CONSENSUS NEED NOT BE EVIDENCE: INFORMATION LIMITS AND ADAPTIVE AUDITING FOR MULTI-LLM EVALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-LLM evaluation is often justified by agreement: independent judges should cancel idiosyncratic errors. The premise fails when evaluators inherit a common hallucination, retrieval artifact, or semantic convention. We separate the external fact from each evaluator's complete private state and from the transcript generated by peer interaction. This yields three limits. First, we derive a phase boundary above which a less accurate model family receives strictly higher agreement, determinant mutual information, and Shannon mutual information. Second, every peer-only transcript is a garbling of the complete private-state experiment, so its full-coverage factual error cannot beat that experiment's Bayes risk. Third, unlimited unlabeled peer observations cannot identify semantic orientation. Over a truth-swap-closed class, the minimax error is one half, even with selective prediction. We then introduce GAVEL, which fits and orients an aggregation policy with verified labels, freezes it, and returns a risk certificate or no certificate using a fresh sequential audit. For heterogeneous task strata, we characterize the information rate of certifying a safe frozen policy by a Chernoff game. An adaptive auditor is anytime-valid and attains the characteristic almost-sure evidence rate. A scalar beta-mixture benchmark admits an exact finite-budget recursion. Synthetic and PEG-policy semi-synthetic experiments expose the reversal and evaluate policy construction and heterogeneous audits. With stratified evidence held fixed, GAVEL reduces mean capped audit draws by 10.6% and 21.2% relative to prevalence sampling on two heterogeneous laws. Frozen-policy comparisons on MMLU and HellaSwag improve verification accuracy over matched single evaluators by 1.07 and 1.14 percentage points. The results separate limits of the peer information state from the evidence needed to orient and audit a frozen policy.

1 INTRODUCTION

In *Mata v. Avianca*, a federal court found that lawyers submitted non-existent judicial opinions containing fake quotations and citations created by ChatGPT and continued to stand by them after judicial orders questioned their existence (United States District Court, 2023). The sanctions order identifies six fake opinions in its remedial provisions. The episode isolates a high-stakes evaluation problem: if several language models validate a citation using related training data, the same retrieval layer, or one another's explanations, does consensus constitute evidence?

This question lies at the intersection of factuality, multi-agent language-model systems, and statistical decision theory. Multi-agent debate has improved reasoning and factual validity in controlled studies, including settings where weaker judges assess stronger models (Du et al., 2024; Khan et al., 2024). ChatEval's debate-based referee team improved evaluation accuracy and correlation with human assessments (Chan et al., 2024). Peer elicitation can reward informative reports without labels (Kong, 2020; Chen et al., 2025), and no-regret analysis describes how agents adapt to utilities (Park et al., 2025). These paradigms address different links: reporting guarantees concern communication of private judgments, whereas factual guarantees concern whether the group's complete information distinguishes the external state. Systematic and interaction-amplified evaluator biases make the second link nontrivial (Wang et al., 2024; Panickssery et al., 2024; Okawa, 2026).

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

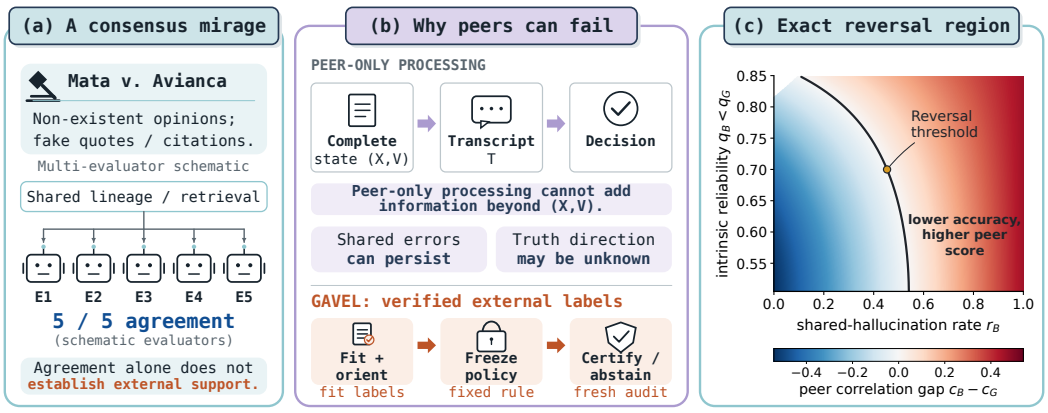


Figure 1: **Shared errors, information limits, and verified audits. a**, *Mata v. Avianca* motivates the schematic evaluation scenario. **b**, Shared errors and semantic ambiguity can persist through peer-only processing. GAVEL uses verified fitting labels to orient a policy and fresh audit labels to assess its frozen risk. **c**, Family B has lower accuracy and a higher peer-information score in the reversal region. The black curve marks equality, and color shows the peer-correlation gap.

Research question. *What can peer-only consensus establish about an external fact, and how much fresh verified evidence is needed to certify the risk of a frozen evaluation policy?* Figure 1 illustrates the question.

Comparison of statistical experiments (Blackwell, 1953) identifies the factual information available to peer interaction. Verified labels support semantic orientation and policy selection. After freezing, certification becomes a sequential test of deployment risk (Chernoff, 1959).

Our contributions are:

1. **Information limits under shared errors.** We derive an exact agreement-reversal boundary and a finite-ensemble shared-error floor. Comparison of statistical experiments and truth-swap symmetry distinguish two further obstacles: residual uncertainty in the complete private state and ambiguity in its semantic orientation. Together, these results specify which failures persist through peer-only interaction and which require external information.
2. **Information rate of certification.** For interior safe risks, the aggregate-risk Chernoff game admits a one-dimensional equal-KL solution with a unique, positive optimal allocation. Its rate separates allocation and stratification gains. We prove an expected-label lower bound and construct anytime-valid evidence attaining the same characteristic constant almost surely under convergent designs. The scalar benchmark also attains its matching expected-time rate.
3. **Policy construction and evaluation.** GAVEL combines label-assisted policy construction with independent sequential certification. Exact calculations, synthetic aggregation and audits, a semi-synthetic shared-error intervention, and cached-report verification evaluate these stages. Matched-evidence controls quantify the finite-budget value of allocation.

2 Related Work

Appendix A.1 gives a detailed discussion.

Peer elicitation and evaluation. Peer prediction studies incentives for informative reports (Prelec, 2004; Miller et al., 2005). Determinant mutual information supports truthful elicitation (Kong, 2020), and PEG applies peer elicitation to LLM discriminators (Chen et al., 2025). Debate and referee systems study reasoning and evaluation through interaction (Du et al., 2024; Chan et al., 2024; Khan et al., 2024). We characterize when shared errors reverse peer scores and how complete private states and verified labels determine factual information and semantic orientation.

Uncertainty and evaluator reliability. Evaluator studies examine positional, self-preference, and interaction effects (Wang et al., 2024; Panickssery et al., 2024; Song et al., 2026; Qian, 2026; Okawa, 2026). Semantic uncertainty measures variation in meaning (Farquhar et al., 2024), latent-annotator models infer reliability (Dawid & Skene, 1979), and selective prediction relates error to coverage (El-Yaniv & Wiener, 2010; Geifman & El-Yaniv, 2017). These perspectives motivate our complete-state formulation and frozen-policy audit.

Active risk control and controlled sensing. Confidence sequences provide time-uniform inference (Howard et al., 2021). Active risk control combines adaptive labeling and prediction-assisted betting (Xu et al., 2024). Hsu & Shekhar (2026) establish oracle log-growth optimality for a given querying rule in active LLM evaluation. Controlled sensing relates allocation to sequential testing cost (Chernoff, 1959; Garivier & Kaufmann, 2016; Mukherjee et al., 2022). For interior safe risks under the stratified Bernoulli null $\pi^\top q \geq \alpha$, our outer allocation game has an explicit equal-KL solution and a unique positive optimal allocation. It separates allocation and stratification gains and yields betting evidence attaining the characteristic rate under convergent designs.

3 Problem Setup: Facts, Private State, and Information Boundaries

Peer evaluation combines an external factual target with information held by individual agents and revealed through interaction. We formalize these objects and their information boundary to distinguish uncertainty in factual prediction from ambiguity in semantic orientation (Theorems 2 and 3).

For citation verification, a peer-visible task is $X = (q, c, d) \in \mathcal{X}$: a claim q, purported evidence c, and generated document d. An authoritative record W determines

$$Y = y(X, W) \in \{-1, +1\},$$

where $+1$ means that c exists and supports q. The same binary-verification formulation covers answer checking: a task contains a question and a fixed proposed answer, and $Y = +1$ when the proposal matches the benchmark reference. We assume $\mathcal{X}$ and all private-state spaces are standard Borel, so regular conditional distributions exist.

Evaluator i has a *complete peer-accessible private state* $V_i \in \mathcal{V}_i$. It includes every internal observation, tool return, and pre-sampled random variable that may be exposed during the protocol. A binary initial judgment $S_i = s_i(X, V_i)$ and a submitted report A_i may be derived from this state. Let $V = V_{1:n}$ and let T be the complete interactive transcript, including messages, payments, and the final randomized output.

Assumption 1 (Peer-only closure)**.** *Conditional on X and V, the transcript is independent of the external fact:*

$$Y \longrightarrow (X, V) \longrightarrow T.$$

Equivalently, $T \mid (X, V)$ is generated by a Markov kernel that does not depend on W or Y.

Scope of peer-only closure. Under peer-only closure, the transcript is generated from (X, V) without additional information about Y. A tool call that adds authoritative information changes the experiment to which the Bayes bound applies. A positive barrier requires residual uncertainty about Y after observing (X, V); if Y is measurable from X alone, then $R^*_V = 0$. The bound concerns missing statistical information rather than computational difficulty.

Write $f^{\pm}_x$ for the conditional sub-probability densities defined in Appendix B.2. A full-coverage rule outputs $\widehat{Y} \in \{-1, +1\}$; a selective rule may also output $\perp$. Its coverage and conditional risk are

$$\text{Coverage} = \mathbb{P}(\widehat{Y} \neq \perp), \qquad \text{Risk}_{\text{sel}} = \mathbb{P}(\widehat{Y} \neq Y \mid \widehat{Y} \neq \perp),$$

when coverage is positive.

Two diagnostic channels. Two binary channels illustrate the mechanisms underlying the general information bounds. In the family channel, $Y, M_h, H_h, E_{h,1}, \ldots, E_{h,n_h}$ are mutually independent,

Y, H_h are balanced, $M_h \sim \text{Bernoulli}(r_h)$, and the $E_{h,i}$ are i.i.d. with $\mathbb{P}(E_{h,i} = 1) = q_h > 1/2$. For $h(i) = h$,

$$S_i = \begin{cases} H_h, & M_h = 1, \\ Y E_{h,i}, & M_h = 0. \end{cases} \tag{1}$$

In the shared-flip channel,

$$S_i = Y F E_i, \quad \mathbb{P}(F = -1) = \beta \le \tfrac{1}{2}, \quad \mathbb{P}(E_i = -1) = \varepsilon < \tfrac{1}{2}, \tag{2}$$

where $Y \sim \text{Unif}\{-1, +1\}$, $F \in \{-1, +1\}$ is independent of Y, and $(E_i)_{i=1}^n$ is i.i.d. and independent of (Y, F).

4 Why Peer Consensus Can Fail

This section traces how shared errors affect multi-agent evaluation, from the ranking of peer reports to the factual information retained by interaction. The results connect score reversal, complete-state uncertainty, and semantic orientation, then quantify the remaining error in a growing ensemble (Theorems 1–3 and Proposition 1).

Theorem 1 (Exact agreement reversal)**.** *For two distinct evaluators in family h of* (1)*, let $d_h = 2q_h - 1$. Their factual accuracy a_h and correlation c_h are*

$$a_h = \tfrac{1}{2} + \tfrac{1}{2}(1 - r_h) d_h, \qquad c_h = \mathbb{E}[S_i S_j] = r_h + (1 - r_h) d_h^2. \tag{3}$$

Their agreement, DMI, and mutual information are strictly increasing in c_h:

$$p_h = \tfrac{1}{2}(1 + c_h), \quad \text{DMI}_h = \tfrac{1}{4} c_h, \quad \text{I}_h = \tfrac{1+c_h}{2} \log(1 + c_h) + \tfrac{1-c_h}{2} \log(1 - c_h).$$

If $1/2 < q_B < q_G < 1$ and $r_G < 1$, define

$$r_B^\star = r_G + (1 - r_G) \frac{d_G^2 - d_B^2}{1 - d_B^2}. \tag{4}$$

Then $r_G < r_B^\star < 1$, and every $r_B > r_B^\star$ satisfies $a_G > a_B$ but $p_G < p_B$, $\text{DMI}_G < \text{DMI}_B$, and $\text{I}_G < \text{I}_B$. At $r_B = 1$, peer information is maximal while $\text{I}(S_i; Y) = 0$.

The three score orderings follow from the balanced binary family channel in (1). More precise peer-score estimation can make the wrong ranking more certain when the common component is unrelated to Y. Independent-error ensembles are recovered at $r_h = 0$.

Theorem 2 (Complete-state Bayes barrier)**.** *Under Assumption 1, every randomized full-coverage decision based on (X, T) has risk at least*

$$R_V^* = \mathbb{E}_X \left[\int_{\mathcal{V}} \min\{f_x^+(v), f_x^-(v)\} \, d\lambda_x(v) \right]. \tag{5}$$

The bound is attained when the complete state V is retained and its Bayes rule is used. The balanced-prior total-variation formula appears in Appendix B.2. The bound also covers strategic reports generated from (X, V) through the transcript kernel.

For binary zero–one loss, a transcript preserves the complete-state Bayes risk exactly when it does not mix states requiring opposite Bayes decisions; it need not preserve the full posterior (Appendix B.6).

Theorem 3 (Exact semantic-orientation minimax limit)**.** *Let $P^{\leftrightarrow}$ replace Y by $-Y$ while leaving every peer-only observable unchanged. For any learner trained on arbitrarily many unlabeled peer samples and any fresh full-coverage prediction,*

$$\text{Risk}_P + \text{Risk}_{P^{\leftrightarrow}} = 1. \tag{6}$$

Hence every nonempty truth-swap-closed model class $\mathcal{M}$ satisfies

$$\inf_{\mathcal{A}} \sup_{Q \in \mathcal{M}} \text{Risk}_Q(\mathcal{A}) = \tfrac{1}{2}.$$

For a selective rule, the same risk-sum identity holds conditional on any peer-observable acceptance event having common positive coverage in the two worlds. Abstention can reduce exposure but cannot orient the accepted predictions.

This distinguishes separation from semantics. Peer states may perfectly separate two latent components, but unlabeled data cannot determine which component means "supported by the record." The result concerns classes without external orientation: assuming a designated evaluator is known to be better than chance already supplies such information, so positively oriented diagnostic channels need not be truth-swap closed. Verified anchors can resolve orientation statistically (Appendix B.7), but do not remove overlap in the deployment-time state experiment.

Proposition 1 (Exact shared-error floor)**.** *Under* (2)*, let n be odd and $\delta_n = \mathbb{P}\{\mathrm{Bin}(n, \varepsilon) > n/2\}$. Majority vote is Bayes-optimal and*

$$R_n^* = \beta + (1 - 2\beta)\delta_n. \tag{7}$$

The remainder obeys the exponential bound (16) *in the proof. Thus more evaluators remove idiosyncratic error exponentially but leave the common flip exactly.*

5 GAVEL: GROUNDED AGGREGATION WITH VERIFIED EVIDENCE LABELS

The peer-only results distinguish residual factual uncertainty from semantic-orientation ambiguity (Theorems 2 and 3). GAVEL uses verified labels to orient and select an aggregation rule, freezes the policy, and then acquires fresh evidence to certify its risk (Theorem 4). Unlabeled reports identify an uncentered second moment because, pointwise,

$$(YA)(YA)^\top = AA^\top. \tag{8}$$

They do not in general identify either $\mu = \mathbb{E}[YA]$ or $\Sigma = \mathbb{E}[AA^\top] - \mu\mu^\top$. These are distinct non-identifiability statements: a truth swap changes μ but leaves Σ unchanged (Appendix C).

Using only unlabeled and fit data, construct prediction kernels h_k and a measurable eligibility event E. Restrict the deployment partition to E and reindex its positive-mass strata as $\mathcal{T}_1, \ldots, \mathcal{T}_K$; set $\pi_k = \mathbb{P}(X \in \mathcal{T}_k \mid E, \mathcal{F}_0) > 0$, so $\sum_k \pi_k = 1$. This includes eligibility that selects within an original stratum. One candidate solves

$$(\widehat{M}_k + \lambda I)\widehat{w}_k = \widehat{\mu}_k, \qquad h_k(A) = \mathrm{sign}(\widehat{w}_k^\top A), \quad \lambda > 0,$$

with the population error bound and moment–covariance identity given in Proposition 4 and Lemma 2, respectively. We also use bounded report features and a positive definite penalty matrix P in place of λI. Fitting labels determine the penalty and decision threshold before the policy is frozen (Appendix E.6). Audit validity holds conditionally on any such frozen builder. GAVEL combines predictable betting with an allocation target that approaches the Chernoff solution (Appendix D.2).

Assumption 2 (Frozen audit experiment)**.** *Let $\mathcal{F}_0$ contain all unlabeled/fit data and the frozen policy, eligibility, strata, prevalences, prediction ties, targets, priors, allocation/refresh/exploration rules, numerical fallbacks, check times, and budget/pool termination rules. Conditional on $\mathcal{F}_0$, $r_k := \mathbb{P}(h_k(A) \neq Y \mid X \in \mathcal{T}_k, E, \mathcal{F}_0)$ is fixed. For every k, the fresh audit sequence, including independent tie coins when required, is i.i.d. from the frozen eligible stratum-k deployment law; the sequences are mutually independent across strata. Before each query, the sampling distribution is chosen from past audit data, and the stratum is drawn using independent fresh randomness. If a scheduled finite pool is exhausted, the audit terminates without certification.*

Audit assumptions. Policy/data hashes, source-time holdouts, and audit logs check policy freezing and label isolation (Appendix A.3). Pipeline changes require a fresh audit epoch. Dependent losses require a certificate justified under the corresponding dependence assumptions.

6 THE INFORMATION PRICE OF CERTIFICATION

This section characterizes the verified-label cost of certifying a frozen multi-agent policy under an aggregate-risk constraint. An explicit Chernoff solution identifies the optimal information rate, predictable betting attains it under convergent designs, and the expected-label lower bound involves the same constant (Proposition 2 and Theorems 4–5).

Fix $\alpha, \delta \in (0, 1)$. Write $r = (r_1, \ldots, r_K)$ for the frozen stratum risks and let the deployment risk be $R(r) = \pi^\top r$. The safe and unsafe parameter sets are

$$\mathcal{H}_1 = \{p \in [0,1]^K : \pi^\top p < \alpha\}, \qquad \mathcal{H}_0 = \{q \in [0,1]^K : \pi^\top q \geq \alpha\}.$$

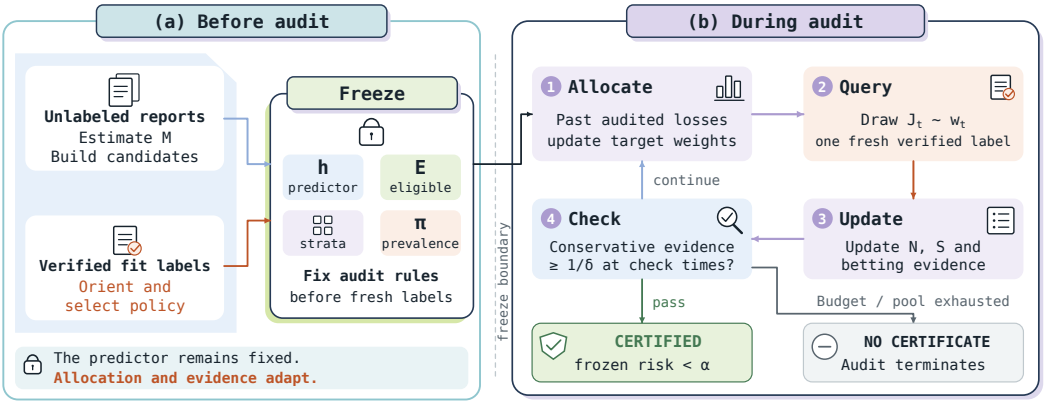


Figure 2: **The fit–freeze–audit procedure.** Fitting labels select a policy. Eligibility and prevalence define the audit population. Fresh labels update allocation and evidence for the frozen policy. The probability of certifying an unsafe risk is at most δ. Algorithm 1 in Appendix D.2 gives the updates.

