VEA-AWARE UNCERTAINTY FOR LLM CHAIN OF ACTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

For the prediction of a binary outcome, Variability–Epistemic–Aleatoric (VEA) uncertainty quantification separates the error of a single model-reported probability into three components: the *variability* (V) of the prediction across repeated runs of the model, the *epistemic* gap (E) between the average prediction and the true conditional probability of the outcome, and the *aleatoric* uncertainty (A) of the outcome itself, which no predictor can impact. This decomposition presents an important distinction that calls for different strategies to improve model performance: averaging over repeated runs of the model of interest reduces variability, and better modeling choices and training procedures reduce the epistemic gap while reporting the intrinsic aleatoric error. We propose an estimation procedure of VEA components in the setting of predictions through a *chain of actions*, namely a single structured generation mechanism that emits an ordered sequence of intermediate judgments before its final action. First, when the true conditional probability is unobserved, we estimate the epistemic gap (E) against a disclosed proxy model, and we provide an aggregate diagnostic that checks the adequacy of that proxy using the observed outcomes. Second, we further decompose the variability (V) into an exact sum of individual action-specific contributions so that action-level instability can be identified and targeted. We apply our VEA framework to evaluate the replicability of published scientific findings, an important but resource-intensive task in practice, using three core benchmark replication datasets (RPP, CB and SSRP). Across 100 runs per study, the VEA decomposition shows that the overall prediction error is dominated by the epistemic component (E) and is concentrated in a small number of specific actions along the chain. This implies that increasing the reasoning effort of the LLM under evaluation improves prediction accuracy, whereas lowering the sampling temperature does not, because the already small variability (V) has no room for further shrinkage.

1 INTRODUCTION

Large language models (LLMs) are increasingly used to produce probability judgments that inform real decisions, but uncertainty quantification (UQ) for these judgments remains an open problem (Kadavath et al., 2022; Lin et al., 2022; Kuhn et al., 2023; Farquhar et al., 2024). A model reports a single uncertainty-like value that may mix several sources of uncertainty, and addressing each source requires a different intervention. What makes UQ even harder is that the model is involved in a *chain of actions (CoA)* to deliver judgment. CoA is a single structured generation based on an ordered sequence of intermediate judgments prior to its final output, so one must know which action contributes which type of uncertainty. The primary objective of this paper is to establish a rigorous UQ framework for CoA-based predictions produced via an LLM.

In practice, instead of relying on a single probability for UQ, we advocate three source-specific UQs. First, re-running the same LLM on the same task can produce a distribution of outcome-predicted probabilities with its spread reflecting the model's *variability* across runs. Second, the mean of this distribution may systematically differ from the task's true replication probability. Such discrepancy leads to *epistemic* uncertainty. Third, even a perfectly calibrated predictor cannot eliminate the irreducible randomness of the replication outcome: even a faithful replication of the same study may succeed or fail by chance. This is *aleatoric* uncertainty. When evaluating an LLM's prediction for a task, these three UQs are complementary and should be reported separately for a fully informed UQ.

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.

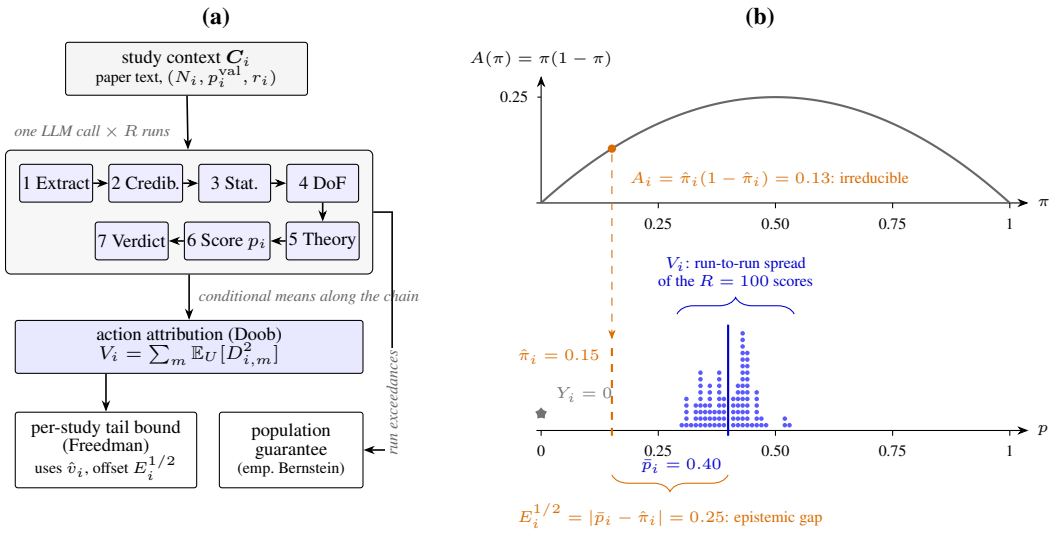

Figure 1: Overview of the chain-of-actions (CoA) and VEA decomposition framework. **(a)** For each study, a structured prompt generates an ordered chain of auditable actions at each run. The conditional means along the generation order form a Doob martingale, which decomposes the run-to-run variance exactly by action; the attribution enters a per-study Freedman tail bound, and per-run exceedances yield a distribution-free empirical-Bernstein population guarantee. **(b)** VEA decomposition for one RPP study ("Where do we look during potentially offensive behavior?," which failed to replicate). Top: the aleatoric variance $A(\pi) = \pi(1-\pi)$; the fitted proxy $\hat{\pi}_i = 0.15$ gives $A_i = 0.13$, irreducible relative to the proxy. Bottom: the same $\hat{\pi}_i$ overlaid on the $R = 100$ LLM scores $p_i^{(r)}$, stacked as dots at 0.01 resolution. The spread of the scores is V_i, the squared distance between their mean $\bar{p}_i$ and $\hat{\pi}_i$ is E_i, and the star marks the observed outcome Y_i.

For convenience, we abbreviate the three types of uncertainties together as VEA, a view of LLM uncertainty first proposed in a recent discussion (Anonymous, 2026). The prediction is typically generated from a designed CoA in which the final verdicts are generated along with a structured sequence of auditable sub-outputs, such as extraction, statistical review, and scoring. This structure motivates decomposing UQ along the CoA, locating which action contributes which type of uncertainty.

The VEA framework in the example of replicability prediction is explained pictorially in Figure 1, including two new key contributions. **(C1) A VEA estimation toolbox under an unknown true probability.** In Section 3, we first decompose the squared prediction risk of task i into the three-component sum $V_i + E_i + A_i$, and then give componentwise unbiased estimation from R repeated runs of the replication success prediction model. To address the issue that the true conditional probability π_i is unknown, we develop a three-tier estimation procedure in Section 3.2. (i) The variability V_i is estimated from the repeated runs alone. (ii) The epistemic UQ E_i is estimated by substituting a disclosed proxy $\hat{\pi}_i$ for the unknown π_i. When $\hat{\pi}_i$ is poorly estimated, a large E_i also absorbs error in the proxy, so we check the adequacy of $\hat{\pi}_i$ by assessing the agreement between the observed squared prediction error and the sum of the three estimated components averaged across studies (Tier 3 of Section 3.2). **(C2) An exact action attribution of variability with bounds at two levels.** Let $h_{i,m}$ denote the conditional mean of the final score of study i given the first m actions of one run. As conditional expectations of the same score under increasing information, the $h_{i,m}$ form a Doob martingale (Doob, 1953), and the total variability splits exactly as $V_i = \sum_m \mathbb{E}_U[(h_{i,m} - h_{i,m-1})^2]$, with no assumption on how the model uses its own intermediate outputs. These per-action contributions yield bounds at two levels. First, for a single study i, a Freedman-type bound controls the probability $\Pr(|p_i - \pi_i| > \tau)$ that one additional run deviates from the true probability by more than a tolerance τ, with the estimated per-action variances in its dominant term. Second, an empirical-Bernstein bound gives a distribution-free, finite-sample upper confidence bound on the proportion of studies whose runs exceed τ. In addition, we develop a cross-fitted estimator $\hat{\theta}_m$ of the

average contribution of action m, built on a Neyman-orthogonal score, which satisfies $\sum_m \hat{\theta}_m = \hat{V}$ exactly for any fitted nuisance functions.

We examine the VEA UQ framework on zero-shot prediction of the replicability of published scientific findings. Direct real-world experiment-based replication is costly, slow, and often unrealistic. Therefore, a growing literature predicts replicability using the original study records through prediction markets (Dreber et al., 2015), supervised models on textual and statistical features (Altmejd et al., 2019; Yang et al., 2020), and most recently large language models. A predicted replication probability $p_i \in [0, 1]$ for study i is interpretable, comparable, and inexpensive at scale, if the prediction is trustworthy. We use three replication datasets, the Reproducibility Project: Psychology (RPP; Collaboration, 2015), the Reproducibility Project: Cancer Biology (CB; Errington et al., 2021), and the Social Sciences Replication Project (SSRP; Camerer et al., 2018), with $R = 100$ structured CoA runs per effect-level replicability claim under a reasoning-effort $\times$ temperature grid (Section 5). Our analysis yields three main findings. (i) The VEA UQ framework can check the adequacy of the proxy replicability probability from the observed outcomes alone. The aggregate diagnostic in Section 3.2 confirms the adequacy of the proxy on RPP and SSRP and detects its failure on CB using only the observed outcomes; on CB the zero-shot LLM is also markedly more accurate than the failed proxy (AUC 0.704 versus 0.581), so the framework directs improvement effort at the proxy rather than at the LLM. (ii) The uncertainty in LLM's zero-shot prediction is dominated by the epistemic component and is concentrated in specific CoA actions. On CB, the extraction action contributes most strongly, whereas on RPP, the dominant contributions arise from the verdict and researcher-degrees-of-freedom actions. Run-to-run variability accounts for a median of 5% of the reducible uncertainty, defined as the sum of the variability and epistemic components. (iii) The decomposition suggests which interventions improve accuracy and at which actions of the chain they act. Raising the reasoning effort of the model acts on the dominant epistemic component and improves the AUC; lowering the sampling temperature acts on the already small variability component and leaves the AUC essentially unchanged. To sum up, the decomposition locates the source of the prediction error and identifies the intervention that reduces it.

2 Related Work

Replicability prediction. Prior work spans prediction markets and expert surveys (Dreber et al., 2015), supervised models on statistical and textual features (Altmejd et al., 2019; Yang et al., 2020), and automated consistency checks (Nuijten & Polanin, 2020). RPP (Collaboration, 2015), CB (Errington et al., 2021), and SSRP (Camerer et al., 2018) supply labeled datasets of direct replications. The replication success is assessed and quantified as a single number per study. As the single label is unable to separate the instability of the predictor from its calibration gap or from the intrinsic difficulty of the claim, the VEA framework fills the gap.

Uncertainty quantification for LLMs. White-box methods read internal signals or elicit verbalized confidence (Kadavath et al., 2022; Lin et al., 2022); black-box methods assess stability across stochastic re-runs (Wang et al., 2022; Manakul et al., 2023); calibration is a persistent concern (Guo et al., 2017), and recent work propagates token-level uncertainty upward (Duan et al., 2025). Instead of proposing another score, we show that run-to-run stability (our V), calibration against a context-conditional proxy (our E), and irreducible task noise (our A) are the three additive terms of one squared-error risk, so existing scores can be read as targeting one component or another.

