level and dual-reported. S1/S2/S4/S6/S7 and S9–S12 are pure PATH-7; S8 is eight PATH-3 pieces plus two PATH-7; S5 carries one PATH-8 piece (the heterogeneous arm); S3 is heterogeneous:

a PATH-7 injection piece and a 20-row PATH-8 smoke piece. That smoke piece diverges 16/20 rows, yet three same-parameter re-runs are byte-identical to one another: the divergence sits in the stack environment below the call, component unidentified. We dropped an earlier frozen-side attribution, and a "binary updated" hypothesis refuted itself (an intermittent build NOTICE on the same binary, uncorrelated with ports). Disposition: FAIL with dual reporting and disclosure, no re-stratification.

Residuals. Named open items, none affecting any quoted number: one comparator cell (HB2 stress risk 0.012313) is bounded rather than bitwise-reproduced ($\in [0.012019,\ 0.0127]$, no flip under any representative rule; merged with the matched-coverage provenance case of Appendix R); one frozen spend ledger carries an unclosed ¥0.125388 residual between its layer-row files and the ledger total (the row-level unit-price identity holds on all 3,620 rows; a key-join closure attempt yields ¥0.265200 $\neq$ ¥0.125388; we make no cost claim); three named missing workpaper rows are not rewritten; the S9 first-halt record was overwritten by the second halt's (its text survives in the registration and the run log; not rebuilt). The full replay manifest (5,685 artifacts with SHA-256) and the forty-entry event register accompany the submission.

Table 9: E2 replay segments. Frozen tokens and frozen wall from the pre-registration (frozen side); replay wall is caliber A. The tally 9/1/3 counts PASS / FAIL / FAIL_DISCLOSED with the S0 gate included (its report artifact is the gate report); FAIL_DISCLOSED marks a gate failure signed off through a registered exemption or disclosure ruling.

Segment	Frozen files	Tokens	Frozen wall (h)	Replay wall (h)	Verdict
S0 gate	`a0_smoke`	2,502	0.02	0.02	PASS
S1	`a1_*`	297,437	2.72	2.84	FAIL
S2	`a2_*`	70,861	2.58	2.60	PASS
S3	`a3*`	2,992	0.03	0.12	FAIL_DISCLOSED
S4	`a4_*`, `a4b`	286,678	0.62	0.66	PASS
S5	`a5_*` (15 files)	3,420,755	23.93	20.65	FAIL_DISCLOSED
S6	`a5c`, `a5g`	38,242	0.34	0.36	PASS
S7	`a5d*` (16 files)	804,114	6.40	2.40	PASS
S8	`a5e*` (10 files)	699,088	4.01	4.08	PASS
S9	`a5h`, `a5i`	1,304,100	7.69	7.23	FAIL_DISCLOSED
S10	`a5j`	297,249	1.97	0.39	PASS
S11	`a6_*`	7,985	0.69	0.71	PASS
S12	`test300_*`	43,856	1.58	1.61	PASS
Total	68 files	7,275,859	52.59	43.66	9/1/3

J EXTRACTION RECALL ON PROCESSBENCH

Setup. We score the production claim extractor (`claims.py`, byte-frozen) against ProcessBench (Zheng et al., 2025) (revision `3bdcd537`; 400 GSM8K-style and 1,000 MATH-style human-curated solutions), offline, zero API calls. Every human step that contains an equation contributes a checkable assertion; steps without equations (38.0%/39.5% of steps) never enter the recall denominator. Predictions P-E1..P-E4 (recall ≥ 0.95; precision ≥ 0.95; MATH recall ≥ 0.90; wrong-solution recall $\leq$ correct-solution recall) were hash-anchored before measurement, along with the escape clause: if the as-is rule leaves >5% of equations undecided, exactly one normalization extension is permitted. Stage-1 undecided rates (94.20%/96.99%) triggered it; the extension (X1: longest maximal numeric substring per side, symmetric, leftmost on ties) was registered before any stage-2 number was computed, and both stages are reported side by side.