When the auditor samples stratum k, it observes a Bernoulli loss with mean r_k. Define

$$C^\star(r) := \sup_{w \in \Delta_K} \inf_{q \in \mathcal{H}_0} \sum_{k=1}^{K} w_k \, \mathrm{kl}(r_k, q_k). \tag{9}$$

This controlled Chernoff information is the largest worst-case log-evidence rate against the closest unsafe world. It reduces to $\mathrm{kl}(r, \alpha)$ for one stratum. Prevalence sampling $w = \pi$ is valid but need not maximize information.

Proposition 2 (One-dimensional Chernoff characterization)**.** *For $r \in (0,1)^K$ with $\pi^\top r < \alpha < 1$, let $u_k(z) \geq r_k$ solve $\mathrm{kl}(r_k, u_k(z)) = z$. There is a unique $z^* > 0$ with $\sum_k \pi_k u_k(z^*) = \alpha$. Then $C^\star(r) = z^*$; the optimal allocation is given explicitly by* (22) *in Appendix D.1. The optimizer is unique, positive, and locally smooth on the interior safe region.*

In $C^\star \geq C_\pi \geq \mathrm{kl}(\pi^\top r, \alpha)$, C_π is the information rate under prevalence sampling with stratum identities retained. The two gaps separate allocation and stratification gains. Appendix D.1 quantifies their dependence on heterogeneity and shows that local allocation matches the optimal rate through cubic order in the safety margin.

Let $N_k(t)$ and $S_k(t)$ count audits and errors in stratum k. Before query t, GAVEL chooses positive sampling probabilities w_t, a reference vector $m_t \in [0,1]^K$, and a stake $\lambda_t \geq 0$ using only past audit data. It draws $J_t \sim w_t$ and observes loss L_t in that stratum. Define

$$z_{t,k}(\ell) = \alpha - \pi^\top m_t + \frac{\pi_k}{w_{t,k}}(m_{t,k} - \ell), \qquad \mathcal{E}_t = \prod_{s=1}^{t} \big[1 + \lambda_s z_{s,J_s}(L_s)\big]. \tag{10}$$

The stake keeps every possible factor nonnegative. Under any unsafe risk q, the conditional mean of the next evidence factor is $1 + \lambda_t(\alpha - \pi^\top q) \leq 1$. This gives an anytime-valid certificate while allowing the design to adapt. For each fixed positive allocation, its oracle reference and stake maximize expected log growth among valid one-step factors (Appendix D.2). The implementation estimates stratum risks, blends local and Chernoff targets with a vanishing local weight, adds prevalence exploration, and optimizes the stake. Updates and conservative numerical checks are given in Appendix D.2.

Theorem 4 (Anytime validity and characteristic evidence rate)**.** *Under Assumption 2, any predictable design in* (10) *with nonnegative factors satisfies*

$$\sup_{q \in \mathcal{H}_0} \mathbb{P}_q\left(\sup_{t \geq 1} \mathcal{E}_t \geq 1/\delta\right) \leq \delta. \tag{11}$$

At an interior safe r, suppose the exact-arithmetic designs have strictly positive factors and converge to the Chernoff allocation, reference vector, and stake specified in (28)*. Then*

$$\frac{1}{t}\log \mathcal{E}_t \longrightarrow C^\star(r) \quad a.s., \qquad \frac{\tau_\delta}{\log(1/\delta)} \longrightarrow \frac{1}{C^\star(r)} \quad a.s.\ as\ \delta \downarrow 0, \tag{12}$$

where $\tau_\delta = \inf\{t : \mathcal{E}_t \geq 1/\delta\}$. The limit varies only the threshold, using one infinite audit stream and design rules independent of δ. Bounded check gaps and sublinear conservative log-evidence error preserve the rate. Proofs and sufficient design conditions appear in Appendices D.2–D.3.

Theorem 5 (Information-theoretic lower bound)**.** *Consider any sequential certifier satisfying Assumption 2 at a safe $r \in \mathcal{H}_1 \cap (0,1)^K$. If it certifies with probability one at r, falsely certifies with probability at most δ for every $q \in \mathcal{H}_0$, and has finite $\mathbb{E}_r\tau$, then*

$$\mathbb{E}_r\tau \geq \frac{\log(1/\delta)}{C^\star(r)}. \tag{13}$$

More generally, let τ denote termination with either a certificate or no certificate, with finite $\mathbb{E}_r\tau$, and let D be the certification event. If $0 < \delta < 1 - \gamma \leq 1$ and $\mathbb{P}_r(D) \geq 1 - \gamma$, replace the numerator by $\mathrm{kl}(1-\gamma, \delta)$; without this separation the universal lower bound is zero. For power-one certification, Theorems 4 and 5 give an expected-time lower bound and an almost-sure characteristic time with the same leading constant $1/C^\star(r)$.

Scalar-mixture benchmark. With i.i.d. prevalence allocation, pooled losses are i.i.d. $\mathrm{Bernoulli}(R)$. The scalar beta-mixture in Appendix D.6 is anytime-valid, has exact finite-budget power, and attains the matching expected-time rate $1/\mathrm{kl}(R, \alpha)$. This stronger expected-time conclusion is specific to the scalar audit.

7 Experiments: Mechanisms, Aggregation, and Audits

We study three parts of the evaluation pipeline: dependence among peer reports, construction of a frozen policy, and the labels needed to audit its risk. Synthetic channels and a PEG-policy MMLU intervention control shared errors. Public model reports support binary verification of fixed proposed answers. Sequential Bernoulli audits isolate the effect of allocation by holding the stratified evidence construction fixed.

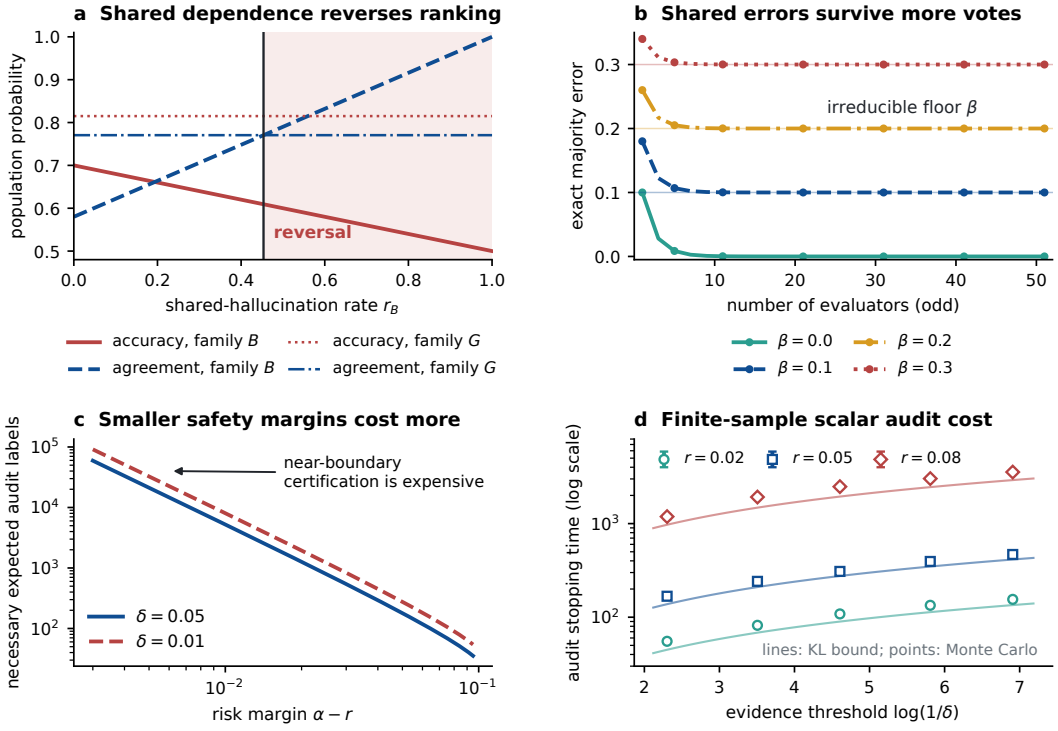


Figure 3: **Synthetic mechanism and audit controls.** a, Exact accuracy/agreement reversal. b, Exact shared-error floor. c, KL information bounds as the risk margin shrinks. d, Mean scalar stopping times over 800 streams.

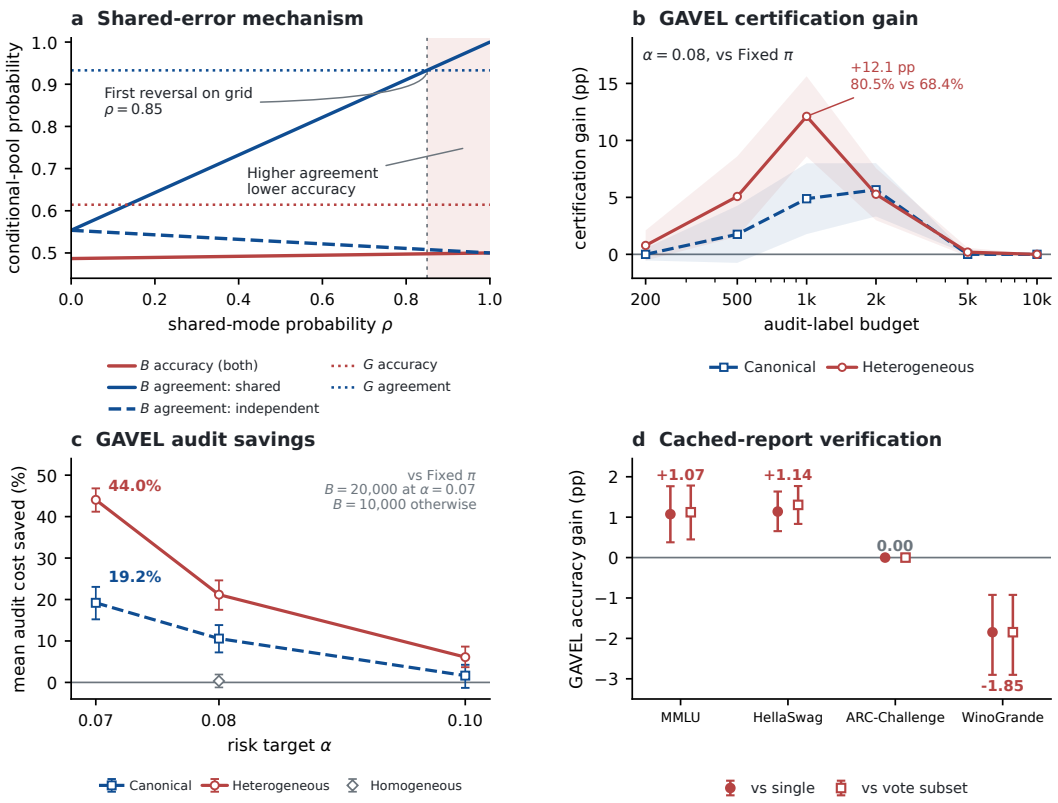


Figure 4: **Dependence, allocation, and verification.** a, Semi-synthetic MMLU agreement reversal under shared signs (shaded). b, GAVEL's certification gain over fixed prevalence at $\alpha = 0.08$, with 95% paired Monte Carlo intervals. c, Relative capped-cost savings across risk targets, with 95% paired bootstrap intervals. Panels b–c hold stratified evidence fixed. d, Verification accuracy gains over fitted single and voting-subset controls, with 95% paired group-bootstrap intervals.

Agreement reversal under shared errors. The family channel gives $r_B^\star = 0.453571$, $a_G = 0.815$, and $a_B = 0.589286$ at $r_B = r_B^\star + 0.10$. Figure 3a shows the population curves, checked with 20,000 simulated tasks per setting. The semi-synthetic MMLU intervention uses the cached PEG policies of `oqwen` and `deepseek_qwen` (Chen et al., 2025). Shared and independent replacement signs have matched individual accuracies. Shared signs first reverse the accuracy and agreement rankings at 0.85 on the 0.01 intervention grid (Figure 4a).

Constructing a policy from verified labels. In the shared-flip channel, risk remains 0.200060 at $(n, \epsilon, \beta) = (13, 0.1, 0.2)$ despite averaging 13 votes (Figure 3b). We then fit policies on the six-judge family channel with two G and four B evaluators. GAVEL uses 10,000 unlabeled reports for its moment and correlation estimates, and 16, 64, or 256 verified labels for fitting. At 64 labels, its mean error is 16.40%, compared with 17.97% for signed weighting, 18.50% for a fitted single evaluator, and 16.57% for cross-validated scalar ridge. The 1.57-point improvement over signed weighting has interval $[1.09, 2.04]$. With 256 fitting labels, GAVEL reaches 15.823% error, about 0.23 percentage points above the exact Bayes risk of 15.590% for this six-evaluator report distribution. Appendix E.6 reports the regularization rule and all three label budgets.

Verifying proposals from real model reports. We evaluate verification on MMLU, HellaSwag, ARC-Challenge, and WinoGrande, covering subject knowledge, event plausibility, science-question reasoning, and commonsense reference resolution (Hendrycks et al., 2021; Zellers et al., 2019; Clark et al., 2018; Sakaguchi et al., 2021). The MMLU panel provides answer-confidence features. RouterBench (Hu et al., 2024) supplies discrete model answers for the other three benchmarks. A fixed proposer supplies each candidate answer, and the benchmark reference determines its correct-

ness. GAVEL and its controls use identical fitting labels and grouped cross-validation before their predictions are frozen. Evaluation measures full-coverage binary verification accuracy. Appendix E.7 gives the data sources, partitions, and controls.

GAVEL improves verification accuracy over the fitted single evaluator by 1.07 percentage points on MMLU and 1.14 points on HellaSwag, reducing errors by 91 and 69 (Figure 4d and Table 3). Its gains over fitted subset voting are 1.12 and 1.30 points. On ARC-Challenge, all three policies make identical decisions and achieve 96.50% accuracy. On WinoGrande, GAVEL achieves 79.95% accuracy, compared with 81.79% for the fitted single. Appendix E.7 decomposes the WinoGrande difference into 15 newly accepted incorrect proposals and one corrected rejection.

Finite-budget value of allocation. The canonical audit has $\pi = (0.40, 0.10, 0.05, 0.45)$, $r = (0.07, 0.22, 0.11, 0.01)$, and $\alpha = 0.08$, giving $C^\star = 0.003779$ and $w^\star = (0.475, 0.173, 0.070, 0.282)$. We compare GAVEL with fixed-prevalence sampling under the same risk estimator, stratified betting evidence, exploration, randomized sampling, and stopping rule. Capped cost $\mathbb{E}\min(\tau, B)$ charges the full budget for noncertification.

Table 1: **Allocation with stratified evidence held fixed.** Mean capped draws at $(\alpha, \delta, B) = (0.08, 0.05, 10000)$. The heterogeneous laws use 512 paired runs each, and the homogeneous control uses 256. The last column is GAVEL minus fixed prevalence with a 95% paired interval.

Risk law	Fixed π	GAVEL	Difference [95% interval]
Canonical	1092.9	977.3	-115.6 $[-155.3, -75.8]$
Heterogeneous	882.8	695.8	-187.0 $[-226.6, -147.3]$
Homogeneous	1307.9	1303.5	-4.4 $[-24.7, 16.0]$

GAVEL saves 10.6% of labels on the canonical law and 21.2% on $r = (0.015, 0.35, 0.25, 0.012)$ relative to fixed prevalence. At 1,000 labels, heterogeneous certification reaches 80.5% versus 68.4% under fixed prevalence (Figure 4b). At $\alpha = 0.07$ and $B = 20000$, the savings increase to 19.2% and 44.0% (Figure 4c). A local-allocation ablation achieves similar costs. Appendix E.9 reports all 25 settings, including parameter and allocation controls, generated risk laws, and null calibration.

8 Discussion and Limitations

Allocation and finite-budget cost. The canonical rates $C^\star = 0.003779$, $C_\pi = 0.003211$, and $C_{\text{local}} = 0.003647$ explain the allocation comparison. Local allocation captures most of the information gain. The Chernoff target saves 3.5% of leading-order cost relative to local allocation and 15.0% relative to prevalence. Risk estimation, exploration, and stake damping also affect finite-budget stopping times. Homogeneous risks make prevalence sampling optimal, linking the control experiment to the equality case in the information-rate analysis.

Scope of the guarantee. GAVEL certifies aggregate frozen-policy risk under known prevalence, fresh independent audit draws, and stable epochs. Unknown prevalence, dependence, drift, and biased reference labels require corresponding changes to the audit.

Empirical scope. Cached reports measure total verification error, with panel-dependent effects on false acceptance. Applications should evaluate their loss weights, subgroup performance, and total fitting and inference costs. Synthetic audits quantify label savings under controlled risk laws.

9 Conclusion

Shared errors can make consensus favor less accurate evaluators. The information bounds distinguish residual uncertainty from semantic orientation, and the Chernoff game quantifies the evidence needed for certification. GAVEL improves frozen-policy verification on MMLU and HellaSwag, matches the fitted single evaluator on ARC-Challenge, and incurs higher error on WinoGrande. Under matched stratified evidence, adaptive allocation reduces synthetic audit costs, with larger savings near the risk boundary.

AI Use Statement

Generative AI tools assisted literature discovery, idea development, drafting and editing, proof development and stress-testing, experiment implementation and execution, numerical and data-provenance checks, figure code and layout, and the finite Lean formalizations described in Appendix G. The authors manually reviewed and verified the AI-assisted content and take responsibility for the final manuscript.

Ethics Statement

The motivating case is drawn from a public judicial order, and the experiments use synthetic draws, public benchmark references, and cached model reports. The proposed system is intended to reduce unsupported automated decisions; it can also increase review burden through abstention. High-stakes deployment requires authoritative-source disclosure, human override, subgroup monitoring, and immediate certificate retirement after a pipeline change.

Reproducibility Statement

The supplementary archive provides runnable experiment code, prepared inputs, configurations, seeds, and recorded outcomes in `experiments/`, figure builders in `figures/`, and the formal proof project in `LeanProofs/`. Appendix E specifies the protocols, data partitions, fitting rules, risk laws, and audit settings. Appendices B–D provide the mathematical proofs, and Appendix G specifies the Lean verification scope and commands.

References

David Blackwell. Equivalent comparisons of experiments. *The Annals of Mathematical Statistics*, 24 (2):265–272, 1953. doi: 10.1214/aoms/1177729032.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. ChatEval: Towards better LLM-based evaluators through multi-agent debate. In *International Conference on Learning Representations*, pp. 9079–9093, 2024.

Baiting Chen, Tong Zhu, Jiale Han, Lexin Li, Gang Li, and Xiaowu Dai. Incentivizing truthful language models via peer elicitation games. In *Advances in Neural Information Processing Systems*, volume 38, pp. 52438–52467, 2025.

Herman Chernoff. Sequential design of experiments. *The Annals of Mathematical Statistics*, 30(3): 755–770, 1959. doi: 10.1214/aoms/1177706205.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? Try ARC, the AI2 Reasoning Challenge. *arXiv preprint arXiv:1803.05457*, 2018. URL https://arxiv.org/abs/1803.05457.

Anirban Dasgupta and Arpita Ghosh. Crowdsourced judgement elicitation with endogenous proficiency. In *Proceedings of the 22nd International Conference on World Wide Web*, pp. 319–330, 2013. doi: 10.1145/2488388.2488417.

A. P. Dawid and A. M. Skene. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 28(1):20–28, 1979.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 11733–11763, 2024. URL https://proceedings.mlr.press/v235/du24e.html.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11:1605–1641, 2010. URL https://jmlr.org/papers/v11/el-yaniv10a.html.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630:625–630, 2024. doi: 10.1038/s41586-024-07421-0.

Aurélien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence. In *29th Annual Conference on Learning Theory*, volume 49, pp. 998–1027. PMLR, 2016. URL https://proceedings.mlr.press/v49/garivier16a.html.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, pp. 4878–4887, 2017.

Peter Grünwald, Rianne de Heide, and Wouter Koolen. Safe testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(5):1091–1128, 2024. URL https://academic.oup.com/jrsssb/article/86/5/1091/7623686.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055–1080, 2021. URL https://arxiv.org/abs/1810.08240.

Chia-Yu Hsu and Shubhanshu Shekhar. Efficient sequential evaluation of large language models. *arXiv preprint arXiv:2607.17409*, 2026. URL https://arxiv.org/abs/2607.17409.

Qitian Jason Hu, Jacob Bieker, Xiuyu Li, Nan Jiang, Benjamin Keigwin, Gaurav Ranganath, Kurt Keutzer, and Shriyash Kaustubh Upadhyay. RouterBench: A benchmark for multi-LLM routing system. *arXiv preprint arXiv:2403.12031*, 2024. URL https://arxiv.org/abs/2403.12031.

