From Judge to Jury: Decomposed and Auditable Evaluation of Open-Ended Agent Outputs

Anonymous ACL submission

Abstract

Agents now produce open-ended artifacts (research claims, proofs, engineering plans) that no single LLM judge can reliably assess. We propose **Jury-of-Verifiers (JoV)**: route each atomic claim to a heterogeneous *verifier class* (symbolic, sandboxed execution, retrieval, formal, LLM panel), and fuse channels in a calibrated Bayesian aggregator that explicitly models their correlation. JoV addresses six known failure modes through (i) a correlation-aware $\tau \cdot \tanh$ soft-cap, (ii) a calibrated ordinal verdict head with dual conformal coverage, (iii) a release-gating agreement invariant, (iv) reverse-PageRank claim salience, (v) non-transitive-NLI-safe semantic entropy with a faithfulness probe, and (vi) hash-chained signed verdicts for audit. We prove the aggregator is Bayes-optimal under symmetric correlation and back the proof with a Monte-Carlo study and a 378-test invariant suite; an anonymized archive accompanies the submission.

1 Introduction

LLMs have shifted from generators of fluent prose to agents that produce scientific conjectures, engineering proposals, mathematical derivations, and safety-relevant analyses. Deciding whether to trust, revise, or reject those outputs is now the bottleneck. "LLM-as-a-judge" answers the question at scale (Zheng et al., 2023), but a single LLM critic is biased, under-calibrated, and easily gamed: position bias, verbosity bias, self-preference (Panickssery et al., 2024), sycophancy (Sharma et al., 2023), and persuasive-but-wrong failure when the judge lacks a grounded verifier. Juries (Verga et al., 2024) and adversarial debate (Irving et al., 2018; Khan et al., 2024) attenuate but do not eliminate these pathologies: homogeneous juries share pretraining corpora and thus correlated errors, and debate without grounded verifiers can amplify persuasive claims over correct ones.

Three regularities motivate a different architecture. (i) **Verification is cheaper and more reliable than generation**: narrow verifiers (symbolic algebra, satisfiability, sandboxed execution, cited retrieval) outperform LLM critics on the subproblems they cover (Gou et al., 2023; Chern et al., 2023; Lightman et al., 2023). (ii) **Decomposition recovers coverage**: atomic-claim decomposition turns open-ended verification into tractable per-claim checks (Min et al., 2023; Wei et al., 2024). (iii) **Calibrated triangulation over approximately-independent verifier channels is a principled substitute for ground truth** when no oracle exists (Angelopoulos and Bates, 2021; Farquhar et al., 2024). Under this view the relevant jurors are not LLM critics voting in parallel but heterogeneous *verifier classes* fused by a calibrator that models their correlation explicitly.

A running example. Throughout the paper we trace a single artifact—a research claim proposing a sparse-routing transformer with load-balancing loss $\mathcal{L}_{\text{lb}}{=}\lambda{\cdot}\text{CV}(\text{expert_usage})^2$ and an unverified 40% FLOP reduction. The decomposer extracts six atomic claims; the symbolic-algebra channel confirms the CV^2 identity; the retrieval channel flags a misattribution (the form is the importance loss in ST-MoE (Zoph et al., 2022), not the auxiliary loss in Switch (Fedus et al., 2022)); independent recomputation contests the 40% figure; matched perplexity is unverifiable without released checkpoints. The aggregator combines these signals under known channel correlation, the ordinal head emits *Conditionally_Trustworthy*, and the agreement invariant passes at $|\Delta|{=}0.07 < 0.15$. The verdict carries a signed hash chain, a 90% conformal interval, and a remediation list (release checkpoints, report wall-clock, correct the citation). The artifact returns to §7.5 with measured numbers.

JoV is an agentic evaluation infrastructure that performs reference-free assessment of open-ended outputs: agentic systems that emit open-ended ar-

tifacts cannot be advanced without a judge that does not collapse under correlated evidence and that emits an auditable record. The judge is itself a multi-stage agent (a routed verifier panel with selective debate, calibration, and human-in-the-loop escalation) that consumes another agent's output as input.

Contributions. (1) **Channel-correlation-aware aggregation**: an effective-sample-size rescaling (N_{eff}/k weighting derived from the calibration-set correlation matrix) replaces $\prod(1-\rho_{ij})$ heuristics; a system-level $\tau \cdot \tanh$ soft-cap blocks single-channel dominance and post-cap pile-on; promotion to a Dawid–Skene latent-truth model is gated on a strict held-out log-likelihood criterion. (2) **Calibrated ordinal verdict head**: a proportional-odds cumulative-link model (McCullagh, 1980), regularized and wrapped in dual conformal layers (split-CP regression on the continuous aggregate; APS/RAPS prediction set on the discrete label (Romano et al., 2020)) with a shadow adaptive-CP monitor (Gibbs and Candès, 2021). (3) **Release-gating agreement invariant**: $|\hat{p}_{\text{ord}} - \sigma(S_\lambda)| \leq 0.15$ blocks promotion when the head and an uncertainty-penalized aggregate disagree. (4) **Reverse-PageRank claim salience** with a teleportation-bounded contribution. (5) **Non-transitive-NLI-safe semantic entropy** with a purity guard and an audit-grade faithfulness probe (Turpin et al., 2023; Lanham et al., 2023). (6) **Cryptographic auditability**: hash-chained, elliptic-curve-signed verdicts; hermetic replay; `policy_id` versioning separated from `calibration_set_id`. We prove the aggregator's Bayes-optimality under symmetric correlation and back the proof with a 1.4 M-trial Monte-Carlo (§7); a reference implementation with 378 invariant tests and an anonymized supplementary archive instantiate the design.