Table 10: Extraction recall on ProcessBench, two stages side by side (per the registration, stage-2 is never quoted alone). Undecided: equations the rule declines to match. ProcessBench text is human-cleaned, so all numbers are best-case upper bounds on production chain-of-thought. Stage-1 rows mix two counting calibers: recall is step-level (GSM8K-style TP=0, FN=1,291; MATH-style TP=5, FN=3,928, FN including undecided steps), precision is claim-level (TP=1, FP=1,112 and TP=40, FP=4,682 respectively).

Stage	Subset	Recall	Precision	Undecided
as-is (production rule)	GSM8K-style	0.0000	0.0009	94.20%
as-is (production rule)	MATH-style	0.0013	0.0085	96.99%
+X1 (one registered extension)	GSM8K-style	0.3989 [0.3721, 0.4262]	0.5606	23.02%
+X1 (one registered extension)	MATH-style	0.2339 [0.2208, 0.2475]	0.5203	34.50%

Verdicts. P-E1 falsified ($0.3989 < 0.95$), P-E2 falsified (pooled precision $0.5280 < 0.95$, both subsets likewise), P-E3 falsified ($0.2339 < 0.90$). P-E4 holds on the pooled set and on GSM8K-style

text (wrong-solution recall 0.2700 $\leq$ 0.2800; 0.3787 $\leq$ 0.4217) and reverses on MATH-style text (0.2347 $>$ 0.2331, a 0.0016 gap with wide CP95 overlap), reported as-is. Consequence for the main text: the extractor is a low-recall component even on clean text, so claim-level machinery audits a visible subset and the certificate's honesty rests on answer-level scoring against gold (§3); this is also why the Track B′ detection band reaches only 18.8% of stress answer errors. Why validity survives low recall: the certificate numerator counts every wrong certified answer against gold, whether or not its claims were ever extracted; concretely, the ten wrong Track B slice answers whose claims fired no channel (e.g., `test-777`) sit inside the certified subset and inside k=10. Recall bounds how much the audit can flag upstream, never the measured risk. Recall shortfall is therefore not hidden risk: a wrong answer whose claims escape extraction is certified and counted in k, so low recall shows up as a worse measured risk level, never as corrupted validity. One deployment-facing caveat remains: an answer with zero extracted claims receives an unqualified opinion by default (356/500 problems on the Track B′ slice). Auditing standards treat unobtainable evidence as a scope limitation meriting a qualified opinion or disclaimer rather than a clean pass; routing zero-claim answers to an inconclusive flag is the conservative variant, and we flag it as a design option rather than enabling it in these runs.

K WIN/LOSS LEDGER (F4)

Main endpoint (risk on certified subset at reported coverage): statistical ties in both domains; McNemar b=c=0 on the paired risk cells. Named losses: two clamp-configured comparators a001 and a010 lose by risk differences 1.7×10^{-4} and 1.3×10^{-4} (single-point comparator degeneracy; disclosed, not generalized). Stress coverage loss: uniform purchasing 85.8% vs. ARA 78.2% at equal zero risk (weak dominance; reported verbatim in §7.2). Cache-reuse metrics for the two clamp variants: 1.000 (a001) and 0.680 (a010); an earlier prediction that the gap localizes to cache behavior was falsified (28.1-token mean-extra gap unlocalized; 106.3 vs. 78.3 tokens).