Akbir Khan, John Hughes, Dan Valentine, Laura Ruis, Kshitij Sachan, Ansh Radhakrishnan, Edward Grefenstette, Samuel R. Bowman, Tim Rocktäschel, and Ethan Perez. Debating with more persuasive LLMs leads to more truthful answers. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 23662–23733, 2024. URL https://proceedings.mlr.press/v235/khan24a.html.

Yuqing Kong. Dominantly truthful multi-task peer prediction with a constant number of tasks. In *Proceedings of the 2020 ACM-SIAM Symposium on Discrete Algorithms*, pp. 2398–2411, 2020. doi: 10.1137/1.9781611975994.147.

Martin Larsson, Aaditya Ramdas, and Johannes Ruf. The numeraire e-variable and reverse information projection. *The Annals of Statistics*, 53(3):1015–1043, 2025. URL https://arxiv.org/abs/2402.18810.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023. URL https://arxiv.org/abs/2211.09110.

Nolan Miller, Paul Resnick, and Richard Zeckhauser. Eliciting informative feedback: The peer-prediction method. *Management Science*, 51(9):1359–1373, 2005. doi: 10.1287/mnsc.1050.0379.

Subhojyoti Mukherjee, Ardhendu S. Tripathy, and Robert Nowak. Chernoff sampling for active testing and extension to active regression. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151, pp. 7384–7432. PMLR, 2022. URL https://proceedings.mlr.press/v151/mukherjee22a.html.

Maya Okawa. Emergence of biased consensus in multi-agent LLM debates. *arXiv preprint arXiv:2608.02827*, 2026. URL https://arxiv.org/abs/2608.02827.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. LLM evaluators recognize and favor their own generations. In *Advances in Neural Information Processing Systems*, volume 37, pp. 68772–68802, 2024.

Chanwoo Park, Xiangyu Liu, Asuman Ozdaglar, and Kaiqing Zhang. Do LLM agents have regret? A case study in online learning and games. In *International Conference on Learning Representations*, pp. 41214–41234, 2025.

Dražen Prelec. A Bayesian truth serum for subjective data. *Science*, 306(5695):462–466, 2004. doi: 10.1126/science.1102081.

Chen Qian. What does multi-agent LLM debate actually change? A layered analysis of disagreement and answer quality. *arXiv preprint arXiv:2609.08016v2*, 2026. URL https://arxiv.org/abs/2609.08016v2.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. WinoGrande: An adversarial Winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106, 2021. doi: 10.1145/3474381. URL https://doi.org/10.1145/3474381.

Victor Shnayder, Arpit Agarwal, Rafael Frongillo, and David C. Parkes. Informed truthfulness in multi-task peer prediction. In *Proceedings of the 2016 ACM Conference on Economics and Computation*, pp. 179–196, 2016. doi: 10.1145/2940716.2940790.

Minsoo Song, Chanwoo Kim, Sugyeong Eo, and Chanjun Park. Beyond consensus: Downward bias and role asymmetry in multi-agent LLM judges for subjective evaluation. *arXiv preprint arXiv:2608.30373*, 2026. URL https://arxiv.org/abs/2608.30373.

United States District Court. Mata v. Avianca, Inc., 678 F. Supp. 3d 443 (S.D.N.Y. 2023). Southern District of New York, 2023. URL https://www.nhd.uscourts.gov/sites/default/files/pdf/Mata-v-Avianca-sanctions-order.PDF. Opinion and order on sanctions, No. 22-cv-1461 (PKC), ECF No. 54, June 22, 2023.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9440–9450, 2024. URL https://aclanthology.org/2024.acl-long.511/.

Ziyu Xu, Nikos Karampatziakis, and Paul Mineiro. Active, anytime-valid risk controlling prediction sets. In *Advances in Neural Information Processing Systems*, volume 37, pp. 60110–60132, 2024.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4791–4800, 2019. URL https://aclanthology.org/P19-1472/.

APPENDIX CONTENTS

A Related Work, Notation, and Assumptions 14

A.1 Additional Related Work 14

A.2 Notation 15

A.3 Assumptions and Scope 15

B Proofs of the Peer-Only Results 15

B.1 Proof of Theorem 1 15

B.2 Proof of Theorem 2 16

B.3 Proof of Theorem 3 17

B.4 Proof of Proposition 1 17

B.5 Reporting error and epistemic error 17

B.6 Decision preservation under peer-only processing 18

B.7 Noisy anchors and semantic orientation 18

C Moment Geometry for Candidate Construction 20

C.1 Regularized moment and covariance directions 20

D Proofs for Grounded Sequential Certification 21

D.1 One-dimensional solution of the aggregate-risk Chernoff game 21

D.2 GAVEL: design, algorithm, and proof of Theorem 4 24

D.3 Checked stopping and conservative evidence 28

D.4 A likelihood lemma under adaptive sampling 29

D.5 Proof of Theorem 5 29

D.6 Pooled Bernoulli reduction and beta-mixture validity 30

D.7 Exact finite-budget power 31

D.8 Matching expected stopping time for the scalar audit 32

D.9 Fresh epochs 33

E Additional Experimental Details 33

E.1 Experiments, data, and scope 33

E.2 Mechanism, orientation, solver, and candidate controls 34

E.3 Semi-synthetic MMLU shared-error control 35

E.4 Scalar regression and assumption-failure stress tests 35

E.5 Reproduction and resource accounting 35

E.6 Moment-based aggregation with limited fitting labels 35

E.7 Verification from cached model reports 36

E.8 Synthetic audit parameters and numerical checks 38

E.9 Allocation, heterogeneity, and finite budgets 39

F Scope of the Peer-Only and Audit Results 41

G Lean Verification Boundary 42

A Related Work, Notation, and Assumptions

A.1 Additional Related Work

Peer prediction and semantic orientation. Peer-prediction mechanisms use reports from other participants to reward informative behavior (Prelec, 2004; Miller et al., 2005; Dasgupta & Ghosh, 2013; Shnayder et al., 2016). Their guarantees concern reporting incentives under assumptions on signals and mechanisms. Determinant mutual information connects truthful elicitation to an information score invariant under information-preserving relabelings (Kong, 2020). PEG develops peer elicitation for LLM discriminators and their online updates (Chen et al., 2025). Our external label Y gives these reports a factual target. The score-reversal calculation relates shared errors to agreement, determinant mutual information, and Shannon mutual information. The truth-swap result then identifies the orientation information needed to interpret an informative report as a factual judgment (Theorems 1 and 3).

Interaction and the complete information state. Debate and referee systems organize communication among models to improve reasoning or evaluation (Du et al., 2024; Chan et al., 2024; Khan et al., 2024). Positional and self-preference effects influence model judgments (Wang et al., 2024; Panickssery et al., 2024), while consensus, prompted disagreement, and conformity affect the outcomes of interaction (Song et al., 2026; Qian, 2026; Okawa, 2026). We analyze the statistical experiment available to such a protocol. Including reasoning states, tool-accessible material, and other peer-accessible information in V makes the bound apply to the full interaction. An independent authoritative source enriches this experiment. Its factual value is measured by the Bayes risk of the enriched experiment (Theorem 2).

Reliability, uncertainty, and selective prediction. Latent-annotator models estimate label distributions and annotator reliability through structural assumptions (Dawid & Skene, 1979). Semantic uncertainty uses variation in meaning to detect confabulation (Farquhar et al., 2024). Selective prediction studies the risk attained at a chosen coverage (El-Yaniv & Wiener, 2010; Geifman & El-Yaniv, 2017). These approaches motivate distinct components of our formulation: labels orient the aggregation rule, report features support policy construction, and the eligible population determines selective risk. The orientation theorem characterizes the information required to resolve truth-swap ambiguity under peer-observable selection, and the audit evaluates the resulting frozen policy.

Active risk control. Time-uniform confidence sequences support inference at adaptive stopping times (Howard et al., 2021). Xu et al. (2024) choose whether to label each incoming example under an expected labeling-budget constraint. Their prediction-assisted betting process yields anytime-valid risk control. They derive a labeling policy by optimizing a quadratic lower bound on log-wealth growth and relate estimation error to regret for that bound. Our audit chooses among task strata and tests the composite null $\pi^\top q \geq \alpha$. Its exact solution identifies both the least-favorable unsafe risk vector and the allocation that maximizes information per verified label (Proposition 2).

Chernoff allocation and optimal evidence. Sequential experimental design allocates observations according to their ability to distinguish competing hypotheses (Chernoff, 1959). Track-and-Stop characterizes the sample complexity of best-arm identification and tracks the corresponding optimal proportions (Garivier & Kaufmann, 2016). Chernoff sampling also supports finite-confidence analyses of active testing and extensions to regression (Mukherjee et al., 2022). For the aggregate-risk null, equal coordinate KL divergences reduce the saddle-point calculation to one scalar equation. The optimal allocation is unique, positive, and locally smooth in the interior safe region. The saddle value gives the expected-label lower-bound constant and separates allocation gains from the gains of retaining stratum identities.

The corresponding evidence construction connects this allocation problem to log-optimal e-variables (Grünwald et al., 2024; Larsson et al., 2025). At each positive fixed allocation, the least-favorable likelihood ratio is the unique valid factor maximizing expected log growth. Predictable estimates converge to this factor and attain its rate almost surely under the stated design assumptions. Thus the aggregate-risk geometry determines an explicit allocation, an optimal one-step evidence factor, and their common characteristic rate (Propositions 2 and 7, Theorem 4).

A.2 NOTATION

For $u, v \in (0,1)$, the binary relative entropy is

$$\mathrm{kl}(u,v) = u \log \frac{u}{v} + (1-u) \log \frac{1-u}{1-v},$$

with the usual lower-semicontinuous extension to the boundary. We use $\Delta_K = \{w \in [0,1]^K : \sum_k w_k = 1\}$. All almost-sure and conditional statements are understood modulo the relevant null sets.

A.3 ASSUMPTIONS AND SCOPE

Complete peer state and external information. Assumption 1 includes all information available to the protocol in V, making every permitted interaction a garbling of that experiment. Tool and provenance records can identify information exposed after the initial vote. Newly accessible material belongs in the complete state, and authoritative evidence places the protocol in the grounded setting. The Bayes risk is then computed for the enriched experiment.

Diagnostic channel independence. The binary channels use mutual independence to separate a common error factor from idiosyncratic noise. Residual dependence and conditional calibration are natural diagnostics for these assumptions. Empirical conditional laws can replace the stylized channels when dependence changes their closed-form phase boundaries. The complete-state Bayes characterization applies to those conditional laws (Theorem 2).

Strata and deployment prevalence. Measurable disjoint strata and their deployment prevalences identify the audit target with selective risk. Comparing the sampling frame with deployment proportions checks this correspondence. Uncertain prevalences call for a certificate based on a prevalence confidence set or appropriate importance weighting.

Policy freezing and audit isolation. Assumption 2 fixes the policy before certification labels are acquired. Immutable policy hashes and label-access records can document this separation. A policy revised using audit labels starts a fresh audit epoch. Reusing such labels requires a validity analysis for the resulting conditioning sigma-field.

Fresh audit observations. Conditionally independent fresh audit arrays support the likelihood-ratio and betting arguments. Source and time holdouts, duplicate checks, and dependence diagnostics help assess this model. Dependent observations require a justified block or predictable-mean e-process, with abstention until the corresponding certificate is available.

Deployment epochs. A stable deployment law transfers the audited risk to deployment within an epoch. Model, prompt, retrieval, and covariate versions define the relevant monitoring record. A change ends the current deployment certificate and calls for recertification on fresh data.

Validity and characteristic rates. Interiority and convergence to the Chernoff design establish the characteristic evidence rate in Theorem 4. Positive allocation coordinates, a positive safety margin, and optimizer stability are useful diagnostics. Predictable betting retains anytime validity whenever its factor conditions hold. Boundary risks and nonunique limiting designs call for their own rate analysis.

B PROOFS OF THE PEER-ONLY RESULTS

B.1 PROOF OF THEOREM 1

Fix a family and suppress h. Write $r = r_h$, $q = q_h$, and $d = 2q - 1$. Mutual independence in (1) gives

$$\mathbb{P}(S_i = Y) = (1-r)q + r\mathbb{P}(H = Y) = (1-r)q + \frac{r}{2} = \frac{1}{2} + \frac{(1-r)d}{2}.$$

On $\{M = 1\}$, $S_i S_j = H^2 = 1$. On $\{M = 0\}$, $S_i S_j = E_i E_j$, and independence gives $\mathbb{E}[E_i E_j] = d^2$. Hence

$$c = \mathbb{E}[S_i S_j] = r + (1 - r)d^2, \qquad \mathbb{P}(S_i = S_j) = \frac{1 + c}{2}.$$

Both signals are balanced. Their joint-probability table, in either common ordering of $\{-1, +1\}$, is

$$P_{ij} = \frac{1}{4}\begin{pmatrix} 1 + c & 1 - c \\ 1 - c & 1 + c \end{pmatrix}.$$

Therefore $|\det P_{ij}| = c/4$. Since each marginal cell has probability $1/2$, direct substitution in the definition of mutual information gives

$$\mathrm{I}(S_i; S_j) = \frac{1 + c}{2}\log(1 + c) + \frac{1 - c}{2}\log(1 - c).$$

For $0 < c < 1$, its derivative is $\frac{1}{2}\log\{(1 + c)/(1 - c)\} > 0$; the endpoints follow by continuity.

For the two families, $c_B > c_G$ is equivalent to

$$r_B + (1 - r_B)d_B^2 > r_G + (1 - r_G)d_G^2,$$

which, because $1 - d_B^2 > 0$, is equivalent to $r_B > r_B^\star$ in (4). Moreover,

$$r_B^\star - r_G = (1 - r_G)\frac{d_G^2 - d_B^2}{1 - d_B^2} > 0$$

and

$$1 - r_B^\star = (1 - r_G)\frac{1 - d_G^2}{1 - d_B^2} > 0.$$

Thus $r_G < r_B^\star < 1$. If $r_B > r_B^\star > r_G$, then $(1 - r_B)d_B < (1 - r_G)d_G$, so (3) also gives $a_B < a_G$. The strict order of all three peer scores follows from their monotonicity in c. Finally, at $r_B = 1$, $S_i = H_B$ is independent of Y and common to every evaluator. Hence $\mathrm{I}(S_i; Y) = 0$, $p_B = 1$, $\mathrm{DMI}_B = 1/4$, and $\mathrm{I}(S_i; S_j) = \log 2$. □

B.2 Proof of Theorem 2

For P_X-almost every x, define sub-probability measures on $\mathcal{V} = \prod_i \mathcal{V}_i$ by

$$\nu_x^{\pm}(B) := \mathbb{P}(Y = \pm 1, V \in B \mid X = x). \tag{14}$$

Let $\lambda_x = \nu_x^+ + \nu_x^-$ and write $f_x^{\pm} = d\nu_x^{\pm}/d\lambda_x$. This avoids undefined class-conditionals when a class has zero conditional probability.

Because the spaces are standard Borel, choose a jointly measurable posterior $\rho(x, v) = \mathbb{P}(Y = +1 \mid X = x, V = v)$ and compose the protocol kernel with all independent decision randomization to obtain the jointly measurable $\varphi(x, v) = \mathbb{P}(\widehat{Y} = +1 \mid X = x, V = v)$. Assumption 1 ensures that the same φ applies under both labels. If μ is the joint law of (X, V), then

$$\begin{aligned}\mathbb{P}(\widehat{Y} \neq Y) &= \int \{(1 - \varphi)\rho + \varphi(1 - \rho)\}\, d\mu \\ &\geq \int \min\{\rho, 1 - \rho\}\, d\mu.\end{aligned}$$

The inequality is pointwise, and equality is attained by the measurable rule that chooses $+1$ on $\{\rho \geq 1/2\}$ and -1 otherwise. Disintegrating μ over X identifies the final integral with the right side of (5); equivalently, the measurable Radon–Nikodym theorem for kernels supplies jointly measurable versions of $f_x^{\pm}$. This also proves attainment rather than merely a pointwise-in-x infimum.

If the conditional prior is balanced, $\nu_x^{\pm} = \frac{1}{2}P_x^{\pm}$, where $P_x^{\pm}$ are the class-conditional probability measures. For any dominating measure,

$$\int \min\{dP_x^+, dP_x^-\} = 1 - \mathrm{TV}(P_x^+, P_x^-).$$

Multiplying by $1/2$ and integrating over X yields

$$R_V^* = \tfrac{1}{2}\mathbb{E}_X[1 - \mathrm{TV}(P_{V|Y=1,X}, P_{V|Y=-1,X})]. \tag{15}$$

The construction of φ_x used only the Markov kernel; reports need not be truthful. □

B.3 Proof of Theorem 3

Couple P and $P^{\leftrightarrow}$ so that the training observations, fresh peer-only observables, and independent learner randomization are identical, while the fresh label is negated. A full-coverage learner then produces the same $\widehat{Y} \in \{-1, +1\}$ in both worlds. Pointwise,

$$\mathbf{1}\{\widehat{Y} \neq Y\} + \mathbf{1}\{\widehat{Y} \neq (-Y)\} = 1.$$

Expectation proves (6). Hence at least one risk is at least $1/2$ for every learner. An independent fair-coin prediction has risk exactly $1/2$ under every distribution, proving equality in the minimax statement for any nonempty truth-swap-closed class.

For a selective rule, its acceptance indicator is a function of the same peer-only observables and randomization, so the coupled acceptance event is identical in both worlds. Conditional on that event, provided its common probability is positive, the same pointwise identity holds and the two selective risks sum to one. □

B.4 Proof of Proposition 1

Set $U = YF$. Since Y is balanced and independent of F, U is balanced and independent of F. Conditional on U, the observations are $S_i = UE_i$ with i.i.d. crossover probability $\varepsilon < 1/2$. For odd n, the MAP rule for U is majority vote and its exact error is

$$\delta_n = \mathbb{P}\{\mathrm{Bin}(n, \varepsilon) > n/2\}.$$

Let $p(S) = \mathbb{P}(U = 1 \mid S)$. Because $Y = UF$,

$$\mathbb{P}(Y = 1 \mid S) = (1 - \beta)p(S) + \beta\{1 - p(S)\} = \beta + (1 - 2\beta)p(S).$$

For $\beta < 1/2$, this posterior is at least $1/2$ exactly when $p(S) \geq 1/2$, so the same majority rule is Bayes-optimal for Y. At $\beta = 1/2$, the posterior equals $1/2$ for every observation; majority vote is still Bayes-optimal, but is no longer unique. Its U-classification error is independent of F, and consequently

$$\begin{aligned}\mathbb{P}(\widehat{Y} \neq Y) &= (1 - \beta)\delta_n + \beta(1 - \delta_n) \\ &= \beta + (1 - 2\beta)\delta_n.\end{aligned}$$

This proves exact Bayes optimality and (7). The Chernoff binomial-tail bound

$$\mathbb{P}\{\mathrm{Bin}(n, \varepsilon) \geq n/2\} \leq \exp\{-n\,\mathrm{kl}(1/2, \varepsilon)\}$$

implies, for $0 < \varepsilon < 1/2$,

$$0 \leq R_n^* - \beta \leq (1 - 2\beta)\exp\{-n\,\mathrm{kl}(1/2, \varepsilon)\}. \tag{16}$$

Using $>$ rather than $\geq$ only tightens the bound. □

B.5 Reporting error and epistemic error

Reporting error admits the following distribution-free decomposition. Let g be a deterministic aggregation rule, define

$$\begin{aligned}\mathrm{ObjErr}_g &= \mathbb{P}\{g(A, X) \neq Y\}, & \mathrm{EpiErr}_g &= \mathbb{P}\{g(S, X) \neq Y\}, \\ \mathrm{StratGap}_g &= \mathbb{P}\{g(A, X) \neq g(S, X)\}.\end{aligned}$$

Then

$$|\mathrm{ObjErr}_g - \mathrm{EpiErr}_g| \leq \mathrm{StratGap}_g \leq \mathbb{P}(A \neq S) \leq \min\left\{1, \sum_i \mathbb{P}(A_i \neq S_i)\right\}.$$

Indeed, the two error indicators can differ only when the two aggregate predictions differ, proving the first inequality. The remaining inequalities follow by event inclusion and the union bound. Thus incentives that eliminate report distortion expose, but do not eliminate, the factual error of the private-state experiment.