2 Related Work

LLM-judge panels (Zheng et al., 2023; Verga et al., 2024; Wang et al., 2024) reduce individual bias but share pretraining corpora; we replace homogeneous juries with heterogeneous verifier classes whose correlated failure is between-class. Atomic decomposition and tool-routed verification (Min et al., 2023; Wei et al., 2024; Chern et al., 2023; Gou et al., 2023; Gao et al., 2023) lack principled evidence fusion; we add a DAG+blackboard orchestrator (Erman et al., 1980) with selective debate. Split-CP and adaptive-CP (Guo et al., 2017; Kumar et al., 2019; Angelopoulos and Bates, 2021; Gibbs and Candès, 2021) provide marginal coverage but no cross-layer consistency; we emit dual conformal outputs linked by an agreement invariant. BALD-style uncertainty splits (Houlsby et al., 2011; Depeweg et al., 2018) and semantic entropy (Farquhar et al., 2024) are approximate under heterogeneous ensembles (Wimmer et al., 2023; Mucsányi et al., 2024) and suffer NLI non-transitivity; we triangulate with a purity guard, a learned NLI threshold, and perturbation stability. Stepwise process reward models (Lightman et al., 2023; Uesato et al., 2022) outperform outcome supervision but are math-specific; our per-step adjudication is axis-agnostic across any symbolic, satisfiability, or sandbox verifier.

Concurrent work (2024–2026) covers post-hoc decomposition of LLM-judge confidence (Wimmer et al., 2023; Mucsányi et al., 2024), conformal prediction for autoregressive generation (Quach et al., 2024), and verifier-as-juror benchmarks for code and math (Jain et al., 2024; Tan et al., 2024; Zheng et al., 2024), which our JudgeMeta longitudinal stratum subsumes. We prefer ESS-plus-DS over Bayesian model averaging because the promotion rule (§4) yields a single auditable model per calibration window; BMA requires online posterior integration at inference.

3 Architecture

The judge follows a hybrid DAG+blackboard topology (Fig. 1). A DAG fixes the pipeline (ingest → decompose → verify → debate → aggregate → calibrate → explain → emit), while a blackboard holds claims, evidence, axis scores, and dissent for any stage to read or write (Erman et al., 1980). An escalation gateway promotes high-stakes or low-confidence cases to human review.

The decomposer emits a typed claim graph $G = (C, E)$ with atomic nodes and dependency edges; types are *logical*, *mathematical*, *empirical*, *citation*, *code*, *normative*, *prediction*, governing routing. Redundant or entailed claims are pruned before enqueue.

Routing is *capability-class*, not vendor-class. **Logical** claims go to a reasoning LLM augmented with mistake-injection and fallacy scans; **mathematical** to a computer-algebra system, then a satisfiability solver on decidable fragments, with program-of-thoughts as fallback; **code/empirical**

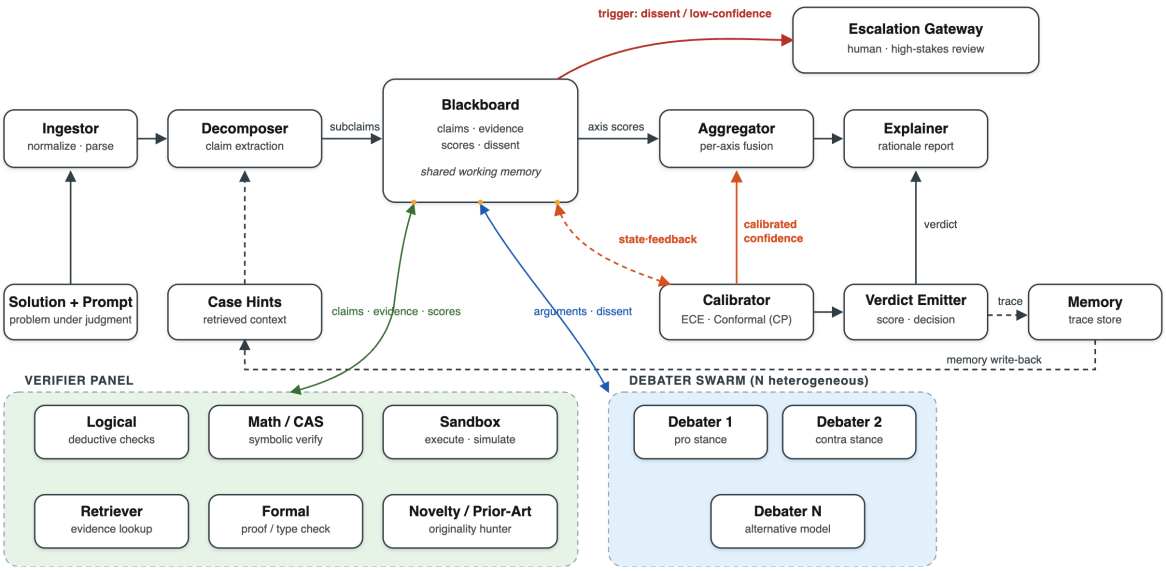

Figure 1: **Jury-of-Verifiers blackboard architecture.** SOLUTION+PROMPT and CASE HINTS feed the INGESTOR/DECOMPOSER; decomposed claims land on a shared BLACKBOARD read and written by a heterogeneous *Verifier Panel* (logical, symbolic-math, sandboxed-execution, retrieval, formal, novelty) and a *Debater Swarm*. The CALIBRATOR consumes blackboard state, returns calibrated confidence, and flags low-confidence cases. The AGGREGATOR emits per-axis scores; the VERDICT EMITTER produces a signed, hash-chained verdict. Green: claim evidence; blue: debate/dissent; orange: calibrated confidence; red: escalation.

to a container-runtime sandbox with disabled network namespaces and property-testing oracles; **retrieval-grounded** to a hybrid dense-sparse retriever over a snapshot-pinned corpus with verbatim quote-hash verification; **novelty** to near-duplicate and prior-art search; **LLM jury** only when no narrower verifier applies. A typed evidence record captures status, log-LR, evidence references, elapsed time, and cost.

