RISK-CONTROLLED DETECTION OF HALLUCINATED REFERENCES WITH AGGREGATED DETECTORS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models frequently fabricate academic references that appear authentic but correspond to no real publication. Existing detectors typically rely on a single aspect of verification and offer no formal guarantees on their error rates. We design a risk-controlled framework that assigns each reference a continuous suspicion score by aggregating heterogeneous verification signals, and then labels it as valid, hallucinated, or deferred to a human, controlling the false positive and false negative rates with as few deferrals as possible. Our guiding principle is that validity comes from calibration and efficiency from aggregation. We calibrate two thresholds with per-class conformal quantiles of the suspicion score, which delivers finite-sample control at user-specified levels. We then show that the deferral rate depends on the score only through its ROC curve, and that the posterior over detector outputs is the optimal aggregator. Experiments on the HallMark benchmark and on StatRef, a new dataset of statistics references, confirm that calibration secures validity and aggregation secures efficiency, with recalibration restoring control under domain shift.

1 INTRODUCTION

Large language models (LLMs) are now used to search the literature and to draft scientific text, and they fabricate references that point to no paper or combine the metadata of real papers with invented fields (Walters & Wilder, 2023; Agrawal et al., 2024). Fabrication rates between 14% and 95% have been measured across 13 LLMs (Xu et al., 2026), and fabricated citations have been found in papers at NeurIPS 2025 (Ansari, 2026) and at ACL venues (Sakai et al., 2026). Checking every reference by hand does not scale, so authors, reviewers, and venues need automated auditors.

To meet this need, a growing range of reference-verification methods has emerged. Some check metadata fields with rules, some query bibliographic databases (Abbonato, 2026; Rao & Callison-Burch, 2026; Reizinger & Brendel, 2026b), and some ask an LLM to judge the reference (Agrawal et al., 2024). However, individual detectors capture different aspects of reference validity, and their verdicts often disagree. On the HallMark benchmark (Reizinger & Brendel, 2026a), the eleven detectors we study are unanimous on 1.9% of the references, and on 59% of them at least three detectors call the reference hallucinated while at least three others call it valid (Table 4). Their errors differ in kind: one verification tool accepts almost no fabricated reference but accuses about a third of the valid ones, whereas one LLM judge does the opposite (Table 3). None of them bounds its error rates, and their confidences are not calibrated.

An auditor of LLM-generated references can be evaluated using three quantities. Its false-positive rate (FPR) is the fraction of valid references it accuses, which matters because an accusation can trigger a misconduct investigation. Its false-negative rate (FNR) is the fraction of fabricated references it accepts. Its deferral rate is the fraction of references it hands to a human for a full check. The first two should stay below tolerances chosen by the user, and the third should be as small as possible. This formulation raises two central questions:

Q1. How can the FPR and FNR of an auditor built from uncalibrated detectors be certified, without parametric assumptions on the detector outputs?

Q2. Given many detectors, should we combine them or select the best one? If we combine them, how, and what does combining buy if any score can be certified?

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.

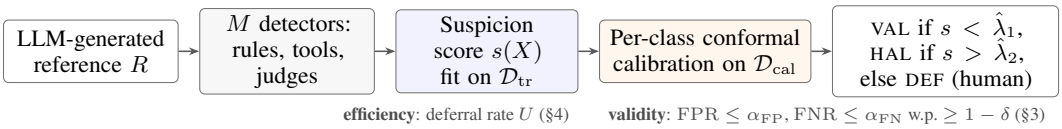

Figure 1: Overview of the proposed framework. Calibration certifies the error rates of any aggregator, and the aggregator determines how many references are deferred.

We answer the first by calibration and the second by aggregation (Fig. 1): an aggregator turns the detector outputs into a suspicion score, two thresholds on this score are calibrated separately on valid and on hallucinated references, and a reference is accepted below the lower threshold, flagged above the upper one, and deferred in between. The contributions are as follows:

- **Validity from calibration** (Section 3). We formulate reference auditing as a three-way decision with class-conditional error constraints. Per-class conformal calibration of the two thresholds controls the FPR and FNR in finite samples in the probably approximately correct (PAC) sense, for any score, including scores trained on another domain. The construction is that of PAC-Wrap (Li et al., 2022), used at the user's tolerances with deferral in place of relaxed tolerances, and randomized tie-breaking removes the conservatism that tied scores would otherwise cause.
- **Efficiency from aggregation** (Section 4). The oracle result for set-valued classification with class-specific error control (Sadinle et al., 2019) shows that the minimum achievable deferral rate is determined by the score's ROC curve, with the posterior probability of hallucination given all detector outputs as the optimal aggregator. Our theoretical contribution is Theorem 3 with Corollary 4, which bound the deferral rate of the calibrated rule by an information term, an aggregation regret, and the cost of the unused error budget, and give the exact distribution of the unused budget.
- **Evidence** (Section 5). On HallMark and on StatRef, a new benchmark of statistics references, the calibration is checked against exact failure probabilities on known finite populations, and every term of Theorem 3 is computed on synthetic populations with known posteriors. Fixed aggregation rules do not reliably beat the best single detector. Boosted stacking of all detectors defers 2.3 to 3.9 times fewer references than the stronger of the raw and re-encoded single detectors.

2 Problem setup

Let $(R, Y) \sim P$, where R is a reference produced by an LLM and $Y \in \{0, 1\}$ is its label. A reference is valid ($Y = 0$) if it corresponds to a real publication whose bibliographic fields are mutually consistent and verifiable, and hallucinated ($Y = 1$) if an essential field is fabricated or materially inconsistent with the true record, including non-existent papers, fake DOIs, wrong author lists or venues, and hybrids of real and fabricated metadata. Let $\pi_1 = P(Y = 1)$ and $\pi_0 = 1 - \pi_1$.

Each reference R is evaluated by M detectors, with detector $m \in [M]$ returning a label $\ell_m(R) \in \{-1, 0, +1\}$ (valid, uncertain, or hallucinated) and a confidence $c_m(R) \in [0, 1]$. We write $X_m = (\ell_m, c_m)$, $X = (X_1, \ldots, X_M)$, $X_S = (X_m)_{m \in S}$ for $S \subseteq [M]$, and $\phi_m = \ell_m c_m \in [-1, 1]$ for the signed evidence of detector m. An aggregator is a measurable map s from detector outputs to $\mathbb{R}$, and $s(R) = s(X(R))$ is the suspicion score of R, with larger values indicating stronger evidence of hallucination. Examples are a single detector ($s = \phi_m$), majority vote ($s = M^{-1} \sum_m \ell_m$), mean evidence ($s = M^{-1} \sum_m \phi_m$), and supervised stacking models that estimate the posterior $\eta(X) = P(Y = 1 \mid X)$ from labeled data (Hastie et al., 2009).

Two thresholds $\lambda_1 \leq \lambda_2$ convert the suspicion score into the following three-way decision rule:

$$h_\lambda(R) = \text{VAL if } s(R) < \lambda_1, \qquad h_\lambda(R) = \text{HAL if } s(R) > \lambda_2, \qquad h_\lambda(R) = \text{DEF otherwise}, \quad (1)$$

with VAL, HAL, and DEF for valid, hallucinated, and defer. The rule has false-positive rate $\text{FPR}(h) = P(h(R) = \text{HAL} \mid Y = 0)$, false-negative rate $\text{FNR}(h) = P(h(R) = \text{VAL} \mid Y = 1)$, and deferral rate $U(h) = P(h(R) = \text{DEF})$. For tolerances $\alpha_{\text{FP}}, \alpha_{\text{FN}} \in (0, 1)$, we seek a rule that satisfies $\text{FPR}(h) \leq \alpha_{\text{FP}}$ and $\text{FNR}(h) \leq \alpha_{\text{FN}}$ with as small a deferral rate $U(h)$ as possible. We fit the aggregator on a labeled set $\mathcal{D}_{\text{tr}}$, which may come from a different domain, and we calibrate the thresholds on a set $\mathcal{D}_{\text{cal}} = \{(R_i, Y_i)\}_{i=1}^n$ that is independent of $\mathcal{D}_{\text{tr}}$, with $n_y = |\{i : Y_i = y\}|$.

Algorithm 1 Per-class conformal calibration of a three-way rule

Require: calibration scores $\{(s(R_i), Y_i)\}_{i=1}^{n}$, tolerances $\alpha_{\mathrm{FP}}, \alpha_{\mathrm{FN}}$, and error level δ

1: **tie-breaking:** attach an independent $\xi \sim \mathrm{Unif}(0,1)$ to every score, here and at deployment, order the pairs (s, ξ) lexicographically, and store each threshold below as such a pair
2: $\delta_0 \leftarrow 1 - \sqrt{1-\delta}$ ▷ per-class budget, since the two classes are independent given the labels
3: sort the valid scores $s^0_{(1)} \le \cdots \le s^0_{(n_0)}$ and set $k_0 \leftarrow \min\{k \le n_0 : \mathbb{P}(\mathrm{Bin}(n_0, 1-\alpha_{\mathrm{FP}}) \ge k) \le \delta_0\}$, $\hat{\lambda}_2 \leftarrow s^0_{(k_0)}$
4: sort the hallucinated scores $s^1_{(1)} \le \cdots \le s^1_{(n_1)}$ and set $k_1 \leftarrow \min\{k \le n_1 : \mathbb{P}(\mathrm{Bin}(n_1, 1-\alpha_{\mathrm{FN}}) \ge k) \le \delta_0\}$, $\hat{\lambda}_1 \leftarrow s^1_{(n_1-k_1+1)}$
5: **if** a minimum is over the empty set, set $\hat{\lambda}_2 = +\infty$ (resp. $\hat{\lambda}_1 = -\infty$)
6: **if** $\hat{\lambda}_1 \le \hat{\lambda}_2$: **return** rule (1) with $(\hat{\lambda}_1, \hat{\lambda}_2)$
7: **else**: **return** the single cut τ at the midpoint of $\hat{\lambda}_2$ and $\hat{\lambda}_1$ (the pair midpoint with tie-breaking, Section C), i.e., VAL if $s \le \tau$ and HAL otherwise ▷ no deferral

3 VALIDITY: CERTIFIED THREE-WAY DECISIONS FOR ANY SCORE

Once the suspicion score is fixed, we use a labeled calibration set to choose the two thresholds so that the resulting rule satisfies the desired class-specific error guarantees. Algorithm 1 sets $\hat{\lambda}_2$ to an upper confidence bound on the $(1-\alpha_{\mathrm{FP}})$-quantile of the valid scores and $\hat{\lambda}_1$ to a lower confidence bound on the α_{FN}-quantile of the hallucinated scores, with no tuning parameters beyond the tolerances, δ, and the equal per-class failure probability δ_0. Tie-breaking is optional for validity, but it removes the conservatism caused by tied scores, and deployment needs only the two stored pairs (Section C). Theorem 1 states the guarantee, and Section A.1 contains the proof.

Theorem 1 (Finite-sample validity for any aggregator)**.** *Let $\delta \in (0,1)$ and let s be any score function fitted independently of $\mathcal{D}_{\mathrm{cal}}$. All statements are conditional on the fitted score. Suppose that, given the labels, the calibration scores are independent, with $s(R_i) \mid Y_i = y$ distributed as $s(R) \mid Y = y$ under P. Let h be the rule returned by Algorithm 1, with or without tie-breaking. Probabilities below are over $\mathcal{D}_{\mathrm{cal}}$ and its tie-breaking noise, and* FPR *and* FNR *average over the noise attached to the evaluated reference. Conditionally on (n_0, n_1):*

(a) (high probability, PAC) $\mathbb{P}\big(\mathrm{FPR}(h) \le \alpha_{\mathrm{FP}} \text{ and } \mathrm{FNR}(h) \le \alpha_{\mathrm{FN}}\big) \ge 1-\delta$.

(b) (in expectation) If the ranks are replaced by $k_0 = \lceil (n_0+1)(1-\alpha_{\mathrm{FP}}) \rceil$ and $k_1 = \lceil (n_1+1)(1-\alpha_{\mathrm{FN}}) \rceil$, with an infinite threshold when $k_y > n_y$, then $\mathbb{E}[\mathrm{FPR}(h)] \le \alpha_{\mathrm{FP}}$ and $\mathbb{E}[\mathrm{FNR}(h)] \le \alpha_{\mathrm{FN}}$.

(c) (label shift) Parts (a) and (b) hold for FPR *and* FNR *under any deployment distribution with the same class-conditional distributions of $s(R)$, whatever its prevalence.*

Our calibration layer builds on PAC-Wrap (Li et al., 2022), which applies the class-specific binomial thresholds of Park et al. (2020), but it is designed for a different objective. When the intersection is empty, for scores between its flagging threshold and a higher acceptance threshold, PAC-Wrap abstains, although any single cut in that interval is also valid. When it contains both labels, the genuinely ambiguous case, PAC-Wrap relaxes both tolerances in steps until the thresholds coincide or swap order. We keep the tolerances chosen by the user and defer on the ambiguous region, combine the two classes by a product bound rather than a union bound, and break ties at random. PAC-Wrap thus reduces ambiguity by loosening the tolerances, whereas we reduce it by aggregation at fixed tolerances, a strategy analyzed in Section 4. Section C will further show how this construction relates to the Neyman–Pearson umbrella (Tong et al., 2018) and Learn-then-Test (Angelopoulos et al., 2025).

Theorem 1 requires no assumptions about the quality, range, continuity, or training domain of the score. With frozen detectors and calibration labels from the deployment distribution, a poor aggregator only raises the deferral rate. The score must be fixed before calibration, and selecting a rule based on its deferral rate on $\mathcal{D}_{\mathrm{cal}}$ therefore requires a fresh calibration set. The guarantee degrades with the Kolmogorov distance between the calibration and deployment score distributions, as quantified in Proposition 6 in Section A.5. Consequently, deploying a new LLM or updating a detector requires recalibration, as examined in Section 5.4.

4 EFFICIENCY: SHOULD WE AGGREGATE, AND HOW?

Having established that calibration controls both error rates for any fixed score, we now turn to efficiency and ask how the choice of score affects the amount of deferral required. To formalize these questions, a three-way rule is a pair $\psi = (\psi_0, \psi_1)$ of functions with values in $[0, 1]$ and $\psi_0 + \psi_1 \leq 1$, giving the probabilities of VAL and HAL. For a score g, let $\beta_g^1(a)$ be the largest true-positive rate $P(\text{HAL} \mid Y = 1)$ of a randomized upper-threshold rule on g with false-positive rate at most a. We refer to the function $a \mapsto \beta_g^1(a)$ as the ROC curve of g. Let $\beta_g^0(b)$ be the largest true-negative rate $P(\text{VAL} \mid Y = 0)$ of a randomized lower-threshold rule with false-negative rate at most b. Let $U_g^\star(\alpha_{\text{FP}}, \alpha_{\text{FN}})$ be the smallest deferral rate of a randomized three-way threshold rule on g that satisfies both constraints in the population. Let $U_S^\star$ be the smallest deferral rate over all rules that depend on R only through X_S. Proposition 2 establishes the relationship between the deferral rate and the ROC curve, and shows that posterior aggregation is optimal. Section A.2 contains the proof.

Proposition 2 (Deferral–ROC identity and optimal aggregation)**.** *(a) For every score g,*

$$U_g^\star = \left[\pi_0\big(1 - \alpha_{\text{FP}} - \beta_g^0(\alpha_{\text{FN}})\big) + \pi_1\big(1 - \alpha_{\text{FN}} - \beta_g^1(\alpha_{\text{FP}})\big)\right]_+ . \tag{2}$$

(b) For every $S \subseteq [M]$, the posterior $\eta_S(X) = P(Y = 1 \mid X_S)$ satisfies $U_{\eta_S}^\star = U_S^\star \leq U_g^\star$ for every aggregator g of X_S.

(c) If $S \subseteq T$, then $U_T^\star \leq U_S^\star$, strictly if $U_S^\star > 0$ and η_T has strictly larger detection power at α_{FP} or validation power at α_{FN} than η_S.

On the ROC plane, $1 - \alpha_{\text{FN}} - \beta_g^1(\alpha_{\text{FP}})$ and $1 - \alpha_{\text{FP}} - \beta_g^0(\alpha_{\text{FN}})$ are the vertical and horizontal shortfalls of the ROC curve of g from the target corner $(\alpha_{\text{FP}}, 1 - \alpha_{\text{FN}})$. The deferral rate is their prevalence-weighted sum, and it vanishes exactly when the curve reaches the corner (Fig. 2b). The problem is binary set-valued classification with class-specific error levels (Sadinle et al., 2019), with DEF as the ambiguous set $\{0, 1\}$, and part (a) is the deferral-rate form of its oracle. Part (b) extends the likelihood-ratio characterization of optimal selective classification to our three-way setting (Heng & Soh, 2026). Part (c) follows directly from the nesting of the available information, because every rule based on X_S is also available when X_T is observed for $S \subseteq T$.

Posterior aggregation weakly dominates fixed aggregation rules, and Proposition 5 in Section A.3 shows that the gap can be strict. Consider a Gaussian model for the standardized detector evidence $Z = (Z_1, \ldots, Z_M)$, where larger coordinates indicate stronger evidence of hallucination and $Z \mid Y = y \sim \mathcal{N}(y\mu, \Sigma)$. Here, μ is the vector of class-conditional mean shifts, and Σ is the common within-class covariance matrix. Posterior aggregation achieves the optimal Mahalanobis separation $(\mu^\top \Sigma^{-1} \mu)^{1/2}$, whereas unweighted averaging of independent detectors achieves only $\|\mu\|_1 / \sqrt{M}$. Consequently, when averaging has a positive deferral rate, adding an uninformative detector leaves the posterior optimum unchanged but strictly increases the deferral of averaging. For M exchangeable detectors with common signal strength d and pairwise correlation ρ, even the optimal separation converges to $d/\sqrt{\rho}$ (Fig. 6), explaining why the gains saturate as similar detectors are added to a correlated pool.

With η estimated from D_{tr} and thresholds calibrated on finite data, the realized deferral exceeds $U_{[M]}^\star$. Theorem 3 splits the excess into three terms, and Section A.4 contains the proof.

Theorem 3 (Deferral bound)**.** *Under the sampling assumptions of Theorem 1, fix a score $\hat{s}$ fitted independently of $\mathcal{D}_{\text{cal}}$, run Algorithm 1 with tie-breaking, and assume both thresholds are finite. Let $\hat{a}_{\text{FP}}$ and $\hat{a}_{\text{FN}}$ be the population error rates of the thresholds $\hat{\lambda}_2$ and $\hat{\lambda}_1$, $a' = \min(\hat{a}_{\text{FP}}, \alpha_{\text{FP}})$, $b' = \min(\hat{a}_{\text{FN}}, \alpha_{\text{FN}})$, and let $t_{\text{FP}}, t_{\text{FN}} \in [0, 1]$ be the posterior values at which the randomized oracle rules on $\eta = \eta_{[M]}$ split at levels a' and b'. Then, almost surely,*

$$U(h) \;\leq\; \underbrace{U_{[M]}^\star(\alpha_{\text{FP}}, \alpha_{\text{FN}})}_{\textit{information}} + \underbrace{\text{Reg}_{\hat{s}}(\hat{a}_{\text{FP}}, \hat{a}_{\text{FN}})}_{\textit{aggregation regret}} + \underbrace{\omega}_{\textit{unused budget}}, \tag{3}$$

$$\omega \;\leq\; \frac{\pi_0(\alpha_{\text{FP}} - \hat{a}_{\text{FP}})_+}{1 - t_{\text{FP}}} + \frac{\pi_1(\alpha_{\text{FN}} - \hat{a}_{\text{FN}})_+}{t_{\text{FN}}}, \tag{4}$$

where $\text{Reg}_{\hat{s}} = U_{\hat{s}}^\star - U_{[M]}^\star \geq 0$ is random through $(\hat{a}_{\text{FP}}, \hat{a}_{\text{FN}})$, $\omega = U_{[M]}^\star(a', b') - U_{[M]}^\star(\alpha_{\text{FP}}, \alpha_{\text{FN}})$ is the exact cost of the unused budget, and a ratio is read as 0 if its numerator is 0 and as $+\infty$ if only its denominator is 0.