Risk decompositions, concentration, and chains of actions. Squared-error risk decompositions have a long history, but their standard forms address the randomness of training data or of an estimator. In contrast, the randomness discussed in this paper is re-execution of a stochastic generator, and the epistemic term is measured against a context-conditional proxy that must itself be estimated and validated (Section 3.2). Our contribution is the estimation theory for these components, together with the action-wise attribution and bounding of the variance term in multi-step generation, based on martingale concentration (Hoeffding, 1963; Freedman, 1975) and the empirical Bernstein inequality (Maurer & Pontil, 2009). This focus on action-level uncertainty is motivated by the growing use of chains and agentic scaffolds in LLM systems (Yao et al., 2022; Shinn et al., 2023; Hong et al., 2024), where reliability across intermediate steps remains an active concern (Zhou et al., 2024; Ferrag et al., 2026). We focus on the single-LLM setting, where the ordered sub-outputs of one structured generation define the actions of the chain.

3 THE VEA DECOMPOSITION

Let $Y_i \in \{0, 1\}$ denote the eventual replication outcome of study i under a pre-specified protocol, and let $\pi_i = \Pr\left(Y_i = 1 \mid \boldsymbol{C}_i\right)$ be its conditional probability given the study context $\boldsymbol{C}_i$ (the original record the model consumes). Let $p_i^{(r)} \in [0, 1]$ be the predicted probability returned by a stochastic LLM on its r-th execution. Each execution is an independent realization $U^{(r)}$ of the model-side randomness U (decoding temperature, sampling seeds, nondeterministic inference), so $p_i = p_i(U)$ is random through U and the R executions are i.i.d. copies. Write $\bar{p}_i = \mathbb{E}_U[p_i]$ for the mean prediction and we put notation conventions for population and sample operators in Appendix A.

Our starting point is the VEA decomposition, which specializes a general foundation decomposition of the prediction risk (Anonymous, 2026) to a binary outcome: under conditional independence of Y_i and U given study i,

$$\mathbb{E}_{U,Y_i|i}\left[(p_i - Y_i)^2\right] = \underbrace{\mathbb{E}_U\left[(p_i - \bar{p}_i)^2\right]}_{V_i \text{ (variability)}} + \underbrace{(\bar{p}_i - \pi_i)^2}_{E_i \text{ (epistemic)}} + \underbrace{\pi_i(1 - \pi_i)}_{A_i \text{ (aleatoric)}}. \tag{1}$$

Measuring the model against the true probability π_i instead of the outcome Y_i removes the aleatoric term: $\mathbb{E}_U[(p_i - \pi_i)^2] = V_i + E_i$ (derivation in Appendix A).

The sum of the first two components of the decomposition, $V_i + E_i = \mathbb{E}_U[(p_i - \pi_i)^2]$, is the quantity our analysis estimates, because a replication outcome is observed only once per study while the true probability π_i is the object the model is asked to match; it is also the event the bounds of Section 4 control. The renaming emphasizes the interventions each term admits: V_i is reducible by repetition (self-consistency, majority voting) or decoding changes; E_i is reducible by recalibration or by improving the model's reasoning; A_i is irreducible.

3.1 ESTIMATORS

Given R repeated executions, the plug-in estimators are

$$\hat{\bar{p}}_i = R^{-1}\sum_{r=1}^{R} p_i^{(r)}, \qquad \hat{V}_i = (R-1)^{-1}\sum_{r=1}^{R}(p_i^{(r)} - \hat{\bar{p}}_i)^2, \qquad \hat{E}_i = (\hat{\bar{p}}_i - \hat{\pi}_i)^2,$$

where $\hat{\pi}_i$ is an external proxy estimate (Section 3.2).

Proposition 3.1 (Finite-sample behavior). *Fix the proxy $\hat{\pi}_i \in [0, 1]$. If the $R \geq 2$ executions are i.i.d. copies of p_i, then $\hat{\bar{p}}_i$ and $\hat{V}_i$ are unbiased for $\bar{p}_i$ and V_i, and $\mathbb{E}[\hat{E}_i] = (\bar{p}_i - \hat{\pi}_i)^2 + V_i/R$, so the corrected $\tilde{E}_i = \hat{E}_i - \hat{V}_i/R$ is exactly unbiased for the plug-in epistemic term $(\bar{p}_i - \hat{\pi}_i)^2$. As $R \to \infty$ with $\hat{\pi}_i$ fixed, $\hat{V}_i \xrightarrow{p} V_i$ and $\tilde{E}_i \xrightarrow{p} (\bar{p}_i - \hat{\pi}_i)^2$.*

The proposition isolates the sensitivity structure: run-to-run variability is estimated robustly regardless of the proxy, whereas the epistemic and aleatoric terms both contain $\hat{\pi}_i$ and inherit any error in it. $\tilde{E}_i$ can be negative in finite samples, as is familiar from variance-component estimation. We retain the negative values in all aggregate calculations, because truncating them at zero would reintroduce the upward bias that the correction removes.

3.2 UNCERTAINTY QUANTIFICATION WHEN π_i IS UNKNOWN

The truth π_i is never observed, so a natural objection is that E_i and A_i are not identifiable at the study level. The framework is designed around this fact in three tiers.

Tier 1: V_i needs no proxy and no labels. The variability term is a pure re-execution functional of the model; $\hat{V}_i$ is unbiased whatever π_i is, and the entire action attribution of Section 4 decomposes V_i alone. All Tier-1 quantities are available for any deployed pipeline without labels.

Tier 2: E_i is proxy-relative, with the proxy disclosed. We estimate a working proxy $\hat{\pi}_i$ from anonymized paper context using a fixed embedding model (`text-embedding-3-large`) that maps $\boldsymbol{C}_i$ to $\boldsymbol{\phi}(\boldsymbol{C}_i) \in \mathbb{R}^d$ with d=6144, and we fit the ℓ_1-penalized logistic working model

$$\hat{\pi}_i = \text{logit}^{-1}\left\{\hat{\alpha} + \boldsymbol{\phi}(\boldsymbol{C}_i)^\top \hat{\boldsymbol{\beta}}\right\},$$

Table 1: What each quantity requires and what applies to a new study at deployment time.

Quantity	Labels needed	Scope	At a new study
V_i, action attribution	none	per instance	computed directly
E_i (proxy-relative)	dataset, to fit $\hat{\pi}$	per instance	via frozen proxy, under exchangeability
A_i (plug-in)	dataset, to fit $\hat{\pi}$	aggregate-validated	$\hat{\pi}(1-\hat{\pi})$ from frozen proxy
Bernstein rate	calibration dataset	population	finite-sample coverage

with nested study-level cross-validation, selecting the penalty and active-set on the inner folds and generating predictions from the outer folds. The embeddings use the same prompt as the zero-shot experiments for comparability. $\hat{\pi}_i$ is a working proxy for π_i, and every reported E_i is explicitly evaluated against this proxy.

Tier 3: an aggregate diagnostic checks the proxy against the observed outcomes. Averaging the left-hand side of the decomposition (1), $\mathbb{E}[(p_i - Y_i)^2]$, over studies gives a diagnostic that uses only the single-draw outcomes:

$$n^{-1}\sum_i R^{-1}\sum_r (p_i^{(r)} - Y_i)^2 \;-\; n^{-1}\sum_i (\hat{V}_i + \tilde{E}_i)$$

The observable aggregate estimates the mean aleatoric term $n^{-1}\sum_i \pi_i(1-\pi_i)$, and agrees with the plug-in $n^{-1}\sum_i \hat{\pi}_i(1-\hat{\pi}_i)$ only if the proxy is well calibrated in aggregate (Appendix A). We report this agreement as a per-dataset calibration check on $\hat{\pi}$. Across all four decoding settings, the aggregate estimate is below the plug-in estimate in every dataset, modestly on RPP but more substantially on CB and SSRP, indicating that the plug-in slightly overestimates the mean aleatoric uncertainty in the latter two datasets. The check is informative in practice. On CB the embedding proxy is a weak predictor of the outcomes (AUC 0.58, Section 5), so a large E_i there measures disagreement between two imperfect predictors rather than error of the LLM alone; indeed, the LLM predicts the CB outcomes better than the proxy does. On RPP and SSRP the proxy passes the check, and E_i can be read as the epistemic error of the LLM. An aleatoric term $\hat{\pi}_i(1-\hat{\pi}_i)$ is reported only where the proxy is a genuine probability of a defined outcome, not for agreement, power, or plausibility scores (Appendix B).

Audit mode vs. deployment mode. The three components have different epistemic statuses, and we report them accordingly (Table 1). V_i, together with its entire action attribution (Section 4), is per-instance, label-free uncertainty computed alongside the prediction for any new study. Because E_i and A_i are not identifiable per study, they operate in *audit mode*, estimated against a disclosed proxy on a labeled dataset and validated in aggregate (Tier 3). At deployment the frozen proxy applies to a new study directly (embed the new context, apply the fitted model), so a proxy-relative $\hat{E}$ can be computed for new studies under an exchangeability assumption, whose failure the diagnostic detects (CB, Section 5.2). The only deployment-time statement with a finite-sample guarantee is the empirical-Bernstein bound on the population failure rate (Section 4.1).

4 Attributing and Bounding Variability along the Chain of Action

For each study i the structured prompt (Appendix B) elicits, inside one LLM call, an ordered sequence of auditable actions indexed $m = 1, \ldots, M$: extraction of basic statistics (sample size, reported p-value, effect size), credibility coding, statistical review, researcher-degrees-of-freedom coding, theory and context, scoring, and verdict. Across R executions this yields $p_{i,m}^{(r)} \in [0,1]$, with the score emitted at action s, so $p_i = p_{i,s}$; each action has a proxy $\hat{\pi}_{i,m}$ (Table 3, Appendix B) with marginal components $V_{i,m}$ and $E_{i,m}$ defined as in Section 3 and the action-level Markov bound $\Pr\left(|p_{i,m} - \hat{\pi}_{i,m}| > \tau_m\right) \le (V_{i,m} + E_{i,m})/\tau_m^2$.

We analyze how action errors compose into the final score by conditioning along the model's own chain of actions. Let $\mathcal{F}_{i,m} = \sigma(\boldsymbol{X}_i, p_{i,1}, \ldots, p_{i,m})$ be the filtration of the chain up to action m, with $\mathcal{F}_{i,0} = \sigma(\boldsymbol{X}_i)$.