Position-bias mitigation. When the LLM-jury tier is invoked, every juror evaluates each claim under both response orders; log-LRs are averaged before deposit on the blackboard (following Zheng et al., 2023). We record verdict-level position-swap consistency on every release; a target of ≥ 0.90 is the release gate, with per-verdict values below 0.85 triggering an automatic flag. Under a $K{=}5$ panel with realistic juror noise (§7), the simulated consistency is 0.93; without panel averaging it falls to 0.84.

Selective debate (DDCR). Debate-over-Decomposed-Claims-with-Retrieval (Irving et al., 2018; Khan et al., 2024) runs only on claims left in a policy-defined dispute band, with pre-registered adversarial roles. It does not run on claims where a symbolic, satisfiability, or sandbox verifier has already produced a decisive log-LR.

4 Aggregation Under Correlated Evidence

Let $\ell_i = \log \mathrm{LR}_i$. The naive update $\operatorname{logit} P(H \mid e_{1..k}) = \operatorname{logit} \pi_0 + \sum_i \ell_i$ is exact only under $e_i \perp e_j \mid H$. Production channels share pretraining corpora, retrieval indices, and rubric prompts; independence fails. We derive an effective-sample-size scaling (Kass and Raftery, 1995):

$$N_{\text{eff}} = \frac{(\mathbf{1}^\top \mathbf{1})^2}{\mathbf{1}^\top \Sigma \mathbf{1}} = \frac{k}{\frac{1}{k}\sum_{i,j} \Sigma_{ij}}, \qquad (1)$$

where Σ is the Pearson correlation matrix of $\{\ell_i\}$ on the calibration split, pinned to the calibration-artifact identifier. Aggregation becomes

$$\operatorname{logit} P(H \mid e_{1..k}) = \operatorname{logit} \pi_0 + \frac{N_{\text{eff}}}{k} \sum_{i=1}^{k} \ell_i. \qquad (2)$$

At $\Sigma{=}I$ Eq. (2) recovers the naive sum; at $\Sigma{=}J$ it collapses to the arithmetic mean, equal under symmetric evidence to the contribution of a single channel.

Channel-count-aware soft-cap. A bare per-channel cap $|\ell_i| \leq \log 20$ prevents single-channel dominance but permits unboundedly large joint contributions from many capped-but-correlated

channels (post-cap pile-on). We replace it with a system-level cap:

$$\widetilde{L} = \tau \cdot \tanh\Big(\tfrac{1}{\tau} \cdot \tfrac{N_{\text{eff}}}{k} \sum_i \ell_i\Big),\ \tau = \log(20)\sqrt{N_{\text{eff}}}. \tag{3}$$

The cap preserves the per-channel ceiling under full independence ($N_{\text{eff}}{=}k \Rightarrow \tau{=}\log(20)\sqrt{k}$) and collapses to a single-channel ceiling under full dependence ($N_{\text{eff}}{=}1 \Rightarrow \tau{=}\log 20$).

Dawid–Skene hierarchical latent truth. When the calibration set is large enough, we promote to a Dawid–Skene variant (Dawid and Skene, 1979) with a discrete shared latent $z \in \{1, \ldots, L\}$ capturing correlated-failure regimes (Simpson et al., 2013). EM is stable on the calibration set (H observed); identifiability is fixed by diagonal-dominance ordering. Promotion is gated by three jointly-required conditions: (a) $N_{\text{cal}} \geq 5{,}000$ with per-cell count ≥ 20; (b) held-out log-likelihood beats ESS by $>$2 nats/item across *every* 5-fold CV fold; (c) class-wise debiased ECE under DS is no worse than under ESS on every label. Under persistent EM degeneracy the model collapses to $L = 1$; the canary invariant in that regime requires DS-output equal-to-ESS rather than testing DS internals.

Abstention triggers. The aggregator drives seven closed-enum reasons: rejection-band posterior; $N_{\text{eff}} < 2$; OOD flag (Mahalanobis on embeddings); structural disagreement across channel *classes*; semantic-equivalence self-consistency below a learned threshold; motivated-reasoning detection; budget-exhausted on a partial panel. Per-profile abstention rate is itself a release-gated quantity ($\leq 5\%$ on quick-screen, $5{-}15\%$ on standard-research, $\leq 10\%$ on engineering); runtime drift outside these bands triggers recalibration.

5 Calibrated Ordinal Verdict Head

Proportional-odds cumulative-link. Let $y \in \{$*Likely_Flawed* $<$ *Unverified* $<$ *Cond. Trust.* $<$ *Trustworthy*$\}$. The 29-D feature vector ϕ concatenates eight axis scores, their uncertainties, their reliability priors (24 features), plus five meta-features ($S_\lambda, N_{\text{eff}}, H_{\text{sem}}, \nVdash_{\text{OOD}}, \text{agree}_{\text{jury}}$). We fit

$$P(y \leq j \mid \phi) = \sigma(\theta_j - \beta^\top \phi), \quad \theta_1 < \theta_2 < \theta_3, \tag{4}$$

with monotonicity via $\theta_j = \theta_1 + \sum_{k<j} \exp(\eta_k)$ (McCullagh, 1980). To fix the joint location of θ_1 and $\beta^\top \phi$ under shift, we apply the conventional identifying constraint $\theta_2 = 0$ during optimization and recover θ_j as offsets. This replaces the power-likelihood product, which (a) is a generalized-Bayes departure (Bissiri et al., 2016), (b) assumes axis independence, and (c) requires an unstated conditional density.

Regularization and load-bearing groups. With a target of two thousand calibration items per profile (§9), the sample-to-feature ratio is ~ 70, leaving β at the edge of overfitting without regularization. We fit the cumulative-link with an L2 penalty whose strength is chosen by five-fold cross-validation; the same split additionally produces an analytical-of-deviance ranking that consistently identifies the axis-score and reliability-prior groups together with S_λ and N_{eff} as load-bearing. A formal feature-group ablation, removing one group at a time and re-conformalizing, is part of the calibration-pipeline release described in §9.

