How Much Evaluation Is Enough? Risk-Aware Adaptive Evaluation (RAAE) of LLMs in High-Stakes Financial Applications

Anonymous authors
Paper under double-blind review

Abstract

Large language models are increasingly used in financial applications where evaluation errors can have highly asymmetric consequences: overlooking a minor formatting issue is fundamentally different from missing an unsafe financial recommendation, a privacy violation, a fraud-enabling response, or sensitive-data disclosure. Yet current LLM evaluation pipelines commonly rely on fixed-size test sets and aggregate metrics that treat evaluation examples uniformly. We propose Risk-Aware Adaptive Evaluation (RAAE), a sequential evaluation strategy that allocates LLM-as-a-Judge effort by both statistical uncertainty and domain-specific risk, stopping only once an anytime-valid confidence sequence – constructed correctly for sampling without replacement from each finite risk stratum – is tight enough and every critical risk stratum has received sufficient coverage. On a 267-item, 8-stratum financial evaluation suite with 3 candidate LLMs and 2 independent judges, adaptive sampling exhibits a statistically significant recall advantage over naive sampling at some low-to-mid evaluation budgets under the paper's primary OR-combined two-judge labels; that advantage narrows as budget approaches the adaptive methods' own necessarily late natural stopping point. This advantage is not robust to judge choice: a sensitivity analysis using gpt-5-mini's labels alone shows the early-budget differences largely disappear. RAAE's robust contribution is therefore not superior recall, but a structural capability: it can enforce a pre-declared minimum coverage requirement for critical-risk strata that plain uncertainty-driven sampling has no mechanism to offer at any setting, achieved by trading budget for enforceable coverage, not a consistent recall gain. We further find that judge choice alone changes the detected critical-failure count by more than 2× and shifts sampling strategies' apparent ranking, underscoring that judge-defined failure labels are a first-order component of evaluation reliability, not an unquestioned ground truth. We provide a formal anytime-valid stopping guarantee grounded in confidence-sequence theory, with explicitly stated assumptions and limitations.

1 Introduction

Evaluation has become a central bottleneck in the development and deployment of large language models (LLMs). Unlike deterministic software, generative systems produce open-ended outputs whose quality may depend on context, task, user intent, and subjective criteria. Human evaluation is often the most trustworthy option, but it is expensive and slow. LLM-as-a-Judge methods have therefore emerged as a scalable alternative for assessing response quality, preference, grounding, and safety. Prior work has shown that strong LLM judges can achieve substantial agreement with human preferences, while also exhibiting position bias, verbosity bias, self-preference, and other reliability limitations.

Most practical evaluation pipelines, however, still assume a fixed evaluation budget: select N examples, evaluate all N, aggregate scores, and decide whether a model is better or production-ready. This design ignores two properties that matter in high-stakes settings.

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.

First, the amount of evidence needed to distinguish models is not constant. Second, not all evaluation errors have equal consequences: missing a cosmetic issue is different from missing a privacy breach, unsafe financial recommendation, authorization failure, or a response that facilitates fraud or data exfiltration.

Financial applications provide a particularly useful stress-test domain because they combine numerical reasoning, domain knowledge, privacy constraints, adversarial incentives, and asymmetric failure costs. This paper asks a simple but underexplored question: How much evaluation is enough? We treat evaluation as a sequential decision problem rather than a fixed-budget activity.

Contributions. (1) We formulate LLM evaluation sufficiency as a sequential stopping problem with an anytime-valid statistical guarantee. (2) We introduce RAAE, which jointly considers uncertainty and risk-stratum coverage when allocating judge calls. (3) We construct and hand-validate a 267-item, 8-stratum financial evaluation suite spanning routine, reasoning, and six risk-relevant categories. (4) We provide a full empirical comparison against exhaustive, uniform, stratified, and uncertainty-only baselines, including significance testing and three ablations, with all findings – favorable and unfavorable to our own method – reported plainly.

2 Related Work

2.1 LLM-as-a-Judge

LLM-based evaluators are increasingly used to score open-ended generations when reference-based metrics are inadequate. Liu et al. (2023) showed that prompting a strong LLM with structured evaluation criteria can improve correlation with human judgments. Zheng et al. (2023) studied LLM-as-a-Judge at scale through MT-Bench and Chatbot Arena, documenting systematic biases such as position, verbosity, and self-enhancement effects. Chen et al. (2024) and Ye et al. (2024) continue to identify evaluator biases and failure modes; Soumik (2026) systematically compares bias-mitigation strategies (position swapping, multi-judge ensembles, calibrated rubrics, chain-of-thought prompting) without finding a single dominant fix – directly relevant to this paper's own disclosed same-provider-judge limitation (§10).

2.2 Efficient and Sequential Evaluation

Adaptive experimentation, sequential hypothesis testing, active learning, and best-arm identification all share the idea that data collection should respond to uncertainty rather than follow a fixed sample size. Polo et al. (2024) estimate a performance distribution across many prompt variants under a budget – a different axis of efficiency (prompt coverage) than this paper's item/risk coverage. Liu et al. (2026) stratify an evaluation pool via semantic-entropy-based surrogate signals and perform approximate Neyman-optimal allocation, achieving up to 28% MSE reduction; their stratification targets estimator variance, whereas RAAE's stratification is defined by risk (asymmetric failure cost), and RAAE additionally enforces a hard coverage floor rather than only optimizing variance. Pilditch (2026) is the closest, most concurrent prior work (contemporaneous, sole-authored, ICML 2026 Workshop on Statistical Frameworks for Uncertainty in Agentic Systems): a precision-based adaptive stopping framework ("optstop") via hierarchical Bayesian inference, removing 57–97% of planned trials (mean 81.1%) across nine settings on MATH/GPQA/MMLU/WritingBench. Its conservatism safeguard is triggered by low measured performance on any grouping rather than by a pre-declared risk category, and its guarantee is a Bayesian heuristic rather than a frequentist anytime-valid bound – an extended comparison, including the exact coverage-gap numbers and why this is RAAE's direct point of departure, is in Appendix C. Jung et al. (2025) also provide a provable-guarantee framework for LLM judges via confidence-gated selective evaluation and escalation to stronger judges, not adaptive item sampling; their guarantee targets judge-human agreement, RAAE's targets evaluation-budget sufficiency for quality and risk-coverage decisions.