Conformal head-to-head (post-hoc). We calibrate split-conformal selective classification on the same two score signals as the A5b curves (resample-agreement entropy, keep low; mean token log-probability, keep high): 50/50 calibration/test splits over 20 seeds, threshold = largest coverage on a pre-set 20-point grid whose one-sided CP95 calibration margin meets the target, in two risk calibers (conditional: D1's certified-subset caliber; joint: Track A's pre-registered full-set caliber). Table 11 reports medians. Three findings. (i) On GSM8K the entropy baseline lands behind ARA on both margins, (0.867, 3.95%) against (0.890, 3.83%), with seed instability (one vacuous seed, coverage range 0–0.885). (ii) On stress the conditional rule collapses to abstain-all at every $\alpha \leq 0.15$ (20/20 seeds at $\alpha \leq 0.10$; 18/20 at 0.15): with 300 calibration points and a 0.61 base rate, no finite-sample margin can certify any subset at the levels ARA achieves; the joint rule operates and trades coverage for risk: (0.307, 12.9%) against (0.278, 9.58%), i.e. higher coverage at higher full-set risk (3.96% against 2.67%), both under the 5% target. (iii) A threshold calibrated on GSM8K at α=0.05 and deployed on stress realizes 18.0% conditional risk against the 5% target: conformal's marginal guarantee requires calibration/deployment exchangeability, which the stress domain violates by construction; D1 is measured on the realized subset and needs none. The entropy score costs 8 resamples per problem; ARA's evidence purchases average 150.2 tokens. The honest asymmetry: on near-i.i.d. GSM8K the baseline carries a distribution-free marginal guarantee that ARA does not (§3.5), and that guarantee exists only where exchangeability holds. A deterministic v1-only arm (frozen 200-problem calibration split $\rightarrow$ v1 eval 300, disjoint, no seeds) removes split variance: the joint rule lands at (0.313, 10.6%) against ARA-v1's (0.283, 10.6%)—a near-tie, reported as such—while the conditional rule still collapses at $\alpha \leq 0.10$ and operates only at α=0.15 ((0.383, 14.8%)). Distribution-free risk-control wrappers built for user-specified targets (Angelopoulos et al., 2024; Bates et al., 2021) bound expected bounded losses marginally; the stress failure modes above are exchangeability violations, which these wrappers do not repair, and D1's object is a realized-subset empirical bound rather than a marginal expectation. A deterministic GSM8K dev$\rightarrow$test arm (frozen dev500, base rate 5.4%) lands at (0.871, 4.18%) at α=0.05—behind ARA on both margins—but overshoots at looser targets (10.85% realized against 0.10; 5.23% full-set against 0.05 joint): the dev split's 2$\times$ lower base rate breaks the realized marginal guarantee even within one benchmark, intermediate between the i.i.d. case and the 3.6$\times$ cross-domain violation.

Table 11: Conformal head-to-head (post-hoc comparator, designed after the frozen results). Split-conformal selective classification on the two A5b score signals; 50/50 calibration/test splits, 20 seeds, median (range). Conditional rule targets certified-subset risk (D1's caliber); joint rule targets full-set risk (Track A's pre-registered caliber). Vacuous: no pre-set grid point passes the calibration margin (abstain-all). stress-v1 det: deterministic arm, frozen 200-problem calibration split → v1 eval 300 (disjoint, no seeds). gsm8k-dev det: frozen dev500 → test1319 (disjoint); its det-cond 0.10 and det-joint 0.05 rows realize above target (conditional 10.85% vs. 0.10; full-set 5.23% vs. 0.05) under the dev/test base-rate skew (5.4% vs. 10.9%). ARA rows report the realized-subset conditional risk of the frozen run (0.0958 on stress; the pre-registered Track A target is the full-set caliber, 16/600=0.0267, Table 3).

Domain	Score	Rule	α	Coverage	Cond. risk	Vacuous
GSM8K	entropy	cond	0.05	0.867 (0–0.885)	0.0395	1/20
GSM8K	entropy	joint	0.05	0.869 (0.844–0.912)	0.0400	0/20
GSM8K	logprob	cond	0.05	0.460	0.0417	4/20
GSM8K	entropy	cond	0.01	vacuous		20/20
GSM8K	entropy	cond	0.10	0.939 (0.906–0.964)	0.0695	0/20
stress	entropy	cond	0.05	vacuous		20/20
stress	entropy	cond	0.15	vacuous		18/20
stress	entropy	joint	0.05	0.307 (0.267–0.330)	0.1290	0/20
stress-v1	entropy	det-cond	0.05	vacuous		det.
stress-v1	entropy	det-cond	0.15	0.383	0.1478	det.
stress-v1	entropy	det-joint	0.05	0.313	0.1064	det.
gsm8k-dev	entropy	det-cond	0.05	0.871	0.0418	det.
gsm8k-dev	entropy	det-cond	0.10	0.999	0.1085	det.
gsm8k-dev	entropy	det-joint	0.05	0.909	0.0575	det.
GSM8K	ARA	—	0.05	0.890	0.0383	—
stress	ARA	—	0.05	0.278	0.0958	—
stress-v1	ARA	—	0.05	0.283	0.1059	—