B.6 Decision preservation under peer-only processing

Proposition 3 (Exact Bayes-risk loss and equality condition)**.** *Under Assumption 1, let* $\rho_V = \mathbb{P}(Y = 1 \mid X, V)$ *and* $\rho_T = \mathbb{P}(Y = 1 \mid X, T)$*. The optimal full-coverage binary risks satisfy*

$$\rho_T = \mathbb{E}[\rho_V \mid X, T], \tag{17}$$

$$R_T^* - R_V^* = \mathbb{E}|\rho_V - 1/2| - \mathbb{E}\big|\mathbb{E}[\rho_V - 1/2 \mid X, T]\big| \geq 0. \tag{18}$$

Equality holds if and only if

$$\mathbb{P}(\rho_V > 1/2 \mid X, T)\,\mathbb{P}(\rho_V < 1/2 \mid X, T) = 0 \quad \textit{a.s.} \tag{19}$$

States at posterior $1/2$ *may coexist with either side of the decision boundary.*

Proof. The Markov condition gives $\mathbb{P}(Y = 1 \mid X, V, T) = \rho_V$; the tower property then gives (17). For any observation U, the optimal binary zero–one risk is $\mathbb{E}\min(\rho_U, 1 - \rho_U) = 1/2 - \mathbb{E}|\rho_U - 1/2|$. Subtracting the two expressions proves the identity in (18), and conditional Jensen proves nonnegativity.

To characterize equality, set $D = \rho_V - 1/2$, $D_+ = \max(D, 0)$ and $D_- = \max(-D, 0)$. Pointwise in (X, T),

$$\mathbb{E}[|D| \mid X, T] - |\mathbb{E}[D \mid X, T]| = 2\min\{\mathbb{E}[D_+ \mid X, T], \mathbb{E}[D_- \mid X, T]\}.$$

This nonnegative gap has zero expectation exactly when its conditional minimum vanishes almost surely. A nonnegative variable has zero conditional mean on an event exactly when it is zero there conditionally almost surely. Thus one of the two strict-sign conditional probabilities must vanish, which is precisely (19). □

Why preserving the full posterior is unnecessary. Take X constant, let V have two equiprobable states with posteriors 0.6 and 0.9, and make T constant. Then $R_V^* = (0.4 + 0.1)/2 = 0.25$ and $R_T^* = 1 - 0.75 = 0.25$. The optimal decision is always $+1$, even though the transcript cannot recover the original posterior. Full posterior preservation is sufficient, but is not necessary for equality for binary zero–one loss. This statement concerns optimal risks; an arbitrary implemented decision rule can still perform worse.

A nonzero barrier with a deterministic authoritative label. Let X be constant, W balanced Rademacher, $Y = W$, and $V = WE$ with independent $\mathbb{P}(E = -1) = \epsilon \in (0, 1/2)$. The label is a deterministic function of the hidden authoritative record (X, W), yet the complete accessible experiment has $R_V^* = \epsilon > 0$. By contrast, if Y is measurable with respect to X alone, its complete-state Bayes risk is zero. The positive risk in this example arises from uncertainty about the hidden record given the accessible state.

For an unknown orientation $\theta \in \{-1, +1\}$, replacing V by θWE gives the same visible law in both worlds. Each known-orientation world still has Bayes risk ϵ; minimax risk over the two unknown orientations is $1/2$. Within-world overlap and across-world semantic ambiguity are different objects.

B.7 Noisy anchors and semantic orientation

Lemma 1 (Orientation from independent anchors)**.** *Freeze a binary rule* g *independently of the anchors. Suppose the* m *anchors are i.i.d. and* $\mathbb{P}\{Y = \theta g(X, V)\} = (1 + \kappa)/2$ *for a fixed unknown* $\theta \in \{-1, +1\}$ *and* $0 < \kappa \leq 1$*. For a fixed tie rule, let* $\widehat{\theta} = \operatorname{sign} \sum_{j=1}^m Y_j g(X_j, V_j)$*. Then*

$$\mathbb{P}(\widehat{\theta} \neq \theta) \leq e^{-m\kappa^2/2}; \qquad m \geq \frac{2}{\kappa^2} \log \frac{1}{\delta_{\text{orient}}} \quad \Longrightarrow \quad \mathbb{P}(\widehat{\theta} \neq \theta) \leq \delta_{\text{orient}}.$$

On an independent deployment draw from the same law, the expected risk of $\widehat{\theta} g$ *is* $(1 - \kappa)/2 + \kappa \mathbb{P}(\widehat{\theta} \neq \theta)$*.*

Proof. The independent variables $B_j = \theta Y_j g(X_j, V_j)$ lie in $[-1, 1]$ and have mean κ. Orientation error, including an unfavorable tie, implies $\sum_j B_j \leq 0$. Hoeffding's inequality gives $\mathbb{P}(\sum_j (B_j - \kappa) \leq -m\kappa) \leq \exp(-m\kappa^2/2)$. With the correct orientation deployment error is $(1 - \kappa)/2$, and with the opposite orientation it is $(1 + \kappa)/2$. Independence of the new draw gives the stated mixture of these two risks. □

The bound gives a sufficient label budget for this orientation model. Total grounding cost depends on the joint information provided by each label. Independent fitting and audit samples separate policy construction from certification.

Corollary 1 (Exact orientation risk for odd anchor budgets)**.** *Under Lemma 1, let m be odd and define*

$$p_m(\kappa) = \mathbb{P}\left\{ \mathrm{Bin}\left(m, \frac{1-\kappa}{2} \right) > \frac{m}{2} \right\}.$$

The majority orientation rule has error $p_m(\kappa)$ under either orientation. In the reduced experiment that observes only the signs $U_j = Y_j g(X_j, V_j)$,

$$\inf_{\widetilde{\theta}} \max_{\theta \in \{-1,+1\}} \mathbb{P}_\theta(\widetilde{\theta} \neq \theta) = p_m(\kappa),$$

where the infimum includes randomized rules based on $(U_1, \ldots, U_m)$. Moreover,

$$p_m(\kappa) \leq (1-\kappa^2)^{m/2} \leq e^{-m\kappa^2/2}.$$

For each fixed $0 < \kappa < 1$, along odd $m \to \infty$,

$$p_m(\kappa) = \Theta_\kappa\left(m^{-1/2}(1-\kappa^2)^{m/2} \right).$$

Proof. **Step 1. Identify the optimal sign decision.** For $0 < \kappa < 1$, the two sign laws have likelihood ratio

$$\log \frac{dP_+^{(m)}}{dP_-^{(m)}}(U_1, \ldots, U_m) = \log \frac{1+\kappa}{1-\kappa} \sum_{j=1}^{m} U_j.$$

Thus majority is the Bayes rule under the uniform prior on θ. Under either orientation, the number of signs opposite to θ is $\mathrm{Bin}(m, (1-\kappa)/2)$. Since m is odd, majority has the same error $p_m(\kappa)$ in both worlds. The maximum risk of any rule is at least its uniform-prior Bayes risk, which is at least $p_m(\kappa)$. Majority attains this lower bound. At $\kappa = 1$, every sign identifies θ exactly.

Step 2. Bound the binomial tail. Put $e = (1-\kappa)/2 \in (0, 1/2)$. The exponential Markov bound for a binomial variable, optimized over positive exponents, gives

$$p_m(\kappa) \leq e^{-m\,\mathrm{kl}(1/2, e)} = [4e(1-e)]^{m/2} = (1-\kappa^2)^{m/2}.$$

The inequality $\log(1-x) \leq -x$ gives the stated Hoeffding comparison.

Step 3. Match the polynomial prefactor. Write $m = 2n+1$ and let

$$b_m = \binom{m}{n+1} e^{n+1}(1-e)^n$$

be the first mass in the error tail. For every $j \geq n+1$, the ratio of successive binomial masses is

$$\frac{m-j}{j+1}\frac{e}{1-e} \leq \frac{e}{1-e} < 1.$$

Consequently,

$$b_m \leq p_m(\kappa) \leq \frac{1-e}{1-2e}\, b_m.$$

Stirling's formula yields

$$b_m = \sqrt{\frac{2}{\pi m}}\sqrt{\frac{e}{1-e}}[4e(1-e)]^{m/2}\{1 + O(m^{-1})\}.$$

The two tail bounds establish the claimed order for each fixed κ. □

C MOMENT GEOMETRY FOR CANDIDATE CONSTRUCTION

Let $Z = YA \in \{-1, +1\}^n$. Since $Y^2 = 1$,

$$ZZ^\top = Y^2 AA^\top = AA^\top$$

pointwise, proving (8). Hence unlabeled reports identify $M = \mathbb{E}[ZZ^\top] = \mathbb{E}[AA^\top]$. They do not identify $\mu = \mathbb{E}[Z]$ or $\Sigma = M - \mu\mu^\top$ in general, but these are distinct failures of identification. A global truth swap reverses μ while preserving $\mu\mu^\top$ and thus Σ. For a separate counterexample, let the visible scalar A be balanced Rademacher in both worlds. With $Y = A$, $(\mu, \Sigma) = (1, 0)$; with independent balanced Y, $(\mu, \Sigma) = (0, 1)$. Both worlds have the same visible law and $M = 1$.

Proposition 4 (Oracle correlated-error certificate)**.** *For any $w \in \mathbb{R}^n$ with $m = w^\top \mu > 0$ and $v = w^\top \Sigma w$, the weighted rule $g_w(A) = \operatorname{sign}(w^\top A)$ obeys*

$$\mathbb{P}\{g_w(A) \neq Y\} \leq \frac{v}{m^2 + v}. \tag{20}$$

If $\Sigma \succ 0$ and $\mu \neq 0$, this upper bound is minimized over its nonempty domain by any $w = c\Sigma^{-1}\mu$, $c > 0$, giving

$$\mathbb{P}\{g_w(A) \neq Y\} \leq \frac{1}{1 + \mu^\top \Sigma^{-1} \mu}.$$

Proof. The error event is contained in $\{w^\top Z \leq 0\}$. Cantelli's inequality applied to $w^\top Z$, whose mean and variance are m and v, gives (20). Minimizing it is equivalent to maximizing the Rayleigh quotient $(w^\top \mu)^2/(w^\top \Sigma w)$. Cauchy–Schwarz in the Σ inner product gives

$$(w^\top \mu)^2 = \{(\Sigma^{1/2} w)^\top (\Sigma^{-1/2} \mu)\}^2 \leq (w^\top \Sigma w)(\mu^\top \Sigma^{-1} \mu),$$

with equality exactly on positive multiples of $\Sigma^{-1}\mu$ when $w^\top \mu > 0$. □

If, for each n, $\mu_i = \mu > 0$, $\operatorname{Var}(Z_i) = v$, and $\operatorname{Corr}(Z_i, Z_j) = \rho$ for $i \neq j$, then uniform voting has $m = n\mu$ and $v_n = nv\{1 + (n-1)\rho\}$. Substitution in (20) yields

$$\mathbb{P}(\text{uniform vote errs}) \leq \frac{v\{1 + (n-1)\rho\}}{n\mu^2 + v\{1 + (n-1)\rho\}}.$$

With fixed μ, v, ρ, the bound is $O(1/n)$ at $\rho = 0$ and has a positive limit at every fixed $\rho > 0$. This is an upper certificate, whereas Proposition 1 gives an exact lower floor in a specific generative channel.

C.1 REGULARIZED MOMENT AND COVARIANCE DIRECTIONS

Lemma 2 (A common population direction)**.** *If $M = \Sigma + \mu\mu^\top$, $\Sigma \succeq 0$, and $\Lambda \succ 0$, then*

$$(M + \Lambda)^{-1}\mu = \frac{(\Sigma + \Lambda)^{-1}\mu}{1 + \mu^\top (\Sigma + \Lambda)^{-1} \mu}. \tag{21}$$

Thus both directions give identical zero-threshold predictions under the same rule for ties. Writing $D = 1 + \mu^\top (\Sigma + \Lambda)^{-1}\mu > 0$, a threshold t for the moment score corresponds to threshold Dt for the covariance score. The choice $\Lambda = \lambda I$ recovers scalar ridge regularization. For $\Lambda = 0$, the same identity holds if $\Sigma \succ 0$ and connects the moment rule to Proposition 4.

Proof. Write $H = \Sigma + \Lambda \succ 0$. The rank-one inverse identity gives $(H + \mu\mu^\top)^{-1} = H^{-1} - H^{-1}\mu\mu^\top H^{-1}/(1 + \mu^\top H^{-1}\mu)$. Multiplication by μ yields (21). The denominator is at least one, so scaling preserves signs and zero scores, including $\mu = 0$. Scaling the threshold by the same positive factor preserves nonzero-threshold decisions. Positive definiteness justifies the same calculation at $\Lambda = 0$ when stated. □

The identity holds for population moments. Finite-sample policies are frozen before risk auditing. Implementations compute the direction by solving a linear system.

D PROOFS FOR GROUNDED SEQUENTIAL CERTIFICATION

D.1 ONE-DIMENSIONAL SOLUTION OF THE AGGREGATE-RISK CHERNOFF GAME

For $q_k^* = u_k(z^*)$, the explicit solution is

$$C^\star(r) = z^*, \qquad w_k^\star(r) = \frac{\pi_k q_k^*(1-q_k^*)/(q_k^* - r_k)}{\sum_j \pi_j q_j^*(1-q_j^*)/(q_j^* - r_j)}. \tag{22}$$

Proof of Proposition 2. **Step 1. Construct the equal-KL alternative.** Fix an interior safe r and positive π. For $q \in (r_k, 1)$,

$$\partial_q \,\mathrm{kl}(r_k, q) = \frac{q - r_k}{q(1-q)} > 0, \qquad \partial_q^2 \,\mathrm{kl}(r_k, q) = \frac{r_k}{q^2} + \frac{1 - r_k}{(1-q)^2} > 0.$$

The upper branch increases continuously from zero to infinity, so $u_k(z)$ is well-defined, continuous, and strictly increasing. Since $\sum_k \pi_k u_k(0) = \pi^\top r < \alpha$ and the sum tends to one, its unique crossing $z^* > 0$ exists. All $q_k^* = u_k(z^*)$ are strictly between r_k and one.

Step 2. Establish the saddle value. For every $w \in \Delta_K$, feasibility of q^* and equality of its coordinate divergences imply

$$\inf_{q \in \mathcal{H}_0} \sum_k w_k \,\mathrm{kl}(r_k, q_k) \le \sum_k w_k \,\mathrm{kl}(r_k, q_k^*) = z^*.$$

Thus $C^\star \le z^*$. Write $d_k = (q_k^* - r_k)/(q_k^*(1-q_k^*))$ and $c = (\sum_j \pi_j/d_j)^{-1} > 0$. Formula (22) says $w_k^* = c\pi_k/d_k$, so the gradient of $f(q) = \sum_k w_k^* \,\mathrm{kl}(r_k, q_k)$ at q^* is $c\pi$. Convexity gives, for every feasible q,

$$f(q) \ge f(q^*) + c\pi^\top (q - q^*) \ge z^*.$$

The inequality extends to the cube boundary by lower semicontinuity. Hence q^* is the inner minimizer and $C^\star \ge z^*$.

Step 3. Identify the unique optimal allocation. If w is any outer optimizer, q^* must minimize its inner objective since its value there is z^*. The cube constraints are inactive at q^*, and the aggregate constraint is active. The convex KKT conditions (Slater holds) force $w_k d_k = \lambda \pi_k$ for some $\lambda \ge 0$. If $\lambda = 0$, every w_k vanishes, contradicting $\sum_k w_k = 1$. Normalization therefore gives exactly w^* and proves uniqueness and positivity, even when competing w are allowed zero coordinates.

Step 4. Establish local smoothness. For $z > 0$, the inverse-function theorem applies because $d_k > 0$. The derivative of $\sum_k \pi_k u_k(z)$ with respect to z is $\sum_k \pi_k/d_k > 0$. The implicit-function theorem therefore gives local smoothness of z^*, q^*, w^* in the interior safe parameters (with π restricted to the positive simplex). □

A deterministic bracket. Let $R = \pi^\top r$ and $\bar{q}_k = r_k + (\alpha - R)(1 - r_k)/(1 - R)$. Then $\pi^\top \bar{q} = \alpha$ and $r_k < \bar{q}_k < 1$. The root lies in

$$0 < z^* \le z_{\mathrm{hi}} := \max_k \,\mathrm{kl}(r_k, \bar{q}_k),$$

because $u_k(z_{\mathrm{hi}}) \ge \bar{q}_k$ for every coordinate. One outer scalar root search and monotone coordinate inversions therefore suffice. Equivalently, minimizing z subject to $\pi^\top q \ge \alpha$ and $\mathrm{kl}(r_k, q_k) \le z$ for every k has value z^*: for $z < z^*$, each feasible coordinate is at most $u_k(z)$, whose weighted sum is too small; q^* attains equality.

Corollary 2 (Local allocation and information-rate decomposition)**.** *Fix interior r, π, let $v_k = r_k(1 - r_k)$, and take $\alpha = R + \Delta$ with $\Delta \downarrow 0$. Then*

$$C^\star(r) = \frac{\Delta^2}{2(\sum_k \pi_k \sqrt{v_k})^2}\{1 + o(1)\}, \qquad w_k^*(r) = \frac{\pi_k \sqrt{v_k}}{\sum_j \pi_j \sqrt{v_j}} + o(1). \tag{23}$$

For any interior safe instance, with $C_\pi(r) = \inf_{q \in \mathcal{H}_0} \sum_k \pi_k \,\mathrm{kl}(r_k, q_k)$,

$$C^\star(r) \ge C_\pi(r) \ge \mathrm{kl}(R, \alpha). \tag{24}$$

If all $r_k = r_0 < \alpha$, then $q_k^* = \alpha$, $w^* = \pi$, and all three rates equal $\mathrm{kl}(r_0, \alpha)$.

Proof. As $\Delta \downarrow 0$, the root tends to zero. At the fixed interior point, $\mathrm{kl}(r_k, r_k + h) = h^2/(2v_k) + O(h^3)$, hence $q_k^* - r_k = \sqrt{2v_k z^*} + O(z^*)$. The aggregate constraint gives $\Delta = \sqrt{2z^*}\sum_k \pi_k \sqrt{v_k} + O(z^*)$. Solving yields the first expansion; substitution into (22) yields the second. These expansions hold locally at fixed interior risks.

The choice $w = \pi$ proves the first inequality in (24). Joint convexity (the log-sum inequality) gives $\sum_k \pi_k \mathrm{kl}(r_k, q_k) \geq \mathrm{kl}(R, \pi^\top q)$. Since $s \mapsto \mathrm{kl}(R, s)$ is increasing for $s \geq \alpha > R$, taking the infimum proves the second. In the homogeneous case all inverse branches coincide, the constraint gives $q_k^* = \alpha$, and (22) gives $w^* = \pi$. The feasible constant alternative then forces equality of all rates. □

The gaps $C^\star - C_\pi$ and $C_\pi - \mathrm{kl}(R, \alpha)$ quantify the information-rate gains from active allocation and retaining stratum identities, respectively. Finite-budget savings also depend on mixing costs and stopping overhead. Homogeneous risks imply prevalence allocation, which is uniform only when prevalences are uniform.

Higher-order allocation and stratification gains. The local formula can be sharpened to quantify the rate loss of the square-root allocation. For a positive allocation w, write

$$C(w; r) = \min_{q \in \mathcal{H}_0} \sum_k w_k \,\mathrm{kl}(r_k, q_k), \qquad q^w(r) = \arg\min_{q \in \mathcal{H}_0} \sum_k w_k \,\mathrm{kl}(r_k, q_k).$$

At an interior safe instance, strict convexity gives a unique interior minimizer. Its aggregate constraint is active, and its multiplier $c_w > 0$ satisfies

$$\pi^\top q^w = \alpha, \qquad w_k \frac{q_k^w - r_k}{q_k^w(1 - q_k^w)} = c_w \pi_k.$$

In particular, $q_k^w > r_k$. These relations characterize the fixed-allocation projection used below.

Proposition 5 (Higher-order local rates)**.** *Fix $r \in (0,1)^K$ and positive $\pi \in \Delta_K$. Let $R = \pi^\top r$, $\alpha = R + \Delta$, and $\Delta \downarrow 0$. Define*

$$v_k = r_k(1 - r_k), \qquad s_\pi = \sum_k \pi_k \sqrt{v_k}, \qquad \bar{v}_\pi = \sum_k \pi_k v_k, \qquad w_k^{\mathrm{loc}} = \frac{\pi_k \sqrt{v_k}}{s_\pi}.$$

For each fixed positive w, put

$$A_w = \sum_k \frac{\pi_k^2 v_k}{w_k}, \qquad B_w = \sum_k \frac{(2r_k - 1)\pi_k^3 v_k}{w_k^2}.$$

Then

$$C(w; r) = \frac{\Delta^2}{2A_w} + \frac{B_w \Delta^3}{3A_w^3} + O(\Delta^4),$$

with a remainder uniform over compact subsets of the positive allocation simplex. Both the optimal and local-allocation rates satisfy

$$C^\star(r) = \frac{\Delta^2}{2s_\pi^2} + \frac{(2R - 1)\Delta^3}{3s_\pi^4} + O(\Delta^4),$$
$$C(w^{\mathrm{loc}}; r) = \frac{\Delta^2}{2s_\pi^2} + \frac{(2R - 1)\Delta^3}{3s_\pi^4} + O(\Delta^4).$$

Consequently, there are finite constants $A, \Delta_0 > 0$, depending on the fixed r, π, such that

$$0 \leq C^\star(r) - C(w^{\mathrm{loc}}; r) \leq A\Delta^4 \qquad (0 < \Delta < \Delta_0).$$

The relative rate loss is $O(\Delta^2)$.