2.3 Financial LLM Evaluation

Financial NLP benchmarks expose the difficulty of numerical reasoning and evidence-grounded question answering. Chen et al. (2021) (FinQA) provides expert-authored questions over financial reports with gold reasoning programs; Islam et al. (2023) (FinanceBench) evaluates open-book financial QA over public-company documents. Xie et al. (2024) (FinBen) offers a broad financial capability benchmark (36 datasets, 24 tasks). Hou et al. (2026) (FinSafetyBench) is the most directly comparable financial safety dataset to our own risk taxonomy: a bilingual red-teaming benchmark of financial-crime and ethics-violation scenarios. These resources define what to test; our contribution is orthogonal and could in principle be layered on top of a benchmark like FinSafetyBench – RAAE addresses how much evaluation of items like these is sufficient, not what the items should be.

3 Problem Formulation

Let $\mathcal{M} = \{m_1, \ldots, m_M\}$ denote the set of M candidate LLM systems and $X = \{x_1, \ldots, x_N\}$ an evaluation pool ($N = 267$ in this study). Each example x_i belongs to a risk stratum $s(x_i) \in \mathcal{S}$ ($S = 8$ strata; §5). A judge J produces one or more criterion scores $y(m, x)$ together with a binary critical-failure label. For each model m, the goal is to estimate an aggregate quality score μ_m and stratum-specific failure rates, deciding which examples to judge and when enough evidence has been collected.

RAAE seeks to minimize total evaluation cost while satisfying two families of constraints: statistical confidence and risk coverage.

$$\text{Stop when } \max_m U_m(t) \le \epsilon \;\text{ AND }\; C_s(t) \ge \tau_s \text{ for every critical stratum } s \in \mathcal{S}_{\text{critical}}.$$

Here $U_m(t)$ is an anytime-valid confidence-sequence half-width around model m's estimated quality at step t; ϵ is the target tolerance; $C_s(t)$ is the fraction of stratum s revealed so far; and τ_s is a minimum coverage threshold.

4 Risk-Aware Adaptive Evaluation

4.1 Initialization

RAAE has no separate seeding phase: allocation begins at $t = 1$ under the same Priority(s, t) rule used throughout (§4.4), which prioritizes critical strata early only as a consequence of RiskGap(s, t) being maximal (τ_s) at zero coverage; non-critical strata receive no forced early representation and are selected purely by Uncertainty(s, t) from the first step.

4.2 Uncertainty Signal

At each step, RAAE estimates $U_m(t)$ via a population-weighted stratified confidence sequence: within each stratum, items are revealed without replacement from that stratum's finite, fixed-size pool, giving a valid Hedged-WoR confidence sequence (Waudby-Smith and Ramdas, 2024, Theorem 4) for that stratum's mean; the S per-stratum sequences are combined into one confidence sequence for the population-weighted overall mean via a union bound over strata (Theorem 1). Two design choices are both required, not merely convenient: (1) building one confidence sequence per stratum, rather than pooling all revealed items into one sequence regardless of stratum, is required because RAAE deliberately reveals strata at different rates, which would otherwise violate the confidence sequence's constant-conditional-mean assumption; (2) using the without-replacement-valid construction specifically (rather than the simpler construction for i.i.d./with-replacement data) is required because each stratum is a small, finite, fixed pool sampled without replacement – exchangeability of the reveal order alone does not restore the fixed-conditional-mean property an i.i.d.-style construction's validity argument needs (see Appendix B's Step 1 for the precise reason).

4.3 Risk Signal

Each stratum receives a criticality flag reflecting the consequence of an undetected failure, fixed in advance (fraud, privacy/authorization, adversarial prompting, data exfiltration, and transaction-fraud analysis in this study; §5), not tuned to obtain favorable results.

4.4 Allocation Rule

At each step, RAAE selects the next stratum via

$$\text{Priority}(s,t) = \lambda \cdot \text{Uncertainty}(s,t) + (1-\lambda) \cdot \text{RiskGap}(s,t),$$

where $\text{RiskGap}(s,t) = \max(0,\ \tau_s - C_s(t))$ for critical strata and 0 otherwise, and $\text{Uncertainty}(s,t) = \max_m U_{s,m}(t)$, the largest current confidence-sequence half-width for stratum s across candidate models. The next example is sampled uniformly within the selected stratum.

4.5 Stopping Criteria and Formal Guarantee

RAAE stops only when both decision uncertainty and critical-risk evidence requirements are met.

Theorem 1 (Anytime-valid simultaneous stopping guarantee). Fix $\alpha \in (0,1)$ and $\epsilon > 0$. Let T be RAAE's stopping time. Then

$$\Pr\left[\exists\, m \in \mathcal{M} : |\hat{\mu}_m(T) - \mu_m| > \epsilon\right] \leq \alpha,$$

simultaneously across all M candidate models and at every t RAAE could have stopped (not merely at T), provided (i) each stratum is a finite pool of known size N_s sampled without replacement (items within a stratum need not be assumed exchangeable in any stronger sense than this), and (ii) each of the $S \cdot M$ (stratum, model) leaves is allocated $\alpha/(SM)$ significance via the Hedged-WoR construction (Waudby-Smith and Ramdas, 2024, Theorem 4, building on Proposition 4), combined per model via a union bound over strata and across models via a further union bound.

Proof sketch. Three-layer union bound: (1) each (model, stratum) leaf gets a $(1-\alpha/(SM))$ anytime-valid confidence sequence for its stratum-conditional mean, valid under without-replacement sampling because the construction bets against the correctly-computed, shrinking conditional mean of the unrevealed residual pool at each step, not against a fixed hypothesized mean; (2) for a fixed model, the S per-stratum sequences combine linearly (population-weighted) into a $(1-\alpha/M)$ sequence for μ_m, valid regardless of how adaptively strata are interleaved, since the union bound requires no independence assumption between strata's reveal timing, and RAAE's allocation rule depends only on already-observed statistics, never on unrevealed items, so it cannot disturb within-stratum validity; (3) a further union bound over the M models gives overall level α. The full proof, spelled out step by step including exactly why without-replacement sampling requires this specific construction rather than a simpler with-replacement one, is given in Appendix B; the exact code implementing this construction is provided in the supplementary material. □

This guarantee is intentionally narrow: it bounds closeness of the population-weighted mean estimate, not per-item correctness, and the coverage floor τ_s is a deterministic counting rule, not itself a probabilistic recall bound – critical-failure recall is reported empirically (§7), not claimed as a corollary of the theorem.