L SCOPE VARIANTS AND THE CPU-ONLY FLOOR

CPU-evidence-only scope: GSM8K slice 495 certified, 13 wrong, coverage 0.990, risk 2.63%, CP95 4.14%; stress 593 certified, 0 wrong, coverage 0.988, risk 0, CP95 0.50%. Shared error: 10 of 16 wrong GSM8K-slice answers are shared with the 7B model (49 7B-wrong problems in the 500-problem scope); stress shared-error analysis unavailable because the local model cannot attempt the magnitudes. The error-bank entropy feature is degenerate on the slice ($q_{75} = 0$), disclosed because it bounds what the risk model could have learned.

Porting beyond arithmetic. Three components are domain-specific: the claim extractor, the evidence-channel roster, and gold labels for measurement; the opinion rule, the budget equation, and the certificate machinery are domain-agnostic. Where no deterministic channel exists, the hierarchy degrades to isolated-LLM evidence, DR estimates weaken, and floors bind earlier; the certificate then reports lower usable α or coverage rather than hiding the loss. A structural boundary follows: rule v2 gates adverse opinions on exact machine counter-evidence, so in such domains the adverse tier is unreachable and the taxonomy collapses to unqualified versus qualified. Our HotpotQA probe (Appendix O) is this regime: the pipeline runs on LLM-only evidence. A trained verifier ports by re-labeling and re-training per domain; ARA ports by swapping the extractor and the channel roster. We expect domains with executable or retrievable ground truth (code tests, database question answering, citation checking) to port first; open-ended generation remains outside the claim model.

M COST LEDGER

Total spend ¥37.787328: audit engagement ¥29.654915, root-cause and verification ¥8.093236, two pilot arms ¥0.037922 and ¥0.001256. Tokens: cache-hit 364,288 at ¥0.15/Mtok, cache-miss 2,127,002 at ¥4.50/Mtok, completion 2,086,013 at ¥13.50/Mtok. Unit-price assertion passes with discrepancy 3.6×10^{-13}. Per-problem cost ¥0.0343. Cache-free equivalent ¥39.37. Budget gates 300/400/800 tokens per problem tier were pre-registered and unchanged. Purchased evidence units total 7,715, where one unit is one execution of an evidence procedure against one claim (one re-computation, one CRED query, one reperform sample, or one invariant check). The ledger records call-level cost rows without a call-type field, so a generation-versus-audit split is not recoverable from the ledger and we do not estimate one; all figures above are all-in engagement costs (itemized rows are rounded ledger entries; the components sum to ¥37.787329). Local auditor GPU compute (the 3B and 8B audit procedures) is not monetized in this ledger. Wall-clock latency was not instru-