Proof. **Step 1. Expand the fixed-allocation projection.** At each fixed interior risk, Taylor expansion gives

$$\mathrm{kl}(r_k, r_k + h) = \frac{h^2}{2v_k} + \frac{(2r_k - 1)h^3}{3v_k^2} + O(h^4).$$

Writing $h_k = q_k^w - r_k$, the multiplier equations become

$$w_k \left\{ \frac{h_k}{v_k} + \frac{(2r_k - 1)h_k^2}{v_k^2} + O(h_k^3) \right\} = c_w \pi_k.$$

Their Jacobian in (h, c_w) at $(0, 0)$, together with $\pi^\top h = \Delta$, has blocks $\mathrm{diag}(w_k/v_k)$, $-\pi$, $\pi^\top$, and 0. Its Schur complement is $A_w > 0$. The implicit-function theorem therefore gives smooth expansions in Δ. The same argument gives bounded derivatives when w ranges over a compact subset of the positive simplex. Inverting each coordinate yields

$$h_k = \frac{\pi_k v_k}{w_k} c_w - \frac{(2r_k - 1)\pi_k^2 v_k}{w_k^2} c_w^2 + O(c_w^3).$$

The constraint gives $\Delta = A_w c_w - B_w c_w^2 + O(c_w^3)$, and hence

$$c_w = \frac{\Delta}{A_w} + \frac{B_w \Delta^2}{A_w^3} + O(\Delta^3).$$

The derivative of the minimized objective with respect to Δ equals c_w. Integrating from $\Delta = 0$, where the value is zero, proves the expansion of $C(w; r)$.

Step 2. Invert the equal-KL branches. The same Taylor expansion, now with $\mathrm{kl}(r_k, r_k + h_k) = z$, gives

$$h_k = \sqrt{2v_k z} + \frac{2}{3}(1 - 2r_k)z + O(z^{3/2}).$$

For the optimal alternative, summing the aggregate constraint gives

$$\Delta = s_\pi \sqrt{2z^*} + \frac{2}{3}(1 - 2R)z^* + O((z^*)^{3/2}).$$

Set $t = \sqrt{2z^*}$. Series inversion yields

$$t = \frac{\Delta}{s_\pi} - \frac{(1 - 2R)\Delta^2}{3s_\pi^3} + O(\Delta^3).$$

Squaring and using $C^\star = z^*$ gives the asserted expansion.

Step 3. Evaluate the local allocation. Substitution of w^{loc} gives $A_{w^{\mathrm{loc}}} = s_\pi^2$ and $B_{w^{\mathrm{loc}}} = s_\pi^2(2R - 1)$. Its rate therefore has the same quadratic and cubic terms as $C^\star$. Optimality makes the difference nonnegative, and the two bounded fourth-order remainders give the stated bound. Since $C^\star/\Delta^2 \to 1/(2s_\pi^2) > 0$, division gives the relative rate loss. □

Corollary 3 (Heterogeneity and the stratification equality condition)**.** *Under the fixed-instance limit of Proposition 5,*

$$\frac{C^\star(r)}{C_\pi(r)} \longrightarrow 1 + \frac{\mathrm{Var}_\pi(\sqrt{v_k})}{s_\pi^2}, \qquad \frac{C_\pi(r)}{\mathrm{kl}(R, R + \Delta)} \longrightarrow 1 + \frac{\mathrm{Var}_\pi(r_k)}{\bar{v}_\pi}.$$

For every interior safe instance with positive π,

$$C_\pi(r) = \mathrm{kl}(R, \alpha) \quad \Longleftrightarrow \quad r_1 = \cdots = r_K.$$

Proof. **Step 1. Separate the two local gains.** Taking $w = \pi$ in the fixed-allocation expansion gives $C_\pi = \Delta^2/(2\bar{v}_\pi) + O(\Delta^3)$. The scalar expansion gives $\mathrm{kl}(R, R + \Delta) = \Delta^2/(2R(1 - R)) + O(\Delta^3)$. The identities

$$\bar{v}_\pi = s_\pi^2 + \mathrm{Var}_\pi(\sqrt{v_k}), \qquad R(1 - R) = \bar{v}_\pi + \mathrm{Var}_\pi(r_k)$$

give both limits. The first gain depends on variation in Bernoulli standard deviations, while the second depends on variation in stratum risks.

Step 2. Identify equality under prevalence sampling. For any interior q with $\pi^\top q = \alpha$, define probability vectors

$$a_k^1 = \frac{\pi_k r_k}{R}, \quad b_k^1 = \frac{\pi_k q_k}{\alpha}, \qquad a_k^0 = \frac{\pi_k (1 - r_k)}{1 - R}, \quad b_k^0 = \frac{\pi_k (1 - q_k)}{1 - \alpha}.$$

The KL chain rule for the joint stratum and loss distribution gives

$$\sum_k \pi_k \,\mathrm{kl}(r_k, q_k) = \mathrm{kl}(R, \alpha) + R\,\mathrm{KL}(a^1 \| b^1) + (1-R)\,\mathrm{KL}(a^0 \| b^0).$$

Apply this identity at $q = q^\pi(r)$. Equality holds exactly when both conditional KL terms vanish. Coordinatewise, these conditions give $q_k = \alpha r_k / R$ and $1 - q_k = (1-\alpha)(1-r_k)/(1-R)$. As $\alpha > R$, solving the two equations gives $r_k = R$ for every k. Conversely, for homogeneous risks the feasible constant alternative $q_k = \alpha$ attains $\mathrm{kl}(R, \alpha)$. □

Sensitivity to allocation error. The next expansion measures the local information-rate cost of perturbing the optimal allocation at a fixed safety margin.

Lemma 3 (Curvature at the optimal allocation)**.** *Fix an interior safe instance. Let*

$$d_k = \frac{q_k^* - r_k}{q_k^*(1-q_k^*)}, \qquad h_k = \frac{r_k}{(q_k^*)^2} + \frac{1-r_k}{(1-q_k^*)^2}, \qquad H = \mathrm{diag}(w_k^* h_k).$$

For $\mathbf{1}^\top u = 0$ and sufficiently small u, set $a = \mathrm{diag}(d_k)u$. Then

$$C^\star(r) - C(w^* + u; r) = \frac{1}{2}\left\{ a^\top H^{-1} a - \frac{(\pi^\top H^{-1} a)^2}{\pi^\top H^{-1} \pi} \right\} + O(\|u\|^3).$$

The displayed quadratic form is strictly positive for each nonzero simplex-tangent perturbation u.

Proof. Differentiate the KKT equations along $w(\epsilon) = w^* + \epsilon u$ at $\epsilon = 0$. With dots denoting derivatives of the projection and its multiplier,

$$H\dot{q} + a = \dot{c}\,\pi, \qquad \pi^\top \dot{q} = 0, \qquad \dot{c} = \frac{\pi^\top H^{-1} a}{\pi^\top H^{-1} \pi}.$$

Positive definiteness of H makes the constrained KKT Jacobian invertible, giving smooth dependence on positive allocations and a third-order Taylor remainder.

The value derivative $\sum_k u_k \,\mathrm{kl}(r_k, q_k^w)$ vanishes at w^* because all coordinate divergences equal $C^\star$ and $\mathbf{1}^\top u = 0$. Its second derivative is

$$a^\top \dot{q} = -a^\top H^{-1} a + \frac{(\pi^\top H^{-1} a)^2}{\pi^\top H^{-1} \pi}.$$

Taylor expansion proves the formula. Weighted Cauchy–Schwarz gives nonnegativity of the quadratic form, with equality exactly when a is proportional to π. Since $w_k^* d_k = c\pi_k$ for a positive c, equality would make u proportional to w^*. The condition $\mathbf{1}^\top u = 0$ then gives $u = 0$. □

D.2 GAVEL: DESIGN, ALGORITHM, AND PROOF OF THEOREM 4

GAVEL combines randomized Chernoff allocation with predictable betting evidence for the aggregate-risk target in Assumption 2. The construction uses active risk-control and prediction-assisted betting ideas (Xu et al., 2024; Hsu & Shekhar, 2026).

Evidence and allocation. Before query t, choose $w_t \in \Delta_K$ with positive coordinates, $m_t \in [0,1]^K$, and $\lambda_t \geq 0$ from the previously disclosed observations. Independently draw $J_t \sim w_t$, then reveal the next fresh Bernoulli loss L_t in that stratum. Write

$$A_t = \alpha - \pi^\top m_t, \qquad z_{t,k}(\ell) = A_t + \frac{\pi_k}{w_{t,k}}(m_{t,k} - \ell), \tag{25}$$

$$F_{t,k}(\ell) = 1 + \lambda_t z_{t,k}(\ell), \qquad \mathcal{E}_t^{\mathrm{bet}} = \prod_{s=1}^{t} F_{s,J_s}(L_s). \tag{26}$$

Require $F_{t,k}(\ell) \geq 0$ for every k and $\ell \in \{0,1\}$. Since $z_{t,k}(1) \leq z_{t,k}(0)$, it suffices to choose

$$0 \leq \lambda_t \leq \frac{1-\eta_t}{-\min_k z_{t,k}(1)}, \qquad 0 < \eta_t < 1. \tag{27}$$

The denominator is positive because $\sum_k w_{t,k} z_{t,k}(1) = \alpha - 1 < 0$. The correction A_t is retained even when m_t approximates a vector on the null boundary.

Proposition 6 (Validity of randomized betting evidence)**.** *Under Assumption 2, for every fixed risk vector q_0 with $\pi^\top q_0 \geq \alpha$, the process in* (26) *is a nonnegative supermartingale starting at one. In particular,*

$$\Pr_{q_0}\left(\sup_{t\geq 1} \mathcal{E}_t^{\mathrm{bet}} \geq 1/\delta\right) \leq \delta.$$

Proof. Condition on all observations and randomization preceding query t. The design is then fixed, so

$$\begin{aligned}\mathbb{E}_{q_0}[F_{t,J_t}(L_t) \mid \mathcal{F}_{t-1}] &= 1 + \lambda_t \left\{\alpha - \pi^\top m_t + \sum_k \pi_k (m_{t,k} - q_{0,k})\right\} \\ &= 1 + \lambda_t(\alpha - \pi^\top q_0) \leq 1.\end{aligned}$$

Nonnegativity gives the supermartingale property, and Ville's inequality gives the bound. Both conclusions hold conditionally on any frozen fit data. They depend on the actual randomized allocation probabilities, rather than on a deterministic target schedule. □

Predictable design used in the experiments. At each refresh we form

$$\widetilde{r}_{k,t-1} = \frac{S_k(t-1) + \kappa r_k^0}{N_k(t-1) + \kappa}, \qquad \kappa > 0.$$

The prior r^0 is fixed before auditing. We use $r_k^0 = 3\alpha/4$ and $\kappa = 64$. For a safe estimate, Proposition 2 supplies the exact target $w^*(\widetilde{r})$. Let

$$w_k^{\mathrm{loc}}(\widetilde{r}) = \frac{\pi_k \sqrt{\widetilde{r}_k(1-\widetilde{r}_k)}}{\sum_j \pi_j \sqrt{\widetilde{r}_j(1-\widetilde{r}_j)}}, \qquad a_t = \frac{1000}{t+1000}.$$

The allocation uses a vanishing local component and prevalence exploration:

$$\begin{aligned} v_t &= (1-a_t) w^*(\widetilde{r}) + a_t w^{\mathrm{loc}}(\widetilde{r}), \\ w_t^{\mathrm{raw}} &= (1-\varepsilon_t) v_t + \varepsilon_t \pi, \qquad \varepsilon_t = \min\{1, 2/\sqrt{t+16}\}. \end{aligned}$$

The implementation rounds these probabilities to positive integer multiples of 2^{-40} summing exactly to one. Both sampling and evidence use the resulting w_t. The reference is its null projection,

$$m_t = q_{w_t}(\widetilde{r}) \in \operatorname*{arg\,min}_{q:\, \pi^\top q \geq \alpha} \sum_k w_{t,k}\, \mathrm{kl}(\widetilde{r}_k, q_k).$$

This design is predictable, and $a_t \to 0$ retains the limiting Chernoff target. Given w_t and m_t, a scalar concave optimization chooses

$$\lambda_t^{\mathrm{opt}} \in \operatorname*{arg\,max}_{0 \leq \lambda \leq \overline{\lambda}_t} \sum_k w_{t,k} \left\{(1-\widetilde{r}_k)\log(1+\lambda z_{t,k}(0)) + \widetilde{r}_k \log(1+\lambda z_{t,k}(1))\right\},$$

where $\overline{\lambda}_t$ is the bound in (27) with $\eta_t = 1/(t+10)$. We use $\lambda_t = \rho_t \lambda_t^{\mathrm{opt}}$, where

$$\rho_t = 1 - (1-\rho_0)\min\{1, \sqrt{100/t}\}, \qquad \rho_0 = 0.75.$$

An unsafe estimate or failed solver gives prevalence allocation and $\lambda_t = 0$ for the adaptive auditor. With a fixed allocation, the sampling target is retained when the stake is zero. Designs are refreshed every 25 queries and remain fixed between refreshes. Certification is checked at multiples of 25. Allocation controls replace v_t by their declared target and use its corresponding null projection. Appendix E.8 gives the parameter comparisons.

Optimal evidence at a fixed allocation. The likelihood-ratio form extends to every positive allocation. This gives an explicit instance of log-optimal e-variable construction for the aggregate-risk null (Grünwald et al., 2024; Larsson et al., 2025). Fix interior safe r and $w \in \operatorname{int}\Delta_K$, and use $C(w;r)$ and its minimizing alternative $q^w(r)$ from Appendix D.1. Write P_q^w for the distribution of (J, L) with $J \sim w$ and $L \mid J = k \sim \mathrm{Bernoulli}(q_k)$.

Algorithm 1 GAVEL: fit, freeze, and audit with predictable betting

Require: Fit/unlabeled data, fresh audit access, (α, δ), budget B, check times Γ.
Fit and freeze $(h, E, \mathcal{T}, \pi)$ and the design rules in Assumption 2; set $\mathcal{E}_0 = 1$.
for $t = 1, \dots, B$ **do**
  if $t = 1$ or t is a scheduled refresh **then**
    Estimate $\widetilde{r}$ from past losses and solve for (q^*, w^*) when safe.
    if the estimate is safe and the solver succeeds **then**
      Form v_t, add prevalence exploration, and compute $m_t = q_{w_t}(\widetilde{r})$.
      Optimize the constrained stake λ_t.
    else
      Use $w_t = \pi$, $m_t = \alpha \mathbf{1}$, $\lambda_t = 0$.
    end if
  else
    Retain the last design (w_t, m_t, λ_t).
  end if
  Draw $J_t \sim w_t$ using fresh randomness.
  Query the next loss L_t in stratum J_t; if the pool is exhausted, return no certificate.
  Update N, S and $\mathcal{E}_t = \mathcal{E}_{t-1}[1 + \lambda_t z_{t,J_t}(L_t)]$.
  if $t \in \Gamma$ and the conservative full-path check certifies $\mathcal{E}_t \geq 1/\delta$ **then**
    return CERTIFIED, with the frozen policy.
  end if
end for
return no certificate at budget exhaustion.

Proposition 7 (Log-optimal factor for a positive allocation)**.** *The minimizer q^w is unique and interior. It satisfies $q^w_k > r_k$, $\pi^\top q^w = \alpha$, and*

$$w_k \frac{q^w_k - r_k}{q^w_k(1 - q^w_k)} = c_w \pi_k \quad \textit{for a unique } c_w = c_w(r) > 0.$$

Among factors $F : \{1, \dots, K\} \times \{0, 1\} \to [0, \infty)$ satisfying $\sup_{q \in \mathcal{H}_0} \mathbb{E}_{P^w_q} F \leq 1$, the unique maximizer of $\mathbb{E}_{P^w_r} \log F$ is

$$F^\circ_w(k, \ell) = 1 + c_w \frac{\pi_k}{w_k}(q^w_k - \ell) = \frac{r^\ell_k (1 - r_k)^{1-\ell}}{(q^w_k)^\ell (1 - q^w_k)^{1-\ell}}.$$

Its expected log factor is $C(w; r)$.

Proof. **Step 1. Identify the null projection.** The weighted divergence is strictly convex in q and diverges at every cube boundary because w is positive and r is interior. A finite feasible value exists, so its unique minimizer is interior. The aggregate constraint is active because the unconstrained minimizer r is safe. Slater's condition and the KKT equations give

$$w_k \frac{q^w_k - r_k}{q^w_k(1 - q^w_k)} = c_w \pi_k, \qquad c_w \geq 0.$$

A zero multiplier would imply $q^w = r$, which is infeasible. Thus $c_w > 0$, each $q^w_k > r_k$, and the displayed equations determine the multiplier uniquely.

Step 2. Verify the factor over the composite null. Substitution at $\ell = 0, 1$ yields the two likelihood ratios in the statement. They are strictly positive. For every $q \in \mathcal{H}_0$,

$$\mathbb{E}_{P^w_q} F^\circ_w = 1 + c_w \sum_k \pi_k (q^w_k - q_k) = 1 + c_w(\alpha - \pi^\top q) \leq 1.$$

Taking logarithms under P^w_r gives $\mathbb{E}_{P^w_r} \log F^\circ_w = C(w; r)$.

Step 3. Establish optimality and uniqueness. Any factor vanishing at an outcome has expected log value $-\infty$ because P^w_r has full support. For a strictly positive feasible factor, Jensen's inequality and change of measure give

$$\mathbb{E}_{P^w_r} \log \frac{F}{F^\circ_w} \leq \log \mathbb{E}_{P^w_r} \frac{F}{F^\circ_w} = \log \mathbb{E}_{P^w_{q^w}} F \leq 0.$$

Equality requires F/F_w° to be constant on the full support and $\mathbb{E}_{P_{q^w}^w} F = 1$. The constant is therefore one, proving uniqueness. □

Lemma 4 (Rate under a convergent positive allocation)**.** *Under Assumption 2, fix interior safe r and a deterministic $\bar{w} \in \operatorname{int} \Delta_K$. Suppose the exact-arithmetic factors in* (26) *are strictly positive and, almost surely,*

$$w_t \to \bar{w}, \qquad m_t \to q^{\bar{w}}(r), \qquad \lambda_t \to c_{\bar{w}}(r).$$

Then $t^{-1} \log \mathcal{E}_t^{\mathrm{bet}} \to C(\bar{w}; r)$ almost surely.

Proof. **Step 1. Obtain a bounded martingale difference.** Let $f_{t,k}(\ell) = \log F_{t,k}(\ell)$, $Y_t = f_{t,J_t}(L_t)$, and

$$\mu_t = \sum_k w_{t,k}\{(1 - r_k) f_{t,k}(0) + r_k f_{t,k}(1)\}.$$

The converging design makes each of the finitely many $f_{t,k}(\ell)$ converge to $\log F_{\bar{w}}^\circ(k, \ell)$. Choose a deterministic H exceeding the largest absolute limiting log factor, and put

$$B_t = \mathbf{1}\{\max_{k,\ell} |f_{t,k}(\ell)| \le H\}, \qquad D_t = B_t(Y_t - \mu_t).$$

The indicator B_t is measurable before query t. Conditional on that history, the joint law of (J_t, L_t) is $P_r^{w_t}$. Consequently $\mathbb{E}[D_t \mid \mathcal{F}_{t-1}] = 0$ and $|D_t| \le 2H$. The martingale with increments D_t/t converges almost surely because $\sum_t \mathbb{E}[D_t^2]/t^2 < \infty$. Kronecker's lemma yields $t^{-1} \sum_{s \le t} D_s \to 0$.