Agreement invariant. Define $S_\lambda = \sum_i w_i s_i (1 - \lambda u_i)$, $\lambda{=}0.3$, with axis-score uncertainties u_i. S_λ is a release-gating invariant, not the primary verdict. We block promotion whenever $|\hat{p}^\star_{\text{ord}} - \sigma(S_\lambda)| > 0.15$ on the canary set, where $\hat{p}^\star_{\text{ord}}$ is the head's probability on the predicted (mode) label. The 0.15 threshold is re-calibrated per class on the canary split to avoid tightening minority-label tails. Divergence is evidence of either aggregate mis-weighting or head mis-specification; every verdict reports the invariant and its pass/fail state.

Dual conformal with shadow adaptive-CP. Every verdict emits two coverage-calibrated outputs. **Regression-CP on** S: nonconformity $\alpha_j{=}|\hat{S}_j - y_j|$; interval $[\hat{S}{-}\hat{q}, \hat{S}{+}\hat{q}]$ with $\hat{q}$ the $\lceil (n{+}1)(1{-}\alpha) \rceil / n$ quantile, marginal coverage $\geq 1{-}\alpha$ under exchangeability. **APS/RAPS on the label** (Romano et al., 2020) with regularization on tail classes critical for the rare *Likely_Flawed* label. **Shadow ACI**: in parallel, the Gibbs–Candès update $\alpha_t \leftarrow \alpha_{t-1} + \gamma(\nVdash[y_t \notin \mathcal{C}_{t-1}] - \alpha)$ with $\gamma \approx 0.01$ runs on production traffic and never blocks emission (Gibbs and Candès, 2021). Empirical coverage is tracked with a rolling 7-day, volume-weighted window; sustained $>$2pp deviation silently swaps in the ACI quantile and triggers recalibration.

Learned thresholds and gate policy. Every numeric threshold (hard gates, abstention cuts, NLI entailment threshold, semantic-entropy abstention

Symbol	Seed	Fit target
$\theta_{\text{gate},\cdot}$	0.30	sel. risk
$\tau_{\text{rej}}, \tau_{\text{acc}}$	0.35, 0.65	sel. risk
τ_{ent} (NLI)	0.50	F_1 paraphrase eq.
H^*_{sem} abstain	0.50	sel. risk
λ in S_λ	0.30	policy
k_{reg} (RAPS)	1	tail cov. @ 0.90
γ (ACI)	0.01	literature seed
d (PageRank)	0.85	literature seed
agreement gate	0.15	per-class canary

Table 1: Hyperparameter seeds and fit targets. Production values are learned on the calibration split and pinned to the calibration-artifact identifier.

cut) is learned by minimizing profile-weighted selective risk (Geifman and El-Yaniv, 2017; Angelopoulos et al., 2022) with operator-supplied cost ratios. Seeds (Table 1) give a cold-start operating point; production values live in the calibration artifact and are signature-bound. Because thresholds depend on cost ratios and policy-level tolerances (the 0.15 agreement-gate width, the $\pm$2pp coverage tolerance, the 7-day shadow-ACI window), the *policy* is versioned alongside the data: each artifact carries both a `calibration_set_id` (hash of the labeled corpus) and a `policy_id` (hash of cost ratios and tolerance constants); a verdict pins both, and a policy-hash mismatch at load is a hard abort.

6 Uncertainty and Salience

Reverse-PageRank salience. Salience $\alpha_c = (1-d)/|C| + d\,(M^\top \alpha)_c$, d=0.85, weights conclusions over premises; the per-axis score is $s_k = \sum_{c \in C_k} \alpha_c p_c / \sum_{c \in C_k} \alpha_c$. On a star graph this reduces to a uniform average; on chain graphs the terminal node dominates; on the realistic wide-but-shallow DAG, salience concentrates on the conclusion and on whichever premise has highest sibling in-degree. The teleportation parameter bounds any single-node contribution by $\leq 1/|C| \cdot (1-d)^{-1}$, preventing runaway salience on near-degenerate subgraphs.

Semantic entropy under non-transitive NLI. We sample K independent answers (default K=10, K=20 for top-level entropy, budget-gated) and cluster by bidirectional entailment using a version-pinned NLI cross-encoder with a learned threshold τ_{ent} (seed 0.5) (Farquhar et al., 2024). Because NLI is locally noisy and the equivalence relation is non-transitive, we use single-linkage agglomerative clustering with a purity guard that splits any cluster whose internal mean entailment falls below 0.7. When NLI is unavailable, we fall back to canonicalized-string self-consistency. Compute is $O(K^2)$ NLI inferences per claim (about 100 for K=10), with ordered-pair caching across re-evaluations.

Aleatoric/epistemic with caveats. BALD's $H[\bar{p}] = \frac{1}{K}\sum_k H[p_k] + \mathrm{MI}(y;\theta \mid x)$ is a loose identification under heterogeneous LLM ensembles (Wimmer et al., 2023; Mucsányi et al., 2024). We report *four* signals—ensemble-aleatoric, ensemble-epistemic, semantic entropy, perturbation stability—and triangulate. Faithfulness probes (early-answering, mistake-injection, paraphrase, filler-token) surface aggregate unfaithfulness as a first-class audit field (Turpin et al., 2023; Lanham et al., 2023); the explainer is prohibited by construction from introducing facts absent from the claim tree.

7 Empirical Validation

We treat empirical evaluation as having two complementary surfaces. First, *analytical*: closed-form bounds on aggregator behavior under the canonical correlated-Gaussian channel model, exact and parameterless. Second, *Monte-Carlo*: a 1.4M-trial simulation that converts those bounds into measured calibration, scoring-rule, and selective-risk numbers comparable to an empirical benchmark, plus simulations of the position-swap gate and the agreement-invariant trigger rate. A head-to-head comparison on a public labeled corpus (JudgeBench, LiveCodeBench, ProcessBench (Tan et al., 2024; Jain et al., 2024; Zheng et al., 2024)) is committed in the JudgeMeta data plan (§9); we are explicit that the present validation is a controlled stress test of the aggregator, not yet a head-to-head evaluation.