Corollary 4 (Law of the unused budget). *In the setting of Theorem 3, conditionally on* (n_0, n_1) *with* $k_0 \le n_0$ *and* $k_1 \le n_1$, $\hat{a}_{\mathrm{FP}} \sim \mathrm{Beta}(r_0, k_0)$ *and* $\hat{a}_{\mathrm{FN}} \sim \mathrm{Beta}(r_1, k_1)$ *independently, with* $r_y = n_y - k_y + 1$. *Hence, with* I_x *the regularized incomplete Beta function and* B_q^{-1} *the Beta q-quantile,* $\mathbb{E}[(\alpha_y - \hat{a}_y)_+] = \alpha_y I_{\alpha_y}(r_y, k_y) - r_y I_{\alpha_y}(r_y + 1, k_y)/(n_y + 1)$, *and with probability at least* $1 - \delta'$ *both* $\alpha_y - \hat{a}_y \le \gamma_y := \alpha_y - B^{-1}_{\delta'/2}(r_y, k_y)$ *(subscripts* y *index FP with class* 0 *and FN with class* 1*). These laws describe the raw thresholds, and after a crossing the final rule defers nothing. With Theorem 1(a), validity and the* γ *bounds hold jointly with probability at least* $1 - \delta - \delta'$.

Each term is tied to one design choice. The detector pool fixes the information term, which never increases as detectors are added (Proposition 2c). The aggregator fixes the regret, which more data cannot remove for a fixed rule. We do not bound the regret for a specific learner, and Section A.4 gives only a sufficient condition for it to be small. The number of calibration labels fixes the unused budget, the price of certifying at level δ, which for fixed tolerances is of order $\sqrt{\alpha(1-\alpha)\log(1/\delta)/n_y}$ as n_y grows (Corollary 4 gives its exact expectation at any n_y), and whose deferral cost depends on the ROC slopes through (4). For our calibration sizes, Corollary 4 puts the expected unused budget on each side at about two to three percentage points, which is what we observe in Table 12. On two synthetic populations with known posteriors, one with dependent discrete detectors and one with correlated Gaussian detectors, we computed every term exactly in Table 9: the bound (3) held in every replication with a slack of about a tenth of a point, the slope bound (4) exceeded ω by about two points, and the regret ranged from 34.7 points for the best single detector to under one point for a well-specified learner.

5 Experiments

We evaluate the proposed framework on two benchmarks, HallMark and StatRef. For HallMark[1] (Reizinger & Brendel, 2026a), we use 2,071 entries of its public splits with the corrected labels released by its authors, 825 valid references verified by them and 1,246 hallucinated references built by perturbation, LLM generation, adversarial crafting, and real-world collection (Section B). StatRef is a new benchmark of 2,180 statistics references: 1,280 verified records from four statistics journals and statistics arXiv preprints, and 900 hallucinated references generated from disjoint source records with 14 perturbation types (Section B). The benchmarks differ in field, citation conventions, and prevalence.

Our experiments use eleven detectors and compare seven methods for aggregating their outputs. The detectors are DOI presence and DOI resolution (rules), bibtex-updater[2] and verify-citations[3] (tools), and seven LLM judges queried zero-shot without retrieval (Gemini 2.5 Flash, DeepSeek-V3.2, DeepSeek-R1, Qwen3-235B-A22B, Mistral Large 3, GPT-5.1, and Claude Sonnet 4.6, abbreviated in the tables). The pool is correlated and includes a near-chance check (Section C, Fig. 7). The aggregators are the best single detector (largest training AUC of ϕ_m), the best learned single detector (the boosting learner below fitted to one detector's outputs, chosen by cross-fitted deferral on the training split), the fixed rules majority vote and mean evidence, and the learned naive Bayes, logistic stacking on ϕ, and boosted stacking (depth-3 gradient-boosted trees on labels and confidences, Hastie et al., 2009). Table 5 adds tools-only and additive boosting, feature ablations, and a Dawid–Skene label model (Dawid & Skene, 1979).

Our evaluation uses 1,000 random, unstratified 40/30/30 train/calibration/test splits. Aggregators are fit on the training split, Algorithm 1 with tie-breaking is applied to the calibration split ($\delta = 0.1$ unless stated otherwise), and FPR, FNR, and deferral are measured on the test split. To check Theorem 1(a) itself, we adapt the known-finite-population protocol of Li et al. (2022): the empirical distribution of a held-out pool (60% of the data) is the population, calibration sets are drawn i.i.d. from it without fixing the class ratio, the population risks are computed exactly, and the exact rank-based failure probability $1 - (1 - p_0)(1 - p_1)$ with $p_y = \mathbb{P}(\mathrm{Bin}(n_y, 1 - \alpha_y) \ge k_y)$ serves as reference. This checks the calibration step, not the labels.

[1] https://github.com/rpatrik96/hallmark

[2] https://github.com/rpatrik96/bibtexupdater

[3] https://github.com/vishakhpk/verify_citations

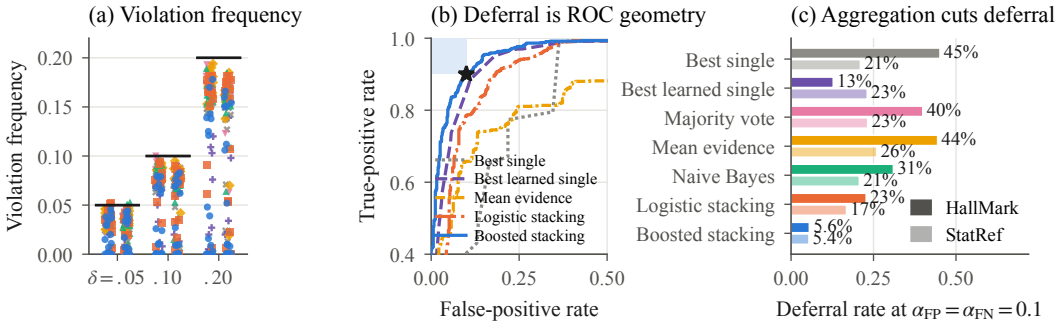

Figure 2: **(a)** Frequency with which FPR $> \alpha_{\mathrm{FP}}$ or FNR $> \alpha_{\mathrm{FN}}$ in the finite-population check (1,000 replications). Each point is one of the seven aggregators of Table 1 at one of 16 tolerance pairs in $\{.05, .10, .15, .20\}^2$ (HallMark left, StatRef right in each group), and the bars mark δ. **(b)** ROC curves on HallMark of one fitted score for each of five aggregators, fitted on a random 40% of the references and evaluated on the rest. The star is the target corner $(\alpha_{\mathrm{FP}}, 1 - \alpha_{\mathrm{FN}}) = (0.1, 0.9)$, and a curve entering the shaded region permits zero deferral. **(c)** Mean test deferral rate at $\alpha_{\mathrm{FP}} = \alpha_{\mathrm{FN}} = \delta = 0.1$.

Table 1: In-domain results (1,000 random splits, $\delta = 0.1$). FPR and FNR are mean test rates at $\alpha_{\mathrm{FP}} = \alpha_{\mathrm{FN}} = 0.1$, Def. is the mean deferral rate at the stated tolerances, W is the mean fraction of references flagged or deferred (Section 5.2), and AUC is the mean test AUC. "Raw": thresholds on ϕ_m.

		HallMark						StatRef					
			α_{FP}=α_{FN}=.10				.05		α_{FP}=α_{FN}=.10				.05
Aggregator	Kind	AUC	FPR	FNR	Def. (%)	W (%)	Def. (%)	AUC	FPR	FNR	Def. (%)	W (%)	Def. (%)
Best single	raw	.845	.070	.076	45.2	69.3	62.6	.929	.075	.071	21.0	59.9	46.0
Best learned single	learned	.946	.069	.074	12.8	61.4	37.0	.924	.075	.070	23.1	61.5	46.9
Majority vote	fixed	.863	.070	.075	39.9	77.7	61.3	.919	.075	.072	23.2	59.4	48.6
Mean evidence	fixed	.844	.070	.075	44.5	81.2	66.9	.913	.075	.072	26.0	61.3	51.8
Naive Bayes	learned	.899	.070	.075	31.0	68.7	52.0	.930	.075	.072	20.7	58.0	41.2
Logistic stacking	learned	.926	.070	.076	22.8	65.3	43.3	.946	.076	.071	16.8	56.7	33.6
Boosted stacking	learned	.966	.069	.075	**5.6**	**60.0**	**22.4**	.965	.075	.070	**5.4**	**46.8**	**25.0**

5.1 The guarantees hold for every aggregator

In the finite-population check, over both benchmarks, the 13 scores of Table 5, 16 tolerance pairs, and three values of δ, the violation frequency of the raw calibrated thresholds matches the exact rank-based failure probability (0.080 against 0.082 on average at $\delta = 0.1$, with a largest standardized deviation of 3.1, below what as many independent comparisons produce by chance, Table 11). The final rule is more conservative, because crossings at loose tolerances shrink both error regions. No configuration has a 95% Clopper–Pearson interval for its violation frequency entirely above δ, and the largest violation frequency at $\delta = 0.1$ is 0.100 (Fig. 2a, Table 10). Mean test error rates are below the tolerances in every configuration (Table 1, Fig. 4). On the same scores, per-class Bentkus and Hoeffding bounds (Bates et al., 2021; Bentkus, 2004) and our grid-based Learn-then-Test implementations are valid but defer more, up to 5.6 times as much for the Bonferroni pair grid, whereas in-expectation split conformal and plug-in thresholds exceed a tolerance in a third to three quarters of the replications, as their weaker guarantee permits (Table 13).

5.2 Aggregation reduces deferral

Tables 1 and 5 show that fixed aggregation does not reliably help. Majority vote and mean evidence stay within about five points of the best single detector on HallMark and defer more than it on StatRef, as does the unsupervised Dawid–Skene model. Additionally, re-encoding one detector explains much of the gain on HallMark: the best learned single detector, in every replication bibtex-updater, whose confidences are not ordered by its error rate, defers 12.8% of the HallMark references against 45.2% for the raw detector, whereas on StatRef re-encoding does not help.

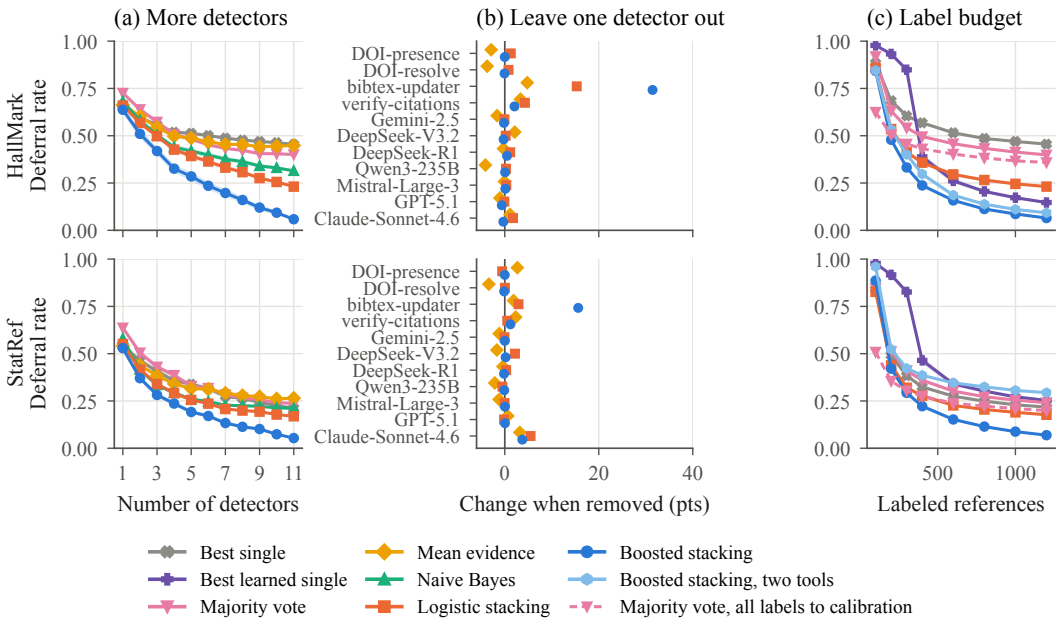

Figure 3: Which detectors and how many labels ($\alpha_{\mathrm{FP}} = \alpha_{\mathrm{FN}} = \delta = 0.1$, 300 replications). **(a)** Deferral rate for random, non-nested detector subsets of each size, with bands of ± 2 Monte Carlo standard errors. **(b)** Change in deferral when one detector is removed from the full set. Paired ± 2 standard errors are below one point and hidden by the markers. **(c)** Deferral rate when N labels are split equally between training and calibration, with bands as in (a). The dashed line gives all N labels to calibrating majority vote. At $N = 100$ a class has fewer than the 29 calibration references needed for a finite threshold, so the rule never flags on HallMark and never accepts on StatRef.

Combining detectors helps beyond that once the learner is flexible enough. At $\alpha_{\mathrm{FP}} = \alpha_{\mathrm{FN}} = \delta = 0.1$, boosted stacking of all eleven detectors defers 5.6% of the references on HallMark and 5.4% on StatRef, against 12.8% and 21.0% for the stronger single-detector baseline (learned on HallMark, raw on StatRef). This is a 2.3 to 3.9-fold reduction at the same certified error rates, and boosted stacking defers less than that baseline in 96% of the HallMark splits and in 100% of the StatRef splits, so on HallMark the re-encoded tool alone occasionally does as well (Table 7). Fig. 8 illustrates one split. Naive Bayes and logistic stacking sit in between, and additive boosting recovers most of the gap between logistic and boosted stacking, so nonlinear transformations of individual outputs account for most of the remaining gain (Table 5). PAC-Wrap would remove deferral differently, by loosening both tolerances in steps of 0.01 until they strictly cross, to 0.13 for boosted stacking and to between 0.16 and 0.28 for single detectors on HallMark (Table 14).

On the finite population, the oracle deferral of boosted stacking at the nominal levels is close to zero on both benchmarks, so almost all of its realized deferral is a calibration residual, whereas for the raw best detector the regret relative to the best score in the pool is at least as large (Table 8), the pattern that Theorem 3 describes.

Deferral counts the references that need a full check. If flagged references are also checked in full, the reviewed fraction is $W = U + P(\textsc{hal}) = 1 - P(\textsc{val})$. Because a large fraction of the references is hallucinated (60% and 41%), whenever the FNR constraint holds at least $\pi_1(1 - \alpha_{\mathrm{FN}})$ of them, about 54% on HallMark and 37% on StatRef, must be flagged or deferred whatever the score. Against the stronger single-detector baseline, W therefore falls by only 2% on HallMark and 22% on StatRef, against 57% and 74% for deferral. If confirming a flag costs half a full check, the cost falls by 12% and 35%, and at a prevalence of 5% W falls by 18% and 54% (Table 6).

5.3 Which detectors, and how many labels?

The aggregation gain depends heavily on bibtex-updater. Without bibtex-updater, boosted stacking defers 37.3% of the HallMark references, well above the re-encoded detector though still below the raw one, and on StatRef it defers about as much as the best single detector (Fig. 3b). On HallMark

the two verification tools alone, which need no LLM queries, come within a few points of the full pool, whereas the seven judges alone defer several times more, and on StatRef neither family alone comes close to the full pool (Table 18).

Tables 15 to 17 and Section 5.4 examine four potential explanations for the HallMark gain. First, HallMark's valid entries were verified with database lookups similar to those the tools perform. The 52 entries that its authors relabeled from hallucinated to valid, mostly LLM-generated citations of real papers, are flagged by verify-citations and bibtex-updater far more often than other valid entries (94% and 54% against 41% and 34%). Under the corrected labels these verdicts are false positives that the FPR budget must absorb, which is why the correction raises the deferral of boosted stacking (by 0.7 points on average in the paired comparison of Table 15), and it still flags 20% of these entries at test time. Second, a pre-screening step shared by the tools makes DOI resolution, verify-citations, and bibtex-updater emit identical outputs on 266 entries (Section C). It catches every future-dated entry, and it flags all 35 valid arXiv preprints of 2026, whose versioned DOIs do not resolve, so boosted stacking falsely flags 47% of this small cohort against 4% of the older valid references, a disparity that the class-wide guarantee permits. Third, 33% of the test references share a title with a training reference, because perturbed variants of one publication can fall on both sides of a random split. With splits drawn over title groups instead, boosted stacking defers 5.4% against 12.4% for the best learned single detector, so the advantage survives, although title groups only approximate independent publications. Fourth, the HallMark-trained stacker does not transfer to StatRef (Section 5.4). The gains are therefore within-benchmark gains for this detector pool, and they are sensitive to label noise: flipping 52 further hallucinated labels to valid at random raises the deferral of boosted stacking by 12.2 points, although boosted stacking of all detectors or of the two tools defers least in every scenario.

More detectors keep helping learned aggregators, whereas the gains of the fixed rules flatten after about six (Fig. 3a), removing the near-chance DOI check lowers the deferral of mean evidence by three to four points, as Proposition 5 predicts, while no removal lowers that of the stackers by more than one point, and among the LLM judges alone the returns diminish quickly (Fig. 5).

Fig. 3c shows that aggregation needs a few hundred labels. When N labels are shared between training and calibration, calibrating majority vote on all of them is best at the smallest budgets, and boosted stacking defers least of all the methods in Fig. 3c, including the best learned single detector and the two-tool stacker, from $N = 200$ on HallMark and from about $N = 300$ on StatRef. Strict tolerances are costly: at $\alpha_{\mathrm{FP}} = 0.01$ a finite flagging threshold needs more valid calibration references than a HallMark calibration split contains, so the rule never flags there, whereas on StatRef boosted stacking defers about a quarter (Table 20).

5.4 Deployment under domain shift

Table 19 moves from computer-science to statistics references and back. Calibrated on HallMark and deployed on StatRef, boosted stacking has an exact FPR of .186 at a tolerance of 0.1 and violates the constraint in 98% of the replications, and in the reverse direction every aggregator violates it in 40% to 100% of the replications. Recalibration on representative target labels restores control for every aggregator and every label budget, with violation frequencies at or below δ up to Monte Carlo error. The efficiency of flexible models does not transfer, however. After recalibration on 800 StatRef labels, the HallMark-trained boosted stacker defers more than naive Bayes, which is consistent with a mismatch between the source and target posteriors, the regret term of Theorem 3 (its AUC on StatRef is 0.923 against 0.965 for the score trained there, and Table 19 reports empirical estimates of the Kolmogorov distances of Proposition 6). Pooling source and target labels also hurts the learned stackers, and giving the two equal total weight removes most of the loss.