Step 2. Remove the truncation and identify the rate. Convergence of all log factors gives $B_t = 1$ eventually, almost surely. Thus $\sum_{s \le t}(Y_s - \mu_s)$ differs from $\sum_{s \le t} D_s$ by a finite pathwise sum for all sufficiently large t. Its average also tends to zero. Continuity gives $\mu_t \to \mathbb{E}_{P_r^{\bar{w}}} \log F_{\bar{w}}^\circ = C(\bar{w}; r)$, and Cesaro averaging finishes the proof. □

Allocation, evidence, and realized growth. The fixed-allocation optimum also separates two sources of evidence loss. For each predictable positive w_t, form the analytical reference $q_t^\circ = q^{w_t}(r)$ and the conditional distributions

$$P_t = P_r^{w_t}, \qquad Q_t = P_{q_t^\circ}^{w_t}, \qquad \widetilde{P}_t(k, \ell) = Q_t(k, \ell) F_{t,k}(\ell).$$

Here F_t is the actual strictly positive factor in (26). Let Y_t and μ_t denote its log and the conditional mean defined in the preceding proof. Since $\pi^\top q_t^\circ = \alpha$, its expectation under Q_t is exactly one, so $\widetilde{P}_t$ is a probability distribution. Expanding the relative entropy gives the conditional identity

$$\mu_t = C(w_t; r) - \mathcal{K}_t, \qquad \mathcal{K}_t = \mathrm{KL}(P_t \| \widetilde{P}_t) \ge 0.$$

With $\mathcal{M}_t = \sum_{s \le t}(Y_s - \mu_s)$, summation yields

$$t C^\star(r) - \log \mathcal{E}_t^{\mathrm{bet}} = \sum_{s \le t} \{C^\star(r) - C(w_s; r)\} + \sum_{s \le t} \mathcal{K}_s - \mathcal{M}_t.$$

The first term measures allocation loss and the second measures the gap between the chosen factor and the best factor for that allocation. The last term is the centered sampling fluctuation. It is a martingale when the increments are integrable, with bounded localization supplied by the preceding proof under its convergence assumptions. The first two terms are nonnegative, while $\mathcal{M}_t$ can have either sign. Each allocation generates its own sampling history and hence its own evidence gap $\mathcal{K}_t$.

Proposition 8 (Ideal Chernoff rate)**.** *Fix an interior safe r. Consider the exact-arithmetic version with strictly positive factors. Suppose its estimates converge to r, and its predictable designs satisfy*

$$w_t \longrightarrow w^*(r), \qquad m_t \longrightarrow q^*(r), \qquad \lambda_t \longrightarrow c(r) := \left\{ \sum_k \frac{\pi_k q_k^*(1 - q_k^*)}{q_k^* - r_k} \right\}^{-1}. \tag{28}$$

Then $t^{-1} \log \mathcal{E}_t^{\mathrm{bet}} \to C^\star(r)$ almost surely.

Proof. The Chernoff solution has $\pi^\top q^* = \alpha$ and

$$\frac{c(r)\pi_k}{w_k^*} = \frac{q_k^* - r_k}{q_k^*(1-q_k^*)}.$$

Consequently the limiting factor is the Bernoulli likelihood ratio:

$$1 + \frac{c(r)\pi_k}{w_k^*}(q_k^* - \ell) = \frac{r_k^\ell (1-r_k)^{1-\ell}}{(q_k^*)^\ell (1-q_k^*)^{1-\ell}}.$$

Interiority makes these factors strictly positive and finite. This is the factor in Proposition 7 at $w = w^*(r)$. Lemma 4 therefore gives

$$\frac{1}{t}\log \mathcal{E}_t^{\mathrm{bet}} \longrightarrow \sum_k w_k^* \,\mathrm{kl}(r_k, q_k^*) = C^\star(r).$$

□

Since $a_t \to 0$, the local component vanishes from the limiting target. For the exact-arithmetic design, vanishing exploration with $\sum_t \varepsilon_t = \infty$ makes every stratum count diverge and hence makes the smoothed estimates consistent. At the limiting Chernoff design, $c(r)$ is the unique maximizer of the expected log factor: its derivative vanishes at the displayed likelihood ratio, and strict concavity gives uniqueness. It lies strictly below the nonnegativity cap because r is interior. Thus the growing cap and $\rho_t \to 1$ yield the required stake limit when optimization errors vanish. Bounded refresh gaps preserve these limits. Exact-rate claims require vanishing clipping, allocation-quantization, and solver errors, or eventually inactive clipping with vanishing remaining errors. The fixed 40-bit implementation is covered by the finite-sample validity result; its measured finite-budget costs are reported separately from this ideal asymptotic statement.

Conservative checked crossings. Floating-point log evidence only screens candidate crossings. At a candidate time we reconstruct the disclosed path, treating the stored binary floating-point design constants as exact rationals, the effective sampling probabilities as dyadic rationals, and the frozen prevalences as their exact count ratios. Interval arithmetic checks all $2K$ possible factors at every historical refresh and computes a lower bound on the accumulated log evidence. Certification requires every design-factor lower bound to be strictly positive and the accumulated lower bound to exceed an upper bound on $\log(1/\delta)$.

This deferred verification has a predictable killed-process interpretation. Let D_t indicate that the interval check on the design chosen before query t proves every possible factor positive; it is measurable before the next stratum and loss are drawn. Define

$$\widehat{F}_t = \begin{cases} F_{t,J_t}(L_t), & D_t = 1, \\ 0, & D_t = 0, \end{cases} \qquad \widehat{\mathcal{E}}_t = \prod_{s=1}^{t} \widehat{F}_s.$$

The conditional expectation of $\widehat{F}_t$ is at most one under the null, so $\widehat{\mathcal{E}}$ is a nonnegative supermartingale. A failed historical design kills this process permanently. Reconstructing its checks only when a crossing is proposed does not change their dependence on past data. Every accepted certificate is a crossing of $\widehat{\mathcal{E}}$; interval rejection or an unresolved numerical comparison gives no certificate. Checking both possible losses is necessary for this argument, even when only one was observed.

D.3 Checked stopping and conservative evidence

On the same infinite process, let $\tau_\delta = \inf\{t : \mathcal{E}_t \geq 1/\delta\}$ and $\tau_\delta^\Gamma = \inf\{t \in \Gamma : \mathcal{E}_t \geq 1/\delta\}$ for a fixed unbounded check schedule Γ, with $\inf \varnothing = \infty$. At budget B, the reported draw cost is $\min(\tau_\delta^\Gamma, B)$ and certification occurs on $\{\tau_\delta^\Gamma \leq B\}$. A predeclared pool exhaustion ends the audit without certification.

Lemma 5 (Checked and conservative evidence). *If $t^{-1}\log \mathcal{E}_t \to C > 0$ almost surely and successive check times have bounded gaps, then $\tau_\delta^\Gamma / \log(1/\delta) \to 1/C$ almost surely. If a numerically certified process satisfies $0 \leq \underline{e}_t \leq \mathcal{E}_t$, certification at its checked crossings remains anytime-valid. If additionally $0 \leq \log \mathcal{E}_t - \log \underline{e}_t = o(t)$ almost surely, its first-order stopping rate is also $1/C$.*

Proof. For every $0 < \epsilon < C$, eventually $(C-\epsilon)t \le \log \mathcal{E}_t \le (C+\epsilon)t$ almost surely. The finite prefix has a finite maximum. For sufficiently large $b = \log(1/\delta)$, no crossing occurs before $b/(C+\epsilon)$, and the first check after $b/(C-\epsilon)$ crosses. Bounded check gaps add a bounded delay to the latter envelope. Divide by b and let $\epsilon \downarrow 0$. Pointwise domination makes the numerical crossing event a subset of the exact crossing event. A sublinear log gap preserves the limiting slope, so the same argument applies to the conservative process. □

The interval reconstruction in Appendix D.2 also checks every possible design factor along the disclosed history. Finite-sample validity covers the implemented finite-precision design. The exact-rate assertion uses the convergence conditions in Theorem 4.

D.4 A LIKELIHOOD LEMMA UNDER ADAPTIVE SAMPLING

For stratum choices made before observing the next loss, let $N_k(t)$ and $S_k(t)$ be the audit and error counts and define

$$\mathcal{L}_t(p) = \prod_{k=1}^{K} p_k^{S_k(t)} (1-p_k)^{N_k(t)-S_k(t)}.$$

Under a parameter $q \in [0,1]^K$ and Assumption 2, for every fixed p the process

$$M_t(p,q) := \frac{\mathcal{L}_t(p)}{\mathcal{L}_t(q)}$$

is a nonnegative martingale under q when $q \in (0,1)^K$, and a nonnegative supermartingale at boundary coordinates under the standard likelihood-ratio convention. To see this, conditional on the past and $J_t = k$, the next multiplier is

$$\left(\frac{p_k}{q_k}\right)^{L_t} \left(\frac{1-p_k}{1-q_k}\right)^{1-L_t},$$

whose conditional expectation under an interior q is $p_k + (1-p_k) = 1$. If $q_k = 0$ or 1, the impossible outcome carries no null mass and the remaining multiplier has expectation at most one. Integrating p against a fixed probability measure η preserves the corresponding martingale or supermartingale property by Tonelli's theorem.

D.5 PROOF OF THEOREM 5

Let τ be the procedure's termination time, allowing certification or non-certification, and let D be its certification event. In the power-one case this agrees with the first certification time almost surely under r. If $\mathbb{E}_r\tau = \infty$, the lower bound follows immediately. Suppose $\mathbb{E}_r\tau < \infty$ and fix $q \in \mathcal{H}_0 \cap (0,1)^K$.

Step 1. Factor the stopped-path likelihood. Include the procedure's independent randomization in its history. For each fixed observed history, the conditional sampling and stopping kernels are identical under r and q. These kernels therefore cancel in the path likelihood ratio. The remaining log-likelihood increment is

$$\xi_t = L_t \log \frac{r_{J_t}}{q_{J_t}} + (1-L_t) \log \frac{1-r_{J_t}}{1-q_{J_t}}.$$

Define the stopped-path space as the disjoint union of finite terminated histories and infinite histories on which the procedure continues. On its component $\{\tau = n\}$, the density of P_r^τ with respect to P_q^τ is $\exp(\sum_{t=1}^n \xi_t)$. Interior r, q give common support to the Bernoulli observations on every finite history. Since $P_r(\tau < \infty) = 1$, setting this density to zero on the infinite-history component completes the Radon–Nikodym derivative. This construction allows $P_q(\tau = \infty) > 0$.

Step 2. Integrate the stopped increments. For fixed interior r, q, there is a finite constant $M_{r,q}$ with $|\xi_t| \le M_{r,q}$, so

$$\mathbb{E}_r \sum_{t \ge 1} \mathbf{1}\{\tau \ge t\} |\xi_t| \le M_{r,q} \mathbb{E}_r \tau < \infty.$$

Write $w_{t,k}$ for the conditional probability of sampling stratum k before observing L_t. The event $\{\tau \geq t\}$ is determined by the history before that draw. Conditional expectation and absolute integrability now give

$$\begin{aligned}\mathrm{KL}(P_r^\tau \| P_q^\tau) &= \mathbb{E}_r \sum_{t=1}^{\tau} \xi_t \\ &= \sum_{t\geq 1} \mathbb{E}_r \left[\mathbf{1}\{\tau \geq t\} \sum_{k=1}^{K} w_{t,k} \,\mathrm{kl}(r_k, q_k) \right] \\ &= \sum_{k=1}^{K} \mathbb{E}_r[N_k(\tau)] \,\mathrm{kl}(r_k, q_k).\end{aligned}$$

Step 3. Include boundary alternatives. For the weighted objective below, adopt $0 \cdot \infty = 0$. The set $\mathcal{H}_0 \cap (0,1)^K$ is dense in $\mathcal{H}_0$: mixing any boundary null with $(\alpha, \ldots, \alpha)$ produces interior nulls. At finite-objective boundary points their weighted divergences converge; infinite-objective boundary points cannot lower the infimum. Hence the infimum over interior nulls equals that over all of $\mathcal{H}_0$.

Step 4. Apply the certification decision. Data processing through the indicator of D gives

$$\mathrm{KL}(P_r^\tau \| P_q^\tau) \geq \mathrm{kl}\{\mathbb{P}_r(D), \mathbb{P}_q(D)\}.$$

For a power-one procedure, this is at least $\mathrm{kl}(1, \delta) = \log(1/\delta)$. Put $w_k = \mathbb{E}_r N_k(\tau)/\mathbb{E}_r \tau$. Then $w \in \Delta_K$ and, for every $q \in \mathcal{H}_0$,

$$\mathbb{E}_r \tau \sum_k w_k \,\mathrm{kl}(r_k, q_k) \geq \log(1/\delta).$$

Taking the infimum over q and using $\inf_q \sum_k w_k \,\mathrm{kl}(r_k, q_k) \leq C^\star(r)$ proves (13). If $\mathbb{P}_r(D) \geq 1 - \gamma$ and $\mathbb{P}_q(D) \leq \delta < 1 - \gamma$, binary relative entropy is minimized at these boundary probabilities, giving the stated numerator $\mathrm{kl}(1 - \gamma, \delta)$. □

Corollary 4 (Certification power and capped audit cost)**.** *Under the interior safe-risk setting of Theorem 5, fix an integer budget $B \geq 1$ and consider a δ-valid audit with fresh draws available through B. Run the audit until certification or budget exhaustion. Let τ be its first certification time, with $\tau = \infty$ if it remains uncertified at B, and define*

$$\tau_B = \min\{\tau, B\}, \qquad p_B = \mathbb{P}_r(\tau \leq B).$$

If $p_B > \delta$, then

$$\mathbb{E}_r \tau_B \geq \frac{\mathrm{kl}(p_B, \delta)}{C^\star(r)}, \qquad BC^\star(r) \geq \mathrm{kl}(p_B, \delta).$$

Proof. Apply the stopped-experiment identity to τ_B and the event $D_B = \{\tau \leq B\}$. For every interior $q \in \mathcal{H}_0$, validity gives $\mathbb{P}_q(D_B) \leq \delta$. Since $p_B > \delta$, binary relative entropy is decreasing in its second argument on $[0, \delta]$, and data processing gives

$$\sum_{k=1}^{K} \mathbb{E}_r[N_k(\tau_B)] \,\mathrm{kl}(r_k, q_k) \geq \mathrm{kl}(p_B, \delta).$$

The sampling rule gives $\sum_k N_k(\tau_B) = \tau_B$. Set $w_k = \mathbb{E}_r[N_k(\tau_B)]/\mathbb{E}_r \tau_B$. Minimizing over interior nulls gives the same infimum as over $\mathcal{H}_0$, by the preceding boundary approximation, and bounds the left side by $\mathbb{E}_r \tau_B \, C^\star(r)$. The first inequality follows, and $\tau_B \leq B$ gives the second. Runs without certification contribute their full budget B to this cost. □

D.6 Pooled Bernoulli reduction and beta-mixture validity

If $J_t \overset{\text{i.i.d.}}{\sim} \pi$, the pooled losses are i.i.d. $\mathrm{Bernoulli}(R)$ and

$$\mathcal{E}_t^{\mathrm{pool}}(\alpha) = \frac{\alpha^{-1} \int_0^\alpha q^{S_t} (1-q)^{t - S_t} \, dq}{\alpha^{S_t} (1-\alpha)^{t-S_t}} \tag{29}$$

is a nonnegative supermartingale under every $R \geq \alpha$. Hence crossing $1/\delta$ is anytime-valid. The dynamic recursion in Appendix D.7 gives the exact crossing probability at any finite budget. For fixed $0 < R < \alpha$, Appendix D.8 also proves $\mathbb{E}_R \tau_\delta / \log(1/\delta) \to 1/\operatorname{kl}(R, \alpha)$. This expected-time conclusion is specific to the scalar audit.

Suppose the schedule is i.i.d. with $\mathbb{P}(J_t = k \mid \mathcal{F}_0) = \pi_k$. Conditional on $\mathcal{F}_0$,

$$\mathbb{P}(L_t = 1 \mid \mathcal{F}_0) = \sum_k \pi_k r_k = R.$$

The mutual independence of the stratum arrays, schedule draws, and tie coins makes the losses mutually independent, so $L_t \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(R)$.

For a fixed $q \in [0, \alpha)$, define

$$M_t(q) = \prod_{s=1}^{t} \left(\frac{q}{\alpha}\right)^{L_s} \left(\frac{1-q}{1-\alpha}\right)^{1-L_s},$$

with the continuous convention at $q = 0$. Under any $R \geq \alpha$, the next multiplier has expectation

$$R\frac{q}{\alpha} + (1-R)\frac{1-q}{1-\alpha} \leq 1.$$

Thus $M_t(q)$ is a nonnegative supermartingale. Its uniform mixture over $[0, \alpha]$ is exactly (29); Tonelli's theorem preserves the supermartingale property. Ville's inequality proves

$$\sup_{R \geq \alpha} \mathbb{P}_R\{\exists t : \mathcal{E}_t^{\text{pool}} \geq 1/\delta\} \leq \delta.$$

If R is the $\mathcal{F}_0$-conditional deployment risk, the precise unconditional statement is

$$\mathbb{P}\{\tau_\delta < \infty,\ R \geq \alpha\} = \mathbb{E}[\mathbf{1}\{R \geq \alpha\}\mathbb{P}(\tau_\delta < \infty \mid \mathcal{F}_0)] \leq \delta.$$

This avoids conflating the random conditional risk with the marginalized risk functional.

For $\widehat{R}_t = t^{-1}\sum_{s=1}^{t} L_s$,

$$\mathbb{E}\widehat{R}_t = R, \qquad \operatorname{Var}(\widehat{R}_t) = \frac{R(1-R)}{t} \leq \frac{1}{4t}.$$

If $R \in (0, 1)$, the Lindeberg–Lévy theorem gives

$$\frac{\sqrt{t}(\widehat{R}_t - R)}{\sqrt{R(1-R)}} \Rightarrow N(0, 1).$$

These moment statements describe estimation; they do not replace the time-uniform certificate.

D.7 Exact finite-budget power

For $s \in \{0, \ldots, t\}$ define the statewise evidence

$$e_t(s) = \frac{\alpha^{-1}\int_0^\alpha q^s(1-q)^{t-s}\,dq}{\alpha^s(1-\alpha)^{t-s}}.$$

Set $u_0(0) = 1$, take out-of-range values to be zero, and recurse

$$u_{t+1}(s) = \mathbf{1}\{e_{t+1}(s) < 1/\delta\}\{(1-R)u_t(s) + Ru_t(s-1)\}. \tag{30}$$

Here $u_t(s)$ is the probability that the audit has not crossed by time t and has accumulated s errors. Induction on t proves this interpretation, so

$$\mathbb{P}_R(\tau_\delta \leq B) = 1 - \sum_{s=0}^{B} u_B(s).$$

The crossing set is coordinatewise decreasing in the loss sequence, hence the power is nonincreasing in R. Evaluating (30) at $R = \alpha - \zeta$ therefore gives a uniform lower power bound over $R \leq \alpha - \zeta$.

A positive gap is necessary for uniform finite-budget power: as $R \uparrow \alpha$, the B-step law converges to the null-boundary law, whose crossing probability is at most δ.

For each fixed $R < \alpha$, the strong law and the elementary Laplace bounds

$$t\,\mathrm{kl}(\widehat{R}_t, \alpha) - O(\log t) \le \log \mathcal{E}_t^{\mathrm{pool}} \le t\,\mathrm{kl}(\widehat{R}_t, \alpha) \tag{31}$$

on a neighborhood of R imply $t^{-1} \log \mathcal{E}_t^{\mathrm{pool}} \to \mathrm{kl}(R, \alpha) > 0$ almost surely. Thus the power-one property holds and

$$\frac{\tau_\delta}{\log(1/\delta)} \to \frac{1}{\mathrm{kl}(R, \alpha)} \quad \text{almost surely.}$$

The lower bound in (31) follows by integrating over an interval of width $1/t$ around the maximizer $q = \widehat{R}_t$; the upper bound replaces the integral by its maximum. This supplies the scalar instance of Theorems 4 and 5.

D.8 Matching expected stopping time for the scalar audit

Proposition 9 (Scalar expected-time asymptotics)**.** *For fixed $0 < R < \alpha < 1$, let τ_δ be the first crossing of $1/\delta$ by* (29)*, checked every step on one infinite i.i.d. Bernoulli stream. Then $\mathbb{E}_R \tau_\delta < \infty$ for every $\delta \in (0, 1)$, and*

$$\lim_{\delta \downarrow 0} \frac{\mathbb{E}_R \tau_\delta}{\log(1/\delta)} = \frac{1}{\mathrm{kl}(R, \alpha)}. \tag{32}$$

The same limit holds for a fixed infinite check schedule with bounded gaps.