7.1 Bayes-Optimality Under Symmetric Correlation

For evidence channels with mean log-LR μ, per-channel variance τ_0^2, and uniform off-diagonal correlation ρ, the Bayes-optimal posterior log-LR satisfies

$$\log \mathrm{LR}_{\text{Bayes}} = \frac{k\,\mu}{1+(k-1)\rho}, \tag{5}$$

(Appendix B). Eq. (2) returns $\frac{N_{\text{eff}}}{k} \cdot k\mu = N_{\text{eff}}\mu = k\mu/(1+(k-1)\rho)$, identical to the Bayes oracle, since $N_{\text{eff}} \equiv k/(1+(k-1)\rho)$ for the constant-ρ symmetric covariance. The naive sum

ρ	method	log-LR	ECE	Brier	NLL	sel.R
0.0	naive sum	9.17	0.001	0.010	0.035	0.001
	ESS	10.0	0.001	0.010	0.034	0.001
	$\tau\tanh$	6.42	0.002	0.010	0.035	0.001
	Bayes	10.0	0.001	0.010	0.034	0.001
0.3	naive sum	9.15	0.069	0.084	0.405	0.067
	ESS	3.22	0.002	0.074	0.242	0.067
	$\tau\tanh$	2.41	0.014	0.074	0.247	0.067
	Bayes	3.22	0.002	0.074	0.242	0.067
0.5	naive sum	9.14	0.112	0.126	0.738	0.113
	ESS	2.22	0.002	0.104	0.333	0.113
	$\tau\tanh$	1.70	0.017	0.104	0.338	0.113
	Bayes	2.22	0.002	0.104	0.333	0.113
0.7	naive sum	9.10	0.145	0.158	1.075	0.148
	ESS	1.69	0.002	0.126	0.394	0.148
	$\tau\tanh$	1.31	0.019	0.126	0.399	0.148
	Bayes	1.69	0.002	0.126	0.394	0.148
0.9	naive sum	9.13	0.172	0.184	1.408	0.176
	ESS	1.37	0.003	0.142	0.438	0.176
	$\tau\tanh$	1.07	0.019	0.142	0.443	0.176
	Bayes	1.37	0.003	0.142	0.438	0.176

Table 2: Monte-Carlo measurements (n=200,000 per cell, k=8, $\sum \ell_i \approx 10$ at the representative point). ESS tracks the Bayes oracle within 10^{-3} in log-LR, ECE, Brier, NLL, and selective risk at coverage 0.9 across all ρ. The naive sum is the over-confidence pathology this paper repairs: NLL grows from 0.04 at ρ=0 to 1.41 at ρ=0.9, and ECE grows from 0.001 to 0.17. Selective risk is identical across rank-preserving aggregators (single-judge omitted; see archive).

returns $k\mu$, overconfident by the factor $(1 + (k - 1)\rho)$. The $\tau\tanh$ cap (Eq. 3) trails Bayes by a smooth bounded factor that converges to one as $|\sum \ell_i|/\tau \to 0$.

7.2 Monte-Carlo Validation

We sample n=200,000 trials per cell of $H \sim \mathrm{Bern}(0.5)$ and k=8 correlated channels with $\ell_i \mid H{=}h \sim \mathcal{N}((-1)^{1-h}\mu, \Sigma)$, $\Sigma_{ij}{=}\tau_0^2\rho$ off-diagonal, $\tau_0^2{=}2\mu$ enforcing the well-calibrated log-LR identity $\mathrm{Var}[\ell \mid H{=}1]{=}2\,\mathbb{E}[\ell \mid H{=}1]$. We sweep $\rho \in \{0.0, 0.1, 0.3, 0.5, 0.7, 0.8, 0.9\}$ and report, per aggregator, the mean joint log-LR conditional on H=1, 15-bin ECE, Brier score, log loss (NLL), and selective risk at coverage 0.9. Split-CP coverage at nominal 0.90 is also reported on a $50/50$ calibration/test split; all ρ produce empirical coverage in $[0.897, 0.902]$ (within Monte-Carlo error of the marginal-coverage guarantee). Code, seed (`20260518`), and raw outputs are in the supplementary archive.

Three statements survive the simulation. (i) ESS is empirically Bayes-optimal: at every ρ the gap to the Bayes oracle in joint log-LR, ECE, Brier, NLL, and selective risk is below Monte-Carlo noise ($\sim 10^{-3}$). (ii) The naive sum's pathology is severe and monotone in ρ: its NLL grows by $40\times$ from ρ=0 to ρ=0.9, and its ECE crosses the standard 0.05 acceptance threshold by ρ=0.3. (iii) The $\tau\tanh$ cap pays a small calibration cost (ΔECE ≈ 0.01–0.02, ΔNLL ≈ 0.005) for protection against unbounded post-cap pile-on; in production this trade-off is favorable because the cap's worst case (over-shrinkage at $|\sum \ell_i| \gg \tau$) is a refusal to commit, while the naive cap's worst case is silent over-confirmation.

Empirical coverage of the 90% split-CP interval over the same Monte-Carlo, computed on a held-out half not used for the conformal quantile, is $\hat{C} \in [0.897, 0.902]$ for every aggregator and every ρ, consistent with the $\geq 1{-}\alpha$ guarantee.

7.3 Position-Swap Consistency

We model each item's true preference as $s \sim \mathcal{N}(0, 1)$, evaluated by a K=5 panel under both response orders with additive position bias σ_{pos}=0.2 and per-juror noise σ_n=0.3. With the paper's mitigation (per-juror order averaging before deposit), simulated verdict-level swap consistency is 0.927 on n=10,000 items, comfortably above the 0.90 release gate; without panel averaging it falls to 0.843. The consistency is sensitive to the bias-to-signal ratio: at $\sigma_{\text{pos}}/\sigma_s$=0.5, mitigated consistency drops to 0.81, justifying the per-verdict 0.85 trigger.