6 Related work

Hallucinated references are documented across LLMs and fields (Walters & Wilder, 2023; Xu et al., 2026; Linardon et al., 2025) and in published papers (Ansari, 2026; Sakai et al., 2026). Detectors check self-consistency (Agrawal et al., 2024), existence and metadata consistency against bibliographic records (Abbonato, 2026; Rao & Callison-Burch, 2026; Shi et al., 2026), or whether cited URLs resolve (Rao et al., 2026). **Hallucination detection** more broadly uses retrieval (Min

et al., 2023; Gao et al., 2023), self-consistency (Manakul et al., 2023), uncertainty (Kuhn et al., 2023; Farquhar et al., 2024), or multiple LLMs (Feng et al., 2024; Verga et al., 2024), typically without error guarantees.

Conformal prediction and risk control (Angelopoulos & Bates, 2023; Bates et al., 2021; Angelopoulos et al., 2024a; 2025) have been applied to LLM factuality (Mohri & Hashimoto, 2024; Cherian et al., 2024), abstention (Abbasi-Yadkori et al., 2024), generation (Quach et al., 2024; Schuster et al., 2022), and trusting model outputs (Gui et al., 2024), with conditional (Gibbs et al., 2025), class-conditional (Ding et al., 2023), and shift-aware (Tibshirani et al., 2019; Barber et al., 2023) variants. Conformal triage (Angelopoulos et al., 2024b) and cascaded LLM judges (Jung et al., 2025) are three-way rules with guarantees on the decisions they do not escalate.

Abstention, deferral, and Neyman–Pearson classification. Classification with a reject option (Geifman & El-Yaniv, 2017) and learning to defer (Mozannar & Sontag, 2020) trade errors against coverage, and conformal versions certify the error rate among accepted predictions (Hallberg Szabadváry et al., 2025). Set-valued classification admits several objectives (Chzhen et al., 2021), among them a bound on the expected set size (Denis & Hebiri, 2017) and class-specific coverage (Sadinle et al., 2019), whose binary case with $\{0, 1\}$ as the deferral set is exactly our problem: we build on its oracle, and its split-calibrated form is Algorithm 1 with the ranks of Theorem 1(b). Selective-risk and expected-size objectives differ from ours, so we do not run them as baselines. PAC-Wrap (Li et al., 2022) certifies both errors of one anomaly detector, and we start from its calibration step. Neyman–Pearson classification (Tong et al., 2018) controls one class-conditional error.

7 Discussion and limitations

We have proposed a framework for detecting hallucinated references in LLM-generated text with statistical guarantees on both the false positive rate and the false negative rate. The ensemble score combines multiple heterogeneous verification signals into a single continuous measure, and the calibrated three-way decision rule certifies a dual-risk guarantee that holds with high probability. The two components do different work: calibration is what makes the error rates valid, and it does so for any score; while aggregation is what makes the rule efficient, because only a well-separated aggregate score can be certified with few deferrals.

Some limitations point to natural directions for future work. First, our framework treats reference validity as a binary variable ($Y = 0$ for valid, $Y = 1$ for hallucinated) and returns a coarse three-way verdict (valid, deferred, or hallucinated) without localizing which bibliographic field is erroneous. In practice, a reference may carry a minor title typo, an incorrect publication year, or a misattributed venue while the underlying work genuinely exists (Rao & Callison-Burch, 2026). Our system cannot distinguish such partially correct citations from wholly fabricated ones. A natural extension is a multi-dimensional risk framework that decomposes the assessment across individual metadata fields (title, authors, year, venue, DOI), each with its own risk tolerance, enabling fine-grained error localization. Second, our method verifies only whether a cited reference exists and does not assess whether the cited work actually supports the claim for which it is invoked. Detecting unsupported or misrepresented citations is a complementary problem that requires natural-language inference over full texts (Byun et al., 2024) and falls outside the scope of the present work.

AI use statement

We generated part of the hallucinated references in the benchmarks with LLMs (Section B) and ran seven LLM judges as detectors whose outputs feed our method (Section C). We used generative AI tools to assist in refining the mathematical proofs, to suggest designs for the experiments, to draft code for the implementations, and to polish the manuscript. All AI-suggested references were manually verified against primary bibliographic records, and all AI-assisted proofs and code were independently checked by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

Reproducibility statement

Algorithm 1 specifies the calibration step, and Section A contains complete proofs. Section C lists aggregator hyperparameters, data splits, random seeds, tie-breaking, baselines, software versions, and evaluation protocols, and Section B describes the construction of StatRef and the HallMark snapshot we used. The supplementary material contains the detector outputs and labels for both benchmarks, and the code used to run the experiments.

References

Yasin Abbasi-Yadkori, Ilja Kuzborskij, David Stutz, András György, Adam Fisch, Arnaud Doucet, Iuliya Beloshapka, Wei-Hung Weng, Yao-Yuan Yang, Csaba Szepesvári, Ali Taylan Cemgil, and Nenad Tomasev. Mitigating LLM hallucinations via conformal abstention. *arXiv preprint arXiv:2405.01563*, 2024.

Diletta Abbonato. CheckIfExist: Detecting citation hallucinations in the era of AI-generated content. *arXiv preprint arXiv:2602.15871*, 2026.

Ayush Agrawal, Mirac Suzgun, Lester Mackey, and Adam Tauman Kalai. Do language models know when they're hallucinating references? In *Findings of the Association for Computational Linguistics: EACL 2024*, pp. 912–928, 2024.

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

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

Anastasios N. Angelopoulos, Stuart R. Pomerantz, Synho Do, Stephen Bates, Christopher P. Bridge, Daniel C. Elton, Michael H. Lev, R. Gilberto González, Michael I. Jordan, and Jitendra Malik. Conformal triage for medical imaging AI deployment. *medRxiv preprint 2024.02.09.24302543*, 2024b.

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

Samar Ansari. Compound deception in elite peer review: A failure mode taxonomy of 100 fabricated citations at NeurIPS 2025. *arXiv preprint arXiv:2602.05930*, 2026.

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816–845, 2023.

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

Vidmantas Bentkus. On Hoeffding's inequalities. *The Annals of Probability*, 32(2):1650–1673, 2004.

Courtni Byun, Piper Vasicek, and Kevin Seppi. This reference does not exist: An exploration of LLM citation accuracy and relevance. In *Proceedings of the Third Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, pp. 28–39, 2024.

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

Evgenii Chzhen, Christophe Denis, Mohamed Hebiri, and Titouan Lorieul. Set-valued classification – overview via a unified framework. *arXiv preprint arXiv:2102.12318*, 2021.

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

Christophe Denis and Mohamed Hebiri. Confidence sets with expected sizes for multiclass classification. *Journal of Machine Learning Research*, 18(102):1–28, 2017.

Tiffany Ding, Anastasios N. Angelopoulos, Stephen Bates, Michael I. Jordan, and Ryan J. Tibshirani. Class-conditional conformal prediction with many classes. In *Advances in Neural Information Processing Systems*, 2023.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024.

Shangbin Feng, Weijia Shi, Yike Wang, Wenxuan Ding, Vidhisha Balachandran, and Yulia Tsvetkov. Don't hallucinate, abstain: Identifying LLM knowledge gaps via multi-LLM collaboration. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14664–14690, 2024.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Y. Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. RARR: Researching and revising what language models say, using language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16477–16508, 2023.

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

Isaac Gibbs, John J. Cherian, and Emmanuel J. Candès. Conformal prediction with conditional guarantees. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 87(4): 1100–1126, 2025.

Yu Gui, Ying Jin, and Zhimei Ren. Conformal alignment: Knowing when to trust foundation models with guarantees. In *Advances in Neural Information Processing Systems*, 2024.

Johan Hallberg Szabadváry, Tuwe Löfström, Ulf Johansson, Cecilia Sönströd, Ernst Ahlberg, and Lars Carlsson. Classification with reject option: Distribution-free error guarantees via conformal prediction. *Machine Learning with Applications*, 20:100664, 2025.

Trevor Hastie, Robert Tibshirani, and Jerome H. Friedman. *The elements of statistical learning: Data mining, inference, and prediction*. Springer, 2nd edition, 2009.

Alvin Heng and Harold Soh. Know when to abstain: Optimal selective classification with likelihood ratios. In *The Fourteenth International Conference on Learning Representations*, 2026.

Jaehun Jung, Faeze Brahman, and Yejin Choi. Trust or escalate: LLM judges with provable guarantees for human agreement. In *The Thirteenth International Conference on Learning Representations*, 2025.

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

Shuo Li, Xiayan Ji, Edgar Dobriban, Oleg Sokolsky, and Insup Lee. PAC-Wrap: Semi-supervised PAC anomaly detection. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 945–955, 2022.

Jake Linardon, Hannah K. Jarman, Zoe McClure, Cleo Anderson, Claudia Liu, and Mariel Messer. Influence of topic familiarity and prompt specificity on citation fabrication in mental health research using large language models: Experimental study. *JMIR Mental Health*, 12:e80371, 2025.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9004–9017, 2023.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12076–12100, 2023.

Christopher Mohri and Tatsunori Hashimoto. Language models with conformal factuality guarantees. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 36029–36047, 2024.

Hussein Mozannar and David Sontag. Consistent estimators for learning to defer to an expert. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7076–7087, 2020.

Sangdon Park, Osbert Bastani, Nikolai Matni, and Insup Lee. PAC confidence sets for deep neural networks via calibrated prediction. In *International Conference on Learning Representations*, 2020.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011.

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

Delip Rao and Chris Callison-Burch. BibTeX citation errors in scientific publishing agents: Evaluation and mitigation. In *Third Conference on Language Modeling*, 2026.

Delip Rao, Eric Wong, and Chris Callison-Burch. Detecting and correcting reference hallucinations in commercial LLMs and deep research agents. *arXiv preprint arXiv:2604.03173*, 2026.

Patrik Reizinger and Wieland Brendel. HALLMARK: Diagnosing three failure modes in LLM citation verifiers. *arXiv preprint arXiv:2607.18360*, 2026a.

Patrik Reizinger and Wieland Brendel. Skills, benchmarks, and verification are what AI-assisted research needs. In *Post-AGI Science and Society Workshop at ICLR 2026*, 2026b.

Mauricio Sadinle, Jing Lei, and Larry Wasserman. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114(525):223–234, 2019.

Yusuke Sakai, Hidetaka Kamigaito, and Taro Watanabe. HalluCitation matters: Revealing the impact of hallucinated references with 300 hallucinated papers in ACL conferences. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 47295–47376, 2026.

Tal Schuster, Adam Fisch, Jai Gupta, Mostafa Dehghani, Dara Bahri, Vinh Q. Tran, Yi Tay, and Donald Metzler. Confident adaptive language modeling. In *Advances in Neural Information Processing Systems*, 2022.

Kaiwen Shi, Weixiang Sun, Zheyuan Zhang, Lichao Sun, Nitesh V. Chawla, and Yanfang Ye. CiteAudit: You cited it, but did you read it? A benchmark for verifying scientific references in the LLM era. *arXiv preprint arXiv:2602.23452*, 2026.

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

Xin Tong, Yang Feng, and Jingyi Jessica Li. Neyman–Pearson classification algorithms and NP receiver operating characteristics. *Science Advances*, 4(2):eaao1659, 2018.

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

William H. Walters and Esther Isabelle Wilder. Fabrication and errors in the bibliographic citations generated by ChatGPT. *Scientific Reports*, 13(1):14045, 2023.

Zuyao Xu, Yuqi Qiu, Lu Sun, Fasheng Miao, Fubin Wu, Xiang Li, Xinyi Wang, Haozhe Lu, Zhengze Zhang, Yuxin Hu, Jialu Li, Luo Jin, Feng Zhang, Rui Luo, Xinran Liu, Yingxian Li, and Jiaji Liu. GhostCite: A large-scale analysis of citation validity in the age of large language models. *arXiv preprint arXiv:2602.06718*, 2026.

A PROOFS

Throughout, for a score g write $G_0(t) = P(g(R) \le t \mid Y = 0)$ and $G_1(t) = P(g(R) < t \mid Y = 1)$, so that the rule (1) has $\mathrm{FPR} = 1 - G_0(\lambda_2)$ and $\mathrm{FNR} = G_1(\lambda_1)$. G_0 is right-continuous and G_1 is left-continuous.

A.1 PROOF OF THEOREM 1

Part (a), valid class. Fix n_0 and let $S_1, \dots, S_{n_0}$ be the valid calibration scores, i.i.d. with distribution function G_0. Let $q = \inf\{t : 1 - G_0(t) \le \alpha_{\mathrm{FP}}\}$. By right-continuity $1 - G_0(q) \le \alpha_{\mathrm{FP}}$, and $1 - G_0(t) > \alpha_{\mathrm{FP}}$ for all $t < q$. Since $\mathrm{FPR}(\lambda) = 1 - G_0(\lambda)$ is non-increasing, $\mathrm{FPR}(\hat{\lambda}_2) > \alpha_{\mathrm{FP}}$ implies $\hat{\lambda}_2 < q$, that is, $S_{(k_0)} < q$, which means that at least k_0 scores are strictly below q. Each S_i is below q with probability $p = \lim_{t \uparrow q} G_0(t) \le 1 - \alpha_{\mathrm{FP}}$. Hence

$$\mathbb{P}\big(\mathrm{FPR}(\hat{\lambda}_2) > \alpha_{\mathrm{FP}}\big) \le \mathbb{P}\big(\mathrm{Bin}(n_0, p) \ge k_0\big) \le \mathbb{P}\big(\mathrm{Bin}(n_0, 1 - \alpha_{\mathrm{FP}}) \ge k_0\big) \le \delta_0.$$

If no k_0 exists, $\hat{\lambda}_2 = +\infty$ and $\mathrm{FPR} = 0$.

Part (a), hallucinated class. Let $T_1, \dots, T_{n_1}$ be the hallucinated scores and $q' = \sup\{t : G_1(t) \le \alpha_{\mathrm{FN}}\}$. By left-continuity $G_1(q') \le \alpha_{\mathrm{FN}}$, and $G_1(t) > \alpha_{\mathrm{FN}}$ for $t > q'$. Since $\mathrm{FNR}(\lambda) = G_1(\lambda)$ is non-decreasing, $\mathrm{FNR}(\hat{\lambda}_1) > \alpha_{\mathrm{FN}}$ implies $\hat{\lambda}_1 = T_{(n_1 - k_1 + 1)} > q'$, i.e., at least k_1 scores are strictly above q'. Each T_i exceeds q' with probability $1 - \lim_{t \downarrow q'} G_1(t) \le 1 - \alpha_{\mathrm{FN}}$, because $G_1(t) > \alpha_{\mathrm{FN}}$ for $t > q'$. As before, the probability is at most $\mathbb{P}(\mathrm{Bin}(n_1, 1 - \alpha_{\mathrm{FN}}) \ge k_1) \le \delta_0$.

Combination. Given the labels, the two samples are independent, so both events $E_0 = \{\mathrm{FPR}(\hat{\lambda}_2) \le \alpha_{\mathrm{FP}}\}$ and $E_1 = \{\mathrm{FNR}(\hat{\lambda}_1) \le \alpha_{\mathrm{FN}}\}$ hold with probability at least $(1 - \delta_0)^2 = 1 - \delta$. If $\hat{\lambda}_1 \le \hat{\lambda}_2$, the rule is (1) and the claim follows. Otherwise, the single cut $\tau \in [\hat{\lambda}_2, \hat{\lambda}_1)$ flags $\{s > \tau\} \subseteq \{s > \hat{\lambda}_2\}$ and accepts $\{s \le \tau\} \subseteq \{s < \hat{\lambda}_1\}$, so its FPR and FNR are at most those in E_0 and E_1.

Part (b). Let S_{n_0+1} be the score of a fresh valid reference. The $n_0 + 1$ scores are exchangeable. The event $\{S_{n_0+1} > S_{(k_0)}\}$ implies that S_{n_0+1} strictly exceeds at least k_0 of the other scores. Breaking ties uniformly at random, the rank of S_{n_0+1} is uniform on $\{1, \dots, n_0 + 1\}$, and strict exceedance of k_0 values forces rank at least $k_0 + 1$. Thus $\mathbb{E}[\mathrm{FPR}(\hat{\lambda}_2)] = \mathbb{P}(S_{n_0+1} > S_{(k_0)}) \le (n_0 + 1 - k_0)/(n_0 + 1) \le \alpha_{\mathrm{FP}}$ for $k_0 = \lceil (n_0 + 1)(1 - \alpha_{\mathrm{FP}}) \rceil$ (and $\mathrm{FPR} = 0$ if $k_0 > n_0$), and $\mathbb{E}[\mathrm{FPR}(h)] \le \mathbb{E}[\mathrm{FPR}(\hat{\lambda}_2)]$ because the single cut only shrinks the flagged region. The FNR side is symmetric.

Part (c). The proofs of (a) and (b) involve the deployment distribution only through the class-conditional distributions of $s(R)$, which are unchanged by assumption.

Tie-breaking. With tie-breaking, apply the argument to the score $\tilde{s}(R, \xi) = (s(R), \xi)$ ordered lexicographically, where ξ is independent of everything else. It does not depend on $\mathcal{D}_{\mathrm{cal}}$, and given the labels the calibration pairs are independent with the class-conditional law of $\tilde{s}$, so (a)–(c) hold for FPR and FNR averaged over the noise of the evaluated reference, with probabilities taken over $\mathcal{D}_{\mathrm{cal}}$ and its noise. A single cut stored as a pair between the two threshold pairs flags a subset of $\{(s, \xi) > \hat{\lambda}_2\}$ and accepts a subset of $\{(s, \xi) < \hat{\lambda}_1\}$ in the lexicographic order, so the combination step applies unchanged. □

A.2 PROOF OF PROPOSITION 2

Write a rule based on X_S as $\psi = (\psi_0, \psi_1)$ with measurable $\psi_0, \psi_1 : \mathcal{X}_S \to [0, 1]$ and $\psi_0 + \psi_1 \le 1$. Then $\mathrm{FPR} = \mathbb{E}_0[\psi_1]$, $\mathrm{FNR} = \mathbb{E}_1[\psi_0]$, and $1 - U = \mathbb{E}[\psi_0] + \mathbb{E}[\psi_1]$, where $\mathbb{E}_y$ denotes expectation given $Y = y$. Recall that $0 < \pi_1 < 1$.

Lower bound. For a class $\mathcal{C}$ of rules, let $\beta^1(a) = \sup\{\mathbb{E}_1[\psi_1] : \mathbb{E}_0[\psi_1] \le a\}$ and $\beta^0(b) = \sup\{\mathbb{E}_0[\psi_0] : \mathbb{E}_1[\psi_0] \le b\}$ over components of rules in $\mathcal{C}$. For any feasible rule in $\mathcal{C}$,

$$\mathbb{E}[\psi_1] = \pi_0\mathbb{E}_0[\psi_1] + \pi_1\mathbb{E}_1[\psi_1] \le \pi_0\alpha_{\mathrm{FP}} + \pi_1\beta^1(\alpha_{\mathrm{FP}}), \qquad \mathbb{E}[\psi_0] \le \pi_1\alpha_{\mathrm{FN}} + \pi_0\beta^0(\alpha_{\mathrm{FN}}).$$

Hence $U \ge 1 - \pi_0\alpha_{\mathrm{FP}} - \pi_1\beta^1(\alpha_{\mathrm{FP}}) - \pi_1\alpha_{\mathrm{FN}} - \pi_0\beta^0(\alpha_{\mathrm{FN}})$, and $U \ge 0$, which gives "$\ge$" in (2).