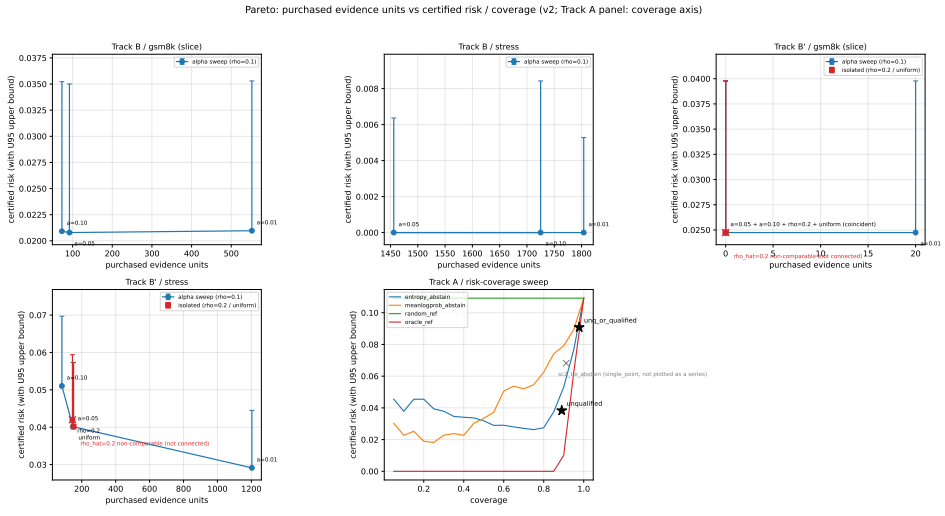


Figure 5: Budget–risk–coverage planes, derived offline from the frozen artifacts with no API access (`pareto_budget_risk_coverage_v2.csv`; per-row provenance in its `source_file` and `source_row` columns; sha-256 logged). Four left panels: residual risk versus evidence purchases, Tracks B and B′, slice and stress regimes. Track B slice arms are flat by construction (pre-set budgets, no sweep); stress arms read k=0; B′ slice arms degenerate to zero purchases except α=0.01 (20 purchases) and coincide at k=12/485; B′ stress arms separate monotonically in α (83/148/1,203 purchases). The $\hat{\rho}$=0.2 cell (isolated marker) is non-comparable (estimator change) and not connected; the uniform baseline coincides with the α=0.05 arm in both B′ regimes. Right panel: Track A coverage–risk plane with four reference sequences (20 thresholds each, connected without interpolation) and the two ARA operating points (stars). B′ purchase counts are derived as upgrades plus credibility-bank runs from the frozen per-arm CSVs; displayed risks use the as-run calibers (rejudged for Track B, D1 for B′); B′ certified U_{95} is recomputed one-sided from the frozen $(k, |S|)$ pairs.

mented. Track B′ is ledgered separately (Appendix F): ¥15.43 all-in, of which the stack-migration probe is ¥4.82; the abandoned earlier arm on another provider (2,742 calls) is archived unmerged. The E3 stack replication (Appendix G) is ledgered separately at ¥90.590184 all-in (TokenHub prices 12/24/1 per Mtok with zero cache billing, including a disclosed ¥1.492512 duplicate-billing defect); an unclosed ¥0.125388 nominal-versus-ledger residual is registered with no cost claim.

What the budget buys. Figure 5 shows where the money goes. In the Track B slice regime the three certified arms are flat at about 2.1% residual risk at coverage 0.95–0.96 on pre-set purchase counts of 552/92/73: budgets were fixed before the runs, not swept, so the panel is three points, not a curve. In the stress regime all three arms read k=0 at 1,456–1,804 purchases; risk is zero and the certified bound tightens only through n_{cert} (0.53%–0.84%). Track B′ is where purchases actually move. In the slice regime the risk floor binds every claim: four of five arms buy nothing, the aggressive α=0.01 arm buys 20, and all five collapse onto one certificate (k=12/485, U_{95}=3.98%). In the stress regime purchases grow from 83 to 1,203 as α tightens from 0.10 to 0.01, residual risk falls from 5.10% to 2.91%, and the certified bound falls from 6.97% to 4.45%; the uniform baseline coincides with the α=0.05 arm on this bank (21/523; 152 versus 148 purchases). The Track A panel plots the coverage–risk plane: four frozen 20-point reference sequences with the two ARA operating points (0.8901, 3.83%) and (0.9773, 9.08%) as stars, sitting between the oracle envelope and the entropy reference at matched coverage. The $\hat{\rho}$=0.2 stress cell (145 purchases) uses an estimator-incompatible caliber and is plotted isolated, not connected.

N MODEL IDENTITY EVIDENCE (TRACK B)

Three legs: (i) model string `deepseek-v4-pro` on all 21,450 cached responses; (ii) one `system_fingerprint` value across the collection; (iii) timestamps 2026-09-09 23:09:36 to 2026-09-10 09:46:45 (UTC+8), inside the provider's documented service window (DeepSeek, 2026a). The archive (EV-20260910-01) contains the response caches and request logs, with SHA-

256 digests recorded in the manifest addendum; digests were re-verified after archiving. Disclosed defect: a configuration freeze timestamp (08:15:00) precedes the true collection end by 1h31m; the documented window bounds it externally.