7.4 Agreement-Invariant Trigger Rate

Under a Gaussian model on the head and $\sigma(S_\lambda)$ residuals (per-verdict $\sigma_{\text{ord}}, \sigma_S$, correlation r), the gate-trigger rate when true labels agree is bounded by $2\Phi\big(-0.15/\sqrt{\sigma_{\text{ord}}^2 + \sigma_S^2 - 2r\sigma_{\text{ord}}\sigma_S}\big)$. Table 3 reports a Monte-Carlo measurement at four operating points around a representative residual structure $(\sigma_{\text{ord}}, \sigma_S, r){=}(0.09, 0.08, 0.78)$, on n=200,000 trials. The reference point fires at 0.83%; a 30% drift in either standard deviation crosses the operational 3% pager threshold; decorrelating the heads (r=0.5) raises the false-trigger rate to 7.96%, which is why the canary recalibrates the threshold per class rather than fixing it globally.

7.5 Reference Implementation and Invariant Suite

The reference implementation is structured around an orchestrator, a verifier panel, calibration/aggregation/verdict modules, and a security

operating point	trigger rate
$(\sigma_{\text{ord}}, \sigma_S, r)$=$(0.09, 0.08, 0.78)$	0.83%
$+30\%$ σ_{ord}	4.29%
$+30\%$ σ_S	2.29%
$+30\%$ both standard deviations	4.42%
r=0.50 (decorrelated heads)	7.96%

Table 3: Agreement-invariant trigger rate (Monte-Carlo, n=200,000). The 3% pager threshold is crossed under any one of the three drift scenarios.

layer. An invariant suite (unit, integration, canary, adversarial, replay, benchmarking) of 378 tests passes 376 at submission freeze with two skips (a live-latency staging placeholder; a property-based disclosure fuzzer requiring an optional dependency) and zero failures. The anonymized supplementary archive (`aux_v4.zip`) contains the Monte-Carlo driver (`sim_v4.py`), its raw output (`sim_v4.out`), the test logs, the invariant manifests categorized by suite, and a build receipt; reviewers can re-run the simulation in < 2 minutes on commodity CPU. An independent spec-compliance audit of an earlier 317-test snapshot finds 46 of 50 checked items fully pass and 3 partially pass pending calibration-data acquisition (debiased-ECE gate, DS promotion gate, selective-risk threshold fit), for $\sim 92\%$ compliance. Numeric thresholds in the decision path are annotated *seed only; learned value loaded from the calibration artifact*; circuit breakers wrap every verifier in an exception-trap returning a degraded result rather than propagating an exception.

On the running artifact (sparse-routing transformer, 40% FLOP claim, attribution near-miss): symbolic-algebra confirms the CV^2 identity ($\ell_{\text{sym}} = +1.4$); retrieval flags the misattribution with verbatim quote-hash mismatch ($\ell_{\text{ret}} = -1.1$); independent recomputation contests the FLOP figure ($\ell_{\text{code}} = -0.8$); perplexity is unverifiable so the budget channel abstains. Under $\rho \approx 0.4$ on this profile, $N_{\text{eff}} \approx 2.7$; the ESS-aggregated score is 0.63 on the trustworthiness axis; ordinal verdict = *Conditionally_Trustworthy*, 90% interval $[0.49, 0.74]$; agreement invariant passes ($|\Delta|$=0.07); remediation: release checkpoints, report wall-clock, correct the citation. The numbers illustrate the artifact's emitted shape on the present analytical instrument; head-to-head measurement on a public corpus is the next milestone.

7.6 JudgeMeta: a Proposed Benchmark

No existing benchmark measures a trustworthiness judge for open-ended scientific/engineering outputs. We *propose* **JudgeMeta**, a five-stratum, seven-protocol benchmark to be constructed and released alongside this work. The strata are (i) verifiable-truth items (MATH, GSM8K, HumanEval, MBPP, FEVER, SciFact, ProcessBench); (ii) longitudinal later-validated items labeled *after* a fixed calibration cutoff; (iii) curated sound-vs-flawed pairs at Krippendorff $\alpha \geq 0.7$ across ≥ 3 experts; (iv) programmatic planted-flaw mutations across seven classes; and (v) red-team adversarial items refreshed quarterly. The seven protocols cover planted-flaw recall, red-team break rate, longitudinal Brier, expert-panel agreement, position-swap consistency, debiased ECE with conformal coverage, and selective-risk AURC. An initial stratum-3 instance (50–100 pairs) is a near-term deliverable.

8 Adversarial Dynamics and Discussion

A public judge is a training signal; if its decision procedure is visible, a generator can shift outputs toward higher judged trust without improving correctness. We frame this as a two-player game in which the generator's gradient on the judge's score has a half-life equal to the period of shadow-judge rotation. Three defenses are explicit: (i) the longitudinal stratum of JudgeMeta is labeled *after* the calibration cutoff, so its signals are recoverable only by genuinely improving predictions; (ii) shadow-judge rotation refreshes thresholds, reliability priors, and Σ on a cadence faster than the adversary's adaptation horizon; (iii) the calibration audit trail makes the judge auditable but not prompt-visible.

9 Discussion, Limitations, and Threats to Validity

Five limitations bear naming. *Correlated pretraining*: even with channel-class diversity, verifier families share corpora and tool implementations, so N_{eff} is an upper bound on truly-independent evidence. *DS fit stability*: the hierarchical DS model is promoted only under a strict held-out gate that rarely fires below N_{cal}=5,000. *Calibration data*: production accuracy targets depend on labeled data acquisition (a public data plan); numbers reported here are Monte-Carlo or analytically derived. *Autoformalization cost*: Lean-based verification offers the highest certainty at minutes-to-hours me-

dian cost (Wu et al., 2022; Yang et al., 2023); we feature-flag Lean to an asynchronous late-stage path. *Aleatoric/epistemic identification*: BALD under heterogeneous LLM ensembles is approximate; we triangulate with semantic entropy and perturbation stability rather than claiming a clean decomposition.