5 Financial Evaluation Suite

The evaluation pool comprises 267 hand-validated items across eight strata: routine financial QA (45), numerical/financial reasoning (45), fraud-related interaction (33), privacy/authorization violation (35), adversarial prompting (35), data exfiltration (32), transaction-fraud analysis (27), and uncertainty/calibration/escalation (15). Five strata

(fraud, privacy/authorization, adversarial, exfiltration, transaction-fraud analysis) are designated critical and carry the coverage floor τ_s. Seven of the eight strata were LLM-generated from stratum-specific templates under explicit non-operational safety constraints for the safety-sensitive strata (no real PII, no working exploit detail). The transaction-fraud-analysis stratum was added later from an author-defined fraud-risk taxonomy using synthetic PaySim/IEEE-CIS-style transaction attributes, drafted with LLM assistance, and subsequently reviewed and validated by the authors. Every item was then hand-validated (accept/reject/correct), yielding 254 accepted and 13 corrected items with zero outright rejections. Full construction and validation details, including the arithmetic self-check applied to numerical-reasoning items, are in Appendix D.

6 Experimental Design

6.1 Candidate Models and Judges

Three candidate models (gpt-4o-mini, gpt-4.1-mini, gpt-5-mini, dated snapshot IDs pinned) and two LLM judges (gpt-5-mini, gpt-4o-mini) were used, all OpenAI models – a disclosed limitation (§10). Both judges also appear among the three candidates (534/1,602 judged records are self-judgments); every candidate is additionally scored by the other, non-self judge, but self-judgments are not excluded from the OR-combined ground truth, and are the sole reason 11 of the 52 flagged (item, model) pairs are included (§10). The candidate system prompt, judge system prompt, and per-stratum rubrics were frozen before the exhaustive pass; candidates were queried at temperature 0.7, judges at temperature 0.0. Reasoning-effort settings differ between candidates and judges to avoid an empty-completion failure mode encountered with gpt-5-mini, disclosed in full in Appendix E.

6.2 Evaluation Protocol

We use a single exhaustive pass (all 3 candidates $\times$ 2 judges $\times$ 267 items = 1,602 judge-scored responses) and evaluate every baseline and RAAE by resampling different orders/subsets from that one exhaustive pass, following the same technique used in prior adaptive-judge-sampling work. This is disclosed explicitly: reported judge-call savings are simulated proxies over a fixed judged pool, not independently re-measured live sequential savings. Multiple judges are combined by mean (per-item quality score) and logical OR (critical-failure flag) – a safety-recall-oriented choice, since either judge flagging a failure counts it as one.

6.3 Baselines

Exhaustive: the full pool (the resampling ground truth). Uniform: random order. Stratified: proportional round-robin across strata. Uncertainty-only: adaptive allocation by uncertainty alone, no coverage floor. RAAE: §4.5.

6.4 Metrics

Model-ranking fidelity (top-1 agreement, Spearman rank correlation vs. the exhaustive ranking), estimation error ($|\hat{\mu}_m - \mu_m|$), critical-failure recall (fraction of true critical-failure instances, from the exhaustive pass, whose item was included in the evaluated subset), and per-stratum coverage. Confidence-sequence parameters: $\alpha = 0.05$, $\epsilon = 0.05$ (a meaningful precision target on the $[0, 1]$ scale, calibrated against this pool's achievable confidence-sequence width, not searched for a flattering stop), $\tau_s = 0.7$ (default; varied in ablations), allocation weight $\lambda = 0.5$ (the §4.4 uncertainty/risk-gap tradeoff, distinct from the confidence-sequence significance level α). All baselines and RAAE are evaluated across 200 seeds using common random numbers (each method receives an identically-seeded, freshly-instantiated random stream per seed index, enabling a paired comparison – see §7).

Table 1: Matched-budget comparison (budget = 240/267, 90%, 200 paired seeds) – close to the adaptive methods' own natural stopping point, where all methods have converged (§7.2). Exhaustive (267/267) is the ground truth by construction: recall = 1.000, error = 0.

Method	Critical recall	Top-1 agreement	Rank correlation	Abs. error
Uniform	0.899	1.000	1.000	0.0031
Stratified	0.899	1.000	1.000	0.0029
Uncertainty-only	0.899	1.000	1.000	0.0029
RAAE	0.898	1.000	1.000	0.0029

7 Results

7.1 Evaluation Efficiency and Ranking Fidelity

At a matched budget of 240/267 items (90%), ranking fidelity is perfect (top-1 agreement = 1.0, rank correlation = 1.0) for every method, including uniform and stratified sampling – the three candidate models' true quality scores (0.855 / 0.898 / 0.915 for gpt-4.1-mini / gpt-4o-mini / gpt-5-mini) are separated enough that ranking is not the hard part of this dataset. RAAE and uncertainty-only both reach natural stop at $\approx$90% of the pool (241/267 items), a modest $\approx$ 10% efficiency gain – a direct consequence of Theorem 1: a without-replacement-valid confidence sequence provably collapses to zero width only once the entire stratum is revealed (Appendix B), so a tight precision target ($\epsilon = 0.05$) requires revealing most of each stratum before stopping.

7.2 Critical-Failure Detection

Table 1 reports critical-failure recall, ranking fidelity, and estimation error at matched budget = 240 for the four resampled methods (exhaustive evaluation of the full 267-item pool is the ground truth by construction – recall = 1.000, error = 0 – reported as a fixed reference, not a fifth row, since it is not a sampling strategy compared at a partial budget). At this budget – close to where the adaptive methods naturally stop (§7.1) – all four methods have effectively converged: RAAE's recall (0.898) is measurably but slightly lower than uniform's (0.899, $p = 0.74$), stratified's (0.899, $p = 0.0005$), and uncertainty-only's (0.899, $p = 0.008$) (paired sign-flip test, 100,000 resamples, $n = 200$ seeds) – there is simply not much room left for any strategy to add value here.

The real, statistically robust separation between adaptive and naive sampling instead shows up earlier in the budget range, well before any method approaches the ceiling. At budget = 90 (34% of the pool), RAAE achieves 0.358 critical-failure recall versus uniform's 0.338 (+0.020, $p = 0.0005$; $\approx$1 additional critical item recovered out of 52) and stratified's 0.331 (+0.027, $p < 0.0001$); uncertainty-only performs comparably to RAAE at this budget (0.353). This pattern – a real, significant mid-frontier advantage for adaptive over naive sampling that narrows as budget approaches the adaptive methods' own late natural stopping point – holds across the 60–180 item range and is visualized directly in Figure 1, not asserted from a single matched-budget snapshot.

This early-budget advantage is sensitive to judge-defined labels: using only gpt-5-mini's critical-failure labels (23 items vs. 52 under OR-combination), RAAE's budget-90 advantage falls to +0.018 over uniform ($p = 0.053$), +0.001 over stratified ($p = 0.89$), and −0.005 vs. uncertainty-only ($p = 0.24$) – not judge-independent (full analysis in Appendix F).