Attainment for threshold rules on g. Let $t = \inf\{u : P(g > u \mid Y = 0) \le \alpha_{\mathrm{FP}}\}$ and $\psi_1^\star = \mathbf{1}\{g > t\} + \theta\mathbf{1}\{g = t\}$ with the *largest* admissible $\theta \in [0,1]$, i.e., $\theta = \min\{1, (\alpha_{\mathrm{FP}} - P(g > t \mid Y = 0))/P(g = t \mid Y = 0)\}$ and $\theta = 1$ when $P(g = t \mid Y = 0) = 0$. By right-continuity $P(g > t \mid Y = 0) \le \alpha_{\mathrm{FP}}$, and $P(g > u \mid Y = 0) > \alpha_{\mathrm{FP}}$ for every $u < t$, so $P(g \ge t \mid Y = 0) \ge \alpha_{\mathrm{FP}}$ and hence $\mathbb{E}_0[\psi_1^\star] = \alpha_{\mathrm{FP}}$. A randomized upper-threshold rule $\mathbf{1}\{g > u\} + \theta_u\mathbf{1}\{g = u\}$ with false-positive rate at most α_{FP} has $u \ge t$, and if $u = t$ then $\theta_u \le \theta$. Its flagged region is therefore contained in that of $\psi_1^\star$, and $\mathbb{E}_1[\psi_1^\star] = \beta_g^1(\alpha_{\mathrm{FP}})$. (Taking the largest θ matters: with $g \mid Y = 0 \sim \mathrm{Unif}(0,1)$, $g \mid Y = 1 \equiv 0.9$, and $\alpha_{\mathrm{FP}} = 0.1$, the choice $\theta = 0$ would have power 0, while $\theta = 1$ has power 1.) Symmetrically, let $t' = \sup\{u : P(g < u \mid Y = 1) \le \alpha_{\mathrm{FN}}\}$ and $\psi_0^\star = \mathbf{1}\{g < t'\} + \theta'\mathbf{1}\{g = t'\}$ with the largest admissible θ', so that $\mathbb{E}_1[\psi_0^\star] = \alpha_{\mathrm{FN}}$ and $\mathbb{E}_0[\psi_0^\star] = \beta_g^0(\alpha_{\mathrm{FN}})$. There are two cases. (i) If $t' < t$, or $t' = t$ and $\theta + \theta' \le 1$, then $\psi_0^\star + \psi_1^\star \le 1$ pointwise and the pair is a feasible rule attaining the lower bound, which is then non-negative. (ii) Otherwise $\psi_0^\star + \psi_1^\star \ge 1$ pointwise (check the regions $g < t$, $g = t$, $t < g < t'$, $g = t'$, $g > t'$). The rule $(1 - \psi_1^\star, \psi_1^\star)$ never defers, has $\mathrm{FPR} \le \alpha_{\mathrm{FP}}$ and $\mathrm{FNR} = \mathbb{E}_1[1 - \psi_1^\star] \le \mathbb{E}_1[\psi_0^\star] \le \alpha_{\mathrm{FN}}$, so it attains $U = 0$, while the bracket in (2) equals $1 - \mathbb{E}[\psi_0^\star] - \mathbb{E}[\psi_1^\star]$ at the binding levels and is at most 0. This proves (a).

Part (b). Let $\nu = P_{0,S} + P_{1,S}$ dominate the class-conditional laws of X_S, with densities p_0, p_1, and define the extended likelihood ratio $L = p_1/p_0 \in [0, \infty]$ (with $L = \infty$ where $p_0 = 0 < p_1$). By the Neyman–Pearson lemma, among all X_S-measurable tests with $\mathbb{E}_0[\psi] \le a$, the power $\mathbb{E}_1[\psi]$ is maximized by a randomized upper-threshold test on L. Since $\eta_S = \pi_1 L/(\pi_0 + \pi_1 L)$ (with $\eta_S = 1$ where $L = \infty$) is a strictly increasing transform of L, the optimal test is an upper-threshold test on η_S, and the argument above applied to $g = \eta_S$ constructs it. The same holds for β^0 with the classes exchanged. Hence $\beta_{\eta_S}^y$ equals the supremum over all X_S-measurable rules, and by (a) the lower bound over all such rules is attained by threshold rules on η_S, so $U_{\eta_S}^\star = U_S^\star$. Threshold rules on any aggregator g of X_S are X_S-measurable, so $U_g^\star \ge U_S^\star$.

Part (c). X_S-measurable rules are X_T-measurable, so the suprema defining β^y are at least as large for T and the bracket in (2) is at most as large. If $U_S^\star > 0$, its bracket is positive, and a strictly larger $\beta^1(\alpha_{\mathrm{FP}})$ (which enters with weight $\pi_1 > 0$) or $\beta^0(\alpha_{\mathrm{FN}})$ (weight $\pi_0 > 0$) strictly decreases the bracket, hence its positive part. □

A.3 Gaussian model of detector evidence

Proposition 5 (Averaging versus posterior aggregation)**.** *Let $\alpha_{\mathrm{FP}} + \alpha_{\mathrm{FN}} < 1$ and let the detector evidence satisfy $Z \mid Y = y \sim N(y\mu, \Sigma)$ in $\mathbb{R}^M$ with $\Sigma \succ 0$ and $\Sigma_{mm} = 1$. A linear aggregator $g_w(Z) = w^\top Z$ ($w \ne 0$) with separation $\Delta_w = w^\top\mu/(w^\top\Sigma w)^{1/2} \ge 0$ has $U_{g_w}^\star = u(\Delta_w)$, where*

$$u(\Delta) = \big[\pi_0\big(1 - \alpha_{\mathrm{FP}} - \Phi(\Delta - z_{\mathrm{FN}})\big) + \pi_1\big(1 - \alpha_{\mathrm{FN}} - \Phi(\Delta - z_{\mathrm{FP}})\big)\big]_+ ,$$

$z_{\mathrm{FP}} = \Phi^{-1}(1 - \alpha_{\mathrm{FP}})$ and $z_{\mathrm{FN}} = \Phi^{-1}(1 - \alpha_{\mathrm{FN}})$, and u is strictly decreasing on $[0, z_{\mathrm{FP}} + z_{\mathrm{FN}}]$ and zero beyond. Moreover:

- *(a) if $\mu \ne 0$, the posterior is a logistic transform of $w^{\star\top}Z$ with $w^\star = \Sigma^{-1}\mu$, which attains $\Delta^\star = (\mu^\top\Sigma^{-1}\mu)^{1/2} = \max_w \Delta_w$ (if $\mu = 0$, the posterior is constant and the smallest feasible deferral is $u(0) = 1 - \alpha_{\mathrm{FP}} - \alpha_{\mathrm{FN}}$).*
- *(b) for independent detectors ($\Sigma = I$) with $\mu \ge 0$, $\mu \ne 0$, averaging attains $\Delta_{\mathrm{avg}} = \|\mu\|_1/\sqrt{M} \le \|\mu\|_2 = \Delta^\star$. Adding an uninformative detector leaves $\Delta^\star$ unchanged but strictly increases the deferral rate of averaging if that rate was positive before, and averaging defers more than the best single detector whenever $\|\mu\|_1 < \sqrt{M}\max_m \mu_m$ and $u(\Delta_{\mathrm{avg}}) > 0$.*
- *(c) for exchangeable detectors with $\mu_m = d > 0$ and pairwise correlation $\rho \in (0,1)$, $\Delta^\star = d\sqrt{M/(1 + (M-1)\rho)}$ increases to $d/\sqrt{\rho}$, so if $d \le \sqrt{\rho}\,(z_{\mathrm{FP}} + z_{\mathrm{FN}})$, no number of such detectors reaches zero deferral.*

Proof of Proposition 5.

For $w \neq 0$ with $w^\top \mu \geq 0$ (so $w^\top \Sigma w > 0$ because $\Sigma \succ 0$), the standardized score $T = g_w(Z)/(w^\top \Sigma w)^{1/2}$ is $N(0,1)$ under $Y = 0$ and $N(\Delta_w, 1)$ under $Y = 1$. Its upper threshold at FPR α_{FP} is z_{FP}, with power $\beta^1 = \Phi(\Delta_w - z_{\mathrm{FP}})$, and its lower threshold at FNR α_{FN} is $\Delta_w - z_{\mathrm{FN}}$, with true-negative rate $\beta^0 = \Phi(\Delta_w - z_{\mathrm{FN}})$. Substituting into (2) gives $u(\Delta_w)$. With Φ' the standard normal density, the bracket has derivative $-\pi_0 \Phi'(\Delta - z_{\mathrm{FN}}) - \pi_1 \Phi'(\Delta - z_{\mathrm{FP}}) < 0$ and vanishes at $\Delta = z_{\mathrm{FP}} + z_{\mathrm{FN}}$, because $1 - \alpha_{\mathrm{FP}} = \Phi(z_{\mathrm{FP}})$ and $1 - \alpha_{\mathrm{FN}} = \Phi(z_{\mathrm{FN}})$. Since $\alpha_{\mathrm{FP}} + \alpha_{\mathrm{FN}} < 1$ gives $z_{\mathrm{FP}} + z_{\mathrm{FN}} > 0$, this proves monotonicity on $[0, z_{\mathrm{FP}} + z_{\mathrm{FN}}]$ and the zero region.

(a) The log-likelihood ratio is $\mu^\top \Sigma^{-1} z - \frac{1}{2} \mu^\top \Sigma^{-1} \mu$, so the posterior is a logistic transform of the linear discriminant $w^{\star\top} z$ with $w^\star = \Sigma^{-1} \mu$, and threshold rules on the posterior are threshold rules on $w^{\star\top} z$. By the Cauchy–Schwarz inequality, $w^\top \mu = (\Sigma^{1/2} w)^\top (\Sigma^{-1/2} \mu) \leq (w^\top \Sigma w)^{1/2} (\mu^\top \Sigma^{-1} \mu)^{1/2}$, so $\Delta_w \leq \Delta^\star$.

(b) With $\Sigma = I$, $w = \mathbf{1}$, and $\mu \geq 0$, $\Delta_{\mathrm{avg}} = \mathbf{1}^\top \mu / \sqrt{M} = \|\mu\|_1 / \sqrt{M} \leq \|\mu\|_2$. Appending $\mu_{M+1} = 0$ leaves $\|\mu\|_2$ unchanged and multiplies $\Delta_{\mathrm{avg}} > 0$ by $\sqrt{M/(M+1)} < 1$. If $u(\Delta_{\mathrm{avg}}) > 0$ before the addition, then $\Delta_{\mathrm{avg}} < z_{\mathrm{FP}} + z_{\mathrm{FN}}$ and, since u is strictly decreasing on $[0, z_{\mathrm{FP}} + z_{\mathrm{FN}}]$, the deferral rate of averaging strictly increases. The best single detector has separation $\max_m \mu_m$, which exceeds Δ_{avg} exactly when $\|\mu\|_1 < \sqrt{M} \max_m \mu_m$.

(c) The vector $\mathbf{1}$ is an eigenvector of $\Sigma = (1-\rho) I + \rho \mathbf{1}\mathbf{1}^\top$ with eigenvalue $1 + (M-1)\rho$, so $\mu^\top \Sigma^{-1} \mu = d^2 M / (1 + (M-1)\rho)$, which increases in M to d^2/ρ for $\rho \in (0,1)$. If $d/\sqrt{\rho} \leq z_{\mathrm{FP}} + z_{\mathrm{FN}}$, then $\Delta^\star < z_{\mathrm{FP}} + z_{\mathrm{FN}}$ for every finite M and $u(\Delta^\star) > 0$. □

A.4 Proof of Theorem 3 and Corollary 4

Condition on the fitted score $\hat{s}$ and on (n_0, n_1), and write $U^\star = U^\star_{[M]}$. Represent the tie-broken score by the real number $\tilde{s} = F(\hat{s}-) + \xi\, p(\hat{s})$, where F is the distribution function of $\hat{s}(X)$ and $p(v) = P(\hat{s}(X) = v)$ (any real embedding of the lexicographic order, such as rank plus ξ, works). Almost surely $\tilde{s}$ orders references as $(\hat{s}, \xi)$ does, its class-conditional distributions are atomless, and threshold rules on $\tilde{s}$ are exactly the randomized threshold rules on $\hat{s}$, so the smallest deferral rate of threshold rules on $\tilde{s}$ at levels (a, b) equals $U^\star_{\hat{s}}(a, b)$ from Proposition 2. Write $G_0(t) = P(\tilde{s} \leq t \mid Y = 0)$ and $G_1(t) = P(\tilde{s} < t \mid Y = 1)$, both continuous, and $t^{\hat{s}}_{\mathrm{FP}}(a) = \inf\{t : 1 - G_0(t) \leq a\}$, $t^{\hat{s}}_{\mathrm{FN}}(b) = \sup\{t : G_1(t) \leq b\}$. For atomless scores the feasible upper thresholds at level a form $[t^{\hat{s}}_{\mathrm{FP}}(a), \infty)$ and the feasible lower thresholds at level b form $(-\infty, t^{\hat{s}}_{\mathrm{FN}}(b)]$, so $U^\star_{\hat{s}}(a, b) = P\big(\tilde{s} \in [t^{\hat{s}}_{\mathrm{FN}}(b), t^{\hat{s}}_{\mathrm{FP}}(a)]\big)$, with $[x, y] = \emptyset$ for $x > y$.

Step 1: the calibrated rule is an oracle rule at its own error levels. If $\hat{\lambda}_1 > \hat{\lambda}_2$, then $U(h) = 0$ and (3) holds because every term on the right is non-negative ($\omega \geq 0$ since $a' \leq \alpha_{\mathrm{FP}}$, $b' \leq \alpha_{\mathrm{FN}}$). Otherwise $U(h) = P(\tilde{s} \in [\hat{\lambda}_1, \hat{\lambda}_2])$. Because $1 - G_0(\hat{\lambda}_2) = \hat{a}_{\mathrm{FP}}$, we have $t^{\hat{s}}_{\mathrm{FP}}(\hat{a}_{\mathrm{FP}}) \leq \hat{\lambda}_2$, and strict inequality would require $P(\tilde{s} \in (t, \hat{\lambda}_2] \mid Y = 0) = 0$ for some $t < \hat{\lambda}_2$. The set of points x with $P(\tilde{s} \in (t, x] \mid Y = 0) = 0$ for some rational $t < x$ is a countable union of null intervals, hence null under the class-0 law, and $\hat{\lambda}_2$ is one of the valid calibration scores, which are drawn from that law. Therefore $t^{\hat{s}}_{\mathrm{FP}}(\hat{a}_{\mathrm{FP}}) = \hat{\lambda}_2$ almost surely, and symmetrically $t^{\hat{s}}_{\mathrm{FN}}(\hat{a}_{\mathrm{FN}}) = \hat{\lambda}_1$. Hence $U(h) = U^\star_{\hat{s}}(\hat{a}_{\mathrm{FP}}, \hat{a}_{\mathrm{FN}}) = U^\star(\hat{a}_{\mathrm{FP}}, \hat{a}_{\mathrm{FN}}) + \mathrm{Reg}_{\hat{s}}(\hat{a}_{\mathrm{FP}}, \hat{a}_{\mathrm{FN}})$ almost surely, with $\mathrm{Reg}_{\hat{s}} \geq 0$ by Proposition 2(b).

Step 2: cost of the unused budget. Since $U^\star$ is non-increasing in each level and $a' \leq \hat{a}_{\mathrm{FP}}$, $b' \leq \hat{a}_{\mathrm{FN}}$, we have $U^\star(\hat{a}_{\mathrm{FP}}, \hat{a}_{\mathrm{FN}}) \leq U^\star(a', b') = U^\star(\alpha_{\mathrm{FP}}, \alpha_{\mathrm{FN}}) + \omega$, which with Step 1 proves (3). For (4), write $\beta^1 = \beta^1_\eta$ and $\beta^0 = \beta^0_\eta$. By (2) and because $x \mapsto [x]_+$ is 1-Lipschitz and non-decreasing, ω is at most the ROC modulus

$$\pi_0(\alpha_{\mathrm{FP}} - a') + \pi_1\big(\beta^1(\alpha_{\mathrm{FP}}) - \beta^1(a')\big) + \pi_1(\alpha_{\mathrm{FN}} - b') + \pi_0\big(\beta^0(\alpha_{\mathrm{FN}}) - \beta^0(b')\big).$$

The function β^1 is the Neyman–Pearson ROC curve of η, which is non-decreasing and concave because randomized tests can be mixed. Its right derivative at a' is at most the extended likelihood ratio at which the level-a' oracle test splits, $L = (\pi_0/\pi_1)\, t_{\mathrm{FP}}/(1 - t_{\mathrm{FP}})$, where t_{FP} is the corresponding posterior value, because increasing the level beyond a' adds class-0 mass only at likelihood ratios

at most L. By concavity, $\beta^1(\alpha_{\mathrm{FP}}) - \beta^1(a') \le L(\alpha_{\mathrm{FP}} - a')$, so the first two terms are at most $\pi_0(\alpha_{\mathrm{FP}} - a')\{1 + t_{\mathrm{FP}}/(1 - t_{\mathrm{FP}})\} = \pi_0(\alpha_{\mathrm{FP}} - \hat{a}_{\mathrm{FP}})_+/(1 - t_{\mathrm{FP}})$, read as $+\infty$ if $t_{\mathrm{FP}} = 1$ and $\alpha_{\mathrm{FP}} > a'$ and as 0 if $\alpha_{\mathrm{FP}} = a'$ (similarly for $t_{\mathrm{FN}} = 0$ below). Symmetrically, the validation power β^0 has right derivative at most $(\pi_1/\pi_0)(1 - t_{\mathrm{FN}})/t_{\mathrm{FN}}$ at b', and the last two terms are at most $\pi_1(\alpha_{\mathrm{FN}} - \hat{a}_{\mathrm{FN}})_+/t_{\mathrm{FN}}$. This argument does not require $\eta(X)$ to be continuous, and it proves (4).

Step 3: distribution of the unused budget. The valid calibration scores $\tilde{S}_1, \dots, \tilde{S}_{n_0}$ are i.i.d. with continuous distribution function G_0, so $G_0(\tilde{S}_i)$ are i.i.d. uniform and $G_0(\hat{\lambda}_2) = G_0(\tilde{S}_{(k_0)}) \sim \mathrm{Beta}(k_0, n_0 - k_0 + 1)$. Hence $\hat{a}_{\mathrm{FP}} = 1 - G_0(\hat{\lambda}_2) \sim \mathrm{Beta}(n_0 - k_0 + 1, k_0)$, and $\mathbb{P}(\alpha_{\mathrm{FP}} - \hat{a}_{\mathrm{FP}} > \gamma_{\mathrm{FP}}) = \mathbb{P}(\hat{a}_{\mathrm{FP}} < B^{-1}_{\delta'/2}(n_0 - k_0 + 1, k_0)) = \delta'/2$. The same holds for $\hat{a}_{\mathrm{FN}} = G_1(\hat{\lambda}_1)$ after reversing the order, and a union bound gives probability at least $1 - \delta'$ for both. Given the labels the two samples are independent, so $\hat{a}_{\mathrm{FP}}$ and $\hat{a}_{\mathrm{FN}}$ are independent as well. With $A \sim \mathrm{Beta}(r, k)$ and density $f_{r,k}$, since $z f_{r,k}(z) = \frac{r}{r+k} f_{r+1,k}(z)$,

$$\mathbb{E}[(\alpha - A)_+] = \int_0^\alpha (\alpha - z) f_{r,k}(z)\, dz = \alpha I_\alpha(r, k) - \frac{r}{r+k}\, I_\alpha(r+1, k), \qquad r + k = n_y + 1.$$

For the rate, $\mathbb{E}[\hat{a}_{\mathrm{FP}}] = (n_0 - k_0 + 1)/(n_0 + 1)$. Bernstein's inequality gives $\mathbb{P}(\mathrm{Bin}(n_0, 1 - \alpha_{\mathrm{FP}}) \ge n_0(1 - \alpha_{\mathrm{FP}} + u)) \le \delta_0$ for $u = \sqrt{2\alpha_{\mathrm{FP}}(1 - \alpha_{\mathrm{FP}}) \log(1/\delta_0)/n_0} + 2\log(1/\delta_0)/(3n_0)$, so $k_0 \le n_0(1 - \alpha_{\mathrm{FP}} + u) + 1$ and $\alpha_{\mathrm{FP}} - \mathbb{E}[\hat{a}_{\mathrm{FP}}] \le u + \alpha_{\mathrm{FP}}/(n_0 + 1) \le u + 1/n_0$. Finally $\mathbb{E}[(\alpha_{\mathrm{FP}} - \hat{a}_{\mathrm{FP}})_+] \le (\alpha_{\mathrm{FP}} - \mathbb{E}[\hat{a}_{\mathrm{FP}}])_+ + \mathbb{E}|\hat{a}_{\mathrm{FP}} - \mathbb{E}[\hat{a}_{\mathrm{FP}}]|$, and the last term is at most $\sqrt{\mathrm{Var}(\hat{a}_{\mathrm{FP}})}$ with the Beta variance $(n_0 - k_0 + 1)k_0/\{(n_0 + 1)^2(n_0 + 2)\} \le 1/(4n_0)$. Hence $\mathbb{E}[(\alpha_{\mathrm{FP}} - \hat{a}_{\mathrm{FP}})_+] \le \sqrt{2\alpha_{\mathrm{FP}}(1 - \alpha_{\mathrm{FP}}) \log(1/\delta_0)/n_0} + \sqrt{\mathrm{Var}(\hat{a}_{\mathrm{FP}})} + 2\log(1/\delta_0)/(3n_0) + 1/n_0$, the order stated in the text.