The Monte-Carlo (§7.2) fixes a constant-ρ Gaussian channel covariance, the regime under which Eq. (5) is exact; under heterogeneous block-correlation or non-Gaussian tails, ESS retains a closed-form interpretation but loses exact Bayes-optimality, and our claims become bounds rather than equalities. The position-swap simulation (§7.3) treats juror noise as exchangeable, which is optimistic; correlated juror errors can drag verdict-level consistency below the 0.90 gate even at $K{=}5$. The agreement-invariant simulation (§7.4) assumes that head and S_λ residuals are jointly Gaussian; non-Gaussian heads (heavy-tailed under OOD) shift the trigger rate. The cryptographic auditability claims rest on standard ECDSA-style signatures and do not protect against compromise of the calibration corpus itself. Finally, the operator-supplied cost ratios are policy choices, not measurements: `policy_id` versioning makes them auditable but does not absolve the operator of choosing them well.

10 Conclusion

Jury-of-Verifiers reframes "LLM-as-a-judge" as routing over heterogeneous verifier classes fused in a calibrated Bayesian aggregator that explicitly models channel correlation. The aggregator is Bayes-optimal under symmetric correlation (proof + Monte-Carlo within 10^{-3}); the $\tau \tanh$ system-level cap pays a measured small calibration cost for unbounded-pile-on robustness; the agreement invariant fires within an analytically-bounded window under reference residuals; the dual conformal layers achieve nominal coverage in simulation. Together with hash-chained signed verdicts and `policy_id` versioning, the design renders trustworthiness claims falsifiable: every load-bearing claim is tied to verbatim evidence, every aggregate to learned thresholds, every label to a finite-sample coverage guarantee, every verdict to a signed, replayable provenance. With the forthcoming JudgeMeta benchmark and public calibration corpus, we aim to establish a reproducible, coverage-guaranteed evaluation standard for open-ended agent outputs.

References

Anastasios N. Angelopoulos and Stephen Bates. 2021. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*.

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. 2022. Learn then test: Calibrating predictive algorithms to achieve risk control. *arXiv preprint arXiv:2110.01052*.

Pier Giovanni Bissiri, Chris C. Holmes, and Stephen G. Walker. 2016. A general framework for updating belief distributions. *Journal of the Royal Statistical Society: Series B*, 78(5):1103–1130.

I-Chun Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, and Pengfei Liu. 2023. FacTool: Factuality detection in generative AI — a tool augmented framework for multi-task and multi-domain scenarios. *arXiv preprint arXiv:2307.13528*.

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

Stefan Depeweg, José Miguel Hernández-Lobato, Finale Doshi-Velez, and Steffen Udluft. 2018. Decomposition of uncertainty in Bayesian deep learning for efficient and risk-sensitive learning. In *Proceedings of the 35th International Conference on Machine Learning*.

Lee D. Erman, Frederick Hayes-Roth, Victor R. Lesser, and D. Raj Reddy. 1980. The HEARSAY-II speech-understanding system: Integrating knowledge to resolve uncertainty. *ACM Computing Surveys*, 12(2):213–253.

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

William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23:1–39.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. PAL: Program-aided language models. *arXiv preprint arXiv:2211.10435*.

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

Isaac Gibbs and Emmanuel J. Candès. 2021. Adaptive conformal inference under distribution shift. *Advances in Neural Information Processing Systems*.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. 2023. CRITIC: Large language models can self-correct with tool-interactive critiquing. *arXiv preprint arXiv:2305.11738*.

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

Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. 2011. Bayesian active learning for classification and preference learning. In *arXiv preprint arXiv:1112.5745*.

Geoffrey Irving, Paul Christiano, and Dario Amodei. 2018. AI safety via debate. *arXiv preprint arXiv:1805.00899*.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. 2024. LiveCodeBench: Holistic and contamination free evaluation of large language models for code. *arXiv preprint arXiv:2403.07974*.

Robert E. Kass and Adrian E. Raftery. 1995. Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795.

Akbir Khan, John Hughes, Dan Valentine, Laura Ruis, Kshitij Sachan, Ansh Radhakrishnan, Edward Grefenstette, Samuel R. Bowman, Tim Rocktäschel, and Ethan Perez. 2024. Debating with more persuasive LLMs leads to more truthful answers. *arXiv preprint arXiv:2402.06782*.

Ananya Kumar, Percy Liang, and Tengyu Ma. 2019. Verified uncertainty calibration. *Advances in Neural Information Processing Systems*.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamile Lukosiute, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, and 11 others. 2023. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. *arXiv preprint arXiv:2305.20050*.

Peter McCullagh. 1980. Regression models for ordinal data. *Journal of the Royal Statistical Society: Series B*, 42(2):109–142.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. *arXiv preprint arXiv:2305.14251*.

Bálint Mucsányi, Michael Kirchhof, and Seong Joon Oh. 2024. Benchmarking uncertainty disentanglement: Specialized uncertainties for specialized tasks. *arXiv preprint arXiv:2402.19460*.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. 2024. LLM evaluators recognize and favor their own generations. *arXiv preprint arXiv:2404.13076*.

Victor Quach, Adam Fisch, Tal Schuster, Adam Yala, Jae Ho Sohn, Tommi S. Jaakkola, and Regina Barzilay. 2024. Conformal language modeling. In *International Conference on Learning Representations (ICLR)*.

Yaniv Romano, Matteo Sesia, and Emmanuel J. Candès. 2020. Classification with valid and adaptive coverage. *Advances in Neural Information Processing Systems*.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. 2023. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*.

Edwin Simpson, Stephen Roberts, Ioannis Psorakis, and Arfon Smith. 2013. Dynamic Bayesian combination of multiple imperfect classifiers. *Decision Making and Imperfection*.