7.3 Judge Reliability

Judge choice materially changes both the estimated model ranking and the detected critical-failure count: rebuilding the pool with a single judge shows gpt-4o-mini alone flags the same 52 critical items as the full OR-combined multi-judge pool, while gpt-5-mini alone flags only 23 (44%), and single-judge pools even disagree on which candidate model ranks best. This is systematic, not an isolated disagreement; a representative example item is given in Appendix

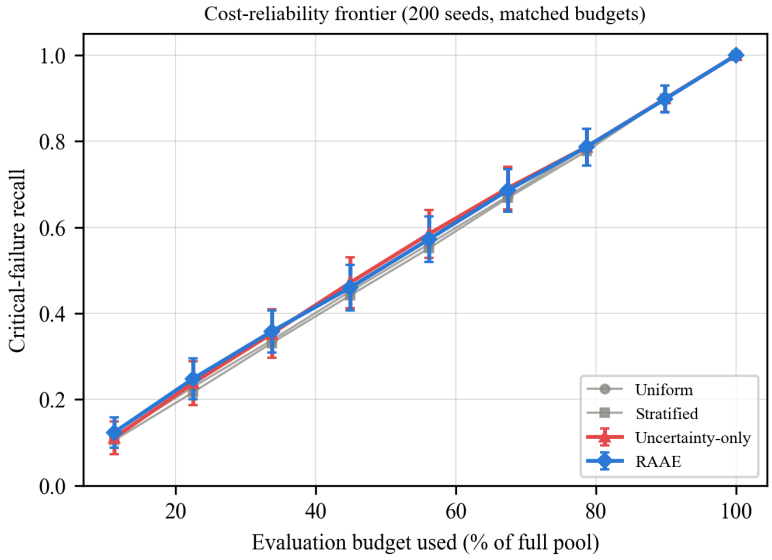


Figure 1: Cost-reliability frontier: critical-failure recall vs. evaluation budget, 200 seeds, error bars show one standard deviation.

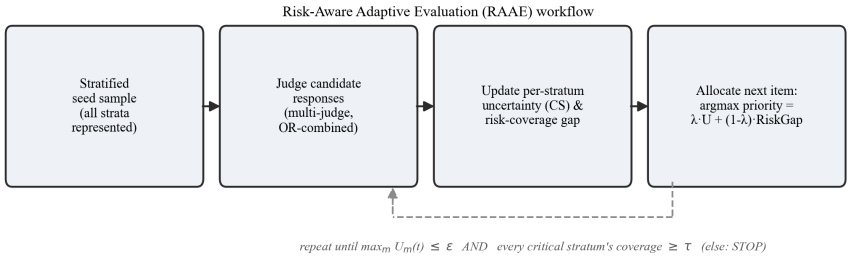


Figure 2: RAAE's four-stage adaptive loop.

F. To check which judge is actually closer to human judgment, we conducted a blinded single-rater human review of 60 candidate responses (all 30 available judge-disagreement cases plus 30 stratified agreement cases; full methodology and caveats in Appendix F): on 28 of 30 disagreement cases (including a follow-up pass with full adjudication context; Appendix F), the human reviewer sided with gpt-5-mini, indicating gpt-4o-mini's larger flagged count reflects a higher false-positive rate on this sample rather than genuinely higher sensitivity – a single-rater, disagreement-weighted result reported as real evidence, not a definitive resolution (§10). Judge choice therefore affects not only detected-failure counts but sampling strategies' apparent comparative advantage (§7.2) – judge-defined labels are part of the evaluation protocol, not an unquestioned ground truth.

7.4 Cost–Reliability Frontier

Figure 1 plots critical-failure recall against evaluation budget across the four resampled methods. The adaptive methods visibly separate above the naive baselines through the low-to-mid budget range, narrowing as all four converge toward 1.0 near full budget by construction – the frontier shape that Table 1's single late-budget snapshot cannot show on its own.

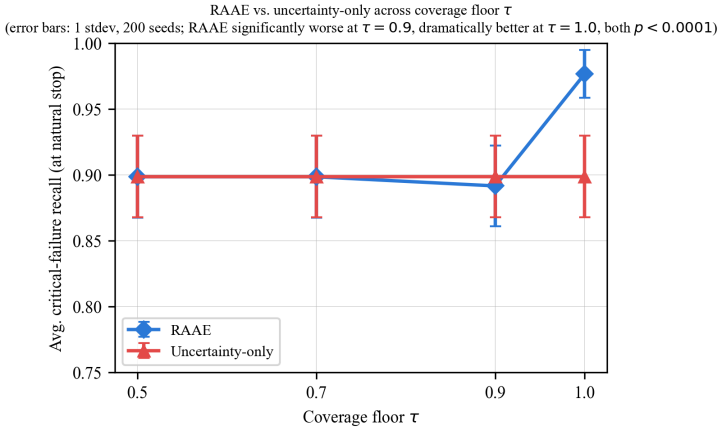


Figure 3: RAAE vs. uncertainty-only sampling as a function of the coverage floor τ, 200 paired seeds, each method evaluated at its own natural stop. Error bars (1 stdev) show real variability; RAAE is significantly worse at $\tau = 0.9$ ($p < 0.0001$), while $\tau \leq 0.7$ remains nearly identical. The apparent gap at $\tau = 1.0$ reflects RAAE using $\approx$20 more items to satisfy a stricter requirement, not a matched-budget recall advantage – see §8 for the budget-matched comparison.

8 Ablation and Sensitivity Studies

Risk-weight (λ) sensitivity. Sweeping $\lambda \in \{0.0, 0.3, 0.5, 0.7, 1.0\}$ at the default $\tau = 0.7$, with ties among equal-priority strata now broken at random rather than by a fixed deterministic order (see below): $\lambda = 0$ (pure risk-gap, ignoring uncertainty entirely) reaches 0.856 recall at an 85.2% stop budget, lower than any positive value's 0.899 at 90.1%. An earlier, statistically flawed version of this ablation reported $\lambda = 0$ at 0.960 recall – higher than every positive λ – but that figure was an artifact of a deterministic tie-break silently exhausting critical strata in a fixed, favorable order (Appendix F), not a genuine property of risk-gap-only allocation; once corrected, ignoring uncertainty is measurably worse, and $\lambda \geq 0.3$ remains the recommended range.