Proposition 4.1 (Chain-of-actions decomposition of variability)**.** *Let $h_{i,m} = \mathbb{E}_U[p_i \mid \mathcal{F}_{i,m}]$ for $m = 0, 1, \ldots, s$, so $h_{i,0} = \bar{p}_i$ and $h_{i,s} = p_i$. Then $D_{i,m} = h_{i,m} - h_{i,m-1}$ is a martingale-difference sequence,*

$$p_i - \bar{p}_i = \sum_{m=1}^{s} D_{i,m}, \qquad \mathbb{E}_U[D_{i,m} D_{i,\ell}] = 0 \ (m \neq \ell), \qquad V_i = \sum_{m=1}^{s} \mathbb{E}_U\big[D_{i,m}^2\big].$$

The sequence $(h_{i,m})_m$ is the *Doob martingale* of the final score with respect to the chain-of-actions filtration, and the decomposition is exact for any square-integrable score and any nested ordering (proof in Appendix A). Each $\mathbb{E}_U[D_{i,m}^2]$ is the run-to-run variance the final score inherits from action m given its predecessors, so the profile ranks actions by injected variance and indicates where repetition or decoding interventions act. The last increment $D_{i,s} = p_i - h_{i,s-1}$ is the part of the score that the earlier sub-outputs do not predict. The sizes of the increments then show where the score actually comes from: if the intermediate judgments determine the score, the upstream increments are large, and if the score fluctuates independently of them, most of the variance falls on $D_{i,s}$. Adding the epistemic gap gives the task-level risk $\mathbb{E}_U[(p_i - \pi_i)^2] = \sum_m \mathbb{E}_U[D_{i,m}^2] + (\bar{p}_i - \pi_i)^2$.

4.1 Tail Bounds at Two Levels

Since both conditional means lie in $[0, 1]$, every increment satisfies $|D_{i,m}| \leq 1$, and no smaller bound holds unless the sub-outputs are genuinely confined to a narrower range of values. The Azuma–Hoeffding inequality (Hoeffding, 1963), which bounds martingale deviations using only the worst-case range of the increments, is therefore *vacuous* with this assumption-free constant over the attainable tolerance range $\tau \leq 1$. The strategy is to move the estimable variance, not the worst-case range, into the dominant term. Write $u = \tau - E_i^{1/2}$ for the variability budget after the epistemic offset $E_i^{1/2} = |\bar{p}_i - \pi_i|$.

Proposition 4.2 (Variance-based per-study tail; Freedman)**.** *Suppose $|D_{i,m}| \leq c_i$ a.s. for all m, and that the predictable quadratic variation satisfies*

$$W_{i,s} = \textstyle\sum_m \mathbb{E}_U[D_{i,m}^2 \mid \mathcal{F}_{i,m-1}] \leq v_i \quad \text{a.s.}$$

Then for $\tau > E_i^{1/2}$,

$$\Pr\left(|p_i - \pi_i| > \tau\right) \ \leq \ 2\exp\Big(-\tfrac{u^2}{2(v_i + \frac{1}{3}c_i u)}\Big).$$

The range enters only the lower-order term, so the always-valid $c_i = 1$ is kept while the dominant term is the estimable variance $v_i = \sum_m \mathbb{E}_U[D_{i,m}^2]$, the action attribution of Proposition 4.1.

Corollary 4.1 (Population-level guarantee; empirical Bernstein)**.** *Let $B_i(\tau) = R^{-1} \sum_r \mathbb{I}\{|p_i^{(r)} - \hat{\pi}_i| > \tau\} \in [0, 1]$ over n independent studies, with sample mean $\bar{B}(\tau)$ and variance $\hat{\sigma}_B^2(\tau)$. Then with probability at least $1 - \delta$, distribution-free,*

$$\pi_{\text{fail}}(\tau) \ = \ \mathbb{E}[B_i(\tau)] \ \leq \ \bar{B}(\tau) + \sqrt{\tfrac{2\hat{\sigma}_B^2(\tau)\ln(2/\delta)}{n}} + \tfrac{7\ln(2/\delta)}{3(n-1)}.$$

The two bounds are the two levels of one guarantee, linked by the law of total expectation: $\pi_{\text{fail}}(\tau) = \mathbb{E}_i[\Pr_U(|p_i - \pi_i| > \tau)]$, so the population failure rate is the average over studies of the per-study failure probability. The Freedman bound controls the inner, per-study probability, and the empirical-Bernstein bound controls the outer average over studies. Both are stated against π_i or its proxy, not the one-shot outcome Y_i; replacing π_i by Y_i breaks the tower because the event then mixes model randomness with a single outcome draw (Appendix A). In practice v_i is replaced by a cross-fitted estimate $\hat{v}_i = \sum_m \hat{\theta}_m$ (Appendix C), making the per-study statement an *estimated* bound reported with study-cluster bootstrap uncertainty, while the empirical-Bernstein bound needs no estimation at all. The two bounds from Freedman and from empirical Bernstein are evaluated empirically in Section 5.5. Our bounds are related to split conformal prediction, the most widely used method for distribution-free uncertainty quantification (Vovk et al., 2005; Lei et al., 2018; Angelopoulos & Bates, 2023). Both are valid in finite samples without distributional assumptions, and both are built from the same scores $|p_i^{(r)} - \hat{\pi}_i|$. They differ in what is fixed and what is solved for. Split conformal prediction fixes a coverage level $1 - \alpha$ and uses a held-out calibration set of studies to find a radius q such that a new study's score falls within q of $\hat{\pi}_i$ with probability at least $1 - \alpha$. We instead fix the tolerance τ and bound the proportion of studies whose scores exceed τ, because this proportion is exactly what our empirical evaluation in Section 5.5 measures.

Table 2: Replication prediction across the three datasets, $R = 100$. AUC $\pm$ Hanley–McNeil SE; accuracy at threshold $0.5 \pm$ Wald SE; Brier is the study-level mean squared error at the default cell, with base-rate benchmarks and bootstrap CIs in Table 9; temperature 0.7 (the temperature-0.2 columns, essentially identical, are in Appendix E). The supervised proxy uses nested study-level CV (20 seeds averaged) and has no effort setting; LLM predictors are zero-shot. The low-effort column is the E-targeting ablation of Section 5.4. RPP outcomes follow the canonical OSF export.

		AUC				
Dataset	**Predictor**	high effort	low effort	**Accuracy**	**Brier**	n
RPP	Text-embedding LR (supervised)	0.768 ± 0.055		0.722 ± 0.047	0.185	90
	LLM vote rate $\bar{v}$ (zero-shot)	0.665 ± 0.062	0.638 ± 0.063	0.678 ± 0.049	0.267	90
	LLM confidence $\bar{q}$ (zero-shot)	0.625 ± 0.063	0.610 ± 0.063	0.678 ± 0.049	0.229	90
CB	Text-embedding LR (supervised)	0.581 ± 0.046		0.589 ± 0.039	0.287	158
	LLM vote rate $\bar{v}$ (zero-shot)	0.704 ± 0.042	0.628 ± 0.045	0.684 ± 0.037	0.248	158
	LLM confidence $\bar{q}$ (zero-shot)	0.698 ± 0.043	0.594 ± 0.046	0.696 ± 0.037	0.219	158
SSRP	Text-embedding LR (supervised)	0.846 ± 0.085		0.762 ± 0.093	0.199	21
	LLM vote rate $\bar{v}$ (zero-shot)	0.827 ± 0.090	0.870 ± 0.078	0.714 ± 0.099	0.229	21
	LLM confidence $\bar{q}$ (zero-shot)	0.788 ± 0.100	0.856 ± 0.083	0.714 ± 0.099	0.220	21

5 Empirical Study

5.1 Setup

Datasets. We use three direct-replication datasets with observed outcomes: **RPP** (Collaboration, 2015), comprising 100 psychology replications, of which 90 effect-level claims from 88 papers met our inclusion criteria of having non-null effect sizes and reported p-values; **CB** (Errington et al., 2021), comprising 158 effect-level claims from preclinical cancer biology; and **SSRP** (Camerer et al., 2018), comprising 21 social-science experiments published in *Nature* or *Science*. Detailed sample accounting is provided in Appendix B.

Pipeline. For each claim, we obtain $R = 100$ zero-shot chain-of-actions responses from the reasoning model `gpt-5.4-mini` through the OpenAI API. We used a 2×2 design that varied reasoning effort $\{\text{high}, \text{low}\}$ and temperature $\{0.7, 0.2\}$. The default setting was high reasoning effort with temperature 0.7. Each response followed the ordered actions defined in Section 4 and used a fixed ten-line schema. The prompt template and actions structuring are listed in Table 3 in Appendix B. For each study, we constructed two summary predictors. The verbalized confidence is computed as $\bar{q}_i = R^{-1} \sum_r p_i^{(r)}$, following Kadavath et al. (2022); Lin et al. (2022). The self-consistency vote rate is computed as $\bar{v}_i = R^{-1} \sum_r \mathbb{I}\{\text{verdict}_i^{(r)} = 1\}$, following Wang et al. (2022).

Proxy. The supervised text-embedding proxy $\hat{\pi}$ of Section 3.2, fit with nested study-level cross-validation, where every $\hat{\pi}_i$ is predicted from a model whose fitting and tuning both exclude study i, then averaged over 20 random fold splits. LLM predictors are zero-shot with no outcome labels.

5.2 Zero-shot prediction performance and the adequacy of the fitted proxy

We compare two LLM summary predictors with the embedding-based proxy for π in Table 2 under the default configuration of high reasoning effort and temperature 0.7. The proxy replicability predictor $\hat{\pi}$ was fitted using ℓ_1-regularized logistic regression. On RPP and SSRP, the embedding-based proxy is strong, with AUCs of 0.768 and 0.846, respectively, allowing the LLM to be evaluated against a well-performing estimate of π. The zero-shot vote rate achieves AUCs of 0.665 on RPP and 0.827 on SSRP, although the SSRP estimate is imprecise ($n = 21$). The pattern differs on CB, where the embedding-based proxy reaches an AUC at only 0.581 while the zero-shot LLM achieves 0.704 (Figure 2a), suggesting that the LLM captures predictive signal not fully recovered by the embedding-based model. When the proxy is strong, the epistemic term has a clear interpretation; when it is weak, as on CB, the diagnostic identifies the proxy itself as the source of the discrepancy. Across all configurations on RPP and CB, the vote rate achieves higher AUC than verbalized confidence but is less well calibrated. On RPP, its expected calibration error is 0.240, compared with