O Out-of-Domain Probes: HotpotQA and MBPP

HotpotQA-500: the audit pipeline transfers, with 250 of 265 sampled certified answers shared-correct with the generator's raw output (94.3%); details in the frozen artifacts. MBPP-100: under the audit-caliber protocol (claims extracted from code and docstrings, CPU-executed tests as deterministic evidence) 0-shot accuracy is 0.35, two-round 0.37, pass@8 0.43; under the standard protocol the same generator scores 0.72. Both calibers are disclosed because they answer different questions: the former measures auditable correctness under our claim model, the latter benchmark skill. Reporting only the larger number would be a caliber switch. Neither probe issues a formal Definition-1 triple against gold references; the dual-caliber certificate is validated on the arithmetic tracks, and these probes measure transfer of the pipeline, not of the guarantee.

P cred Workpaper Semantics

In the frozen workpapers, the field `val` stores eval(`lhs`), the CPU-recomputed value of the claim's left-hand side. A CRED verdict compares the critic's `cred_answer` against `val`. Consequently, on claims already flagged by recomputation, a *match* verdict means the critic agrees with the CPU (and hence contradicts the generator); it does not confirm the generator's claim. This reading was confirmed with the team on 2026-09-11 and underlies the attribution in §7.4: CRED drove 0 of 7 adverse opinions on stress.

Q Behavioral Continuity

Three legs support a stable endpoint across the collection window: (i) response-style statistics are stationary across the window; (ii) capability probes at window edges agree; (iii) the single `system_fingerprint` is constant. Combined with the external timeline (Appendix N), we attribute the measurements to one stable model serving. The freeze-timestamp defect is disclosed in Appendix N and does not affect request-level timestamps used here.

Threat model. The three legs rule out wholesale substitution and silent rerouting to a different model within the window. They do not rule out a provider-side grey launch or A/B experiment that preserves the `system_fingerprint` string and leaves style statistics stationary; continuity evidence of this kind is observational, and certificates issued under it inherit that residual threat.

R Repair Loops and Auditor Scale

H1 (pre-registered recompute-tool ablation): the recompute tool is switched on versus off at matched budget B=5 on two stress batches of 300 problems each (files v1/v2); the switch touches only recompute detection and upgrades, while $\hat{p}_{\text{err}}$ and materiality are still computed and enter planning. On batch v1 the tool-off arm misses 34/300 wrong answers (11.33%) against 9/300 (3.00%) with the tool on; all 25 net-rescued answers are adverse-under-full $\rightarrow$ unqualified-under-notool transitions (paired McNemar on the flag-miss indicator, n_{01}=25, n_{10}=0, $p = 6.0 \times 10^{-8}$), detection recall moving from 81.2% (147/181, tool off) to 95.0% (172/181, tool on), and the adverse-opinion gap (141 versus 0) is entirely the recompute channel. Confirmation against the absolute α=0.05 target: batch v1 9/300, $p = 0.065$ (kept, not hidden); batch v2 7/300 (2.33%), $p = 0.016$; pooled 16/600 (2.67%), 95% CI $[1.53, 4.29]$, $p = 0.003$. The two batches come from separately generated stress files: pid labels collide on 184/300 across batches but question texts differ on all 184, so pooling treats 600 independent problems, 599 of them distinct — the v1 file duplicates one problem (`stress-0086`/`stress-0103`), both rows wrong and both flagged (adverse) in the tool-on arm; counting it once leaves every conclusion intact (batch v1 9/299, $p = 0.067$; pooled 16/599, $p = 0.0033$), consistent with the count-once convention of Appendix F. The same pair also appears as-run in the frozen Track B stress bank and its replay merges; count-once recomputation flips no