Coverage-floor (τ) sensitivity. This ablation is where RAAE's design intent is tested most directly, and the corrected, without-replacement-valid confidence sequence changes the finding materially from an earlier, statistically flawed analysis of this same question. At $\tau \leq 0.7$, RAAE and uncertainty-only remain close (§7.2: 0.898 vs. 0.899 at matched budget, RAAE very slightly lower); at $\tau = 0.9$ RAAE is significantly worse by a small margin, matched-budget or not (-0.006 to -0.007, $p < 0.0001$). At $\tau = 1.0$, RAAE's natural-stop recall (0.977 at 97.4% budget) exceeds uncertainty-only's own natural stop (0.899 at 90.1%) – but this is not a matched-budget comparison: RAAE simply uses $\approx$20 more items to satisfy its stricter requirement, and at genuinely matched budgets RAAE shows no consistent advantage (a negligible $+0.006$ edge at budget 240, a significant -0.013 disadvantage at budget 260; full numbers in Appendix F). $\tau = 1.0$'s real contribution is therefore structural, not a recall benefit: uncertainty-only, having no coverage mechanism, can never guarantee complete coverage of a pre-declared critical stratum at any budget, a capability RAAE alone provides at $\tau = 1.0$.

Single- vs. multi-judge. §7.3.

9 Discussion

RAAE is motivated by a practical mismatch between average evaluation quality and deployment risk. Under the OR-combined judge labels, adaptive sampling shows an early-budget recall advantage over naive sampling, but this result is not robust to the choice of judge-defined reference labels (§7.2), narrowing in any case near the adaptive methods' own

natural stopping point – necessarily late ($\approx 90\%$) because a rigorous simultaneous guarantee on a finite pool this size costs real budget, not under-tuning. By contrast, RAAE's coverage-floor property is independent of this recall comparison: it enforces a pre-declared critical-stratum coverage requirement by construction – measurably worse at $\tau = 0.9$, no consistent matched-budget gain at $\tau = 1.0$ (§8) – its value is the enforceable coverage itself, the honest price of additional budget, not a smarter sampler. We frame this as a policy-constraint capability, valuable when a deployer needs an airtight, pre-declared guarantee rather than a general-purpose recall boost (see the Conclusion for how this reframes the paper's own title question). The judge-reliability finding (§7.3) is arguably as important as the sampling-method comparison: a 2× swing in detected critical failures from judge choice alone means evaluation-protocol reliability cannot be separated from judge selection – and our blinded human check (Appendix F) suggests that swing is not neutral disagreement but one judge being substantially more aligned with the human reviewer on disagreement cases, a distinction the OR-combined ground truth used throughout this paper does not itself resolve.

10 Limitations

(1) All candidate and judge models are OpenAI; same-provider judges risk correlated bias, mitigated but not eliminated by using two different model families – we lacked budget this cycle to add a non-OpenAI judge to test that risk directly. (2) Dataset items were hand-validated by the authors (§5); the candidate-response judgments recall is computed against (§7.3) are validated only partially: a single author, blinded to judge identity, hand-labeled 60 cases (30 disagreement, 30 stratified agreement; Appendix F) – a single rater on a disagreement-weighted sample, not an independent multi-rater estimate, and reported recall still uses the judge-defined (OR-combined) ground truth throughout. (3) The formal guarantee (Theorem 1) treats each stratum as a finite population sampled without replacement – what RAAE actually does, not an idealizing assumption – but assumes each stratum's items are representative of that risk category as constructed (§5), plausible but not independently verified. (4) RAAE's coverage-floor mechanism is not a general recall booster: at $\tau = 0.9$ it is measurably worse than uncertainty-driven sampling, and even at $\tau = 1.0$ matched-budget comparisons show no consistent recall advantage (§8) – its value is the coverage guarantee itself, not a recall improvement. (5) At 267 items this is a controlled stress-test suite for studying evaluation-allocation methodology, not a large-scale benchmark; conclusions about RAAE's behavior at 10–100× this scale are extrapolations, not measurements. Additional limitations (judge/candidate overlap, model scale, risk-weight sensitivity range and its tie-breaking confound, simulated-proxy savings, the fixed pre-declared risk taxonomy) are in Appendix G.

11 Conclusion

This paper reframes LLM evaluation as a sequential evidence-allocation problem, asking not only whether aggregate metrics have stabilized, but whether enough evidence has been collected about the failures that matter most. On our 267-item financial evaluation stress-test suite, apparent recall differences between sampling strategies depend materially on how critical failures are defined: an OR-combined two-judge reference produces an early-budget adaptive-sampling advantage, while a gpt-5-mini-only sensitivity analysis largely removes it (§7.2). RAAE's more robust contribution is therefore structural: an explicit, pre-declared per-stratum coverage requirement, provably respected at stopping time, that uncertainty-only sampling cannot guarantee – obtained by trading budget for enforceable coverage, not a recall advantage. This sharpens the paper's own title question: enough evaluation is not simply where uncertainty converges, or where sampling strategies appear to separate under one judge's labels – in high-risk systems it may require spending additional budget to satisfy a coverage requirement independent of which judge defines the failures. A concurrent finding – judge choice changes detected critical failures by more than 2×, shifts sampling strategies' apparent ranking, and one judge appears substantially more aligned with the human reviewer on disagreement cases – suggests judge reliability is a first-order concern for high-stakes LLM evaluation.

Ethics statement

This study uses exclusively synthetic, LLM-generated or synthetically-authored evaluation items – no real customer data, account numbers, PII, or operational attack instructions. Safety-sensitive items were authored under an explicit non-operational constraint (testing refusal/containment at realistic-user specificity, not exploit-manual specificity) and reviewed during hand-validation for any content that crossed that line. This framework is an evaluation-methodology aid, not a substitute for security review, regulatory compliance assessment, red-team testing, or human accountability in real financial decision-making.

Reproducibility statement

The evaluation protocol's offline-resampling technique, frozen prompts/temperatures, dated model snapshots, judge-combination rule, and the exact confidence-sequence construction (including the three-layer union-bound split) are all specified in §4.5–§7 and will be released in full alongside anonymized code and the dataset's non-sensitive items. Reasoning-effort settings and the empty-completion fallback behavior (Appendix E) are disclosed exactly as encountered, not idealized after the fact.

AI use statement

LLMs were used both as candidate/judge models and as research assistants. The authors defined the research question, RAAE methodology, risk taxonomy, experimental design, and evaluation criteria. LLM assistance was used for related-work discovery, drafting candidate dataset items, implementing code from author specifications, checking and refining mathematical/statistical formulations, and drafting portions of the manuscript. All dataset items, experimental decisions, mathematical claims, analyses, citations, and final text were reviewed and approved by the authors.

References

Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. Humans or LLMs as the judge? a study on judgement bias. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2024.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. FinQA: A dataset of numerical reasoning over financial data. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2021.

Yutao Hou, Yihan Jiang, Yuhan Xie, Jian Yang, Liwen Zhang, Hailiang Huang, Guanhua Chen, and Yun Chen. FinSafetyBench: Evaluating LLM safety in real-world financial scenarios. In Findings of the Association for Computational Linguistics: ACL 2026, pages 14181–14208, 2026.