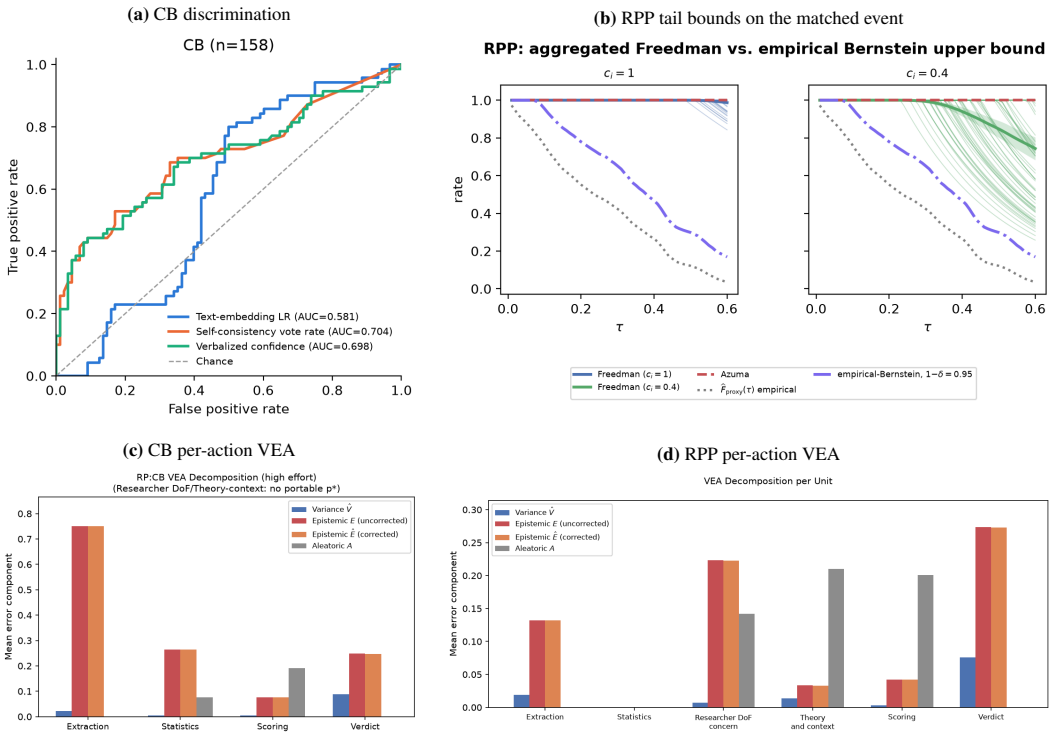

Figure 2: Empirical results at the default cell (high effort, temperature 0.7, $R = 100$). **(a)** On CB, the zero-shot predictors achieve AUCs of about 0.70, outperforming the trained embedding proxy at 0.58. **(b)** Range-based per-study bounds from Azuma–Hoeffding and Freedman with $c_i = 1$ are vacuous, while Freedman with the support restriction $c_i = 0.4$ becomes informative for $\tau \gtrsim 0.3$. The empirical-Bernstein population bound lies above and closely follows the empirical exceedance curve. **(c,d)** Per-action VEA shows uniformly small variability $\hat{V}$ and larger corrected epistemic gaps $\tilde{E}$ concentrated in identifiable actions, including extraction on CB and verdict and researcher-DoF coding on RPP. Aleatoric A is shown only for actions with a probabilistic proxy.

0.105 for verbalized confidence (Appendix D), illustrating the distinction between discrimination and calibration which will be confirmed by the VEA decomposition profile.

5.3 The VEA profile: systematic gaps, not sampling noise

The per-action VEA decomposition for RPP and CB under default configuration is presented in Figure 2c,d. **(i) Variability is small.** Action-level $\hat{V}$ ranges from 0.003 to 0.09. Across RPP claims, run-to-run variability accounts for a median of 5% of the reducible error $\sum_m (\hat{V}_{i,m} + \tilde{E}_{i,m})$, and 11% among claims with measurable variability. Thus, run-to-run variation contributes only a small share of the total reducible error. **(ii) The epistemic gap is concentrated in identifiable actions.** On CB, the extraction action has the largest gap, with $\tilde{E} = 0.75$, indicating frequent errors in extracting (N, p, r) from cancer-biology papers. Note that only 106 of 158 claims produced a scorable extraction match. On RPP, the largest gaps arise at the verdict action, with $\tilde{E} \approx 0.27$, and then in the coding of researcher degrees of freedom. **(iii) The scoring action is close to its proxy.** This action has the smallest epistemic gap, with $\tilde{E}$ approximately 0.08 on CB, 0.05 on RPP, and 0.03 on SSRP, and most of the remaining error is aleatoric. That is, the reported probability is close to the context-conditional proxy, and the residual error is mostly irreducible.

5.4 From decomposition to intervention

A useful decomposition should indicate which interventions are likely to improve performance. The 2×2 grid provides two such tests. First, changes in temperature primarily affect V, which accounts

for only about 5% of the reducible error in the full-grid results reported in Appendix E and suggests that temperature changes should have little effect on discrimination. Second, reasoning effort primarily affects the systematic component E through processes such as extraction and evidence evaluation. Lower reasoning effort should therefore reduce predictive performance. In particular, when reasoning effort is reduced from high to low, AUC of $\bar{v}$ on RPP is reduced from 0.665 to 0.638 and the AUC of $\bar{q}$ on CB from 0.698 to 0.594 (Table 2). A reversed ordering is observed in SSRP, possibly due to its small sample size. The action attribution further indicates where interventions may be most effective. On CB, it identifies extraction as a major source of error, suggesting that retrieval or structured parsing may be more useful than additional sampling. The action-level profile can also guide the allocation of a per-claim audit budget across actions according to $\tau^*_{i,m} \propto (\hat{V}_{i,m} + \tilde{E}_{i,m})^{1/3}$, as detailed in Appendix E. Finally, the aggregate diagnostic in Section 3.2 can assess the proxy itself. On CB, the diagnostic flags the embedding proxy as inadequate, separating efforts to improve $\hat{\pi}$ through larger datasets or better features from efforts to improve the LLM.

5.5 Tail bounds in practice

Figure 2b evaluates the guarantees on RPP using the matched event $\{|p^{(r)}_{i,s} - \hat{\pi}_i| > \tau\}$. CB and SSRP show qualitatively similar patterns (Appendix E). As Proposition 4.2 anticipates, Azuma–Hoeffding is vacuous throughout when using the distribution-free increment bound. Freedman's variance remains limited at $c_i = 1$ because the epistemic offset $E_i^{1/2}$ leaves little of the tolerance τ available for the variability term, particularly where empirical failures are concentrated at $\tau \lesssim 0.3$. With a support-restricted increment constant, however, Freedman becomes informative for $\tau \gtrsim 0.3$. The empirical-Bernstein bound is the only bound that is both valid and close to the empirical curve. We therefore use empirical Bernstein as the primary finite-sample guarantee and Freedman to translate the per-action attribution into a per-study guarantee.

6 Discussion

Summary. VEA supplements a reported probability with three separately interpretable statistics: how much the prediction varies across runs (V, label-free), how far its mean lies from the best context-conditional predictor available (E, proxy-relative), and how much error no predictor could remove (A). Only the first is claimed as per-instance, inference-time uncertainty; the latter two are audit-mode quantities validated in aggregate, with the empirical-Bernstein bound as the guarantee available at deployment (Table 1). The action attribution then identifies which action to modify and which intervention to apply, and Section 5.4 shows these implications hold on real interventions.

Limitations. First, "zero-shot" does not preclude pretraining contamination: the replication outcomes of all three datasets are published, and the model may have seen them. We therefore conducted an experiment to assess how accurately the zero-shot predictor could recall study replicability outcomes without knowing paper content (Appendix E). Five-hit recall rates were 2.2% for RPP, 29.7% for CB, and 19.0% for SSRP, suggesting greater potential contamination in CB and SSRP, consistent with those projects' greater online exposure. Stratifying prediction accuracy by probe recognition confirms the exposure matters: accuracy rises with the recognition count in all three datasets (Table 12). The CB margin over the trained proxy persists within every recognition stratum, including the unrecognized effects, but narrows there: contamination inflates the absolute numbers without fully explaining the CB advantage. Second, all results use one model family; the grid varies decoding and effort, not architecture. Third, the proxy $\hat{\pi}$ is an estimate, not an oracle; on CB it fails its own validation, and E there is a disagreement measure. Fourth, the orthogonal attribution estimator (Appendix C) attaches inference only to the study-averaged action contributions $\hat{\theta}_m$; the per-claim attribution profiles derived from the same fitted nuisances remain descriptive. The framework is descriptive: it does not improve the forecaster; it makes the forecaster auditable and indicates where improvement effort should be directed.

REPRODUCIBILITY STATEMENT

All predictions, proxies, decomposition outputs, and bound computations are produced by the companion repository[1]; stored prediction CSVs are the artifacts of record, and figures/tables are regenerated deterministically from them.

AI USE STATEMENT

In this work, we used generative AI tools (LLM assistants) in five ways. (i) *Research ideation and execution*: LLM assistants were used to discuss and refine the formulation of the VEA framework and the chain-of-actions attribution, and to assist in implementing and debugging the experimental pipeline. (ii) *Proving mathematical claims*: LLM assistance was used to check steps in the proofs of the propositions and corollary; every statement and proof was subsequently re-derived and verified manually by the authors. (iii) *Drafting*: LLM assistance was used to summarize the experimental results and to produce an initial draft of the Empirical Study section (Section 5). (iv) *Writing polish*: LLM assistants were used to edit the manuscript for readability and consistency. (v) *Code refactoring*: We wrote the original experiment scripts. LLM assistance was used to refactor them and to generate figures and tables from stored results, with all changes reviewed by the authors. We did not use generative AI to generate any data reported in this paper: the only LLM-generated outputs are the study-level predictions that constitute the object of study, produced by the evaluated model under the protocol of Section 5.1. The remaining subtasks in the ICLR disclosure list (e.g., translation, survey design, qualitative data analysis) are not applicable to this work. We reviewed all AI-assisted content: mathematical claims were checked independently of the assistant, AI-assisted code was tested and its outputs verified against the stored prediction files, and all numerical claims in the text were checked against the computed results. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Adam Altmejd, Anna Dreber, Eskil Forsell, Juergen Huber, Taisuke Imai, Magnus Johannesson, Michael Kirchler, Brian Nosek, and Thomas Pfeiffer. Predicting the replicability of social science lab experiments. *PLOS ONE*, 14(12):e0225826, 2019.

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

Anonymous. Comments on "an overview of large language models for statisticians". *The American Statistician*, 2026. To appear.

Colin F Camerer, Anna Dreber, Felix Holzmeister, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Gideon Nave, Brian A Nosek, Thomas Pfeiffer, et al. Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, 2(9):637–644, 2018.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018.

Open Science Collaboration. Estimating the reproducibility of psychological science. *Science*, 349 (6251):aac4716, 2015.

Joseph L Doob. *Stochastic Processes*. Wiley, New York, 1953.

Anna Dreber, Thomas Pfeiffer, Johan Almenberg, Siri Isaksson, Brad Wilson, Yiling Chen, Brian A Nosek, and Magnus Johannesson. Using prediction markets to estimate the reproducibility of scientific research. *Proceedings of the National Academy of Sciences*, 112(50):15343–15347, 2015.

[1] https://anonymous.4open.science/r/UQ-LLM-Replicability-47DB

Jinhao Duan, James Diffenderfer, Sandeep Madireddy, Tianlong Chen, Bhavya Kailkhura, and Kaidi Xu. Uprop: Investigating the uncertainty propagation of llms in multi-step agentic decision-making. *arXiv preprint arXiv:2506.17419*, 2025.

Timothy M Errington, Maya Mathur, Courtney K Soderberg, Alexandria Denis, Nicole Perfito, Elizabeth Iorns, and Brian A Nosek. Investigating the replicability of preclinical cancer biology. *eLife*, 10:e71601, 2021.

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

Mohamed Amine Ferrag, Norbert Tihanyi, and Merouane Debbah. From llm reasoning to autonomous ai agents: A comprehensive review. *IEEE Access*, 2026.

David A. Freedman. On tail probabilities for martingales. *The Annals of Probability*, 3(1):100–118, 1975.