Step 4: a sufficient condition for small regret. Suppose $\eta(X)$ has a density bounded by C and $\sup_x |m(\hat{s}(x)) - \eta(x)| \le \varepsilon$ for a strictly increasing m. We show $\mathrm{Reg}_{\hat{s}} \le 4C\varepsilon$ at every level. This is not a learning guarantee for any fitted learner. Threshold rules on $\hat{s}$ and on $m(\hat{s})$ coincide for strictly increasing m, so assume $\sup_x |\hat{s}(x) - \eta(x)| \le \varepsilon$. With a bounded density, $\eta(X)$ is atomless, so no randomization is needed for η. Fix levels (a, b) and let $\tau_2 = t_{\mathrm{FP}}(a) + \varepsilon$ and $\tau_1 = t_{\mathrm{FN}}(b) - \varepsilon$, where $t_{\mathrm{FP}}(\cdot), t_{\mathrm{FN}}(\cdot)$ are the oracle thresholds on η. Then $\{\hat{s} > \tau_2\} \subseteq \{\eta > t_{\mathrm{FP}}(a)\}$ and $\{\hat{s} < \tau_1\} \subseteq \{\eta < t_{\mathrm{FN}}(b)\}$, so (τ_1, τ_2) is feasible for $\hat{s}$ at (a, b) (if $\tau_1 > \tau_2$, the single cut at τ_2 is feasible and defers nothing) and $U^\star_{\hat{s}}(a, b) \le P(\tau_1 \le \hat{s} \le \tau_2) \le P\big(t_{\mathrm{FN}}(b) - 2\varepsilon \le \eta \le t_{\mathrm{FP}}(a) + 2\varepsilon\big)$. Since $[t_{\mathrm{FN}}(b) - 2\varepsilon, t_{\mathrm{FP}}(a) + 2\varepsilon] \subseteq [t_{\mathrm{FN}}(b) - 2\varepsilon, t_{\mathrm{FN}}(b)) \cup [t_{\mathrm{FN}}(b), t_{\mathrm{FP}}(a)] \cup (t_{\mathrm{FP}}(a), t_{\mathrm{FP}}(a) + 2\varepsilon]$, the density bound gives $U^\star_{\hat{s}}(a, b) \le U^\star(a, b) + 4C\varepsilon$. □

A.5 Validity under distribution shift

Proposition 6. *Let the calibration data be drawn from P and deployment references from Q, and let $\kappa_0 = \sup_t |P(s > t \mid Y = 0) - Q(s > t \mid Y = 0)|$ and $\kappa_1 = \sup_t |P(s < t \mid Y = 1) - Q(s < t \mid Y = 1)|$ be the Kolmogorov distances between the class-conditional score distributions. Under the conditions of Theorem 1(a), with probability at least $1 - \delta$ over $\mathcal{D}_{\mathrm{cal}}$, $\mathrm{FPR}_Q(h) \le \alpha_{\mathrm{FP}} + \kappa_0$ and $\mathrm{FNR}_Q(h) \le \alpha_{\mathrm{FN}} + \kappa_1$.*

Proof. For every t, $Q(s > t \mid Y = 0) \le P(s > t \mid Y = 0) + \kappa_0$, and in particular at $t = \hat{\lambda}_2$. A single cut flags a subset of $\{s > \hat{\lambda}_2\}$ and accepts a subset of $\{s < \hat{\lambda}_1\}$ under both P and Q. The FNR side is symmetric. With tie-breaking, apply the same argument to (s, ξ). Since ξ is independent with the same law under P and Q, a randomized boundary tail is a convex combination of the two one-sided tails of s, so κ_y computed on s still bounds the difference. Combining this with Theorem 1(a) yields the stated bounds with probability at least $1 - \delta$. Since κ_y is at most the total-variation distance between the class-conditional distributions of the references, the bound is tighter than a total-variation bound. A pure prevalence shift gives $\kappa_0 = \kappa_1 = 0$. □

B Benchmarks

HallMark. We use the public splits of HallMark (development, test, and stress-test) with the outputs of all eleven detectors, 2,071 BibTeX entries in total, one fewer than the 2,072 public labels of

the current release (a stress-test canary entry is absent from our files). The labels are the corrected labels released by the benchmark authors, whose ground-truth audit relabeled 52 of these entries from hallucinated to valid (Reizinger & Brendel, 2026a): 825 valid and 1,246 hallucinated entries. According to the benchmark documentation, valid entries are verified publications, mostly DBLP records of machine learning papers from 2021–2023 together with recent arXiv preprints, and hallucinated entries come from systematic perturbation, LLM generation, adversarial crafting, and real-world collection. According to its documentation, the current release contains 2,526 entries including a hidden test split. The 825 valid entries are 738 verified publications of 2021–2023, the 52 relabeled entries (48 LLM-generated citations of real 2021–2023 papers and 4 perturbations that turned out to be real), and 35 arXiv preprints of 2026. Our detector-output files identify references by the benchmark's public keys. For the subgroup and title-disjoint analyses (Tables 16 and 17) we join them to the hallucination type, generation method, publication year, and title of the public release at commit `ddd6297` of 21 September 2026, whose labels agree with ours on all 2,071 entries (`make_metadata.py` in the supplementary code). Calibration controls errors with respect to the labels used, and Table 15 reports the results under the original labels and under additional random label flips. 54% of the valid entries have no DOI field in their BibTeX input, so missing identifiers are incomplete metadata rather than evidence of fabrication.

StatRef. We collected 2,180 verified statistics references published between 1990 and 2026: 1,379 records from *Biometrika*, *Journal of the American Statistical Association*, *The Annals of Statistics*, and *Journal of the Royal Statistical Society: Series B*, obtained through OpenAlex, and 801 arXiv preprints whose primary category is `stat.AP`, `stat.CO`, `stat.ME`, `stat.ML`, `stat.OT`, or `stat.TH`. All records were normalized to a common schema (title, authors, year, venue, DOI, identifiers), deduplicated, and validated by DOI or arXiv identifier resolution, field completeness, venue consistency, and cross-database agreement. We kept 1,280 records unchanged as valid references. The other 900 records served as sources for hallucinated references, one per source, and were removed from the valid set. With a fixed random seed, the 900 hallucinated references were allocated equally to three tiers (300 each) and, within tiers, to 14 perturbation types following the HallMark taxonomy, with the counts in Table 2. **Label semantics.** As in HallMark, a reference counts as hallucinated if an essential field is fabricated or materially inconsistent with its source record. Some types, such as partial author lists, near-miss titles, preprints cited as journal articles, and future years, can resemble benign bibliographic variation, and our labels follow the generation procedure rather than independent expert adjudication. The per-record type and tier assignment of the hallucinated entries was not retained with the detector-output files and cannot be recovered from them, so we report StatRef only in aggregate. Breakdowns by type, tier, and publication year for StatRef, and an evaluation on independently adjudicated LLM-generated references, are left to future work.

C Experimental details

Detector outputs. Every detector returns a label in $\{\text{valid, hallucinated, uncertain}\}$ and a confidence in $[0, 1]$ for every reference, and there are no missing outputs. *DOI presence* returns hallucinated (confidence 0.8) when the entry has no DOI and valid (0.6) otherwise. *DOI resolution* follows the benchmark's public baseline (`hallmark/baselines/doi_only.py`, identical in the releases up to v1.2.3): a HEAD request to doi.org that follows redirects, returning valid (0.85) on a 200 response, hallucinated (0.75) on any other status, and valid (0.5) when the DOI is missing or the request fails. Publisher landing pages that apply bot mitigation answer with other statuses after a successful doi.org redirect, which the maintainers later documented for a sample of valid entries and now treat as indeterminate. This is the likely reason why 726 of the 956 valid StatRef references with a DOI, mostly journal articles, and 187 of 380 on HallMark "fail" resolution. Our files match this mechanism (the only confidences are 0.5, 0.75, and 0.85, plus the pre-screening values below), but we do not have the execution traces and report the detector as given. *bibtex-updater* is a multi-source verification pipeline whose valid verdicts carry confidences 0.58–0.67 and hallucinated verdicts 0.29–0.95. *verify-citations* cross-references arXiv, ACL Anthology, Semantic Scholar, DBLP, and Google Scholar and returns 0.9 for valid and 0.8–0.95 for hallucinated verdicts. The LLM judges receive the BibTeX entry with a zero-shot verification prompt and return a label and a verbalized confidence, without tool use or retrieval. The signed evidence $\phi_m = \ell_m c_m$ discards the confidence of uncertain outputs, and the boosting features retain it. It is an encoding rather than a probability, because confidences need not have a common meaning across detectors. Table 3 reports native error rates, mean confidence and accuracy of decisive verdicts, and the expected calibration error

Table 2: Perturbation types used to construct the 900 hallucinated StatRef references.

Tier	Type	Count	Operation
1: explicit	Fabricated DOI	75	Replace the DOI with a non-resolving synthetic DOI.
	Non-existent venue	75	Replace the venue with a fictitious venue and keep the other fields.
	Placeholder authors	75	Replace the author list with synthetic names.
	Future date	75	Replace the year with a future year.
2: cross-record	Chimeric title	43	Keep authors and DOI and take the title from another record.
	Wrong venue	43	Keep title and authors and assign a real but incorrect venue.
	Swapped authors	43	Keep title and DOI and take the authors from another record.
	Preprint as published	43	Remove arXiv fields and cite a preprint as a journal article.
	Hybrid fabrication	43	Keep a real DOI, alter the title, and replace the authors.
	Merged citation	43	Combine title, DOI, authors, and venue from different records.
	Partial author list	42	Truncate a multi-author list.
3: subtle	Near-miss title	100	Change one phrase of the title and keep the other fields.
	Plausible fabrication	100	Generate a synthetic title and authors with a real venue and remove the DOI.
	arXiv version mismatch	100	Remove arXiv fields and add inconsistent publication metadata.

(ECE, ten equal-width bins). The detectors are not independent. DOI resolution, verify-citations, and bibtex-updater emit identical hallucinated outputs with identical confidences on 266 HallMark entries (231 hallucinated) and 143 StatRef entries (143 hallucinated), and DOI resolution returns hallucinated with confidence 0.85 or 0.95 only on them. The benchmark's public baseline code (`hallmark/baselines/prescreening.py`, as in release v1.1.1) contains a pre-screening step shared by these wrappers that overrides a tool's valid verdict: hallucinated with confidence 0.85 when doi.org answers 404 to the DOI, and 0.95 when the publication year lies in the future. From release v1.2.0 the pre-screen strips arXiv version suffixes before resolving and flags only doi.org's own 404 or 410 responses, so it would no longer flag the preprints below. Our files match the earlier mechanism exactly. The 59 entries with confidence 0.95 are the benchmark's future-date entries, and the 35 valid entries among the 266 are the 35 arXiv preprints of 2026, whose DOI strings carry a version suffix that doi.org does not resolve. Such a check responds to identifiers and dates rather than to the publication itself, so part of this signal may not carry over to naturally occurring fabrications. We cannot trace the executions themselves, so this is an inference from the public code and the release metadata, and all aggregation results are results for this correlated pool. Table 15 repeats the main comparison without these entries, which removes the whole 2026 cohort and changes the population, and Table 16 reports the calibrated decisions on that cohort.

Aggregators. All aggregators use the eleven detectors unless stated otherwise. *Best single*: the detector whose signed evidence ϕ_m has the largest AUC on the training split (bibtex-updater in 99.7% of the HallMark replications and Claude Sonnet 4.6 in 100.0% of the StatRef replications). *Majority vote*: $M^{-1}\sum_m \ell_m$. *Mean evidence*: $M^{-1}\sum_m \phi_m$. *Naive Bayes*: each detector output is mapped to one of six categories (label $\times$ confidence above or below its training median), and the score is the sum of per-detector log-likelihood ratios with add-one smoothing. *Logistic stacking*: ℓ_2-regularized logistic regression on $(\phi_1, \ldots, \phi_M)$ with inverse regularization strength 1. *Boosted stacking*: scikit-learn's `HistGradientBoostingClassifier` (Pedregosa et al., 2011) on the $4M$ features $(\mathbf{1}\{\ell_m = 1\}, \mathbf{1}\{\ell_m = -1\}, c_m, \phi_m)_{m \le M}$ with maximum depth 3, learning rate 0.05, 200 iterations, at least 20 samples per leaf, and ℓ_2 regularization 1. *Best learned single*: for each detector m, the boosted-stacking learner with the same hyperparameters is fitted to the four features of detector m. The detector is selected on the training split only, by two-fold cross-fitting: the learner is fitted on one half, and on the other half it is calibrated on a random half at $\alpha_{\mathrm{FP}} = \alpha_{\mathrm{FN}} = \delta = 0.1$ with tie-breaking and its deferral is measured on the rest, in both directions. The selected learner is then refitted on the whole training split. *Ablations* (Table 5): boosted stacking on the two verification tools only, the best raw single detector selected by certified deferral estimated on the training split

(two-fold cross-calibration at $\alpha_{\mathrm{FP}} = \alpha_{\mathrm{FN}} = \delta = 0.1$ with tie-breaking, leaving the calibration split untouched), boosted stacking with depth-one trees (additive in the features, 400 iterations), boosted stacking on ϕ only, logistic stacking on the $4M$ boosting features, and the Dawid–Skene label model (Dawid & Skene, 1979), fitted by 100 EM iterations on the unlabeled six-category outputs of the training split (initialized at the majority-vote posterior and oriented to agree with majority vote) and scored by its posterior log-odds. All hyperparameters were fixed before any experiment and were not tuned. Category medians are computed on the training split with ties assigned to the upper category.

Tie-breaking. Each score s receives an independent $\xi \sim \mathrm{Unif}(0,1)$, and pairs (s,ξ) are compared lexicographically. Every sampled occurrence of a reference receives fresh noise. Each threshold is stored as a pair (v,u) of a score value and a noise value, and a new reference is flagged if $(s,\xi) > (v_2,u_2)$ and accepted if $(s,\xi) < (v_1,u_1)$. If the pairs cross, the single cut is the pair midpoint, $((v_1+v_2)/2, 1/2)$ if $v_1 \neq v_2$ and $(v_1,(u_1+u_2)/2)$ otherwise, projected onto the closed certified interval between the two pairs in the lexicographic order, which guards against floating-point rounding of a midpoint onto an endpoint value, so deployment needs only the two stored pairs. At the upper endpoint the accepted set differs from the certified one only on the event of equality with the threshold pair, which has probability zero because every evaluated reference receives fresh continuous noise. A regression test in the supplementary code checks this containment on adversarial cases (near-equal distinct values, exact ties, endpoint rounding, and positive rescaling of the scores on a case that necessarily crosses). For computation, scores are mapped to (rank of value)$+\xi$, which preserves this order, and FPR, FNR, and U are averaged exactly over the noise of the evaluated references. A threshold (v,u) flags a reference with score v with probability $1-u$.

Calibration baselines (Table 13). Because the FPR depends only on λ_2 and the FNR only on λ_1, the pairs certified by the two per-class binomial tests of Algorithm 1 form a product set, of which $(\hat\lambda_1, \hat\lambda_2)$ is the element with the smallest deferral. This is optimality within that rectangle rather than among all PAC procedures, and unequal per-class budgets fixed in advance are also valid. All rows except the first are run on the same raw scores, without tie-breaking, so that they are paired with "Algorithm 1 (raw scores)". *Per-class Bentkus and Hoeffding bounds* (RCPS-style, Bates et al., 2021; Bentkus, 2004): for each class, the smallest order statistic whose empirical exceedance rate is certified below the tolerance by the Hoeffding upper confidence bound, or by the Bentkus p-value $e\,\mathbb{P}(\mathrm{Bin}(n_y,\alpha_y) \le n_y\hat r)$, at level δ_0. *Union bound*: Algorithm 1 with $\delta/2$ per class instead of $1-\sqrt{1-\delta}$. *LTT, pair grid + Bonferroni*: the candidate thresholds are the 1,001 grid points $\{0, 0.001, \ldots, 1\}$, and pairs (λ_1,λ_2) are tested with the maximum of the two binomial p-values at level $\delta/|\Lambda|$, $|\Lambda|$ being the number of ordered pairs. Since the certified pairs again form a product set, the procedure returns the smallest certified λ_2 and the largest certified λ_1 (a single cut if they cross). *LTT, multi-start fixed sequence*: $K = 10$ centres $c_k = (k-0.5)/10$. Around each centre, bands $[c_k - r, c_k + r] \cap [0,1]$ with $r = 0, 0.001, 0.002, \ldots$ are tested in order of increasing width with the maximum of the two binomial p-values at level δ/K, the first certified band of each chain is kept (valid because the maximum p-value is monotone along the chain), and the kept band with the smallest empirical deferral is returned. *Split conformal (in expectation)*: the ranks of Theorem 1(b). *Plug-in*: $k_0 = \lceil n_0(1-\alpha_{\mathrm{FP}})\rceil$ and $k_1 = \lceil n_1(1-\alpha_{\mathrm{FN}})\rceil$. *PAC-Wrap* (Li et al., 2022) (Table 14): per-class thresholds at level $\delta/2$ each, the equal split of its union bound. While the intersection of its two prediction sets contains both labels, which with the order statistics of Algorithm 1 means that the thresholds have not crossed, both tolerances are increased by Δ (0.1 as in their Algorithm 1, or 0.01) and the thresholds are recomputed, and the midpoint is then used as a single cut. Their Algorithm 1 relaxes while $\hat\tau_{\mathrm{fn}} < \hat\tau_{\mathrm{fp}}$, so it also stops when the two thresholds coincide, which on the raw discrete scores of Table 14 can happen earlier than the strict crossing $\hat\lambda_1 > \hat\lambda_2$, by one step at $\Delta = 0.1$ and by up to about a dozen steps at $\Delta = 0.01$, because tied order statistics can persist over a range of tolerances (for tie-broken scores the two stopping rules coincide). Table 14 reports the returned tolerance under both rules, with the tied value deferred by the three-way rule under the equality stop, and the violation frequency of the original tolerances under the strict rule. The exact-binomial per-class nested version of Learn-then-Test returns the thresholds of Algorithm 1 and is not listed separately. *NP threshold*: the single cut $\hat\lambda_2$ (FPR only) or $\hat\lambda_1$ (FNR only) of Algorithm 1 computed with the full level δ for that class and no deferral. *Fixed band* and *cut at 0.5*: thresholds on the predicted probability without calibration. Grid-based procedures apply only to scores in $[0,1]$.