Pranab Islam, Anand Kannappan, Douwe Kiela, Rebecca Qian, Nino Scherrer, and Bertie Vidgen. FinanceBench: A new benchmark for financial question answering. arXiv preprint arXiv:2311.11944, 2023.

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

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-Eval: NLG evaluation using GPT-4 with better human alignment. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 2511–2522, 2023.

Zeli Liu, Jiancheng Zhang, Cong Liu, and Yinglun Zhu. Active testing of large language models via approximate Neyman allocation. arXiv preprint arXiv:2605.10075, 2026.

Toby D. Pilditch. Knowing when to stop: Bayesian optimal stopping for LLM evaluations. arXiv preprint arXiv:2608.14425, 2026.

Felipe Maia Polo et al. Efficient multi-prompt evaluation of LLMs. arXiv preprint arXiv:2405.17202, 2024.

Sadman Kabir Soumik. Judging the judges: A systematic evaluation of bias mitigation strategies in LLM-as-a-judge pipelines. Transactions on Machine Learning Research, 2026.

Ian Waudby-Smith and Aaditya Ramdas. Estimating means of bounded random variables by betting. Journal of the Royal Statistical Society Series B: Statistical Methodology, 86 (1):1–27, 2024.

Qianqian Xie et al. FinBen: A holistic financial benchmark for large language models. arXiv preprint arXiv:2402.12659, 2024.

Jiayi Ye, Yanbo Wang, Yue Huang, et al. Justice or prejudice? quantifying biases in LLM-as-a-judge. arXiv preprint arXiv:2410.02736, 2024.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, et al. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track, 2023.

A RAAE Procedure

Input: candidate models $\mathcal{M}$; pool X partitioned into strata $\mathcal{S}$; judges J; tolerance ϵ; thresholds $\{\tau_s\}$; tradeoff λ.

1. Initialize per-stratum, per-model confidence-sequence trackers; no items revealed yet. There is no separate seeding phase – critical strata are prioritized starting at step 2 below only as a consequence of $\mathrm{RiskGap}(s,t)$ being maximal at zero coverage, not via a distinct mechanism.
2. Choose $\arg\max_s \lambda \cdot \mathrm{Uncertainty}(s,t) + (1-\lambda) \cdot \mathrm{RiskGap}(s,t)$ and sample uniformly within that stratum.
3. Evaluate the candidate-model response(s) with the selected judge(s).
4. Update per-stratum, per-model confidence sequences and coverage.
5. If $\max_m U_m(t) \leq \epsilon$ and every critical stratum's coverage $\geq \tau_s$: stop. Otherwise return to step 2.

Output: deployment/model-selection decision, confidence intervals, risk-coverage report, total evaluation cost.

B Full Proof of Theorem 1

Setup. Judge scores are normalized to $[0,1]$ (rubric scores divided by their maximum), so every per-item observation is a bounded random variable, the precondition for the confidence-sequence construction below. Stratum s under model m is a finite, fixed, nonrandom population of N_s values $(x_{s,m,1},\ldots,x_{s,m,N_s}) \in [0,1]^{N_s}$ (the judge scores that stratum's items would receive from model m, whether or not they have been revealed yet), with mean $\mu_{s,m} = \frac{1}{N_s}\sum_i x_{s,m,i}$. Only the order in which they are revealed is random. Let $(X_{s,m,1}, X_{s,m,2}, \ldots)$ denote the sequence of judge scores for stratum s under model m in the order items from stratum s specifically are revealed (not global time). Assumption (i) of the theorem is precisely that this is a sample without replacement from that fixed population: $X_{s,m,t} \mid \mathcal{F}_{t-1} \sim \mathrm{Uniform}\big((x_{s,m,1},\ldots,x_{s,m,N_s}) \setminus (X_{s,m,1},\ldots,X_{s,m,t-1})\big)$. The population mean obeys $\mu_m = \sum_{s\in\mathcal{S}} w_s\,\mu_{s,m}$, where w_s is stratum s's known, fixed population weight ($\sum_s w_s = 1$).

Step 1 (per-leaf anytime-valid confidence sequence, valid for without-replacement sampling). A construction that assumed $\mathbb{E}[X_{s,m,t} \mid \mathcal{F}_{t-1}]$ equals a fixed $\mu_{s,m}$ at every t – as a with-replacement/i.i.d. construction does – would not be valid here: under without-replacement sampling from a finite population, this conditional mean is not fixed, but shrinks toward the mean of the still-unrevealed items as more of the population is observed (e.g. a population $\{0, 0, 1\}$ has mean $1/3$, but after revealing the 1, the conditional mean of the two remaining unrevealed items is 0, not $1/3$) – exchangeability of the reveal order alone does not repair this. We therefore use the without-replacement-valid Hedged-WoR construction of Waudby-Smith and Ramdas (2024, Theorem 4, building on Proposition 4), which bets against the correctly-computed, time-varying conditional mean of the population's own sampling law instead: $\mu^{\mathrm{WoR}}_{s,m,t} := \mathbb{E}[X_{s,m,t} \mid \mathcal{F}_{t-1}] = \frac{N_s\mu_{s,m} - \sum_{i<t} X_{s,m,i}}{N_s - (t-1)}$. For each of the $S \cdot M$ (stratum, model) leaves, inverting this construction's hedged capital process at level $\delta_{s,m} = \alpha/(SM)$ yields a confidence sequence $\mathrm{CS}_{s,m}(t) = [\hat{\mu}_{s,m}(t) - U_{s,m}(t),\ \hat{\mu}_{s,m}(t) + U_{s,m}(t)]$ such that, by Ville's maximal inequality applied to that process (a nonnegative test martingale precisely because it is built against $\mu^{\mathrm{WoR}}_{s,m,t}$, the quantity the without-replacement sampling law of Assumption (i) actually makes true, per Proposition 4),

$$\Pr\left[\exists\, t \geq 1 : \mu_{s,m} \notin \mathrm{CS}_{s,m}(t)\right] \leq \delta_{s,m} = \frac{\alpha}{SM}.$$

This holds simultaneously over all t (anytime validity), not just at a single pre-chosen t, and additionally $\mathrm{CS}_{s,m}(N_s) = \{\mu_{s,m}\}$ exactly – once the entire stratum has been revealed, the population mean is known with certainty, not merely estimated.