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

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30, 1963.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Steven Yau, Zijuan Lin, Liyang Zhou, et al. Metagpt: Meta programming for a multi-agent collaborative framework. In *International Conference on Learning Representations*, volume 2024, pp. 23247–23275, 2024.

Saurabh Kadavath et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

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

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113 (523):1094–1111, 2018.

Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words. *arXiv preprint arXiv:2205.14334*, 2022.

Potsawee Manakul, Adian Liusie, and Mark Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896*, 2023.

Andreas Maurer and Massimiliano Pontil. Empirical bernstein bounds and sample variance penalization. In *Proceedings of the 22nd Conference on Learning Theory (COLT)*, 2009.

Michèle B Nuijten and Joshua R Polanin. "statcheck": Automatically detect statistical reporting inconsistencies to increase reproducibility of meta-analyses. *Research synthesis methods*, 11(5): 574–579, 2020.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in neural information processing systems*, 36:8634–8652, 2023.

Aad W van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.

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

Xuezhi Wang et al. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

Yang Yang, Wu Youyou, and Brian Uzzi. Estimating the deep replicability of scientific findings using human and artificial intelligence. *Proceedings of the National Academy of Sciences*, 117 (20):10762–10768, 2020.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022.

Zhanke Zhou, Rong Tao, Jianing Zhu, Yiwen Luo, Zengmao Wang, and Bo Han. Can language models perform robust reasoning in chain-of-thought prompting with noisy rationales? *Advances in Neural Information Processing Systems*, 37:123846–123910, 2024.

A NOTATION AND PROOFS

Notation. $\mathbb{E}$ always denotes a population expectation, its subscript naming the randomness averaged over ($\mathbb{E}_U$: model re-execution; $\mathbb{E}_i$: the draw of a study from the study population). No sample mean is written with a bare $\mathbb{E}$: empirical averages appear as explicit sums, and quantities computed from data carry hats or tildes ($\hat{V}_i$, $\tilde{E}_i$, $\hat{\theta}_m$), following the population/empirical separation of the empirical-process literature (van der Vaart, 1998).

Derivation of the decomposition (1). Decompose the error as $p_i - Y_i = (p_i - \bar{p}_i) + (\bar{p}_i - \pi_i) + (\pi_i - Y_i)$ and expand the square. The three squared terms have conditional expectations $\mathbb{E}_U[(p_i - \bar{p}_i)^2] = V_i$, $(\bar{p}_i - \pi_i)^2 = E_i$ (a constant given study i), and $\mathbb{E}_{Y_i}[(\pi_i - Y_i)^2] = \mathrm{Var}(Y_i) = \pi_i(1-\pi_i) = A_i$, the last because $Y_i \sim \mathrm{Bernoulli}(\pi_i)$ given the context. Each cross term vanishes: $\mathbb{E}_U[p_i - \bar{p}_i] = 0$ kills the two terms carrying that factor (for the term also carrying $\pi_i - Y_i$, conditional independence of U and Y_i given study i factorizes the expectation into a product with the zero factor), and $\mathbb{E}_{Y_i}[\pi_i - Y_i] = 0$ kills the remaining term. The second identity is the same expansion without the $(\pi_i - Y_i)$ term. □

Proof of Proposition 3.1. All execution-level expectations are over U. The sample-mean and $(R-1)$-normalized sample-variance identities give $\mathbb{E}[\hat{\bar{p}}_i] = \bar{p}_i$ and $\mathbb{E}[\hat{V}_i] = V_i$. For fixed $\hat{\pi}_i$,

$$\mathbb{E}[\hat{E}_i] = \mathbb{E}[(\hat{\bar{p}}_i - \hat{\pi}_i)^2] = \mathrm{Var}(\hat{\bar{p}}_i) + (\bar{p}_i - \hat{\pi}_i)^2 = V_i/R + (\bar{p}_i - \hat{\pi}_i)^2,$$

and subtracting $\mathbb{E}[\hat{V}_i]/R$ proves exact unbiasedness of $\tilde{E}_i$. Because $0 \le p_i \le 1$ all moments exist; the strong law applied to p_i and p_i^2 yields $\hat{\bar{p}}_i \to \bar{p}_i$ and $\hat{V}_i \to V_i$ almost surely, and the probability limits follow by continuity. □

Proof of Proposition 4.1. Since p_i is $\mathcal{F}_{i,s}$-measurable, $h_{i,s} = p_i$; and $h_{i,0} = \mathbb{E}_U[p_i \mid \mathcal{F}_{i,0}] = \bar{p}_i$. The sum telescopes: $\sum_{m=1}^{s} D_{i,m} = h_{i,s} - h_{i,0} = p_i - \bar{p}_i$. For the martingale-difference property, the tower property gives $\mathbb{E}_U[D_{i,m} \mid \mathcal{F}_{i,m-1}] = \mathbb{E}_U[h_{i,m} \mid \mathcal{F}_{i,m-1}] - h_{i,m-1} = 0$. For $m < \ell$, $D_{i,m}$ is $\mathcal{F}_{i,\ell-1}$-measurable, so $\mathbb{E}_U[D_{i,m}D_{i,\ell}] = \mathbb{E}_U[D_{i,m}\,\mathbb{E}_U[D_{i,\ell} \mid \mathcal{F}_{i,\ell-1}]] = 0$. Finally $V_i = \mathbb{E}_U[(\sum_m D_{i,m})^2] = \sum_m \mathbb{E}_U[D_{i,m}^2]$ by the orthogonality just shown. □

Proof of Proposition 4.2. By Proposition 4.1, $(D_{i,m})_{m=1}^{s}$ is a martingale-difference sequence under $\Pr_U$ with partial sum $p_i - \bar{p}_i$. Under the stated conditions, Freedman's inequality (Freedman, 1975) gives, for every $u \ge 0$, $\Pr(|\sum_m D_{i,m}| > u) \le 2\exp\{-u^2/(2(v_i + \frac{1}{3}c_i u))\}$. For $\tau > E_i^{1/2}$ the triangle inequality $|p_i - \pi_i| \le |\sum_m D_{i,m}| + E_i^{1/2}$ yields $\{|p_i - \pi_i| > \tau\} \subseteq \{|\sum_m D_{i,m}| > u\}$ with $u = \tau - E_i^{1/2}$; substituting proves the claim. □

Proof of Corollary 4.1. The variables $B_1(\tau), \ldots, B_n(\tau)$ are independent across studies and take values in $[0,1]$ with common mean $\pi_{\mathrm{fail}}(\tau)$. The empirical Bernstein inequality (Maurer & Pontil, 2009) applied to $\{B_i(\tau)\}$ gives the display. Independence across studies is the only distributional input; dependence among the R runs within a study is absorbed into the bounded average $B_i(\tau)$. □

Why both bounds are stated against π_i, not Y_i. Define the per-study exceedance probability $B_i^*(\tau) = \Pr_U(|p_i - \pi_i| > \tau)$; the population failure rate obeys the tower $\pi_{\mathrm{fail}}(\tau) = \mathbb{E}_i[B_i^*(\tau)]$. Freedman bounds the inner factor per study; empirical Bernstein bounds the outer average from the run averages $B_i(\tau)$, which are unbiased for $B_i^*(\tau)$ because each run is a fresh draw of U with the proxy fixed. Replacing the proxy by the single Bernoulli outcome Y_i breaks the tower: the event $\{|p_i - Y_i| > \tau\}$ mixes model randomness with the one-shot outcome draw, so there is no clean inner factor for Freedman to bound and B_i no longer estimates a well-defined per-study probability.

The aggregate proxy diagnostic (Tier 3). Averaging the left-hand side of the decomposition (1) over studies and runs, $\mathbb{E}[n^{-1}\sum_i R^{-1}\sum_r (p_i^{(r)} - Y_i)^2] = n^{-1}\sum_i (V_i + E_i + A_i)$, while $\mathbb{E}[n^{-1}\sum_i (\hat{V}_i + \tilde{E}_i)] = n^{-1}\sum_i (V_i + E_i^{\mathrm{plug}})$ with $E_i^{\mathrm{plug}} = (\bar{p}_i - \hat{\pi}_i)^2$. The difference of the two observables therefore estimates $n^{-1}\sum_i \pi_i(1-\pi_i) + n^{-1}\sum_i (E_i - E_i^{\mathrm{plug}})$; the second term vanishes exactly when the proxy matches the truth and is small when $\hat{\pi}$ is well calibrated, in which case the

difference should agree with the plug-in $n^{-1}\sum_i \hat{\pi}_i(1-\hat{\pi}_i)$. Disagreement flags the proxy. This diagnostic uses each Y_i only through the observable aggregate, so it requires no per-study knowledge of π_i.

B Datasets, Actions, and Proxies

Sample accounting. RPP recruited replications of 100 studies from three 2008 psychology journals; 97 yielded a valid replication outcome; 90 effect-level claims (from 88 papers; two papers contribute two effects each, sharing one set of model runs) also carry the original-study effect sizes required by the statistical-review action and constitute the analysis set. CB contributes 158 effect-level claims from Errington et al. (2021); SSRP contributes 21 studies from Camerer et al. (2018). The text-embedding proxy for RPP is evaluated on the 88 papers; classification tables use the 90 claims.

Table 3: Auditable actions of the chain of actions, their outputs, and their proxies. Action 2 elicits prose only and is not scored: RPP-era studies (2008–2011) predate preregistration and open-data norms, so no ground truth exists. Actions 4–5 have expert-coded proxies only on RPP. An aleatoric term is reported only for actions whose proxy is a genuine probability (action 6; action 3's power proxy approximately); proxies that are constants (action 1) or realized outcomes (action 7) give $A \equiv 0$ by construction.

action m	Output $p_{i,m}^{(r)} \in [0,1]$	Proxy $\hat{\pi}_{i,m}$	Datasets
1. Extraction	Match indicator for extracted $(N_i, p_i^{\text{val}}, r_i)$ against the dataset record	Constant 1 (correct extraction is the unique target)	all
2. Credibility	Prose assessment of transparency and preregistration	*not scored*	—
3. Statistical review	`STAT_SCORE`: power/plausibility on $[0,1]$	Power-analytic reanalysis from original (N_i, d_i)	all
4. Researcher DoF	`QRP_SEVERITY` rescaled to $[0,1]$	Expert-coded expectancy-bias and diligence codes	RPP
5. Theory & context	`SURPRISE_RATING` rescaled to $[0,1]$	Expert-coded "surprising result"	RPP
6. Scoring	p_i: the replication probability	Embedding-LR proxy $\hat{\pi}_i$ (Section 3.2)	all
7. Verdict	Binary verdict $\in \{0,1\}$	Observed outcome Y_i	all

Prompt and parsing. The structured system prompt is a seven-step template with a fixed ten-line output schema, shared byte-identically across datasets by a single prompt module; responses are parsed into the action fields above, and runs are repeated until $R = 100$ valid (parseable, non-refused) responses per claim are collected in every grid cell.

All three datasets are scored with one prompt. The system prompt below is byte-identical across RPP, CB and SSRP except for the calibration clause marked ⟨PRIOR⟩; the user turn is identical except for the locator block marked ⟨LOCATOR⟩. Both substitutions are listed in Table 4.