Protocols. *In-domain*: 1,000 replications with independent random 40/30/30 train/calibration/test partitions (Tables 1, 5 to 7 and 20, Fig. 2c, Fig. 4). The "no tie-breaking" column of Table 5 repeats

this run with the same partitions and Algorithm 1 applied to the raw scores. *Finite-population check*: 1,000 replications. 40% of the data trains the aggregator, the remaining 60% forms a pool whose empirical distribution plays the role of P, a calibration set of half the pool size is drawn i.i.d. with replacement from the pool, and population FPR, FNR, and U are computed exactly on the pool (Fig. 2a, Tables 8 and 10 to 14). *Detector subsets*: 300 replications of the in-domain split. In each replication and for each size, one detector subset (and one subset of LLM judges) is drawn uniformly at random (Fig. 3a,b, Table 18, Fig. 5a,b). *Label budget*: 300 replications. The references after the first 1,200 of a random permutation form the test set, and the first N are split equally into training and calibration (Fig. 3c). The best learned single detector and the two-tool stacker are evaluated on the same labels as the six basic aggregators. *Shift*: 500 replications. The aggregator is trained on a random 70% of the source benchmark and source calibration uses the other 30%. For deferral rates, the target benchmark is split into a pool of 800 references that supplies target labels and a test set with the remaining references. For violation frequencies, the whole target benchmark serves as the population, target calibration labels (and, for target-only rules, a separate set of training labels) are drawn i.i.d. from it independently of each other, and risks are computed exactly (Table 19). A training half that contains a single class cannot be used to fit a supervised learner, and such draws are recorded rather than dropped silently (0 and 0 occurrences in the two directions). *Robustness*: 300 replications of the in-domain protocol for each scenario of Table 15, with the corrected HallMark labels, the original labels, the corrected labels with 52 further hallucinated entries flipped to valid uniformly at random in every replication, and both benchmarks without the entries on which the three detectors above emit identical outputs. The label scenarios of a benchmark share the partition and tie-breaking streams of each replication, and the flips come from a separate stream, so the label comparisons are paired. The scenarios without the shared entries use a different population and are not paired. *Subgroups*: under the corrected labels, the expected test decision of every reference (flag, defer, accept, averaged over its tie-breaking noise) is recorded in the same replications and pooled over the test appearances of the references of each subgroup (Table 16), with subgroup fields taken from the public release (Section B). *Title-disjoint splits*: 300 replications on HallMark. References with the same normalized title form a group, the 40/30/30 partition is drawn over groups so that no title appears on both sides of a split, and it is compared with a row-level partition of the same realized sizes drawn in the same replication (Table 17). *Known posteriors*: 1,000 replications per population, with 1,000 training and 600 calibration draws (Table 9, with details in its caption). All random partitions are unstratified, so class counts vary across replications, and the reported calibration counts are averages. Monte Carlo errors of the 300- and 500-replication sweeps are shown as bands or are below one point.

Reproducibility. Replication r of each experiment draws all of its randomness from a NumPy generator seeded with the sequence (s, r), where $s = 2027$ for the in-domain run and its no-tie-breaking repeat, 4049 for the finite-population check, 777 for detector subsets, 31337 for the label budget, 8128 and 6174 for the two known-posterior populations, $(99, r, \text{direction})$ for shift, $(5252, r)$ for robustness with $(5252, r, 1)$ for the random label flips, and $(6161, r)$ for the title-disjoint splits. Boosted stacking uses random state 0 in the benchmark experiments and the replication index in the known-posterior populations, and the best learned single detector uses the replication index for its learners and the sequence $(r, 17)$ for its selection split. The figures use fixed seeds as well (seed 0 for the held-out ROC curves of Fig. 2b and seed 5 for Fig. 8). The script `run_all.sh` in the supplementary code runs the regression test and every experiment in this order and then regenerates all tables, number macros, and figures from the saved results, in about 45 minutes on a ten-core laptop CPU with Python 3.14.4, NumPy 2.4.4, SciPy 1.17.1, scikit-learn 1.8.0, pandas 3.0.2, joblib 1.5.3, and Matplotlib 3.10.9 (a requirements file lists the package versions). It stops at the first failure so that partial results cannot feed the tables, and its log records the package versions and the SHA-256 hashes of the input files. This excludes the cost of running the detectors, which we did not measure, and it does not rebuild the detector outputs or the benchmarks. The HallMark metadata file used for Tables 16 and 17 is shipped with the code and can be regenerated from the public release with `make_metadata.py`.

D ADDITIONAL RESULTS

The finite-population check evaluates population risks. On a held-out test split, the fraction of replications in which the *test* FPR or FNR exceeds its tolerance at $\alpha_{\mathrm{FP}} = \alpha_{\mathrm{FN}} = \delta = 0.1$ ranges from 0.15 to 0.21 over the seven aggregators of Table 1, above δ, although the population violation

Table 3: Individual detectors on the full benchmarks: native FPR and FNR (uncertain outputs count as neither), fraction of uncertain outputs (Unc.), mean confidence (Conf.) and accuracy (Acc.) of decisive verdicts, expected calibration error of those confidences (ECE, ten bins), and mean test deferral rate (%) when the detector's signed evidence is calibrated alone ($\alpha_{\mathrm{FP}} = \alpha_{\mathrm{FN}} = \delta = 0.1$).

		HallMark							StatRef						
Detector	Family	FPR	FNR	Unc.	Conf.	Acc.	ECE	Def.	FPR	FNR	Unc.	Conf.	Acc.	ECE	Def.
DOI-presence	Rule	.539	.513	.000	.702	.477	.225	86.2	.253	.561	.000	.666	.620	.119	81.6
DOI-resolve	Rule	.227	.674	.000	.666	.504	.165	77.2	.567	.542	.000	.696	.443	.256	78.6
bibtex-updater	Tool	.348	.015	.000	.562	.852	.294	45.2	.538	.026	.000	.529	.673	.155	48.6
verify-citations	Tool	.441	.348	.000	.852	.615	.238	74.9	.050	.426	.000	.880	.795	.116	59.2
Gemini-2.5-Flash	LLM	.068	.551	.014	.942	.637	.312	61.0	.087	.448	.000	.963	.764	.205	61.7
DeepSeek-V3.2	LLM	.761	.079	.001	.930	.649	.303	78.4	.542	.079	.000	.914	.649	.304	62.2
DeepSeek-R1	LLM	.182	.330	.000	.850	.729	.128	57.3	.245	.191	.000	.884	.777	.107	48.7
Qwen3-235B-A22B	LLM	.419	.232	.005	.957	.692	.266	73.1	.393	.102	.000	.969	.727	.242	52.3
Mistral-Large-3	LLM	.116	.486	.009	.956	.658	.298	68.3	.143	.271	.000	.972	.804	.168	49.2
GPT-5.1	LLM	.342	.214	.004	.919	.734	.185	55.4	.129	.197	.000	.950	.843	.108	36.2
Claude-Sonnet-4.6	LLM	.120	.298	.004	.924	.772	.152	52.9	.058	.177	.000	.903	.893	.011	21.1

Table 4: Agreement among the eleven detectors: distribution of the number k of "hallucinated" verdicts per reference, fraction of references on which all decisive verdicts agree (Unan.), fraction with at least three "hallucinated" and at least three "valid" verdicts (Split), and mean pairwise disagreement among decisive verdicts (Disagr.).

	Share of references (%) with k "hallucinated" verdicts among the 11 detectors														
Benchmark	k=0	k=1	k=2	k=3	k=4	k=5	k=6	k=7	k=8	k=9	k=10	k=11	Unan.	Split	Disagr.
HallMark	0.0	2.8	12.4	14.4	12.0	9.0	6.9	7.6	10.1	13.2	9.7	1.9	1.9	58.7	39.7
StatRef	3.4	9.4	19.4	11.7	8.2	6.4	4.5	4.4	5.7	15.7	10.4	0.8	4.2	40.9	35.0

frequencies are at or below δ up to Monte Carlo error (Table 10). The difference is test-sampling noise, since the test estimate targets a different quantity. With about 250 valid test references, the test FPR has a standard deviation of about 0.017 around the population FPR, so a rule whose population FPR is 0.08 exceeds 0.10 on the test split in roughly 12% of replications. For in-expectation calibration, mean test FPR and FNR approach the tolerances, as Theorem 1(b) permits, and the test violation frequency ranges from 0.54 to 0.72.

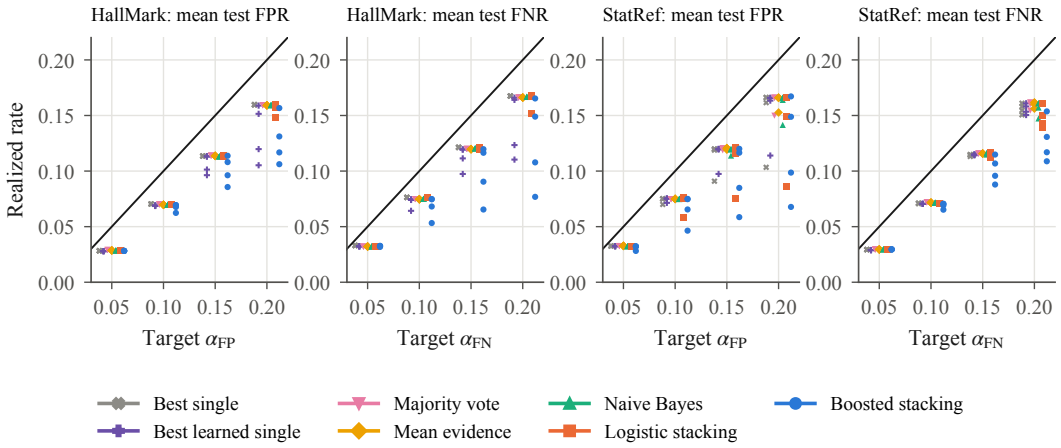

Figure 4: Mean test FPR and FNR against their tolerances over the grid $\{.05, .10, .15, .20\}^2$ ($\delta = 0.1$, 1,000 replications). Each marker is one aggregator and tolerance pair, and all lie below the diagonal. The figure shows average error rates only, and the high-probability event is checked in Table 10.

Table 5: Ablations and tie-breaking (in-domain, $\delta = 0.1$, deferral rates in %). "No t.-b.": Algorithm 1 applied to the raw scores of the same splits. The deferral-selected raw detector coincides with the AUC-selected one in at least 96% of replications, boosting on ϕ alone matches boosting on all features, and additive (depth-1, 400 iterations) boosting recovers most of the gap between logistic and boosted stacking (depth 3, 200 iterations). Because the two learners have different capacity budgets, this split between transformations of individual outputs and interactions is suggestive rather than a decomposition.

	HallMark				StatRef			
Aggregator	AUC	Def.$_{.10}$	no t.-b.	Def.$_{.05}$	AUC	Def.$_{.10}$	no t.-b.	Def.$_{.05}$
Best single	.845	45.2	48.3	62.6	.929	21.0	26.3	46.0
Best learned single	.946	12.8	20.1	37.0	.924	23.1	28.8	46.9
Majority vote	.863	39.9	50.5	61.3	.919	23.2	30.9	48.6
Mean evidence	.844	44.5	44.5	66.9	.913	26.0	26.0	51.8
Naive Bayes	.899	31.0	31.3	52.0	.930	20.7	20.8	41.2
Logistic stacking	.926	22.8	22.8	43.3	.946	16.8	16.8	33.6
Boosted stacking	.966	5.6	5.6	22.4	.965	5.4	5.4	25.0
Best single, selected by deferral	.843	45.6	49.4	63.2	.929	21.1	26.3	45.8
Boosted stacking, two tools	.953	8.8	13.9	33.7	.913	27.8	33.9	43.8
Boosted stacking, depth 1 (additive)	.955	11.2	12.1	28.6	.961	7.9	8.1	24.9
Boosted stacking, ϕ only	.965	6.0	6.1	23.2	.966	5.7	5.7	24.2
Logistic stacking, $4M$ features	.937	19.1	19.1	35.1	.947	15.6	15.6	33.3
Dawid–Skene (unsupervised)	.845	42.9	43.3	71.9	.912	26.3	26.4	50.7

Table 6: Review burden and cost at $\alpha_{\mathrm{FP}} = \alpha_{\mathrm{FN}} = \delta = 0.1$ (in-domain, %). Flag: $P(\mathrm{HAL})$. Defer: $P(\mathrm{DEF})$. $C_{1/2} = P(\mathrm{DEF}) + \frac{1}{2}P(\mathrm{HAL})$ is the cost if confirming a flag costs half a full check, and $W = P(\mathrm{HAL}) + P(\mathrm{DEF}) = 1 - P(\mathrm{VAL})$ is the fraction of references a human checks if flags are also checked in full, so Defer, $C_{1/2}$, and W are the costs when confirming a flag costs 0, $\frac{1}{2}$, and 1 full checks. When the two thresholds do not cross, W depends only on the acceptance threshold, so an accept-or-defer rule with the same acceptance threshold has the same W. The last column of each block reweights the class-conditional rates to prevalence $\pi_1 = 0.05$ (a pure prior-shift calculation, not a measured low-prevalence deployment).

	HallMark					StatRef				
	benchmark prevalence				π_1=.05	benchmark prevalence				π_1=.05
Aggregator	Flag	Defer	$C_{1/2}$	W	W	Flag	Defer	$C_{1/2}$	W	W
Best single	24.2	45.2	57.2	69.3	37.3	38.9	21.0	40.5	59.9	39.6
Best learned single	48.6	12.8	37.1	61.4	18.3	38.4	23.1	42.3	61.5	42.1
Majority vote	37.8	39.9	58.8	77.7	57.2	36.2	23.2	41.3	59.4	38.8
Mean evidence	36.7	44.5	62.8	81.2	65.4	35.3	26.0	43.7	61.3	41.8
Naive Bayes	37.7	31.0	49.9	68.7	35.7	37.3	20.7	39.4	58.0	36.5
Logistic stacking	42.5	22.8	44.1	65.3	27.8	39.9	16.8	36.8	56.7	34.4
Boosted stacking	54.4	5.6	32.8	60.0	14.9	41.4	5.4	26.1	46.8	18.4
Boosted stacking, two tools	50.9	8.8	34.3	59.7	14.3	32.4	27.8	44.0	60.2	40.0
Boosted stacking, depth 1 (additive)	50.2	11.2	36.3	61.4	18.4	40.7	7.9	28.3	48.6	21.2

Table 7: Paired differences between aggregators over the 1,000 in-domain splits at $\alpha_{\mathrm{FP}} = \alpha_{\mathrm{FN}} = \delta = 0.1$ (percentage points): mean [2.5%, 97.5% quantiles] of the difference in deferral rate U and review burden W, and the percentage of splits in which the first aggregator defers less. The quantiles describe variability across repartitions of one benchmark, not uncertainty about other benchmarks.

Benchmark	Comparison	ΔU $[q_{2.5}, q_{97.5}]$	ΔW $[q_{2.5}, q_{97.5}]$	Wins (%)
HallMark	Boosted stacking vs. best single	-39.6 $[-49.8, -31.9]$	-9.3 $[-13.5, -5.5]$	100
	Boosted stacking vs. best learned single	-7.2 $[-16.7, +0.7]$	-1.4 $[-5.0, +2.1]$	96
	Boosted stacking vs. boosted stacking, two tools	-3.3 $[-10.2, +3.5]$	$+0.3$ $[-2.6, +2.9]$	81
	Boosted stacking vs. logistic stacking	-17.2 $[-27.3, -8.5]$	-5.3 $[-9.5, -1.1]$	100
	Best learned single vs. best single	-32.4 $[-43.4, -20.4]$	-7.9 $[-12.1, -3.9]$	100
StatRef	Boosted stacking vs. best single	-15.6 $[-23.3, -7.9]$	-13.1 $[-20.5, -5.5]$	100
	Boosted stacking vs. best learned single	-17.7 $[-31.4, -9.3]$	-14.7 $[-24.7, -6.7]$	100
	Boosted stacking vs. boosted stacking, two tools	-22.4 $[-30.2, -13.7]$	-13.4 $[-20.8, -6.1]$	100
	Boosted stacking vs. logistic stacking	-11.4 $[-20.0, -3.3]$	-9.9 $[-18.0, -2.3]$	100
	Best learned single vs. best single	$+2.1$ $[-3.4, +15.9]$	$+1.6$ $[-3.4, +12.3]$	32

Table 8: Empirical counterpart of Theorem 3 on the finite population (1,000 replications, $\delta = 0.1$, %). $U(h)$: realized deferral. $U^{\star}_{\hat{s}}$: oracle deferral of the fitted score at the nominal levels, computed exactly on the population. Regret: $U^{\star}_{\hat{s}}$ minus the smallest oracle deferral among the 13 scores of Table 5, which is 0.2% and 0.1% at $\alpha = 0.1$ and 11.6% and 12.6% at $\alpha = 0.05$ (HallMark, StatRef). That smallest value is at least $U^{\star}_{[M]}$, so this proxy understates the regret relative to the true posterior and is evaluated at the nominal rather than the realized levels. It is also near zero for flexible scores because the empirical posterior of a finite population can nearly separate its references. Calib.: $U(h) - U^{\star}_{\hat{s}}$, a calibration residual that combines ω with the change of regret between nominal and realized levels and is not the unused-budget term itself. True terms are computed in Table 9. For strong scores the residual dominates, and for weak scores the regret does.