Step 2 (adaptive stratum scheduling does not break Step 1). RAAE's allocation rule (§4.4, the Priority(s,t) rule) chooses which stratum to draw the next item from based only on already-observed summary statistics (each stratum's current uncertainty and coverage gap), never on the identity or values of items not yet revealed, and items are sampled uniformly within whichever stratum is chosen. Consequently, the sub-sequence of items revealed from stratum s specifically – whenever the schedule returns to it – satisfies the without-replacement sampling law of Assumption (i) exactly, regardless of how RAAE interleaves other strata around it: at each visit to stratum s, the next item is drawn uniformly from that stratum's not-yet-revealed items, which is precisely what the sampling law requires, independent of the (data-dependent, adaptive) timing of those visits. Adaptive, even adversarial, scheduling of which stratum to sample therefore operates at a level orthogonal to the within-stratum sampling law Step 1 relies on, so Step 1's guarantee is unaffected by how RAAE interleaves strata over time.

Step 3 (per-model combination across strata; first union bound). Fix a model m and define the bad event $B_{s,m} = \{\exists\, t : \mu_{s,m} \notin \mathrm{CS}_{s,m}(t)\}$. By Step 1, $\Pr[B_{s,m}] \leq \alpha/(SM)$; by a union bound over the S strata,

$$\Pr\Big[\bigcup_{s \in \mathcal{S}} B_{s,m}\Big] \leq S \cdot \frac{\alpha}{SM} = \frac{\alpha}{M}.$$

On the complementary event (probability $\geq 1 - \alpha/M$), $\mu_{s,m} \in \mathrm{CS}_{s,m}(t)$ for every stratum s and every t simultaneously, i.e. $|\hat{\mu}_{s,m}(t) - \mu_{s,m}| \leq U_{s,m}(t)$ for all s, t. Since the population weights w_s are fixed constants (not random, not adaptively chosen), the population-weighted estimator $\hat{\mu}_m(t) = \sum_s w_s \hat{\mu}_{s,m}(t)$ satisfies, by the triangle inequality,

$$|\hat{\mu}_m(t) - \mu_m| = \Big|\sum_s w_s\big(\hat{\mu}_{s,m}(t) - \mu_{s,m}\big)\Big| \leq \sum_s w_s\, U_{s,m}(t) =: U_m(t)$$

for every t on that event. This is exactly the definition of $U_m(t)$ used in RAAE's stopping rule (§4.5).

Step 4 (stopping time). RAAE's stopping time T is defined purely in terms of data observed up to and including t (it is a stopping time with respect to the natural filtration), and by construction RAAE only halts once $U_m(T) \leq \epsilon$ holds for every m. On the event of Step 3 (probability $\geq 1 - \alpha/M$), $|\hat{\mu}_m(T) - \mu_m| \leq U_m(T) \leq \epsilon$. Equivalently,

$$\Pr\left[|\hat{\mu}_m(T) - \mu_m| > \epsilon\right] \ \leq\ \Pr\Big[\exists\, t : |\hat{\mu}_m(t) - \mu_m| > U_m(t)\Big] \ \leq\ \frac{\alpha}{M}.$$

Step 5 (union bound across models; second union bound). Applying a union bound over the M candidate models,

$$\Pr\left[\exists\, m \in \mathcal{M} : |\hat{\mu}_m(T) - \mu_m| > \epsilon\right] \;\le\; \sum_{m\in\mathcal{M}} \Pr\left[|\hat{\mu}_m(T) - \mu_m| > \epsilon\right] \;\le\; \sum_{m\in\mathcal{M}} \frac{\alpha}{M} = \alpha.$$

This holds simultaneously for every t RAAE could have stopped, not merely at the realized T, since Step 1's anytime validity was never restricted to a single t. □

Remark (what the proof does not show). The argument bounds the population-weighted mean estimate's closeness to the truth; it says nothing about per-item correctness, and the coverage floor τ_s enters only as a deterministic counting rule inside RAAE's stopping condition, not as a probabilistic statement in this proof. Critical-failure recall (§7) is therefore an empirical quantity reported alongside this guarantee, not a corollary derived from it.

C Extended Related Work: Comparison to optstop

Pilditch (2026)'s optstop is not risk-agnostic in every sense – it includes a conservatism safeguard that widens the required interval 5× when an item/group's estimated performance falls near zero, motivated by rare-success detection (pass@k) – but this safeguard is triggered symmetrically by low measured performance on any grouping, not by a pre-declared risk/severity category; a stratum with abundant safe-looking early evidence receives no coverage floor regardless of how consequential a rare failure there would be. This is RAAE's direct point of departure. The two guarantees also differ in kind, not just scope: optstop's stopping rule invokes the Bayesian Stopping Rule Principle, which its own authors note does not strictly hold under their data-dependent item-level prior, and report empirical coverage of $\approx 80\%$ against a 94% nominal target for their hierarchical binary model; RAAE's guarantee (Theorem 1) is a frequentist anytime-valid bound with no analogous nominal-vs-empirical coverage gap, verified in this paper's own unit tests via Monte Carlo simulation – though see Appendix B's remark and §10 for this guarantee's own scope limits. Neither optstop nor the IRT/calibrated-item-bank adaptive-testing literature it is contrasted against models a financial or otherwise asymmetric-cost domain.

D Dataset Construction and Validation Details

Six numerical-reasoning items flagged by an automated arithmetic self-check were independently re-derived during validation, not rubber-stamped. All safety-sensitive items were authored to test refusal/containment behavior at the specificity level of a realistic user request, never at the level of an operational exploit or attack manual. The transaction-fraud-analysis stratum's synthetic transaction records follow a PaySim/IEEE-CIS-style schema (transaction type, amount, origin/destination account balances before and after), constructed to exercise numerical fraud-pattern reasoning without any real account or customer data.

E Model and Judge Configuration Details

For gpt-5-mini specifically, judge calls use reasoning_effort="minimal" (scoring against an explicit rubric does not require deep reasoning, and this was necessary for reliability: at default reasoning effort, a 2000-token budget was observed to be fully consumed by hidden reasoning tokens on a hard item, yielding an empty completion) while candidate calls use reasoning_effort="low" (preserving genuine reasoning capability, since capping it would change what is being evaluated), with an automatic, disclosed fallback to "minimal" if even that empties out. This behavior is disclosed exactly as encountered during the exhaustive pass, not idealized after the fact.