System prompt

```
You assess whether a focal effect in a scientific study will replicate.
    Important calibration: across <REFERENCE POPULATION>, roughly 40% <
    OUTCOME PHRASING>. Use this as your prior -- do not assume
    replication is the default.

Work through these seven steps in order, each in 1-3 sentences.

1. Extraction: report only facts stated in the paper -- focal claim,
    sample size N, the focal effect size expressed as a correlation r (
    convert from d, eta-squared, or f if the paper reports one of those;
    write NA if no conversion is possible), p-value, study design (within
    - or between-subjects), primary endpoint, and effect type (main
```

```
    effect, interaction, or other). Write 'not reported' for anything
    absent.
2. Credibility: rate transparency, preregistration, open data/code,
    measurement validity, and design strength.
3. Statistics: judge statistical power, multiplicity, model
    appropriateness, and effect-size plausibility; summarize as a single
    [0,1] statistical-soundness score. Small N (<50), large effect sizes
    (r>.5), or p just above .01 should lower this score.
4. QRP severity: score researcher-degrees-of-freedom concerns on a 0-3
    scale. 0 = no concerns (preregistered, open data, single primary
    outcome, adequate N). 1 = minor (one or two of: no preregistration,
    no open data, modest N). 2 = moderate (multiple concerns: no
    preregistration AND no open data AND either many outcomes/conditions
    or p just under .05 or flexible covariates). 3 = severe (most of the
    above, or evidence of optional stopping or HARKing). State the
    specific concerns that drove your score, not a generic list.
5. Theory and context: give a surprise rating from 1 to 5 (1 =
    theoretically expected, 5 = highly surprising or counterintuitive);
    then judge contextual or temporal fragility (effects tied to a
    specific population, cultural moment, or sensitive topic) as low,
    medium, or high -- and state the specific reason if medium or high.
6. Score: start from the 40% base rate. Adjust upward for high
    statistical soundness (step 3), low QRP severity (step 4 = 0 or 1),
    and low surprise/fragility (step 5). Adjust downward for low
    soundness, QRP severity >= 2, or high fragility. Give your final
    replication probability in [0,1].
7. Verdict: replicable or unreplicable, consistent with whether Q_HAT is
    above or below 0.5, with a one-sentence rationale referencing the
    dominant factor.

After the reasoning, end with EXACTLY these ten lines and nothing after
    them. Use NA if a value is not reported in the paper; numbers only (
    no actions):
EXTRACT_N: <integer sample size>
EXTRACT_P: <p-value as a number, e.g. 0.03>
EXTRACT_R: <effect size as a correlation r, e.g. 0.30>
EFFECT_TYPE: <main|interaction|other>
STAT_SCORE: <0-1 statistical-soundness score>
QRP_SEVERITY: <integer 0-3>
SURPRISE_RATING: <integer 1-5>
CONTEXT_FRAGILITY: <low|medium|high>
Q_HAT: <0-1 replication probability>
VERDICT: <replicable|unreplicable>
```

User turn

```
Here is the original study paper (OCR text; author/affiliation names are
    masked):

<PAPER_TEXT>

<LOCATOR>

Using the text above as the source material, locate this focal study/
    effect and complete the seven-step replicability assessment.
```

The locator identifies only the effect being evaluated. The original study's test statistic, sample size, p-value, and effect size are withheld, because our goal is to assess what can be inferred from the paper's text alone. This requires the model to interpret statistical information in its original textual context instead of rely on preprocessed tabular inputs. In addition, the same ten-line output contract is used across all three datasets, allowing a single parser and a common set of checklist-action definitions to be applied throughout. Third, the paper text provided to the model is extracted from each PDF's text layer and anonymized before evaluation. We use `gpt-4.1-nano` to identify mentions of authors and institutional affiliations in the extracted text, followed by a deterministic

Table 4: The three dataset-specific substitutions. The 40% base rate itself never varies – only the named reference population, the short outcome phrasing around it, and the focal-action locator do. The substitution strings are reproduced verbatim; the prompts sent to the model used the replication project's original name RP:CB, which this paper abbreviates as CB.

Dataset	`<REFERENCE POPULATION>`	`<OUTCOME PHRASING>`	`<LOCATOR>`
RPP	a large sample of psychology experiments	replicate successfully	`Focal study to evaluate: study {n}.` followed by the description of the focal effect, the type of effect, and the type of analysis, each line omitted when the record does not carry it.
CB	completed RP:CB preclinical cancer biology replication attempts	of tested effects replicated successfully	`Focal study to evaluate: {title} (RP:CB paper {n}).`, then `Focal effect: Experiment {e}, Effect {f}.`, then `Description of the focal effect: ...` The 30 of 53 papers with no completed CB effect get the title line alone.
SSRP	large-scale systematic replication attempts in the social and behavioral sciences	of tested effects replicated successfully	`Focal study to evaluate: {title} (SSRP study {n}).` and nothing further; see the note below.

text-replacement step that masks all identified mentions. This ensures that no author- or affiliation-identifying information is provided to the model under evaluation which could bias the replicability prediction..

Table 5: Refusal and parse-failure counts across the reasoning-effort × temperature grid, R=100 per cell, all three datasets (collected n: RPP 9000, CB 15800, SSRP 2100 per cell). "Refuses to answer" requires explicit refusal language AND no `q_hat` produced; a refusal phrase that still yields an answer is hedging, shown separately and not counted as a failure. "Fails to parse" requires every prompted field null on the same row – a single null field is usually the source paper not reporting that value, not a model failure.

		Temperature 0.7		Temperature 0.2	
Dataset	**Failure type**	**high effort**	**low effort**	**high effort**	**low effort**
RPP	Null response	0	0	0	0
	Refuses to answer	0	0	0	0
	Fails to parse (all fields null)	0	0	0	0
	(hedging language, still answered)	2	1	1	1
CB	Null response	0	0	0	0
	Refuses to answer	0	0	0	0
	Fails to parse (all fields null)	0	0	0	0
	(hedging language, still answered)	5	20	4	9
SSRP	Null response	0	0	0	0
	Refuses to answer	0	0	0	0
	Fails to parse (all fields null)	0	0	0	0
	(hedging language, still answered)	0	1	0	0

Proxy estimation detail. The embedding-LR proxy uses nested cross-validation: within each of 5 outer folds, the ℓ_1 penalty and active-set cap are chosen by an inner 5-fold loop; predictions are taken on the held-out outer fold; the procedure is repeated over 20 fold-split seeds and averaged. The 20-seed averaged prediction is materially stronger and stabler than any single seed (ensemble AUC 0.768 vs. per-seed mean 0.689, SD 0.045 across seeds); stored predictions are the artifacts of record. In-sample fits reach AUC 1.0, quantifying the overfitting that nested CV removes.

C Orthogonal Cross-Fitted Action Attribution

The per-action attribution $\mathbb{E}_U[D^2_{i,m}]$ of Proposition 4.1 is defined through the unknown conditional means $h_{i,m}$; a naive plug-in $(\hat{h}_m - \hat{h}_{m-1})^2$ inherits first-order regularization bias from the fitted nuisances. We target the population-average action contribution with a Neyman-orthogonal score and cross-fitting, following the double/debiased machine-learning template (Chernozhukov et al., 2018).

Population target. Pool the (i, r) observations. Let $P = p^{(r)}_{i,s}$ denote the scoring output and condition on the upstream sub-outputs through $\boldsymbol{W}_m = (\boldsymbol{X}_i, p^{(r)}_{i,1:m})$ for $m = 0, \ldots, s-1$, with $\boldsymbol{W}_0 = \boldsymbol{X}_i$; the score is held out of every $\boldsymbol{W}_m$. Write $h_m(\boldsymbol{W}_m) = \mathbb{E}[P \mid \boldsymbol{W}_m]$ and define

$$\theta_m = \mathbb{E}\big[(h_m(\boldsymbol{W}_m) - h_{m-1}(\boldsymbol{W}_{m-1}))^2\big] \ (m = 1, \ldots, s-1), \qquad \theta_{\text{res}} = \mathbb{E}\big[(P - h_{s-1}(\boldsymbol{W}_{s-1}))^2\big],$$

so $\mathbb{E}[(P - h_0(\boldsymbol{X}))^2] = \sum_{m<s} \theta_m + \theta_{\text{res}}$.

Sampling model. All expectations in this appendix are with respect to the law of one pooled observation, generated in two stages: a study is drawn from the study population, which fixes its context $\boldsymbol{X}_i$, and a run is drawn as a fresh realization of the model randomness U, which generates the actions and the score given that context. Three consequences fix the notation. (i) $\boldsymbol{W}_m$ is a *random vector*; its realization for observation (i, r) is $(\boldsymbol{X}_i, p^{(r)}_{i,1:m})$, so it varies over both i and r. (ii) $h_m(\cdot)$ is a *fixed measurable function* of $\boldsymbol{w}$, indexed by m only, not by i or r: the conditional expectation $h_m(\boldsymbol{w}) = \mathbb{E}[P \mid \boldsymbol{W}_m = \boldsymbol{w}]$ integrates the randomness remaining after $\boldsymbol{W}_m$ is fixed, namely the actions generated after action m and the score. Because the study enters the pooled law only through its context, conditioning on $\boldsymbol{W}_m \supseteq \boldsymbol{X}_i$ recovers the per-study Doob mean of Proposition 4.1: evaluated at observation (i, r)'s realization, $h_m(\boldsymbol{W}_m) = h^{(r)}_{i,m}$ almost surely, and in particular $h_0(\boldsymbol{X}_i) = \bar{p}_i$. (iii) Each θ_m is a *scalar*, a function of m alone: the outer expectation averages over both stages, so

$$\theta_m = \mathbb{E}_i\big\{\mathbb{E}_U[D^2_{i,m}]\big\}, \qquad \sum_{m<s} \theta_m + \theta_{\text{res}} = \mathbb{E}_i[V_i].$$

Thus the θ's are the study-averaged attribution and their sum is the study-averaged V_i; pooling across (i, r) is an estimation device, not a change of estimand. Per-claim profiles are obtained separately, by evaluating the fitted $\hat{h}_m$ on claim i's runs and averaging within the claim; these are estimates, not exact decompositions. Under cross-fitting the fitted nuisances are one function per fold, $\hat{h}^{(-k)}_m$, still indexed by neither i nor r.

Orthogonal score and exact telescoping. For fitted nuisances $\hat{h}_m$ define

$$\phi_m = 2(\hat{h}_m - \hat{h}_{m-1})P - \hat{h}^2_m + \hat{h}^2_{m-1}, \qquad \phi_{\text{res}} = (P - \hat{h}_{s-1})^2.$$

At the truth $\mathbb{E}[\phi_m] = \theta_m$, and the score is Neyman-orthogonal: nuisance error enters only quadratically.

Proposition C.1 (Exact empirical telescoping)**.** *For every observation and any fitted* $\hat{h}_0, \ldots, \hat{h}_{s-1}$, $\sum_{m=1}^{s-1} \phi_m + \phi_{\text{res}} = (P - \hat{h}_0)^2$.