		HallMark				StatRef			
α	Aggregator	$U(h)$	$U^{\star}_{\hat{s}}$	Regret (proxy)	Calib.	$U(h)$	$U^{\star}_{\hat{s}}$	Regret (proxy)	Calib.
0.10	Best single	45.8	40.0	39.8	5.8	21.0	10.8	10.7	10.2
0.10	Best learned single	12.8	2.0	1.9	10.8	23.1	13.3	13.2	9.8
0.10	Majority vote	39.8	27.7	27.5	12.1	23.5	13.3	13.2	10.2
0.10	Mean evidence	44.6	34.7	34.5	9.9	26.2	15.0	14.9	11.2
0.10	Naive Bayes	31.3	19.2	19.0	12.1	20.8	10.8	10.7	10.0
0.10	Logistic stacking	23.1	11.7	11.5	11.4	16.9	7.3	7.2	9.6
0.10	Boosted stacking	5.7	0.2	0.0	5.5	5.5	0.1	0.0	5.4
0.10	Boosted stacking, two tools	8.7	0.4	0.2	8.3	27.5	18.4	18.2	9.2
0.10	Boosted stacking, depth 1 (additive)	11.3	2.8	2.6	8.5	7.9	1.2	1.1	6.8
0.05	Best single	62.0	52.5	40.8	9.5	45.5	30.4	17.8	15.2
0.05	Best learned single	37.0	21.7	10.1	15.2	46.7	33.0	20.4	13.7
0.05	Majority vote	61.0	49.8	38.1	11.3	48.3	34.1	21.5	14.2
0.05	Mean evidence	66.5	53.1	41.5	13.4	51.5	35.9	23.3	15.6
0.05	Naive Bayes	51.8	42.5	30.8	9.3	41.7	30.2	17.6	11.5
0.05	Logistic stacking	43.0	32.9	21.2	10.2	33.8	24.9	12.3	8.9
0.05	Boosted stacking	22.6	11.6	0.0	11.0	25.3	12.9	0.3	12.4
0.05	Boosted stacking, two tools	33.6	16.1	4.5	17.5	43.9	35.6	23.0	8.2
0.05	Boosted stacking, depth 1 (additive)	28.2	17.9	6.3	10.3	25.3	14.5	1.9	10.8

Table 9: Known-posterior check of Theorem 3 at $\alpha_{\mathrm{FP}} = \alpha_{\mathrm{FN}} = \delta = 0.1$ ($\pi_1 = 0.5$, 1,000 replications, 1,000 training and 600 calibration draws). *Discrete*: six detectors with five ordinal output levels, conditionally independent given the label and a latent difficulty that makes four detectors uninformative on hard references (a population of 5^6 atoms), with tie-breaking. *Gaussian*: six detectors with separations $1.3 \times (1.2, 1.0, 0.8, 0.5, 0.3, 0.1)$ and within-class correlation 0.3, with linear scores. True terms are computed exactly at the realized levels: $\mathrm{Reg}_{\hat{s}}(\hat{a})$, the exact unused-budget cost ω, and its slope bound (4). "Bound $- U(h)$" is the mean slack of (3) with ω and with the slope bound. The largest violations of (3) over all replications are 1×10^{-14} with ω and 6.9×10^{-16} with the slope bound, i.e., floating-point error, and the bound was never vacuous. Benchmark proxies are those of Table 8 (information: best nominal oracle deferral among the fitted scores, regret: nominal oracle deferral minus that, and the calibration residual).

		True terms (%)					Bound $- U(h)$		Benchmark proxies (%)		
Population	Score	$U(h)$	$U^{\star}_{[M]}$	$\mathrm{Reg}_{\hat{s}}(\hat{a})$	ω	Slope	with ω	with slope	Info.	Regret	Calib.
Discrete	Best single	62.4	18.1	34.7	9.7	11.6	0.1	2.0	18.5	32.9	11.0
	Mean evidence	32.1	18.1	4.2	9.9	12.0	0.1	2.1	18.5	3.3	10.3
	Naive Bayes	29.2	18.1	1.2	10.0	12.1	0.1	2.1	18.5	0.7	10.0
	Logistic stacking	28.4	18.1	0.4	10.0	12.1	0.1	2.1	18.5	0.0	9.9
	Boosted stacking	32.2	18.1	4.1	10.0	12.1	0.1	2.1	18.5	3.4	10.4
	True posterior	28.0	18.1	0.0	10.0	12.0	0.1	2.1	–	–	–
Gaussian	Best single	39.2	14.6	15.4	9.3	11.2	0.1	2.0	15.0	14.1	10.2
	Mean evidence	49.1	14.6	25.3	9.3	11.2	0.1	2.0	15.0	23.9	10.2
	Naive Bayes	34.2	14.6	10.4	9.3	11.2	0.1	2.0	15.0	9.3	10.0
	Logistic stacking	24.2	14.6	0.4	9.3	11.2	0.1	2.0	15.0	0.0	9.2
	True posterior	23.8	14.6	0.0	9.3	11.2	0.1	2.0	–	–	–

Table 10: Finite-population check: largest violation frequency of the final rule over the 13 scores of Table 5 for each tolerance pair and δ, with the average exact rank-based failure probability of the raw thresholds in parentheses (1,000 replications). The largest of many noisy frequencies is biased upward. The pointwise Monte Carlo standard error at the nominal level is 0.007, 0.009, and 0.013 for $\delta = 0.05, 0.1, 0.2$. †: thresholds cross in more than 30% of runs, so the final rule is much more conservative than the raw thresholds.

		HallMark			StatRef		
α_{FP}	α_{FN}	δ=.05	δ=.10	δ=.20	δ=.05	δ=.10	δ=.20
0.05	0.05	.043 (.035)	.085 (.074)	.174 (.156)	.038 (.035)	.084 (.075)	.170 (.157)
0.05	0.10	.045 (.037)	.091 (.077)	.179 (.162)	.042 (.038)	.080 (.079)	.175 (.164)
0.05	0.15	.045 (.038)	.091 (.079)	.182 (.164)	.040 (.038)	.090 (.081)	.184 (.168)
0.05	0.20	.051 (.038)	.092 (.079)	.182 (.165)	.038 (.039)	.076 (.081)	.185 (.169)
0.10	0.05	.040 (.037)	.086 (.078)	.171 (.163)	.046 (.037)	.090 (.078)	.178 (.162)
0.10	0.10	.045 (.039)	.094 (.081)	.187 (.169)	.053 (.040)	.090 (.082)	.182 (.169)
0.10	0.15	.047 (.040)	.090 (.082)	.183 (.171)	.046 (.040)	.095 (.083)	.197 (.172)
0.10	0.20	.052 (.040)	.095 (.083)	.194 (.172)	.042 (.041)†	.086 (.084)†	.183 (.173)†
0.15	0.05	.043 (.038)	.092 (.079)	.177 (.166)	.047 (.038)	.100 (.079)	.192 (.165)
0.15	0.10	.044 (.040)	.099 (.083)	.185 (.172)	.051 (.040)	.090 (.083)	.181 (.172)
0.15	0.15	.044 (.041)	.096 (.084)†	.189 (.174)†	.047 (.041)	.099 (.085)	.194 (.175)†
0.15	0.20	.048 (.041)†	.094 (.085)†	.193 (.175)†	.045 (.042)†	.084 (.086)†	.167 (.176)†
0.20	0.05	.047 (.039)	.087 (.080)	.178 (.167)	.043 (.038)	.088 (.080)	.185 (.166)
0.20	0.10	.041 (.041)	.094 (.083)†	.185 (.173)†	.049 (.041)	.092 (.084)	.178 (.173)
0.20	0.15	.042 (.042)†	.094 (.085)†	.192 (.176)†	.038 (.042)†	.080 (.086)†	.179 (.176)†
0.20	0.20	.043 (.042)†	.100 (.086)†	.192 (.177)†	.017 (.042)†	.036 (.086)†	.056 (.177)†

Table 11: Finite-population check aggregated over the 13 scores of Table 5 and the 16 tolerance pairs (1,000 replications each). Final viol.: violation frequency of the final rule. Raw viol.: violation frequency of the raw thresholds $(\hat{\lambda}_1, \hat{\lambda}_2)$ before resolving a crossing. Exact ref.: the exact rank-based failure probability $1-(1-p_0)(1-p_1)$, computed at each replication's realized class counts and averaged. Crossing: percentage of runs with $\hat{\lambda}_1 > \hat{\lambda}_2$, on average over the grid and at $\alpha_{\mathrm{FP}} = \alpha_{\mathrm{FN}} = 0.1$ and 0.2.

Benchmark	δ	Final viol.	Raw viol.	Exact ref.	Crossing (%)	Cross. at .10 (%)	Cross. at .20 (%)
HallMark	0.05	.030	.038	.039	13	1	48
	0.10	.065	.081	.081	16	1	51
	0.20	.133	.170	.169	18	3	53
StatRef	0.05	.028	.038	.040	17	1	75
	0.10	.058	.080	.082	20	1	82
	0.20	.119	.167	.170	24	3	90

Table 12: Unused error budget $(\alpha - \hat{a})_+$ of the raw calibrated thresholds in the finite-population check ($\delta = 0.1$, 1,000 replications). Real.: realized mean [90% quantile]. Exact: $\mathbb{E}[(\alpha - \hat{a})_+]$ from Corollary 4. γ: its high-probability bound at $\delta' = 0.1$. Both are averaged over the realized class counts.

		HallMark						StatRef					
		FP side			FN side			FP side			FN side		
α	Aggregator	real. [q_{90}]	exact	γ	real. [q_{90}]	exact	γ	real. [q_{90}]	exact	γ	real. [q_{90}]	exact	γ
0.10	Best single	.031 [.051]	.031	.055	.025 [.041]	.025	.046	.025 [.041]	.025	.045	.029 [.047]	.029	.053
0.10	Best learned single	.031 [.052]	.031	.055	.025 [.041]	.025	.046	.025 [.042]	.025	.045	.029 [.046]	.029	.053
0.10	Mean evidence	.032 [.052]	.031	.055	.025 [.042]	.025	.046	.025 [.041]	.025	.045	.030 [.048]	.029	.053
0.10	Logistic stacking	.032 [.052]	.031	.055	.025 [.042]	.025	.046	.025 [.041]	.025	.045	.029 [.047]	.029	.053
0.10	Boosted stacking	.032 [.052]	.031	.055	.025 [.041]	.025	.046	.025 [.041]	.025	.045	.029 [.049]	.029	.053
0.05	Best single	.022 [.034]	.022	.037	.017 [.028]	.018	.031	.018 [.029]	.018	.031	.021 [.033]	.021	.035
0.05	Best learned single	.022 [.034]	.022	.037	.018 [.029]	.018	.031	.018 [.029]	.018	.031	.021 [.033]	.021	.035
0.05	Mean evidence	.022 [.035]	.022	.037	.018 [.029]	.018	.031	.018 [.029]	.018	.031	.021 [.033]	.021	.035
0.05	Logistic stacking	.022 [.034]	.022	.037	.018 [.029]	.018	.031	.018 [.029]	.018	.031	.021 [.033]	.021	.035
0.05	Boosted stacking	.022 [.035]	.022	.037	.018 [.028]	.018	.031	.018 [.028]	.018	.031	.021 [.033]	.021	.035

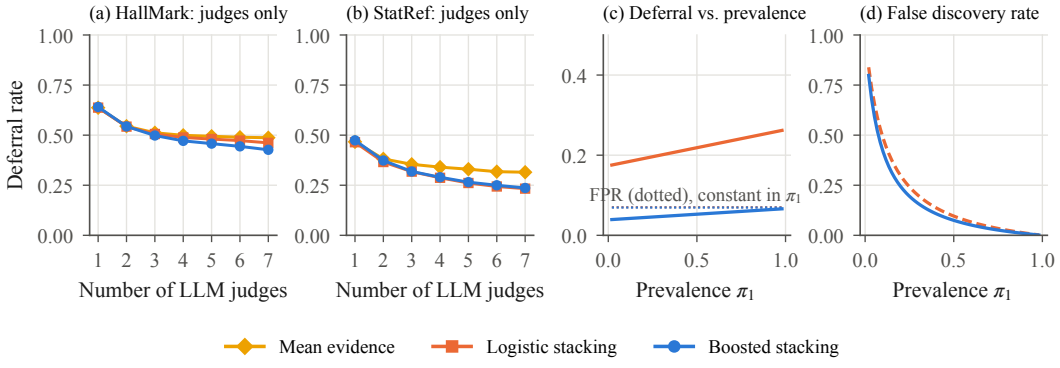

Figure 5: **(a,b)** Deferral rate when only a random subset of LLM judges is aggregated ($\alpha_{\mathrm{FP}} = \alpha_{\mathrm{FN}} = \delta = 0.1$, 300 replications). **(c,d)** Class-conditional guarantees under prevalence shift for logistic (orange) and boosted (blue) stacking calibrated on HallMark at $\alpha_{\mathrm{FP}} = \alpha_{\mathrm{FN}} = 0.1$. FPR (dotted) and FNR do not depend on the prevalence π_1 (Theorem 1c), whereas the deferral rate $(1-\pi_1)U_0 + \pi_1 U_1$ with $U_y = P(\mathrm{DEF} \mid Y = y)$ and the fraction of flagged references that are valid (false discovery rate) change with π_1, so a guarantee on the latter would not transfer across LLMs with different hallucination rates.

Table 13: Calibration procedures applied to the same logistic-stacking scores (top) and boosted-stacking scores (bottom) in the finite-population check ($\alpha_{\mathrm{FP}} = \alpha_{\mathrm{FN}} = \alpha$, $\delta = 0.1$, 1,000 replications). Viol.: frequency of $\mathrm{FPR} > \alpha_{\mathrm{FP}}$ or $\mathrm{FNR} > \alpha_{\mathrm{FN}}$. Def.: mean deferral rate. The first row uses tie-breaking, and all other rows use the same raw scores. The exact-binomial nested Learn-then-Test procedure returns the thresholds of Algorithm 1 and is therefore identical to the first two rows, so the grid-based rows measure the cost of a Bonferroni or fixed-sequence search over pairs, not of Learn-then-Test as such. The upper block of each part targets joint high-probability control. For rows marked * the violation frequency is a diagnostic rather than a failure, because those methods do not claim that guarantee (split conformal controls the errors in expectation, and NP rules control one error). The single-cut rules (NP thresholds and the cut at 0.5) defer nothing by construction. Definitions are in Section C.

	HallMark				StatRef			
	α=.05		α=.10		α=.05		α=.10	
Calibration	Viol.	Def.	Viol.	Def.	Viol.	Def.	Viol.	Def.
Algorithm 1 (tie-breaking)	.072	.430	.076	.231	.074	.338	.090	.169
Algorithm 1 (raw scores)	.063	.431	.067	.232	.067	.339	.081	.169
union bound $\delta/2$	.061	.432	.065	.232	.064	.340	.078	.170
Per-class Bentkus bound	.020	.459	.027	.263	.022	.370	.027	.195
Per-class Hoeffding bound	.000	1.00	.000	.440	.000	1.00	.000	.336
LTT, pair grid + Bonferroni	.000	.873	.000	.480	.000	.748	.000	.380
LTT, multi-start fixed sequence	.002	.500	.005	.319	.001	.443	.004	.225
Split conformal (in expectation)*	.626	.337	.664	.131	.628	.258	.644	.083
Plug-in empirical quantiles*	.732	.321	.742	.121	.742	.245	.713	.073
NP threshold, FPR only*	1.00	.000	1.00	.000	1.00	.000	1.00	.000
NP threshold, FNR only*	1.00	.000	1.00	.000	1.00	.000	1.00	.000
Fixed band $[0.2, 0.8]^*$	.999	.278	.023	.278	1.00	.124	.756	.124
Cut at 0.5^*	1.00	.000	1.00	.000	1.00	.000	1.00	.000

	HallMark				StatRef			
	α=.05		α=.10		α=.05		α=.10	
Calibration	Viol.	Def.	Viol.	Def.	Viol.	Def.	Viol.	Def.
Algorithm 1 (tie-breaking)	.071	.226	.062	.057	.073	.253	.062	.055
Algorithm 1 (raw scores)	.059	.228	.056	.058	.065	.254	.060	.056
union bound $\delta/2$	.056	.229	.055	.058	.062	.255	.058	.056
Per-class Bentkus bound	.023	.261	.017	.078	.027	.293	.017	.077
Per-class Hoeffding bound	.000	1.00	.000	.236	.000	1.00	.000	.259
LTT, pair grid + Bonferroni	.000	.844	.000	.283	.000	.724	.000	.313
LTT, multi-start fixed sequence	.007	.324	.004	.106	.006	.342	.003	.108
Split conformal (in expectation)*	.602	.132	.402	.010	.642	.144	.371	.008
Plug-in empirical quantiles*	.717	.120	.454	.007	.750	.129	.400	.005
NP threshold, FPR only*	1.00	.000	.812	.000	1.00	.000	.494	.000
NP threshold, FNR only*	1.00	.000	.599	.000	1.00	.000	.763	.000
Fixed band $[0.2, 0.8]^*$	.671	.164	.003	.164	.939	.136	.004	.136
Cut at 0.5^*	1.00	.000	.969	.000	1.00	.000	.950	.000

Table 14: PAC-Wrap versus Algorithm 1 on the same scores in the finite-population check ($\alpha_{\mathrm{FP}} = \alpha_{\mathrm{FN}} = \delta = 0.1$, 1,000 replications). Ours: deferral rate of Algorithm 1 with tie-breaking at the requested tolerances. PAC-Wrap (raw scores, union bound, relaxation step $\Delta = 0.1$ as in its Algorithm 1 or 0.01): mean tolerance ε at which its relaxation removes the ambiguity, when the relaxation stops at the first strict crossing of the two thresholds ("strict", our adaptation) or as soon as they coincide ("equality", the weak inequality of its Algorithm 1, which on discrete raw scores can stop earlier, by one step at $\Delta = 0.1$ and by several steps at $\Delta = 0.01$), and the frequency with which its final single cut under the strict rule violates the original tolerance 0.1. The two stopping rules differ by at most 0.06 in mean ε, and its violation frequency of the relaxed tolerance is at most .060 for every aggregator.

	HallMark				StatRef			
	Ours	PAC-Wrap, $\Delta = 0.1$ / 0.01			Ours	PAC-Wrap, $\Delta = 0.1$ / 0.01		
Aggregator	Def. (%)	ε (strict)	ε (equality)	Viol. (orig. 0.1)	Def. (%)	ε (strict)	ε (equality)	Viol. (orig. 0.1)
Best single	45.8	0.32 / 0.28	0.31 / 0.27	1.00 / 1.00	21.0	0.21 / 0.18	0.20 / 0.17	1.00 / 1.00
Best learned single	12.8	0.20 / 0.16	0.19 / 0.14	.989 / .985	23.1	0.23 / 0.19	0.21 / 0.17	1.00 / 1.00
Majority vote	39.8	0.32 / 0.28	0.28 / 0.23	1.00 / 1.00	23.5	0.27 / 0.21	0.21 / 0.18	1.00 / 1.00
Mean evidence	44.6	0.30 / 0.26	0.30 / 0.26	1.00 / 1.00	26.2	0.24 / 0.20	0.24 / 0.20	1.00 / 1.00
Naive Bayes	31.3	0.30 / 0.24	0.30 / 0.24	1.00 / 1.00	20.8	0.22 / 0.19	0.22 / 0.19	1.00 / 1.00
Logistic stacking	23.1	0.23 / 0.19	0.23 / 0.19	1.00 / 1.00	16.9	0.20 / 0.17	0.20 / 0.17	1.00 / 1.00
Boosted stacking	5.7	0.19 / 0.13	0.19 / 0.13	.653 / .563	5.5	0.19 / 0.13	0.19 / 0.13	.664 / .450

Table 15: Robustness of the in-domain deferral rate (%, $\alpha_{\mathrm{FP}} = \alpha_{\mathrm{FN}} = \delta = 0.1$, 300 replications per column with seeds distinct from those of Table 1, so the corrected-label column differs from its 1,000-replication values by Monte Carlo error). HallMark with the corrected labels, with the benchmark's original labels (52 entries labeled hallucinated that its authors later relabeled as valid), with 52 further hallucinated labels flipped to valid uniformly at random in each replication, and without the entries on which DOI resolution, verify-citations, and bibtex-updater emit identical outputs (a selected subset that removes all 35 valid 2026 preprints). StatRef as given and without those entries. The three HallMark label columns share partitions and tie-breaking streams, so their differences are paired (last row). The other rows give the largest mean test FPR/FNR over the aggregators and the resulting sample sizes.