Proof. **Step 1. Obtain a uniform evidence bound.** Choose $\epsilon > 0$ with $I_\epsilon = [R - \epsilon, R + \epsilon] \subset (0, \alpha)$. For $u = \widehat{R}_t \in I_\epsilon$, integrate the numerator over $[u, u + 1/t]$. For all sufficiently large t, this interval lies in a fixed compact subset of $(0, \alpha)$. Write $\ell_t(q) = t\{u \log q + (1 - u) \log(1 - q)\}$. Since $\ell_t'(u) = 0$ and $|\ell_t''(q)| \le K_\epsilon t$ uniformly on that compact set, Taylor's theorem bounds the loss from the maximum by $K_\epsilon/(2t)$ on the interval. Its width is $1/t$, so for a constant C_ϵ independent of u, t,

$$\log \mathcal{E}_t^{\mathrm{pool}} \ge t\,\mathrm{kl}(u, \alpha) - \log t - C_\epsilon \quad (u \in I_\epsilon). \tag{33}$$

Step 2. Bound the stopping-time tail. Set $c_\epsilon = \inf_{u \in I_\epsilon} \mathrm{kl}(u, \alpha) > 0$ and $b = \log(1/\delta)$. For any fixed $\zeta > 0$, let $n_b = \lceil (1 + \zeta) b / c_\epsilon \rceil$. For sufficiently large b, $t c_\epsilon - \log t - C_\epsilon \ge b$ for every $t \ge n_b$: it holds at n_b because ζb dominates $O(\log b)$, and the expression is increasing for $t > 1/c_\epsilon$. Therefore

$$\{\tau_\delta > t\} \subseteq \{\mathcal{E}_t^{\mathrm{pool}} < e^b\} \subseteq \{|\widehat{R}_t - R| > \epsilon\}, \qquad \mathbb{P}_R(\tau_\delta > t) \le 2e^{-2t\epsilon^2}$$

for $t \ge n_b$, by Hoeffding. Summing tails gives

$$\mathbb{E}_R \tau_\delta \le n_b + \frac{2e^{-2n_b\epsilon^2}}{1 - e^{-2\epsilon^2}}.$$

For an arbitrary fixed b, the same argument starts at a sufficiently large deterministic n, proving finite expectation without assuming uniform integrability.

Step 3. Identify the expected-time constant. Divide by b, let $b \to \infty$, then $\zeta \downarrow 0$ and $\epsilon \downarrow 0$. Continuity of KL gives the upper limit $1/\mathrm{kl}(R, \alpha)$. Theorem 5 applied to this power-one, finite-mean scalar certifier gives the matching lower bound.

Step 4. Extend to scheduled checks. For checks at $\Gamma = \{\gamma_1, \gamma_2, \ldots\}$ with gaps at most g, let $s(t)$ be the last check at or before t. Eventually $s(t) \ge t - g$. If $s(t) \ge n_b$, failure to have certified by t implies $\mathcal{E}_{s(t)}^{\mathrm{pool}} < e^b$, whose probability is at most $2e^{-2(t-g)\epsilon^2}$. Summing from $n_b + g$ proves the same upper limit and finite expectation. The lower bound again applies to the checked certifier. This proof does not assume evidence is monotone after an unobserved crossing. □

Extending this expected-time result to heterogeneous adaptive auditing requires integrable control of the time needed for allocation, estimation, and numerical computation to enter their stable regimes.

Lemma 6 (Uniform scalar evidence expansion)**.** *Fix a compact interval $I \subset (0, \alpha)$. Uniformly over integer counts $s \in \{0, \dots, t\}$ with $u = s/t \in I$, the statewise evidence in Appendix D.7 satisfies*

$$\log e_t(s) = t \operatorname{kl}(u, \alpha) - \frac{1}{2} \log t + \frac{1}{2} \log\{2\pi u(1-u)\} - \log \alpha + O_I(t^{-1}).$$

In particular, a finite constant K_I gives

$$\log e_t(s) \geq t \operatorname{kl}(u, \alpha) - \frac{1}{2} \log t - K_I \qquad (u \in I).$$

Proof. Put

$$h_u(q) = u \log q + (1-u) \log(1-q), \qquad B_t(u) = \int_0^1 e^{t h_u(q)}\, dq.$$

The beta identity gives $B_t(u) = s!(t-s)!/(t+1)!$. Stirling's formula, uniformly for $u \in I$, yields

$$B_t(u) = e^{t h_u(u)} \sqrt{\frac{2\pi u(1-u)}{t}} \{1 + O_I(t^{-1})\}.$$

For $q \geq \alpha > u$, monotonicity gives

$$h_u(q) - h_u(u) \leq h_u(\alpha) - h_u(u) = -\operatorname{kl}(u, \alpha).$$

Since $d_I = \min_{u \in I} \operatorname{kl}(u, \alpha) > 0$, the preceding expansion gives, uniformly on I,

$$0 \leq \frac{\int_\alpha^1 e^{t h_u(q)}\, dq}{B_t(u)} \leq C_I \sqrt{t}\, e^{-t d_I}$$

for a finite C_I. Truncation therefore changes the logarithm by $O_I(\sqrt{t}\, e^{-t d_I}) = O_I(t^{-1})$.

Finally,

$$\log e_t(s) = \log \int_0^\alpha e^{t h_u(q)}\, dq - \log \alpha - t h_u(\alpha).$$

Substitution and $h_u(u) - h_u(\alpha) = \operatorname{kl}(u, \alpha)$ give the expansion. Boundedness of $\log\{2\pi u(1-u)\}$ on I gives the lower bound, with the finitely many smaller admissible counts absorbed into K_I. □

D.9 Fresh epochs

Let $\mathcal{G}_e$ be the complete history immediately before epoch e opens its audit. Freeze $(h^{(e)}, E_e, \pi^{(e)}, \alpha_e, \delta_e)$ as $\mathcal{G}_e$-measurable objects. Conditional on $\mathcal{G}_e$, suppose a fresh non-reused audit stream satisfies Assumption 2 for the new deployment law and $\sum_e \delta_e \leq \omega$ almost surely. If $F_e = \{\tau_e < \infty, R_e \geq \alpha_e\}$, conditional validity and a union bound give

$$\mathbb{P}\left(\bigcup_e F_e\right) \leq \mathbb{E}\left[\sum_e \mathbb{P}(F_e \mid \mathcal{G}_e)\right] \leq \mathbb{E}\left[\sum_e \delta_e\right] \leq \omega.$$

No independence across epochs is required. Matching one-dimensional marginals or exchangeability alone does not suffice. For example, if $L_t = B$ for every t with $B \sim \text{Bernoulli}(\alpha)$, the stream is exchangeable with marginal risk α, but on $B = 0$ the beta-mixture evidence diverges and false certification occurs with probability $1 - \alpha$.

E Additional Experimental Details

E.1 Experiments, data, and scope

The experiments combine controlled report channels, moment-based aggregation, public cached model reports, and sequential audits. PEG's probability policies (Chen et al., 2025) supply the semi-synthetic MMLU channel. HELM (Liang et al., 2023) and RouterBench (Hu et al., 2024) supply discrete model responses for fixed-proposal verification. Synthetic Bernoulli losses give known stratum risks for allocation comparisons and calibration.

E.2 Mechanism, orientation, solver, and candidate controls

The family-channel Monte Carlo uses 20,000 independent tasks per setting, five evaluators per family, and the original $(q_G, q_B, r_G) = (0.85, 0.70, 0.10)$. The six r_B values are $0, 0.25, r_B^\star, r_B^\star + 0.10, 0.8, 1$. The control retains the common M_h while making H_h evaluator-specific; its accuracy is unchanged and correlation is $(1 - r_h)(2q_h - 1)^2$. Accuracy/agreement errors are clustered by task. DMI and MI are computed from the across-task joint table of one predetermined evaluator pair.

Shared-flip calculations cover odd $n \leq 101$, $\epsilon \in \{0.05, 0.10, 0.20\}$, and $\beta \in \{0, 0.10, 0.20, 0.30, 0.50\}$, using binomial survival/log-survival functions. At $\beta = 0.5$ the risk is exactly 0.5 for every n. Anchor experiments cover $\kappa \in \{0.1, 0.3, 0.6, 0.8\}$ and $m \in \{1, 3, 5, 9, 17, 33, 65, 129, 257\}$, compare exact binomial tails with the Hoeffding bound, and check both orientations with 100,000 simulated repetitions per cell. The deployment risk is $(1 - \kappa)/2 + \kappa p_{\text{orient}}$, so successful orientation leaves the baseline overlap error.

The Chernoff study includes arbitrary-prevalence homogeneous laws (where $w^\star = \pi$), a heterogeneous uniform-prevalence counterexample to always allocating uniformly, and margins $0.001, 0.002, 0.005, 0.01, 0.02, 0.04$. Its local approximation and numerical residuals are archived separately from actual stopping-label results.

Synthetic aggregation uses the family channel above, with the fitting and regularization rules in Appendix E.6. Figure 5 places these results alongside the orientation, local-information, and dependence controls.

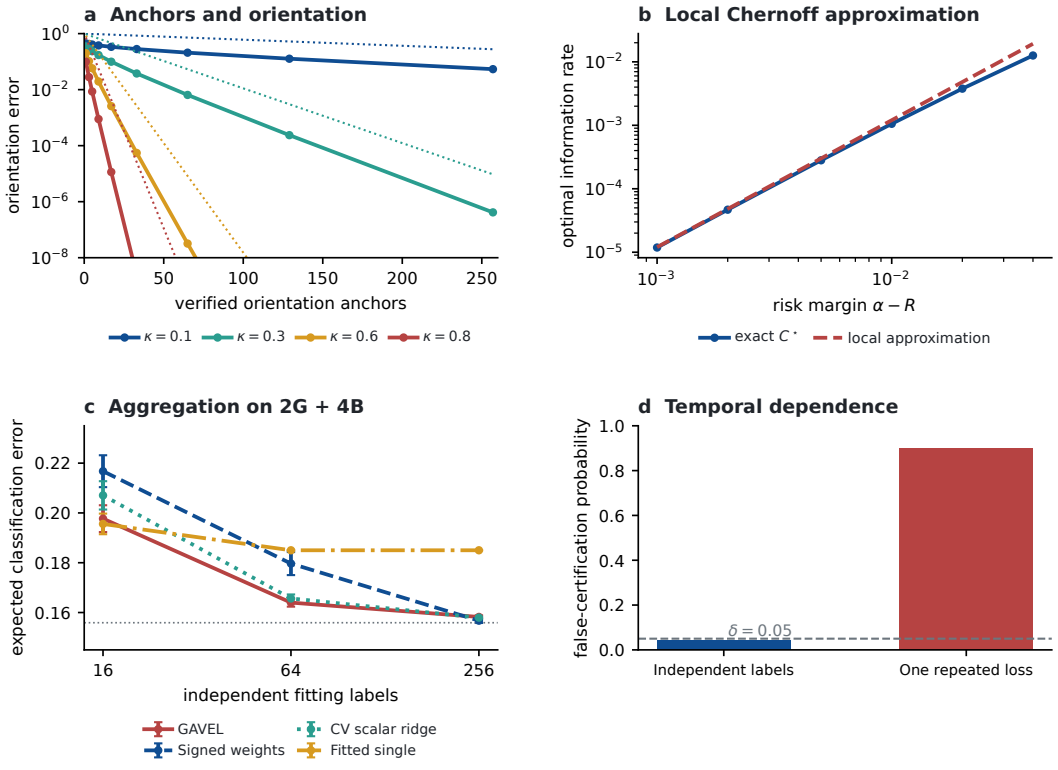


Figure 5: **Supporting controls.** a, Calculated anchor error (solid) and Hoeffding bound (dotted). b, Local versus exact Chernoff rates. c, Exact synthetic aggregation risk averaged over 512 independent fits per label budget, with 95% mean intervals. d, Scalar null crossing under independent losses and the repeated-loss construction.

E.3 SEMI-SYNTHETIC MMLU SHARED-ERROR CONTROL

The MMLU control in Figure 4a uses the initial cached probability policies of `oqwen` and `deepseek_qwen`. At each shared-mode probability $\rho = 0, 0.01, \ldots, 1$, the exact report-pattern calculation compares a common fair replacement sign against independent fair replacement signs, with the same mode indicator and base probabilities. Both arms have individual accuracy $(1-\rho)a + \rho/2$. The displayed curve uses five stochastic replicas per policy. The shared-sign arm first reverses agreement and factual accuracy at $\rho = 0.85$; the independent-sign arm has no corresponding reversal. This intervention isolates the effect of dependence in the specified report channel.

E.4 SCALAR REGRESSION AND ASSUMPTION-FAILURE STRESS TESTS

The independently rebuilt scalar recursion matches the archived rounded values: at $(\alpha, \delta, B) = (0.10, 0.05, 5000)$, crossing probabilities are 0.0431631 for $R = 0.10$, 0.00741282 for $R = 0.12$, and 0.977039 for $R = 0.08$. At $R = 0.05$, the computed noncertification probability is approximately 3.99×10^{-28}, so the displayed success probability rounds to one. The recurrence keeps every state and stores survival mass and truncated mean time.

For the out-of-assumption stress $L_t = B_0$, $B_0 \sim \text{Bernoulli}(0.10)$, the all-zero trajectory crosses within the budget, giving exact false-certification probability 0.90, compared with 0.0432 under i.i.d. Bernoulli(0.10) losses. Changing canonical prevalence to $(0.30, 0.25, 0.05, 0.40)$ changes target risk from 0.06 to $0.0855 > 0.08$; an old certificate cannot transfer to that target. Directed false-negative audit-label noise with probability 0.20 changes true risk 0.09 to observed risk 0.072. These perturbations violate the theorem's conditional independence, prevalence, or reference-label assumptions and measure the consequences for certification.

The synthetic mechanism summary is shown in main-text Figure 3.

E.5 REPRODUCTION AND RESOURCE ACCOUNTING

The synthetic supplement uses root seed 20260916. Independent streams are keyed by experiment, scenario, replicate, and stratum, so batching and worker scheduling do not change the paired comparisons. Numba accelerates the numerical loops; interval checks determine certification. Fixed source versions, configurations, per-run counts, and paired summaries accompany the tables. The sequential audit studies below specify their separate random seeds, shared settings, and evaluation laws.

E.6 MOMENT-BASED AGGREGATION WITH LIMITED FITTING LABELS

Common estimator. For a bounded report vector ϕ, write $\widehat{M}$ for its empirical second moment and $\widehat{\mu}$ for the fitted mean of $Y\phi$. We solve

$$(\widehat{M} + P)\widehat{\beta} = \widehat{\mu}, \qquad P \succ 0, \qquad h(\phi) = \text{sign}(\widehat{\beta}^\top \phi - b).$$

The fitting data determine P and b before freezing. The synthetic experiment uses raw signed reports, $b = 0$, and an unlabeled estimate of M. The cached-data experiments use bounded report features and estimate both moments on the fitting partition. Appendix C.1 gives the covariance interpretation for $P = \lambda I$.

Covariance regularization. The synthetic law has two G evaluators with $(q_G, r_G) = (0.85, 0.10)$ and four B evaluators with $(q_B, r_B) = (0.70, 0.5535714286)$. From 10,000 independent unlabeled report vectors, we estimate correlations C_{ij} and set

$$A_{ij} = (C_{ij} - 0.3)_+ \mathbf{1}\{i \neq j\}, \quad D_{ii} = \sum_j A_{ij}, \quad L = \frac{D - A}{\max\{\max_i D_{ii}, 10^{-12}\}}, \quad P = \lambda(I + \gamma L).$$

The Laplacian penalty shrinks differences between the coefficients of correlated reports. Four-fold fitting-label validation selects $\lambda \in \{0.03, 0.1, 0.3, 1, 3, 10, 100, 1000\}$ and $\gamma \in \{0, 0.25, 1, 4, 16, 64, 256\}$ by classification error. Ties prefer larger γ, then larger λ. A 128-replicate development phase fixed this rule before evaluation on 512 independent fits at each of $m = 16, 64, 256$ labels.

Comparators and evaluation. All rules share the same labeled and unlabeled reports within a replicate. Signed weighting sets $\widehat{\beta} = \widehat{\mu}$, retaining negative coefficients. The fitted single evaluator maximizes $|\widehat{\mu}_i|$ and uses its estimated orientation. Scalar ridge uses $P = \lambda I$. Diagonal ridge additionally permits $P = \lambda \operatorname{diag}(\max\{1 - \widehat{\mu}_i^2, 0.05\})$. Supervised ridge estimates M within each fitting fold. Each cross-validated ridge control uses the same four folds and lambda grid. The constant rule uses the majority fitting label. Score ties incur half an error.

Evaluation sums exact probabilities over the 64 report patterns and two truth values, averaged over independent fitting samples. Thus uncertainty comes from fitting samples. Paired intervals use their empirical standard errors.

Table 2: **Synthetic aggregation error (%).** Each column averages 512 independent fits. Every rule has the same fitting-label budget.

Policy	16 labels	64 labels	256 labels
GAVEL	19.770	16.399	15.823
Signed weighting	21.676	17.966	15.673
Fitted single evaluator	19.558	18.500	18.500
Fixed ridge (0.1)	21.539	18.005	17.282
Cross-validated scalar ridge	20.705	16.565	15.795
Cross-validated diagonal ridge	20.724	16.700	15.867
Supervised scalar ridge	20.362	16.579	15.833
Majority vote	33.064	33.064	33.064
Fitted constant	50.000	50.000	50.000

At 64 fitting labels, GAVEL improves accuracy over signed weighting by 1.567 points (95% interval $[1.095, 2.040]$), over the fitted single by 2.101 points ($[1.938, 2.264]$), and over cross-validated scalar ridge by 0.166 points ($[0.064, 0.269]$). At 16 labels the fitted single has the lowest mean error. Exact enumeration gives Bayes risk $174611/1120000 = 0.15590268$, attained by the population signed-weight direction. With 256 fitting labels, GAVEL and signed weighting achieve errors of 15.823% and 15.673%, respectively, corresponding to excess risks of approximately 0.23 and 0.08 percentage points. Their paired error difference is 0.150 percentage points, with a 95% interval of $[0.039, 0.261]$.

The evaluation uses seed 2026092511 and separate development and evaluation stream keys. Independent checks recompute all 13,824 saved risks and 1,536 selection decisions, with complete Cholesky refits for twelve fitting grids.

E.7 VERIFICATION FROM CACHED MODEL REPORTS

Panels and task. We report verification on MMLU, HellaSwag, ARC-Challenge, and WinoGrande. MMLU model reports come from HELM's cached MMLU runs (Liang et al., 2023). The other three panels use the RouterBench zero-shot cache (Hu et al., 2024). WinoGrande provides a further commonsense task based on reference resolution (Sakaguchi et al., 2021). Each task supplies a question, a fixed proposed answer, and the other models' responses. Its binary label records whether the proposal matches a standard reference answer. The resulting metric is full-coverage verification accuracy. Each panel's methods share the same reports, fitting labels, and evaluation questions.

The MMLU panel has Qwen2.5-7B-Instruct, Llama-3.1-8B-Instruct, Gemma-2-9B, and Mistral-Nemo, with Qwen as proposer. RouterBench has GPT-3.5-turbo-1106 as proposer and ten peers: GPT-4-1106-preview, Claude-instant-v1, Claude-v1, Claude-v2, WizardLM-13B, CodeLlama-34B, Llama-2-70B, Mistral-7B, Mixtral-8x7B, and Yi-34B. The cached outputs support a range of model families and task types within the same verification formulation.

References and partitions. MMLU reference options come from the source scenario records. RouterBench references are aligned with the public HellaSwag release and the martian-evals ARC/WinoGrande records by question and option content. Correct options are remapped after any displayed-order permutation. Parsing accepts a single option letter, exact option text, or explicit final-answer statement. Unmapped peer responses are unknown. Every method rejects an invalid proposal.

The MMLU panel has 5,571 fitting and 8,471 evaluation questions. Normalized question stems inherit fixed source/evaluation membership from the cached benchmark pool, with source precedence for repeated stems, and share a fitting fold. ARC-Challenge uses validation for fitting and test for evaluation, giving 300 and 1,170 rows after source precedence for repeats. HellaSwag and WinoGrande use a 40/60 hash split with salt 20260925. HellaSwag groups questions from the same original source document or video, yielding 3,984 fitting and 6,058 evaluation rows. WinoGrande groups matching normalized sentence templates and candidate pairs, yielding 509 and 758 rows. A group stays in one partition and one of five fitting folds.

Features and fitting. For judge j, let b_j indicate a valid report and $a_j \in \{-1, 0, 1\}$ encode disagreement, unknown, or agreement with the proposal. A supported own-answer log score ℓ_j gives $c_j = \exp(\max\{-30, \ell_j\})$ and availability indicator $u_j = 1$. Other confidence entries are zero. Candidate maps contain an intercept and (a_j, b_j), optionally adding $(c_j, c_j a_j, c_j^2, c_j^2 a_j, u_j)$ and pairwise report relations. RouterBench supplies discrete answers, so its maps use agreement, availability, and pairwise terms.