Proof. Summing and telescoping, $\sum_{m<s}(2(\hat{h}_m - \hat{h}_{m-1})P - \hat{h}^2_m + \hat{h}^2_{m-1}) = 2P(\hat{h}_{s-1} - \hat{h}_0) - (\hat{h}^2_{s-1} - \hat{h}^2_0)$. Adding $\phi_{\text{res}} = P^2 - 2P\hat{h}_{s-1} + \hat{h}^2_{s-1}$ cancels the $\hat{h}_{s-1}$ terms and leaves $(P - \hat{h}_0)^2$. No property of the fitted nuisances is used. That $\mathbb{E}[\phi_m] = \theta_m$ at the truth follows from $\mathbb{E}[h_m P] = \mathbb{E}[h^2_m]$ (tower property). □

Thus the estimated action contributions telescope *exactly* to the estimated total variability regardless of nuisance quality: the analogue, at the estimator level, of the exact population decomposition.

Cross-fitting and inference. Partition the n studies (not the nR runs) into K folds; all runs of a study stay together because n is the effective sample size. Fit $\hat{h}_m^{(-k)}$ off-fold, evaluate scores on-fold, and average: $\hat{\theta}_m = n^{-1}\sum_i R_i^{-1}\sum_r \hat{\phi}_{ir,m}$. Cluster-robust standard errors use study-level mean scores $\bar{\phi}_{i,m}$: $\widehat{\mathrm{se}}(\hat{\theta}_m) = \{n(n-1)\}^{-1/2}\{\sum_i(\bar{\phi}_{i,m} - \hat{\theta}_m)^2\}^{1/2}$, with simultaneous bands by a multiplier bootstrap over studies. A root-n normal limit holds under independent study clusters, exchangeable runs, fixed s, bounded outputs, and $\|\hat{h}_m - h_m\|_2 = o_p(n^{-1/4})$. At $n = 90$ we report study-cluster bootstrap intervals alongside, use deliberately simple pre-specified nuisance learners, and do not truncate negative $\hat{\theta}_m$ before inference. The sum $\hat{v}_i = \sum_m \hat{\theta}_m$ is the Freedman plug-in of Section 4.1.

Cross-fitting and inference are performed at the study/paper-cluster level. Among the 158 CB effects, 103 are analyzable. The other 55 lack the two complete runs required for decomposition: 52 have no q_1 values, and three have only one. By comparison, the $\hat{\theta}_m$ estimates for RPP and SSRP use the complete samples, with every repeated runs per effect or study producing a fully parsable q_1, q_3, q_4, q_5, q_6 chain.

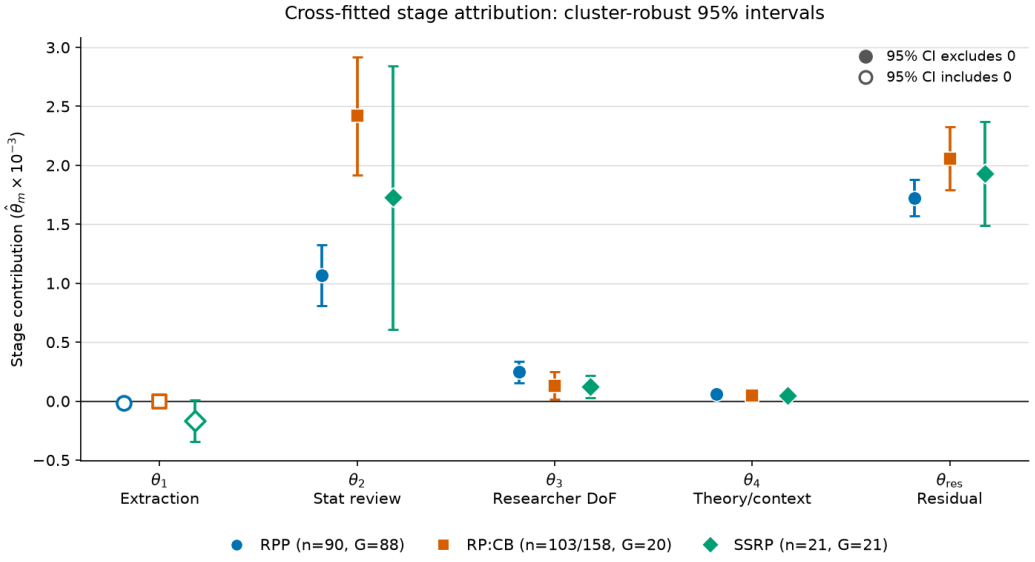

Figure 3: Cross-fitted estimates of action-specific contributions to run-to-run variance $\hat{\theta}_i$ with 95% cluster-robust intervals

Figure 3 shows the point estimates of empirical orthogonal action scores across the three datasets and the corresponding cluster-robust 95% confidence intervals. Statistical review is the largest attributable source of run-to-run variation across all three datasets, with the largest estimate observed in CB. Extraction is indistinguishable from zero across datasets, while researcher degrees of freedom and theory/context make smaller positive contributions. The substantial residual component indicates that much of the run-to-run variation remains unexplained by the four modeled actions.

For the coverage simulation, we generated 2000 datasets, each containing $n = 90$ independent effect clusters and $R = 100$ runs per effect. For each effect, we drew $\mu_i \sim U(0.3, 0.7)$; for each run, we independently drew $z_1, \ldots, z_4, z_{\mathrm{res}} \sim N(0,1)$ and generated $p_{i1}^{(r)} = \mu_i + \sqrt{\theta_1} z_1$, $p_{im}^{(r)} = p_{i,m-1}^{(r)} + \sqrt{\theta_m} z_m$ for $m = 2, \ldots, 4$, and $P_i^{(r)} = p_{i4}^{(r)} + \sqrt{\theta_{\mathrm{res}}} z_{\mathrm{res}}$, truncating probabilities to $[0,1]$. We set $\theta_1, \theta_2, \theta_3, \theta_4, \theta_{\mathrm{res}} = (0.0006, 0.0016, 0.0003, 0.0002, 0.0020)$ such that true values are on the same scale as seen in the empirical estimates. Then, we repeated the full analysis for estimating θ_m, bias evaluation, and coverage assessment.

The estimator in simulation experiments exhibited negligible bias across all five components (Table 6), with absolute bias no larger than 0.018×10^{-3}, significantly lower than the true sizes of action-specific θ. The pointwise 95% coverage ranged from 93.8% to 95.7%. The simultaneous bootstrap band achieved 92.9% joint coverage, indicating modest undercoverage relative to the nom-

Table 6: Simulation results for the variance-component estimators.

Component	True θ ($\times 10^{-3}$)	Bias ($\times 10^{-3}$)	95% CI coverage[a]	Bootstrap band width[b] ($\times 10^{-3}$)
θ_1	0.60	0.006	0.943	0.175
θ_2	1.60	0.018	0.938	0.249
θ_3	0.30	0.003	0.953	0.091
θ_4	0.20	0.001	0.954	0.070
θ_{res}	2.00	0.006	0.957	0.155

[a] Pointwise coverage is the proportion of simulations in which the 95% confidence interval for each component contains its true value.
[b] Bootstrap band width is the mean width of the simultaneous 95% bootstrap band using the resampling size $B = 10000$, constructed using a common critical value across all components.

inal 95% level. Across the 2,000 simulated datasets, the mean absolute discrepancy between the sum of the five estimated components and the known total variance was 0.71%, indicating close recovery of the variance decomposition.

D Uncertainty-Quantification Metric Definitions

For completeness we record the exact computations behind the label-free consistency and calibration metrics reported in Tables 8 and 9.

Run-to-run dispersion. For verbalized confidence, dispersion is the standard deviation across runs, $\mathrm{SD}_r(p_i^{(r)})$.

Binary entropy. Let $n_{1,i} + n_{0,i} = R$ count the two verdicts and $\bar{v}_i = n_{1,i}/R$. Binary entropy is

$$H(\bar{v}_i) = -\bar{v}_i \log_2 \bar{v}_i - (1 - \bar{v}_i) \log_2 (1 - \bar{v}_i),$$

with $0 \log_2 0 = 0$.

Pairwise agreement. The agreement statistic is

$$A_i^{\text{agr}} = \frac{\binom{n_{1,i}}{2} + \binom{n_{0,i}}{2}}{\binom{R}{2}}.$$

Let $c_i = 1$ if the majority verdict for study i is correct and $c_i = 0$ otherwise; agreement reliability is $\mathrm{AUC}(\{A_i^{\text{agr}}, c_i\}_{i=1}^n)$, using A_i^{agr} as the prediction score for correctness.

Expected calibration error. ECE uses equal-count bins (five for RPP and CB, three for SSRP, scaled to n), comparing the mean predicted probability with the observed replication rate in each bin and averaging the absolute differences (Guo et al., 2017).

E Additional Results

Per-action VEA values (default cell). CB (n per action in parentheses): extraction $\hat{V} = 0.023$, $\tilde{E} = 0.750$ (106); statistical review $\hat{V} = 0.005$, $\tilde{E} = 0.265$, $A = 0.077$ (90); scoring $\hat{V} = 0.005$, $\tilde{E} = 0.076$, $A = 0.191$ (158); verdict $\hat{V} = 0.089$, $\tilde{E} = 0.247$ (158). SSRP: extraction $0.034/0.527$; statistical review $0.005/0.216/0.002$; scoring $0.004/0.030/0.223$; verdict $0.081/0.229$ ($n = 21$ throughout). The temperature-0.2 profiles are essentially identical (e.g., CB extraction $\tilde{E} = 0.767$), confirming that the epistemic structure is not a decoding artifact.

Variability share. Across the 92 RPP claims with computable profiles, the per-study ratio of $\sum_m \hat{V}_{i,m}$ to $\sum_m (\hat{V}_{i,m} + \tilde{E}_{i,m})$ has median 0.05 overall and 0.11 among the 57 claims with nonzero measured variability; its maximum is 0.52. No study's error is variability-dominated.

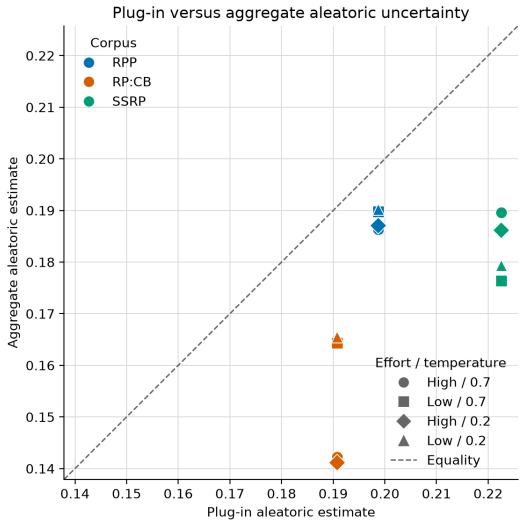

Figure 4: Comparison of plug-in and aggregate estimates of mean aleatoric uncertainty across datasets, reasoning-effort levels, and temperatures; the dashed diagonal denotes exact agreement between the two estimators.