	HallMark				StatRef	
Aggregator	corrected labels	original labels	52 random flips	w/o shared	as given	w/o shared
Best single	45.2	46.5	48.9	47.8	20.8	26.6
Best learned single	12.9	13.2	26.9	7.1	23.1	29.2
Majority vote	39.4	44.0	46.7	39.4	23.2	29.7
Mean evidence	44.0	47.3	51.6	43.4	26.0	32.4
Logistic stacking	22.8	23.2	33.8	16.3	17.1	22.1
Boosted stacking	5.4	4.7	17.6	2.9	5.3	9.3
Boosted stacking, two tools	9.0	8.5	22.8	2.2	27.6	33.1
Largest mean test FPR/FNR	.071/.076	.070/.076	.072/.076	.068/.073	.078/.073	.076/.070
References / hallucinated (%)	2071 / 60	2071 / 63	2071 / 58	1805 / 56	2180 / 41	2037 / 37
Boosted stacking, paired change vs. corrected $[q_{2.5}, q_{97.5}]$	–	-0.7 $[-5.0, +3.6]$	$+12.2$ $[+4.4, +21.9]$	–	–	–

Table 16: HallMark subgroups under the corrected labels (300 in-domain replications, $\alpha_{\mathrm{FP}} = \alpha_{\mathrm{FN}} = \delta = 0.1$). Native: percentage of "hallucinated" verdicts of bibtex-updater, verify-citations, and the seven judges (mean) on the full benchmark. Calibrated columns: expected test decisions of boosted stacking and of the best learned single detector, pooled over the test appearances of the subgroup's references. For valid cohorts Flag is a false-positive rate, and for hallucinated subgroups Accept is a false-negative rate. Subgroup fields come from the public release (Section B). The guarantees of Theorem 1 are class-wide and do not hold within subgroups, as the 2026 cohort shows.

		Native "hallucinated" verdicts (%)			Boosted stacking (%)			Best learned single (%)		
Subgroup	n	bibtex-upd.	verify-cit.	judges	Flag	Defer	Accept	Flag	Defer	Accept
Valid references										
Valid, 2021–2023, unchanged label	738	30	38	27	4	2	94	5	5	91
Valid, relabeled by the benchmark audit	52	54	94	16	20	7	73	19	5	76
Valid, arXiv preprints of 2026	35	100	100	77	47	26	27	38	62	0
Hallucinated, by generation method										
Perturbation	892	98	71	67	84	8	9	72	19	9
LLM-generated	155	99	83	50	83	6	11	80	11	9
Adversarial	119	100	32	92	100	0	0	88	10	2
Real-world	80	100	16	79	94	4	2	94	5	0
Hallucinated, by type										
Plausible fabrication	146	100	18	83	95	3	2	93	7	0
arXiv version mismatch	130	98	59	67	83	11	6	85	13	2
Swapped authors	130	99	89	43	84	6	10	67	17	17
Merged citation	105	100	90	75	97	2	1	92	6	2
Partial author list	102	97	68	32	61	7	32	41	18	41
Near miss title	92	92	52	46	73	12	15	79	13	9
Wrong venue	81	100	53	65	77	15	7	79	21	0
Nonexistent venue	76	96	54	87	93	2	5	73	22	4
Placeholder authors	74	100	95	78	94	5	2	62	28	10
Chimeric title	70	100	67	75	92	5	3	85	15	0
Fabricated DOI	67	97	100	89	88	5	7	38	57	6
Future date	59	100	100	100	100	0	0	90	10	0
Preprint as published	59	100	39	59	71	19	10	91	9	0
Hybrid fabrication	55	100	56	88	100	0	0	84	13	3

Table 17: Title-disjoint splits on HallMark (300 replications, $\delta = 0.1$, deferral in %). In each replication the 40/30/30 partition is drawn over groups of references that share a normalized title (1,344 groups, 438 of them with more than one reference), so that no title appears on both sides of a split, and it is compared with a row-level partition of the same sizes drawn in the same replication, in which 33% of the test references share a title with a training reference. Differences are taken within replications (mean and 2.5%/97.5% quantiles across replications), and because the two partitions of a replication are different random draws, the quantiles are wide. Title groups approximate publication families and are not a source-disjoint evaluation.

	Def. (%), α=.10			Def. (%), α=.05		AUC	
Aggregator	rows	titles	titles − rows [$q_{2.5}$, $q_{97.5}$]	rows	titles	rows	titles
Best single	45.3	45.1	−0.2 [−11.3, +10.7]	62.8	62.5	.845	.847
Best learned single	12.4	12.4	+0.0 [−12.5, +13.3]	36.8	36.9	.946	.947
Majority vote	39.7	39.6	−0.0 [−13.3, +12.3]	61.2	60.8	.865	.865
Mean evidence	44.3	44.1	−0.1 [−13.3, +13.3]	66.7	65.9	.846	.845
Naive Bayes	31.1	30.5	−0.6 [−17.4, +14.0]	52.2	51.6	.900	.901
Logistic stacking	22.9	22.4	−0.5 [−13.9, +13.5]	43.2	43.2	.928	.927
Boosted stacking	5.4	5.4	+0.0 [−6.9, +8.1]	22.0	22.6	.966	.966
Boosted stacking, two tools	8.7	8.5	−0.2 [−9.2, +10.2]	33.7	34.1	.953	.953

Table 18: Deferral rate (%) for combinations of detector families (Rule: DOI presence and DOI resolution. Tool: bibtex-updater and verify-citations. LLM: seven judges) at $\alpha_{\mathrm{FP}} = \alpha_{\mathrm{FN}} = \delta = 0.1$ (300 replications).

	HallMark: deferral (%)						StatRef: deferral (%)					
Families	Best	Vote	Mean	NB	Logit	Boost	Best	Vote	Mean	NB	Logit	Boost
Rule	77.3	83.6	83.2	84.0	78.2	76.6	81.9	82.7	78.3	82.0	80.8	73.2
Tool	45.4	48.5	52.4	37.6	38.4	9.0	58.8	37.3	39.4	37.3	39.9	28.1
LLM	53.7	48.8	48.7	48.0	46.2	42.7	21.3	31.7	31.5	27.5	23.3	23.6
Rule+Tool	45.4	44.7	47.8	38.2	38.9	9.4	58.9	39.2	35.8	32.8	33.1	24.2
Rule+LLM	53.8	51.5	51.0	49.2	43.5	41.3	21.2	32.6	31.2	26.9	22.3	22.9
Tool+LLM	45.3	36.7	41.3	31.7	25.5	5.8	21.2	23.5	26.5	21.2	16.3	5.3
All	45.4	40.0	44.9	31.5	23.1	5.8	21.3	23.6	26.5	21.1	16.9	5.4

Table 19: Transfer between benchmarks ($\alpha_{\mathrm{FP}} = \alpha_{\mathrm{FN}} = \delta = 0.1$, 500 replications). *Source-calibrated*: aggregator and thresholds from the source. *Recalibrated*: source aggregator with thresholds from n_t target labels. *Target-only*: $n_t/2$ target labels to train and $n_t/2$ to calibrate. *Pooled*: the source training split plus 400 target labels to train (Rew.: equal total source and target weight) and 400 to calibrate. FPR, FNR, and Viol. are exact risks with the target benchmark as the population (red: above tolerance, or above δ by more than two Monte Carlo standard errors), and Def. is the test deferral rate (%). AUC is the test AUC of the source-trained score on the target benchmark, and $\hat{\kappa}_0/\hat{\kappa}_1$ are empirical two-sample Kolmogorov statistics between the source calibration and target test scores of the two classes, estimates of the population distances of Proposition 6 rather than certified bounds.

	Source-calibrated						Recalibrated on target				Target-only			Pooled (800)	
							n_t=200		n_t=800		200	800			
Aggregator	FPR	FNR	Viol.	Def.	AUC	$\hat{\kappa}_0/\hat{\kappa}_1$	Viol.	Def.	Viol.	Def.	Def.	Viol.	Def.	Def.	Rew.
Best single	.047	.099	.434	53.3	.734	0.52/0.15	.084	55.4	.090	48.1	49.5	.084	24.6	26.5	–
Majority vote	.061	.063	.008	27.9	.919	0.19/0.22	.070	36.3	.072	22.2	52.8	.064	26.9	26.9	–
Mean evidence	.064	.070	.006	28.8	.913	0.30/0.22	.062	39.1	.082	25.0	54.6	.070	29.6	29.6	–
Naive Bayes	.060	.055	.044	26.3	.928	0.14/0.19	.062	31.0	.058	18.7	49.0	.082	24.1	22.0	–
Logistic stacking	.037	.141	1.00	14.6	.928	0.43/0.11	.082	29.5	.096	20.7	45.5	.066	20.1	21.7	20.4
Boosted stacking	.186	.083	.982	8.9	.923	0.37/0.13	.042	33.6	.082	22.2	44.5	.058	10.9	15.9	11.5

HallMark → StatRef (above), StatRef → HallMark (below)

	Source-calibrated						Recalibrated on target				Target-only			Pooled (800)	
							n_t=200		n_t=800		200	800			
Aggregator	FPR	FNR	Viol.	Def.	AUC	$\hat{\kappa}_0/\hat{\kappa}_1$	Viol.	Def.	Viol.	Def.	Def.	Viol.	Def.	Def.	Rew.
Best single	.130	.229	1.00	9.3	.810	0.29/0.19	.062	66.0	.090	50.9	67.0	.080	48.4	57.4	–
Majority vote	.082	.091	.418	33.3	.864	0.20/0.23	.080	49.8	.104	38.1	63.2	.070	43.0	42.9	–
Mean evidence	.079	.078	.396	39.7	.845	0.31/0.22	.064	54.1	.086	42.7	68.1	.088	47.4	47.4	–
Naive Bayes	.106	.058	.658	31.6	.876	0.32/0.19	.062	47.5	.084	34.1	55.8	.074	34.8	38.7	–
Logistic stacking	.139	.021	1.00	21.2	.907	0.35/0.13	.076	43.6	.086	24.6	51.5	.092	27.4	31.4	29.8
Boosted stacking	.281	.085	1.00	8.0	.879	0.48/0.14	.058	54.1	.064	31.8	47.2	.082	11.7	20.8	14.1

Table 20: Strict tolerances (in-domain, $\delta = 0.1$, 1,000 replications). No flag: fraction of replications in which $\hat{\lambda}_2 = +\infty$, which happens when the calibration split has fewer than $\lceil \log \delta_0 / \log(1 - \alpha_{\mathrm{FP}}) \rceil$ valid references (147 at $\alpha_{\mathrm{FP}} = 0.02$, 296 at 0.01, 593 at 0.005). HallMark calibration splits contain about 247 and StatRef splits about 384 valid references.

		HallMark				StatRef			
$(\alpha_{\mathrm{FP}}, \alpha_{\mathrm{FN}})$	Aggregator	FPR	FNR	Def. (%)	No flag (%)	FPR	FNR	Def. (%)	No flag (%)
(0.02, 0.1)	Best single	.0077	.076	64.7	0	.0090	.071	38.5	0
	Logistic stacking	.0075	.076	50.9	0	.0092	.071	27.8	0
	Boosted stacking	.0074	.075	33.3	0	.0089	.071	15.8	0
(0.01, 0.1)	Best single	.0000	.076	69.3	100	.0027	.071	43.7	0
	Logistic stacking	.0000	.076	65.3	100	.0024	.071	38.0	0
	Boosted stacking	.0000	.075	60.0	100	.0026	.071	24.0	0
(0.01, 0.05)	Best single	.0000	.033	72.1	100	.0027	.029	59.2	0
	Logistic stacking	.0000	.032	71.2	100	.0024	.029	50.2	0
	Boosted stacking	.0000	.033	65.6	100	.0026	.030	39.0	0
(0.005, 0.05)	Best single	.0000	.033	72.1	100	.0000	.029	75.5	100
	Logistic stacking	.0000	.032	71.2	100	.0000	.029	68.9	100
	Boosted stacking	.0000	.033	65.6	100	.0000	.030	61.8	100

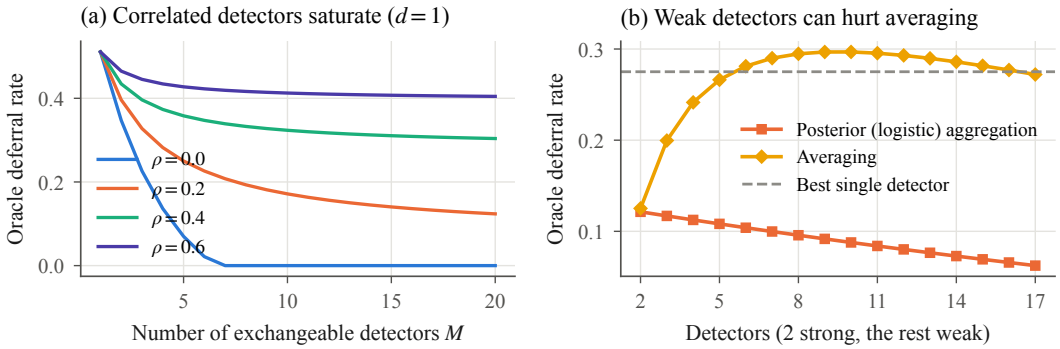

Figure 6: Oracle deferral rates in the Gaussian model of Proposition 5 ($\alpha_{\mathrm{FP}} = \alpha_{\mathrm{FN}} = 0.1$, $\pi_1 = 0.5$). **(a)** Exchangeable detectors with separation $d = 1$. With $\rho = 0$ the deferral rate reaches zero, while for the plotted correlations $\rho \in \{0.2, 0.4, 0.6\}$ it saturates above zero, as it does whenever $d/\sqrt{\rho} \leq z_{\mathrm{FP}} + z_{\mathrm{FN}}$, i.e., $\rho \geq 0.152$ here. Smaller positive correlations still allow zero deferral for large M. **(b)** Two strong independent detectors ($\mu = 1.6, 1.28$) followed by weak ones ($\mu = 0.25$). Posterior aggregation improves monotonically, whereas averaging becomes worse than the best single detector for intermediate numbers of weak detectors.

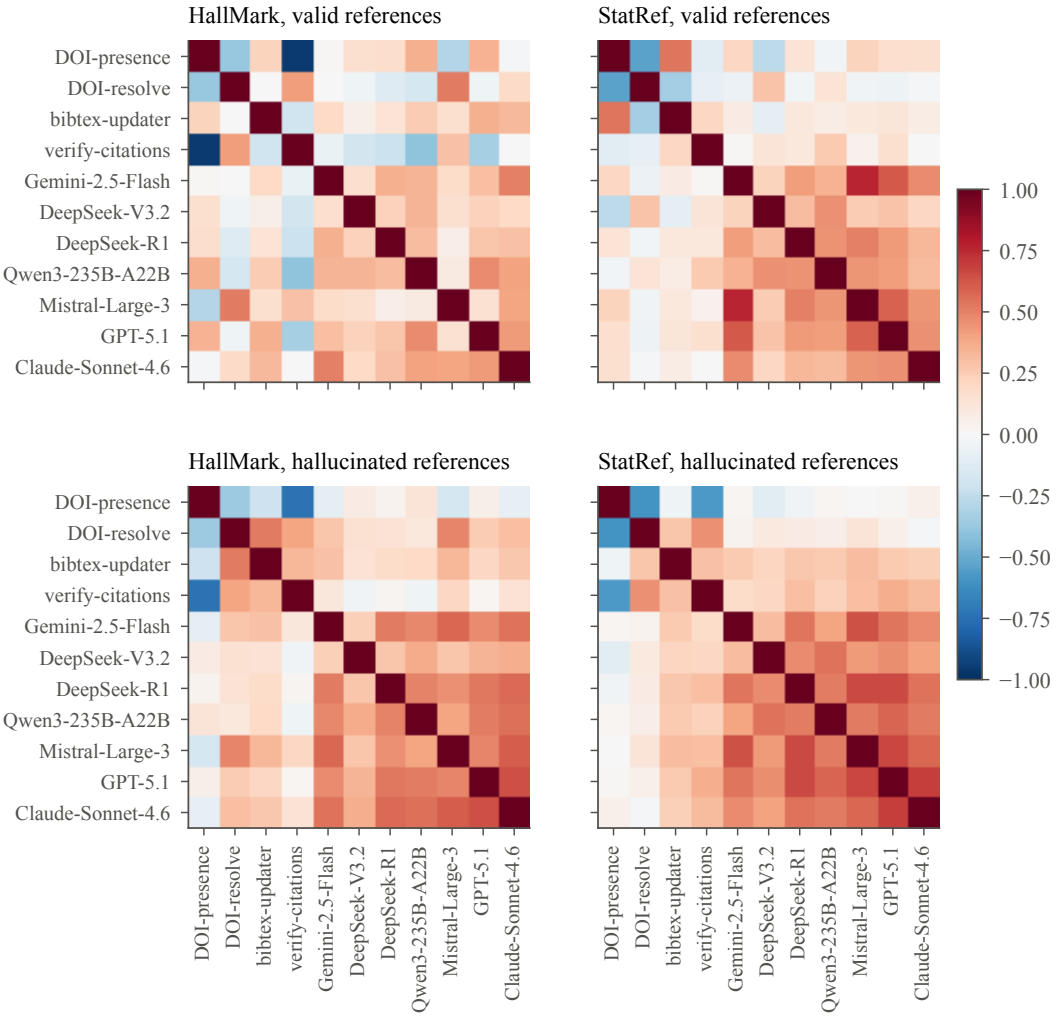

Figure 7: Within-class correlation of the signed evidence ϕ_m for valid (top) and hallucinated (bottom) references.

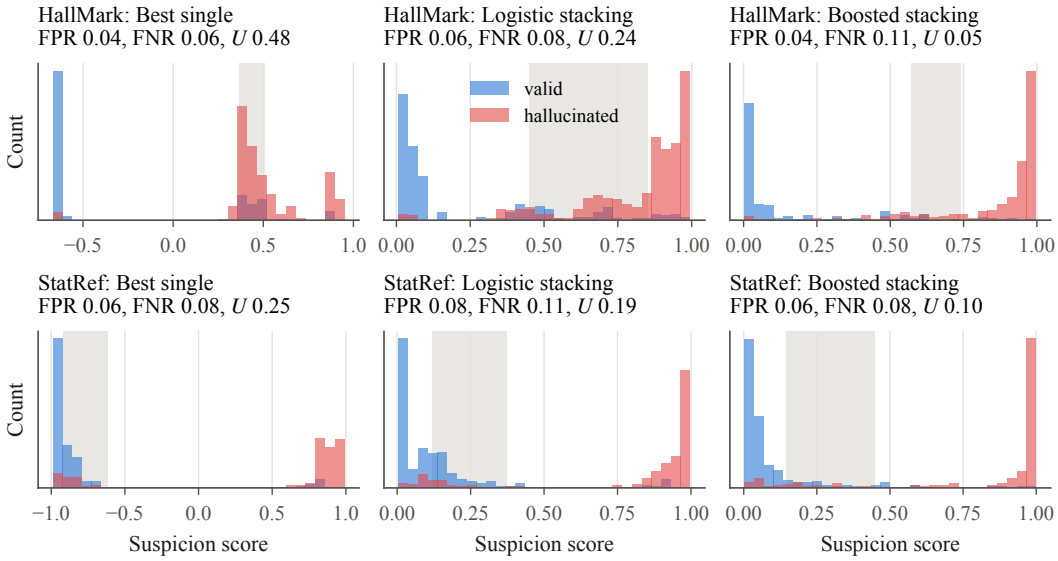

Figure 8: Raw-score illustration: test score distributions for one random 40/30/30 split (drawn with NumPy seed 5, outside the seed sequences of Section C), with the deferral region of Algorithm 1 without tie-breaking shaded ($\alpha_{\mathrm{FP}} = \alpha_{\mathrm{FN}} = \delta = 0.1$). Single-split test rates fluctuate around their averages. The best single detector places many references of both classes inside a narrow deferral region, whereas aggregated scores separate the classes and leave few references there.