recomputed Track B cell (all displacements $\leq 7.2 \times 10^{-4}$); one comparator cell (HB2 stress risk) is bounded rather than bitwise-reproduced, pending matched-coverage provenance. A pre-registered expectation that the tool-off arm's residual adverse opinions would come from invariant checks was not borne out: invariants never fire on this positive-integer domain, and the tool-off arm issues zero adverse opinions. Correction loop (A5c/A5g, GSM8K; disjoint from H1 in problems and pairings): fix/regression by opinion class: adverse-tier fixes 9/27 (33.3%, CI $[16.5, 54.0]$) with regression 2/3 (66.7%); qualified-tier fixes 30/72 (41.7%) with regression 26/43 (60.5%). By channel: recompute 7/8 (87.5%), invariant 1/17 (5.9%), reperform 28/64 (43.8%, regression 68.6%). The two experiments measure different objects: the McNemar counts flag-miss transitions (wrong and unflagged) over the 300 paired stress problems, while the fix/regression rates count answer-level changes among flagged GSM8K answers. Auditor scale (Qwen2.5-3B): tier error 5.0%/53.4%/89.7% (unqualified/qualified/adverse); a disclaimer subpopulation rates 39.3% (11/28). Stress win/loss by objective: OP1 wins against all three comparators (10.6 vs. 11.9/36.9/60.3), OP2 loses to two (42.1 vs. 29.2/39.6); 3B OP1 wins all three (5.0 vs. 7.4/17.3/24.5) with one entropy-objective loss (20.5 vs. 19.6). Self-consistency reference points: stress sc@3 60.0% at 36.1% tokens; 3B sc@2 64.8% at 10.6% tokens.

S Claims, Novelty, and Supporting Evidence

Classical ingredients, not claimed as new. One-sided Clopper–Pearson bounds (Clopper & Pearson, 1934), exact binomial tests, Holm correction (Holm, 1979), greedy knapsack allocation, selective classification (El-Yaniv & Wiener, 2010; Geifman & El-Yaniv, 2017), and distribution-free wrappers (Angelopoulos & Bates, 2023). The audit-risk factorization itself is institutional (International Auditing and Assurance Standards Board, 2019), used as an allocation device (§3.2).

Claimed as new, each with direct support. **C1 (the assembly).** The institutional transplant as one working system: risk-model budgeting (§3.2), independence-ranked evidence (§3.4), CPU-gated adverse opinions (§3), dual-caliber certificate (§3.5). Support: end-to-end operation on three tracks (Table 1; Appendices E, F); opinion tiers separate error on every API engagement (§7.2, §7.6). **C2 (the certificate).** The dual-caliber empirical certificate D1: certified subset, one-sided CP95 bound, stated coverage, stated cost. Support: the twelve-cell pre-registered battery with Holm sensitivities (Table 3) and the sealed reproduction (Table 5); remarks D1–D4 delimit scope. **C3 (budget discipline).** The budget equation as an allocation device whose misspecification costs efficiency, not validity, with the ratio observation (the α-sweep doubles as a reliance-constant probe; floors and costs $\hat{\rho}$-independent). Support: four ratio points across §6 and §7.6; the degenerate-calibrator episode, where the rejudged risk model carries no signal yet the certificate stands (§7, Appendix A). **C4 (independence hierarchy).** Independence from the generator's failure modes as the currency of audit evidence. Support: the isolation effect ($p \approx 10^{-19}$, §5) and the channel-value inversion across stacks (§7.6). **C5 (opinion rule v2).** Adverse opinions gated on exact machine counter-evidence. Support: all 16 Track B′ stress adverse opinions are wrong answers (§7.6); root-causing attributes 98.7% of adverse evidence to our own extraction layer, not arithmetic inability (§7.4). **C6 (the method).** Pre-registered bidirectional falsification with negative results and comparator losses disclosed. Support: three falsifications and one vacuous pass reported alongside the holds (Table 3); named losses (Appendix K); a repair batch kept at $p{=}0.065$ (Appendix R).

Known-weak components, stated. The risk model's resolution floor bounds usable α; Track B's rejudged calibrator is degenerate (Appendix A); extraction recall is at best 0.40 on clean text (§8). None of these touches certificate validity (§3.2). A fourth weak component is the allocator itself: it scores claims independently and never beats uniform purchasing on these workloads (exact ties on Track A and Track B′; a weak-dominance loss of 7.6 coverage points at equal zero risk on Track B stress). Prediction floors and coarse procedure costs leave little risk variance for weighting to exploit, and modeling the sequential cascade structure of multi-step traces (an early premise error invalidates downstream claims) is open. Greedy multiple-choice knapsack carries the classical $1/2$-approximation guarantee; with per-problem budgets of at most 800 tokens over a handful of procedures the feasible-bundle space is small, and we did not measure the residual optimality gap.