Table 7: Reasoning-effort × temperature grid, AUC (Hanley–McNeil SE), $R = 100$ per cell. SSRP's reversal of the effort ordering occurs within wide SEs at $n = 21$ and is not interpreted. Temperature 0.2 cells for CB and SSRP are computed and reported in the repository; classification tables for those cells mirror the 0.7 columns.

		Temperature 0.7		Temperature 0.2	
Dataset	**Predictor**	**high effort**	**low effort**	**high effort**	**low effort**
RPP	$\bar{v}$	0.665 ± 0.062	0.638 ± 0.063	0.665 ± 0.062	0.620 ± 0.063
	$\bar{q}$	0.625 ± 0.063	0.610 ± 0.063	0.624 ± 0.063	0.603 ± 0.063
CB	$\bar{v}$	0.704 ± 0.042	0.628 ± 0.045	0.721 ± 0.041	0.624 ± 0.045
	$\bar{q}$	0.698 ± 0.043	0.594 ± 0.046	0.695 ± 0.043	0.574 ± 0.046
SSRP	$\bar{v}$	0.827 ± 0.090	0.870 ± 0.078	0.808 ± 0.095	0.841 ± 0.087
	$\bar{q}$	0.788 ± 0.100	0.856 ± 0.083	0.808 ± 0.095	0.856 ± 0.083

Audit-budget allocation. Given a per-claim tolerance budget τ, minimizing the summed action-level Markov bounds subject to $\sum_m \tau_m = \tau$ yields $\tau^*_{i,m} = \tau\, a_{i,m}^{1/3} / \sum_\ell a_{i,\ell}^{1/3}$ with $a_{i,m} = \max(\hat{V}_{i,m} + \tilde{E}_{i,m}, 0)$, and the closed-form per-claim bound $\min\{1, (\sum_m a_{i,m}^{1/3})^3/\tau^2\}$. At practically relevant budgets ($\tau = 0.2$) the bound is clamped at 1 for nearly all claims: a Chebyshev-type bound cannot control chains whose epistemic gaps are of the same order as the budget, consistent with the main-text conclusion that the practical guarantee is at the population level, not per claim.

Bound panels for CB and SSRP. The corresponding panels reproduce Figure 2b qualitatively on both datasets, at both effort levels and both temperatures: Azuma–Hoeffding vacuous, Freedman informative only past $\tau \approx 0.3$ under a support-restricted increment constant, empirical Bernstein tracking the empirical curve from above. Illustrative per-study panels (one low- and one high-variance claim per dataset, chosen at the 5th/95th percentile of $\hat{v}_i$) are provided in the repository alongside per-dataset notes.

Calibration metrics. On RPP (default cell), verbalized confidence attains ECE 0.105 and run-to-run dispersion $\mathrm{SD}(p_i^{(r)}) = 0.053$; the vote rate attains ECE 0.240, binary entropy 0.343 bits, and pairwise agreement 0.851. Definitions are in Appendix D.

Table 8: Calibration and dispersion metrics across the reasoning-effort × temperature grid, R=100 per cell across all three datasets.

		Temperature 0.7		Temperature 0.2	
Dataset	**Metric**	**high effort**	**low effort**	**high effort**	**low effort**
RPP	ECE($\bar{v}$)	0.240	0.231	0.264	0.228
	Entropy($\bar{v}$), bits	0.343	0.404	0.340	0.421
	Agreement($\bar{v}$)	0.851	0.829	0.852	0.819
	ECE($\bar{q}$)	0.105	0.137	0.128	0.159
	SD($\hat{q}$)	0.053	0.056	0.052	0.057
CB	ECE($\bar{v}$)	0.195	0.285	0.199	0.267
	Entropy($\bar{v}$), bits	0.400	0.335	0.392	0.382
	Agreement($\bar{v}$)	0.823	0.858	0.825	0.839
	ECE($\bar{q}$)	0.078	0.054	0.083	0.071
	SD($\hat{q}$)	0.066	0.057	0.065	0.056
SSRP	ECE($\bar{v}$)	0.233	0.200	0.244	0.185
	Entropy($\bar{v}$), bits	0.365	0.365	0.323	0.424
	Agreement($\bar{v}$)	0.838	0.840	0.855	0.816
	ECE($\bar{q}$)	0.193	0.194	0.191	0.195
	SD($\hat{q}$)	0.056	0.052	0.055	0.053

Study-level bootstrap intervals. Table 9 adds finite-sample uncertainty to the classification and calibration metrics for the default cell: 95% confidence intervals from a study-level cluster bootstrap that resamples papers with replacement (RPP: 88 papers carrying 90 claims; CB: 23 papers carrying 158 effects; SSRP: 21 studies), B=10,000 resamples, percentile intervals. Accuracy thresholds the predictor at 0.5; the Brier score is the study-level mean of (predictor $- Y_i)^2$; ECE uses the same equal-count bins as Table 8, refit within every resample. The final column benchmarks each Brier score against a constant predictor set to the observed base rate. Consistent with the separation of discrimination and calibration, $\bar{q}$ improves on the base-rate benchmark significantly on CB ($\Delta = -0.028$, CI excluding zero) and matches it on RPP and SSRP, while $\bar{v}$ discriminates better but is worse calibrated, so its Brier score exceeds the benchmark on RPP ($\Delta = +0.038$, CI $[-0.034, 0.119]$). The full effort × temperature grid is in the companion repository.

Table 9: Study-cluster bootstrap estimates for the default LLM cell (high effort, temperature 0.7, R=100). Values are point estimates with percentile 95% intervals from B=10,000 paper-cluster resamples. RPP outcomes use the original OSF export. The final two columns give the Brier score of a constant predictor set to the observed sample base rate and the paired Brier difference from that benchmark (positive is worse).

Dataset	**Predictor**	**Accuracy**	**Brier**	**ECE**	**Base-rate Brier**	**Δ Brier**
RPP	$\bar{v}$	0.678 [0.578, 0.774]	0.267 [0.191, 0.348]	0.240 [0.140, 0.332]	0.229	+0.038 [−0.034, 0.119]
	$\bar{q}$	0.678 [0.580, 0.775]	0.229 [0.205, 0.254]	0.105 [0.072, 0.225]	0.229	−0.000 [−0.024, 0.035]
CB	$\bar{v}$	0.684 [0.576, 0.786]	0.248 [0.170, 0.324]	0.195 [0.114, 0.283]	0.247	+0.002 [−0.065, 0.076]
	$\bar{q}$	0.696 [0.590, 0.807]	0.219 [0.192, 0.244]	0.078 [0.060, 0.188]	0.247	−0.028 [−0.046, −0.003]
SSRP	$\bar{v}$	0.714 [0.524, 0.905]	0.229 [0.088, 0.387]	0.233 [0.068, 0.402]	0.236	−0.007 [−0.150, 0.192]
	$\bar{q}$	0.714 [0.524, 0.905]	0.220 [0.177, 0.268]	0.193 [0.091, 0.373]	0.236	−0.016 [−0.069, 0.100]

Contamination probe To assess whether performance could reflect pretraining leakage, we ran a contamination probe using (R=10) independent queries per study or claim at high reasoning effort. Each query supplied only the study title and a one-line locator identifying the focal finding, with no article text, authors information, journals' title or replication outcome. The model was instructed to answer solely from prior familiarity, without plausibility-based reasoning or field-level priors, and to return a familiarity judgment (recognized, unsure, or unfamiliar), a recalled verdict (replicable, unreplicable, or unknown), and a one-sentence explanation.

For action i, define

$$K_i = \sum_{r=1}^{10} \mathbf{1}\{F_{ir} = \text{recognized},\ D_{ir} = Y_i\},$$

where Y_i is the observed replication outcome, F_{ir} is the reported familiarity category, and D_{ir} is the reported replication verdict. The study-level any-hit and five-hit recall rates are $N^{-1}\sum_i \mathbf{1}\{K_i \geq 1\}$ and $N^{-1}\sum_i \mathbf{1}\{K_i \geq 5\}$, respectively. As a complementary effect-level diagnostic, we stratified all probe responses by dataset and observed outcome and reported the number and proportion that were recognized overall and jointly classified as recognized & replicable, recognized & unreplicable, or recognized & unknown; the outcome-concordant cell represents claimed correct recall. These descriptive tables exclude inferential p-values, and self-reported recognition is treated as a necessary indicator of recall rather than definitive evidence of memorization.

Table 10: Study-level contamination-probe recall.

Dataset	Effects/claims	Any-hit recall rate	Five-hit recall rate
RPP	90	0.233	0.022
CB	158	0.785	0.297
SSRP	21	0.429	0.190

Table 11: Effect-level contamination-probe responses by observed outcome.

Dataset	True outcome	Responses	Recognized & replicable	Recognized & unreplicable	Recognized & unknown
RPP	replicable	320	**14 (4.4%)**	1 (0.3%)	60 (18.8%)
RPP	unreplicable	580	11 (1.9%)	**33 (5.7%)**	130 (22.4%)
CB	replicable	700	**324 (46.3%)**	80 (11.4%)	248 (35.4%)
CB	unreplicable	880	200 (22.7%)	**204 (23.2%)**	402 (45.7%)
SSRP	replicable	130	**5 (3.8%)**	3 (2.3%)	37 (28.5%)
SSRP	unreplicable	80	3 (3.8%)	**35 (43.8%)**	19 (23.8%)

The probe revealed substantial heterogeneity across datasets. CB showed the strongest evidence of possible recall, with a five-hit rate of 29.7%; outcome-concordant recognized responses occurred in 46.3% of runs for replicable effects and 23.2% for unreplicable effects. Recall was less frequent in SSRP and especially inconsistent in RPP. SSRP recall was concentrated among unreplicable studies, whereas the corresponding responses in RPP were uncommon for either outcome. The large decline from any-hit to five-hit recall indicates that much of the apparent familiarity was not stable across repeated queries; therefore, these results identify possible pretraining exposure, particularly for CB, rather than establishing verbatim memorization.

Recognition-stratified accuracy. To connect recognition to prediction quality, Table 12 stratifies default-cell majority-verdict accuracy by the recall count K_i, with the trained embedding proxy evaluated on the same strata. Accuracy rises with recognition in all three datasets (Spearman correlation between K_i and per-study correctness: 0.15 RPP, 0.27 CB, 0.38 SSRP), so pretraining exposure inflates measured performance and the absolute accuracies should be read with that in mind. The CB conclusion nonetheless survives the stratification: the zero-shot pipeline outperforms the trained embedding proxy within *every* recognition stratum, including the unrecognized effects (52.9% vs. 47.1% at K_i=0, n=34), though the margin there is modest and the stratum small.

Table 12: Default-cell majority-verdict accuracy stratified by contamination-probe recall count K_i, with the trained embedding proxy evaluated on the same strata.

Dataset	**Stratum**	n	**LLM $\bar{v}$ accuracy**	**Embedding-LR accuracy**
RPP	$K_i = 0$	69	0.638	0.696
	$K_i \geq 1$	21	0.810	0.810
CB	$K_i = 0$	34	0.529	0.471
	$K_i \geq 1$	124	0.726	0.621
	$K_i \geq 5$	47	0.809	0.660
SSRP	$K_i = 0$	12	0.583	0.750
	$K_i \geq 1$	9	0.889	0.778