The estimator in Appendix E.6 uses diagonal penalties. An anchor model's coordinates and the intercept have penalty λ, and other coordinates have penalty $\lambda\rho$. The fitting grid crosses $\lambda \in \{0.001, 0.01, 0.1\}$, $\rho \in \{1, 10, 100\}$, anchor model, feature map, and pooled or stratum-specific coefficients. A stratum with fewer than 50 valid fitting proposals uses pooled coefficients. Five grouped folds provide out-of-fold scores. Thresholds minimize fitting error within strata having at least 30 valid proposals, with a pooled fallback. One common configuration minimizes total fitting error, after which coefficients are refitted and thresholds are frozen.

The MMLU fit uses 24 coordinates, pooled coefficients, and $P = 0.001I$. HellaSwag uses 133 coordinates, pooled coefficients, $\lambda = 0.01$, and relative penalty $\rho = 100$, anchored at GPT-4. The single-model controls have the same folds and threshold calibration, their own bounded and quadratic features, and penalties in $\{0.001, 0.01, 0.1, 1, 10\}$. Controls include one common fitted single, stratum-specific single selection, each separately calibrated model, full-panel agreement counts and fractions, fitted nonempty peer-subset voting, and a constant. The common single selects Gemma for MMLU and GPT-4 for the other three tasks.

Table 3: **Verification on the four reported benchmarks.** Accuracy (%) and gain over the fitted single in percentage points. Intervals use 8,192 paired group-bootstrap resamples.

Panel	Fit	Eval.	GAVEL	Single	Subset	Gain [95% interval]
MMLU	5571	8471	77.16	76.08	76.04	$+1.07$ $[+0.38, +1.77]$
HellaSwag	3984	6058	89.50	88.36	88.20	$+1.14$ $[+0.65, +1.63]$
ARC-Challenge	300	1170	96.50	96.50	96.50	$+0.00$ $[+0.00, +0.00]$
WinoGrande	509	758	79.95	81.79	81.79	-1.85 $[-2.90, -0.92]$

Accuracy and error composition. The MMLU and HellaSwag comparisons reduce total errors by 91 and 69. Against fitted subset voting, their gains are 1.121 and 1.304 points. ARC-Challenge produces identical GAVEL and single predictions. WinoGrande favors the single by 1.847 points. Table 4 separates acceptance of incorrect proposals from rejection of correct ones. These conditional rates complement the prevalence-weighted total error.

Table 4: **Conditional error rates (%).** False acceptance is conditional on an incorrect proposal; false rejection is conditional on a correct proposal.

Panel	FA: GAVEL	FA: single	FR: GAVEL	FR: single
MMLU	52.40	57.85	10.61	9.88
HellaSwag	16.93	13.92	5.82	9.98
ARC-Challenge	13.33	13.33	1.35	1.35
WinoGrande	33.33	28.33	11.35	11.57

Decision changes on WinoGrande. On WinoGrande, GAVEL changes 16 decisions relative to the fitted single evaluator (Table 5). These changes accept one correct proposal and 15 incorrect

Table 5: **Changes relative to the fitted single evaluator.** Corrected errors count questions on which GAVEL is correct and the single evaluator is wrong. Introduced errors count the reverse. The net reduction is their difference.

Panel	Corrected errors	Introduced errors	Net reduction
MMLU	480	389	91
HellaSwag	151	82	69
ARC-Challenge	0	0	0
WinoGrande	1	15	−14

proposals that the single evaluator rejects. False acceptance rises from 28.33% to 33.33%, while false rejection falls from 11.57% to 11.35%. The resulting verification accuracy is 79.95%, compared with 81.79% for the single evaluator.

Bootstrap intervals condition on the frozen procedures and evaluation groups and are pointwise across tasks. HellaSwag uses 5,044 evaluation groups, and MMLU uses 8,364. Fitting choices, coefficients, thresholds, and predictions are fixed before evaluation scoring in the recorded protocol.

Archive and reconstruction. The RouterBench revision is `784021482c3f320c6619ed4b3bb3b41a21424fcb`. Source run versions and object hashes, reference mappings, group assignments, candidate fitting scores, and frozen predictions are recorded in the experiment archive. Independent checks align 12,779 questions and 140,569 reports across the three RouterBench panels. They reproduce 225 selected fits and verify 151,734 policy predictions. For MMLU, the checks reconstruct source fits, predictions, and paired summaries under the same protocol. The bundled summary files contain the selected policies and paired comparisons for the four reported benchmarks.

Table 6: **MMLU controls.** Frozen-policy errors, including separately calibrated individual models.

Policy	Errors	Accuracy (%)
GAVEL	1935	77.16
Stratum-specific aggregate	1953	76.94
Fitted constant	2453	71.04
Fitted single	2026	76.08
Stratum-specific single	2035	75.98
Calibrated Qwen	2134	74.81
Calibrated Llama	2411	71.54
Calibrated Gemma	2026	76.08
Calibrated Mistral	2305	72.79
Calibrated vote count	2080	75.45
Calibrated vote fraction	2065	75.62
Fitted voting subset	2030	76.04

E.8 SYNTHETIC AUDIT PARAMETERS AND NUMERICAL CHECKS

We compare predictable designs using the evidence in Appendix D.2. The common estimator has mean $3\alpha/4$ and pseudo-count strength κ. Each design is refreshed every 25 queries. A candidate crossing is certified only after full-path interval reconstruction. A path that reaches its cap without a certificate incurs the full cap.

Fitting the design parameters. Separate development streams compare estimator strength, stake damping, exploration, and local-to-Chernoff transitions. The two stages contain 17,280 and 12,160 paths. The resulting GAVEL settings are $(\kappa, \rho_0, c_{\exp}, \tau_0) = (64, 0.75, 2, 1000)$, with $a_t = \tau_0/(t+\tau_0)$. The final evaluation fixes these settings and uses fresh seeds.

The matched fixed-prevalence, local, and exact-Chernoff controls use $(64, 0.75, 2)$ throughout. A fixed-prevalence parameter control uses $(64, 0.5, 2)$, and an exact-Chernoff parameter control uses $(256, 0.5, 4)$. Both provide comparisons for the joint effect of estimator strength, initial stake, and

Table 7: **HellaSwag controls.** Frozen-policy errors, including separately calibrated individual models.

Policy	Errors	Accuracy (%)
GAVEL	636	89.50
Stratum-specific aggregate	636	89.50
Fitted constant	2548	57.94
Fitted single	705	88.36
Stratum-specific single	705	88.36
Calibrated GPT-3.5	2548	57.94
Calibrated WizardLM	2548	57.94
Calibrated Yi	1166	80.75
Calibrated Claude-instant	1800	70.29
Calibrated Claude-v1	1982	67.28
Calibrated Claude-v2	1670	72.43
Calibrated GPT-4	705	88.36
Calibrated CodeLlama	2548	57.94
Calibrated Llama-2	2182	63.98
Calibrated Mistral	2549	57.92
Calibrated Mixtral	2569	57.59
Calibrated vote count	1257	79.25
Calibrated vote fraction	1260	79.20
Fitted voting subset	715	88.20

exploration. The two adaptive targets and fixed prevalence share randomized stratum sampling, so their comparison isolates allocation without changing the count mechanism.

Numerical design and evidence checks. An independent optimizer verifies null projections and composite-null factor expectations on 800 parameter states. In the final evaluation, outward-rounded arithmetic verifies every reported certificate. Independent reconstruction checks all 66,048 paths, comprising 314,069,675 draws and 12,562,787 update blocks. A separate 65-digit implementation checks 450 complete paths. These records also reproduce all stopping counts, capped costs, and paired summaries. The final evaluation uses root seed 2026092502 and disjoint keys for each scenario, replicate, and stratum.

E.9 Allocation, heterogeneity, and finite budgets

Risk laws and comparisons. The canonical prevalence is $\pi = (0.40, 0.10, 0.05, 0.45)$. The canonical risks are $(0.07, 0.22, 0.11, 0.01)$, and the heterogeneous risks are $(0.015, 0.35, 0.25, 0.012)$. Their weighted means are 0.0600 and 0.0589. A homogeneous control has $r_k = 0.06$. We use $\alpha \in \{0.07, 0.08, 0.10, 0.12\}$, a fixed-mean family

$$r_s = 0.06\mathbf{1} + s\big((0.07, 0.22, 0.11, 0.01) - 0.06\mathbf{1}\big), \qquad s \in \{0, 0.5, 1, 1.15\},$$

and confidence-level controls at $\delta = 0.01$. Four additional laws use prevalence counts drawn as $\mathrm{Multinomial}(960, p) + 10\mathbf{1}$, $p \sim \mathrm{Dirichlet}(2\mathbf{1})$, with $\mathrm{beta}(0.7, 2)$ risks rescaled to weighted mean 0.06. Risk vectors are redrawn if any coordinate is below 0.001 or above 0.7. The generation seed is 2026092503. Boundary and unsafe controls have constant risks 0.08 and 0.09 at $\alpha = 0.08$.

The local-allocation ablation uses $w_k \propto \pi_k \sqrt{\widetilde{r}_k(1 - \widetilde{r}_k)}$ throughout. The remaining audit components are shared with GAVEL. Tables and figures denote this ablation by Local.

All six methods share each replicate's potential loss arrays and stratum-selection uniforms. There are 25 settings and 66,048 method-paths. Safe settings use 256 or 512 replicates, and calibration uses 2,048 per law. Monitoring budgets are $200, 500, 1000, 2000, 5000, 10000$ and, at $\alpha = 0.07$, 20000. All reported cost comparisons include noncertification at the full budget. The tables use each setting's maximum budget. Paired differences have normal Monte Carlo intervals across replicates.

Figure 4b displays the paired difference in certification probability at each monitoring budget. Figure 4c reports $100(1 - \overline{C}_{\mathrm{GAVEL}}/\overline{C}_{\pi})$, where $\overline{C}$ is the mean capped cost. Its pointwise 95% percentile intervals use 8,192 paired bootstrap resamples of the replicate indices within each setting. The cost ratio is recomputed in each resample, retaining noncertification at the full budget.

Table 8: **Audit costs across risk laws and margins.** Mean capped draws. Budgets are 20,000 at $\alpha = 0.07$ and 10,000 otherwise. Exact uses the exact target throughout. Daggered controls change parameters as specified in Appendix E.8.

Risk law	N	GAVEL	Fixed π	Local	Exact	Fixed†	Exact†
Canonical (0.07)	512	3623.7	4485.0	3606.9	3732.1	4178.6	3588.9
Canonical (0.08)	512	977.3	1092.9	961.2	988.0	1089.8	1033.1
Canonical (0.10)	256	346.1	351.9	344.3	348.7	407.1	439.2
Heterogeneous (0.07)	512	2081.2	3719.6	2057.8	2191.8	3139.6	2428.2
Heterogeneous (0.08)	512	695.8	882.8	697.6	709.0	878.2	842.3
Heterogeneous (0.10)	256	300.6	320.1	300.2	304.1	356.7	403.7
Homogeneous (0.08)	256	1303.5	1307.9	1313.9	1309.0	1339.4	1248.2
$s = 0$, $\alpha = 0.12$	256	237.2	238.1	236.9	236.0	289.5	316.7
$s = 0.5$, $\alpha = 0.08$	256	1095.5	1111.3	1085.4	1090.4	1149.6	1109.2
$s = 0.5$, $\alpha = 0.12$	256	242.2	243.8	242.2	240.8	294.2	319.4
$s = 1$, $\alpha = 0.12$	256	224.1	228.2	224.4	225.7	270.4	300.5
$s = 1.15$, $\alpha = 0.08$	256	864.3	1058.7	875.8	889.0	1048.7	967.7
$s = 1.15$, $\alpha = 0.12$	256	216.8	221.9	217.2	218.3	263.8	299.2
Canonical, $\delta = 0.01$	256	1368.9	1570.6	1377.4	1391.5	1594.5	1485.5
Heterogeneous, $\delta = 0.01$	256	972.9	1241.9	967.0	1000.1	1209.4	1168.1

Primary comparisons and approximation. At $\alpha = 0.08$, the GAVEL-minus-prevalence differences are -115.6 draws ($[-155.3, -75.8]$) on the canonical law and -187.0 ($[-226.6, -147.3]$) on the heterogeneous law. GAVEL and local allocation have means 977.3 versus 961.2 and 695.8 versus 697.6, respectively. All paths certify by the final budget. At $\alpha = 0.07$, savings over prevalence reach 19.2% and 44.0%. These results identify risk heterogeneity and a small safety margin as favorable conditions for adaptive auditing.

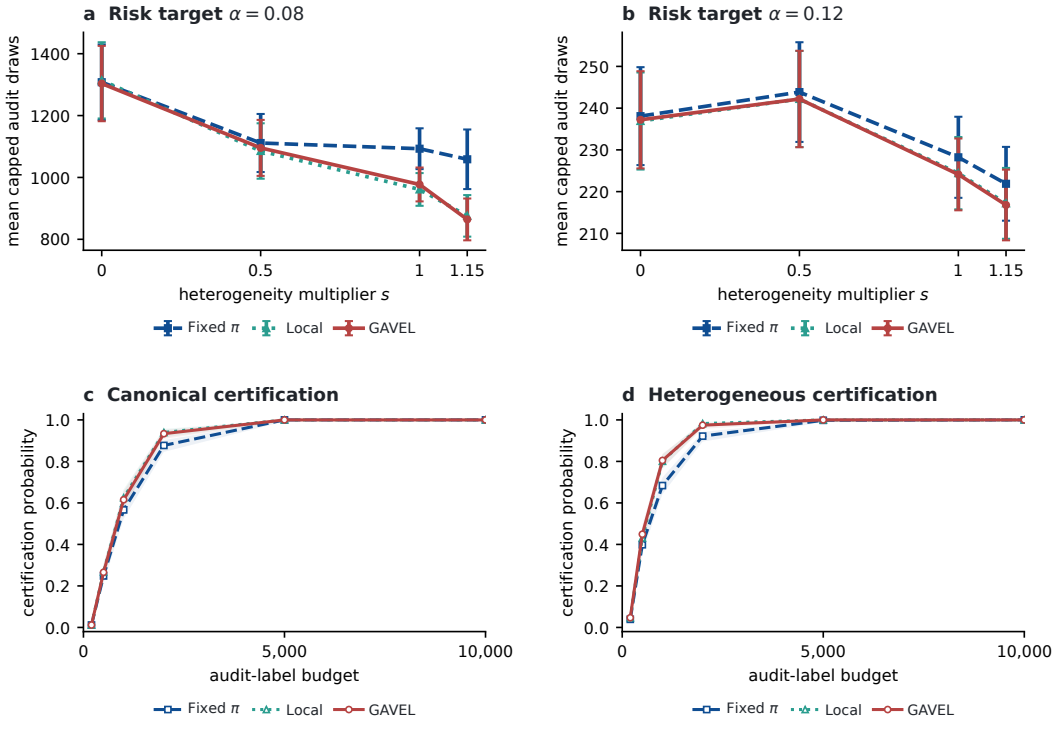


Figure 6: **Allocation across margins and budgets.** a–b, Mean capped cost over the fixed-mean family at $\alpha = 0.08$ and 0.12. Error bars are 95% Monte Carlo mean intervals. c–d, Certification probability in the primary canonical and heterogeneous laws, with 95% Wilson intervals.

For a fixed allocation, let $C(w) = \inf_{q:\,\pi^\top q \geq \alpha} \sum_k w_k \,\mathrm{kl}(r_k, q_k)$. The canonical law has $C^\star = 0.00377924$, $C_\pi = 0.00321112$, and $C_{\mathrm{local}} = 0.00364666$. Their ideal leading-order cost reductions are 15.0% relative to prevalence and 3.5% relative to the local target. The corresponding heterogeneous-law reductions are 20.9% and 4.9%. Finite-budget costs also reflect learning the risks, exploration, and stake damping. The local approximation captures much of the observed allocation benefit, while GAVEL approaches the exact limiting target.

Table 9: **Generated risk laws.** Mean capped draws on four independently generated laws, with α in parentheses. Each row contains 256 paired runs; budgets are 20,000 at 0.07 and 10,000 at 0.08.

Risk law	N	GAVEL	Fixed π	Local	Exact	Fixed†	Exact†
Law 1 (0.07)	256	4751.0	4887.9	4762.1	4813.8	4849.7	4532.1
Law 1 (0.08)	256	1349.4	1369.1	1353.9	1344.0	1437.3	1312.5
Law 2 (0.07)	256	4795.3	5069.4	4726.3	4762.3	5101.0	4557.3
Law 2 (0.08)	256	1176.6	1210.2	1170.9	1179.3	1277.4	1189.1
Law 3 (0.07)	256	4487.5	4957.6	4496.3	4543.2	4839.0	4375.6
Law 3 (0.08)	256	1247.9	1317.2	1247.0	1254.6	1376.1	1296.9
Law 4 (0.07)	256	3104.5	3923.0	3117.3	3116.7	3839.0	3003.2
Law 4 (0.08)	256	881.4	991.7	884.4	870.9	1057.0	969.1

GAVEL's mean cost is lower than matched fixed prevalence in all 23 safe settings. Across the generated laws, local and exact allocation alternate in their ordering. At generated law 1 and $\alpha = 0.07$, GAVEL, local, and exact-common certify 255/256 paths. Both fixed controls and the exact parameter control certify 256/256. The exact parameter control certifies 511/512 paths in the canonical 0.07 setting. All other safe final-budget certification rates are one.

Calibration. At the boundary, GAVEL issues 56/2048 certificates, a frequency 0.02734 with exact binomial 95% interval $[0.02072, 0.03536]$. Fixed prevalence, local, and exact-common issue 62, 57, and 60. At unsafe risk 0.09, each of these four methods issues 5/2048 certificates (0.00244, interval $[0.00079, 0.00569]$). The parameter controls give boundary counts 49 and 74 and unsafe counts 1 and 0. The nominal level is 0.05. Proposition 6 establishes validity under the composite null, and these frequencies check its numerical implementation.

F Scope of the Peer-Only and Audit Results

Complete private states. The Bayes bound applies data processing to the complete peer-accessible state, including information that interaction may reveal after the initial vote. A tool call that adds authoritative information changes that experiment. The agreement-reversal and orientation results describe how informative peer statistics can still fail to identify the external fact.

Selective prediction and orientation. For a peer-observable acceptance event with positive coverage in both swapped worlds, Theorem 3 gives conditional risks summing to one. Abstention changes coverage; resolving orientation requires information that distinguishes the two worlds.

Certification cost after freezing. The stopped-experiment lower bound applies after the policy and strata are fixed. Fit labels can change h_k and hence the risk vector r, so policy selection, evaluator calls, and source acquisition contribute separate pipeline costs.

Aggregate-risk allocation. The allocation game is an instance of Chernoff's controlled sensing (Chernoff, 1959). Its null is an aggregate selective-risk constraint for a frozen evaluator policy, and the optimal allocation specifies which task strata to audit after policy construction.

Auditing a predictor with a shared-error floor. A frozen full-coverage predictor based only on the shared-flip reports S has risk at least R_n^* in (7). If $R_n^* \geq \alpha$, a valid audit certifies it with probability at most δ. Reducing full-coverage risk below this floor requires additional deployment-time information. A revised policy requires a fresh audit. Selecting an eligible population instead changes the target to its conditional risk, to which the same full-coverage bound need not apply.

G LEAN VERIFICATION BOUNDARY

The supplementary archive includes the complete `LeanProofs/` project, containing 30 finite identities and inequalities. They cover family-channel statistics and the reversal threshold, truth-swap indicators, correctness second moments, shared-flip errors, and variance bounds. Further modules check GAVEL's weighted one-step means and factor positivity, fixed-allocation likelihood-ratio and log-growth-gap identities, boundary normalization, and the weighted variance decompositions and coordinate algebra used in the heterogeneity analysis.

From `LeanProofs/`, run `lake exe cache get` followed by `sh verify.sh`. The included README also gives a Windows entry point. The archive supplies the pinned toolchain and dependency manifest, source hashes, and verification records. The scripts check source integrity, reject proof placeholders and additional axioms, run warning-fatal module checks, and audit all 30 theorem dependencies.

The formal verification covers these finite algebraic and logarithmic statements under their declared hypotheses. The probabilistic arguments, optimization existence, uniqueness and log optimality, KL inequalities, full stratification equality conditions, and asymptotic expansions and limits are established by the written proofs in this appendix.