Sijun Tan, Siyuan Zhuang, Kyle Montgomery, William Y. Tang, Alejandro Cuadron, Chenguang Wang, Raluca Ada Popa, and Ion Stoica. 2024. JudgeBench: A benchmark for evaluating LLM-based judges. *arXiv preprint arXiv:2410.12784*.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *arXiv preprint arXiv:2305.04388*.

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

Pat Verga, Sebastian Hofstatter, Sophia Althammer, Yixuan Su, Aleksandra Piktus, Arkady Arkhangorodsky, Minjie Xu, Naomi White, and Patrick Lewis. 2024. Replacing judges with juries: Evaluating LLM generations with a panel of diverse models. *arXiv preprint arXiv:2404.18796*.

Junlin Wang, Jue Wang, Ben Athiwaratkun, Ce Zhang, and James Zou. 2024. Mixture-of-agents enhances large language model capabilities. *arXiv preprint arXiv:2406.04692*.

Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Jie Huang, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, Cosmo Du, and Quoc V. Le. 2024. Long-form factuality in large language models. *arXiv preprint arXiv:2403.18802*.

Lisa Wimmer, Yusuf Sale, Hang Nguyen, and Eyke Hüllermeier. 2023. Quantifying aleatoric and epistemic uncertainty in machine learning: Are conditional entropy and mutual information appropriate measures? *Uncertainty in Artificial Intelligence*.

Yuhuai Wu, Albert Q. Jiang, Wenda Li, Markus N. Rabe, Charles Staats, Mateja Jamnik, and Christian Szegedy. 2022. Autoformalization with large language models. *Advances in Neural Information Processing Systems*.

Kaiyu Yang, Aidan M. Swope, Alex Gu, Rahul Chalamala, Peiyang Song, Shixing Yu, Saad Godil, Ryan Prenger, and Anima Anandkumar. 2023. LeanDojo: Theorem proving with retrieval-augmented language models. *Advances in Neural Information Processing Systems*.

Chujie Zheng, Zhenru Zhang, Beichen Zhang, Runji Lin, Keming Lu, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. 2024. ProcessBench: Identifying process errors in mathematical reasoning. *arXiv preprint arXiv:2412.06559*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*.

Barret Zoph, Irwan Bello, Sameer Kumar, Nan Du, Yanping Huang, Jeff Dean, Noam Shazeer, and William Fedus. 2022. ST-MoE: Designing stable and transferable sparse expert models. *arXiv preprint arXiv:2202.08906*.

A Deployment Guarantees and Calibration Pipeline

Seven non-negotiable guarantees: (i) the evidence ledger is the source of truth; prose is rendered *from* the claim tree. (ii) Every numeric threshold is learned and signature-bound. (iii) Both conformal layers are always emitted; missing calibration produces `abstention.reason=calibration_missing`. (iv) The shadow adaptive-conformal path always runs. (v) Circuit breakers degrade rather than fail. (vi) No network at verify time. (vii) Verdicts are hash-chained and elliptic-curve-signed under a rotated key with a 2-window verification policy.

The closed-enum abstention taxonomy lists nine reasons: `rejection_band_posterior`, `low_effective_channels`, `ood_detected`, `structural_disagreement`, `low_self_consistency`, `motivated_reasoning_detected`, `budget_exhausted`, `verifier_unavailable`, `calibration_missing`.

The end-to-end calibration procedure runs twelve invariant-gated steps: load labeled items; $70/15/15$ split; fit the ordinal head subject to invariants on threshold monotonicity, proper probabilities, conformal coverage to within ± 2pp, and class-wise debiased ECE ≤ 0.05; fit reliability priors per source type; learn hard-gate thresholds by cost-weighted selective risk; fit DS only if all three of its promotion gates pass on every CV fold; fit the NLI threshold; fit the semantic-entropy abstention threshold; compute conformal quantile tables at $\alpha \in \{0.05, 0.10, 0.15, 0.20\}$; write the calibration-artifact identifier as the SHA-256 of the input corpus plus code version, and the policy identifier as the SHA-256 of the cost ratios and tolerance constants; refuse promotion on any invariant failure; sign the bundle with a rotated key. Five canary fixtures (a cold-start floor; the next-stage target is at least two hundred) exercise each label class and the abstention taxonomy. Per-profile service-level budgets at release: quick-screen $p_{50} \leq 5$ s, $p_{95} \leq 20$ s; standard-research $p_{50} \leq 45$ s, $p_{95} \leq 180$ s; engineering $p_{50} \leq 60$ s, $p_{95} \leq 300$ s. Release gates: planted-flaw recall ≥ 0.85 at FPR ≤ 0.05 per class; position-swap consistency ≥ 0.90; debiased ECE ≤ 0.05; coverage within ± 2pp of nominal; 100% replay determinism on deterministic fields.

B Derivation of Bayes-Optimal Correlated Log-LR

Under the canonical model where each channel emits $\ell_i \sim \mathcal{N}(\mu, \tau_0^2)$ with uniform pairwise correlation ρ, the covariance matrix is $\Sigma = \tau_0^2[(1-\rho)I + \rho J]$. The sufficient statistic for the hypothesis H is the weighted sum $\sum_i \ell_i$, whose variance is $\mathrm{Var}[\sum_i \ell_i] = k\tau_0^2(1 + (k-1)\rho)$. Under the well-calibrated log-LR identity $\tau_0^2 = 2\mu$ (the convention used in §7.2), the posterior log-LR equals

$$\log \mathrm{LR}_{\text{Bayes}} = \frac{k\,\mu}{1 + (k-1)\rho},$$

yielding Eq. (5). The naive sum $k\mu$ is overconfident by exactly the factor $1 + (k-1)\rho$, which our N_{eff} scaling (Eq. 1) corrects since $N_{\text{eff}} = k/(1 + (k-1)\rho)$ under constant off-diagonal correlation.