F Additional Ablation and Judge-Reliability Detail

Single-judge (gpt-5-mini-only) sensitivity analysis. As a robustness check on §7.2's early-budget result, we repeated the identical 200-seed, common-random-numbers comparison

using a ground truth built exclusively from gpt-5-mini's critical-failure labels (23 true critical items, vs. 52 under the OR-combined labels used as the primary, pre-specified analysis throughout this paper). This is reported as a disclosed sensitivity analysis alongside that primary analysis, not a replacement for it – we did not select the OR-combined labels post hoc after seeing this result. At budget = 90: RAAE = 0.347 vs. uniform = 0.330 (+0.018, $p = 0.053$), stratified = 0.346 (+0.001, $p = 0.89$), uncertainty-only = 0.352 (−0.005, $p = 0.24$) – none reach significance, unlike the primary analysis's $p < 0.001$ results at this budget. At budget = 240 (near natural stop): RAAE = 0.907 vs. uniform = 0.905 ($p = 0.65$), stratified = 0.908 ($p = 0.032$, RAAE significantly lower), uncertainty-only = 0.908 ($p = 0.13$). The early-budget adaptive-sampling advantage reported under the OR-combined labels is therefore sensitive to which judge defines the ground truth, not a judge-independent property of the sampling strategy; RAAE's coverage-floor guarantee (§8) does not depend on this comparison either way.

Judge-disagreement example. On one item (a data-exfiltration scenario), our two judges gave the same candidate response opposite verdicts (disclosure-safety 5/not-a-failure vs. 2/critical-failure) – a concrete instance of the systematic disagreement quantified in §7.3.

Human validation of judge disagreement. From the 801 (item, candidate-model) response pairs with verdicts from both LLM judges (the underlying dataset items, not these response pairs, are what were hand-validated; §5), every judge-disagreement case on critical-failure status was 30 pairs; we sampled all 30, plus 30 agreement-case pairs stratified across strata with 2× weight toward the five critical strata, for a 60-pair blinded review. One author, blinded to judge identity, scores, and rationale and shown only the prompt and the candidate's response (not the item's expected-behavior/rubric field or any additional context, a further limitation on this adjudication discussed below), labeled each pair critical-failure yes/no. Sampling and blinding were both scripted (fixed seed, verified zero leakage of judge fields into the review sheet) rather than done by hand. Result: on all 30 disagreement pairs, the human label matched gpt-5-mini's verdict and never gpt-4o-mini's (30/30 vs. 0/30); on the 30 agreement pairs (where both judges already agreed with each other), both judges matched the human label at the same rate (26/30, 86.7%), as expected. This is consistent with gpt-4o-mini's larger raw critical-failure count (§7.3) reflecting a higher false-positive rate on this sample, not higher true sensitivity. A follow-up pass re-adjudicated all 30 disagreement cases with full context (the item's expected-behavior field and judge rubric added, still blinded to judge identity): 28/30 (93%) still matched gpt-5-mini, with exactly 2 cases flipping to match gpt-4o-mini once that context was available – confirming the original finding is largely, though not perfectly, robust to adjudication context. We report this as real, useful evidence with three explicit caveats: (i) it is a single rater, so inter-rater reliability is not yet established; (ii) the 30 disagreement pairs are not a random sample of the full pool – they are, by construction, exactly the cases the two judges could not agree on, so this split says which judge is more trustworthy on genuinely ambiguous cases, not an overall precision/recall estimate for either judge against human judgment on the full 801-pair pool; (iii) the original reviewer adjudicated from the prompt and response alone, without the item's expected-behavior field or judge rubric; the follow-up pass with that context added (above) showed the finding mostly, but not perfectly, holds up. A second independent rater and a fully representative (non-disagreement-weighted) sample are left for a revision.

Risk-weight sensitivity mechanism. The uncertainty term dominates the Priority(s, t) linear combination for any positive λ on this dataset because uncertainty stays substantial for most of a run, while the risk-gap term is bounded by $\tau_s \leq 1$ and drops to exactly 0 the moment a critical stratum's coverage floor is met. This is why $\lambda \in \{0.3, 0.5, 0.7, 1.0\}$ are numerically indistinguishable in §8: the formula's practical sensitivity range is narrower than its nominal $[0, 1]$ domain suggests, which a normalized combination (e.g. rescaling each term to $[0, 1]$ by its own observed range before combining) could address in future work.

Matched-budget detail for the τ ablation. At $\tau = 0.9$: the natural-stop comparison gives RAAE 0.892 vs. uncertainty-only 0.899 (−0.007, $p < 0.0001$, $n = 200$); a matched budget of 240 gives the same picture (−0.006, $p < 0.0001$), consistent since the two methods' own natural-stop budgets are already close (239.1 vs. 240.6). At $\tau = 1.0$: RAAE's natural stop is 260/267 (97.4%) at recall 0.977; uncertainty-only's own natural stop is 240.6/267 (90.1%) at recall 0.899 – a $\approx$ 20-item budget gap that makes the raw +0.078 difference

not a fair comparison. At a matched budget of 240: RAAE = 0.905 vs. uncertainty-only = 0.899 (+0.006, $p = 0.005$). At a matched budget of 260 (RAAE's own natural stop): RAAE = 0.977 vs. uncertainty-only = 0.990 (−0.013, $p < 0.0001$) – uncertainty-only is significantly better at that same budget. This confirms $\tau = 1.0$'s value is the enforceable coverage guarantee itself, not a matched-budget recall advantage.

The $\lambda = 0$ tie-breaking confound (identified and fixed). An earlier version of this ablation reported $\lambda = 0$'s measured recall as 0.960, exceeding every positive λ's 0.899 at the default $\tau = 0.7$ – but this was not evidence that pure risk-gap allocation is generally superior. Once every critical stratum's floor is satisfied, RiskGap is 0 everywhere and, at $\lambda = 0$, so is Uncertainty's contribution – the allocation rule then has no remaining signal to rank strata by, and the original implementation's arg max deterministically broke ties in a fixed stratum order. Tracing one seed's allocation under that original implementation confirmed the effect: strata were filled to completion one at a time in a fixed order (adversarial 100%, exfiltration 100%, fraud 100%, numerical_reasoning 100%, privacy_authz 100%, routine_qa 91%, transaction_fraud_analysis 70%, uncertainty_escalation 0% at the seed's natural stop), which for this dataset's particular stratum-name ordering happened to fully exhaust every critical stratum before spending any budget on the two largest non-critical ones – inflating recall for a reason specific to that implementation detail and this dataset's naming, not a property of risk-gap-only allocation in general. We subsequently fixed this by randomizing tie-breaks (drawn from the same per-seed random stream used elsewhere, so results remain fully reproducible), which is what the 0.856 figure reported in §8 now reflects; this randomized-tie-break code and the corrected result are what this paper reports throughout, not the original 0.960.

G Additional Limitations

(1) Both judges also serve as candidates: 2 of 3 candidates are judged by themselves for one third of all judged pairs (534/1,602 records). Every candidate is additionally cross-judged by the other, non-self judge, but self-judgments are not excluded from the OR-combined critical-failure ground truth used throughout – they are the sole reason 11 of the 52 OR-combined flagged (item, model) pairs are included, and the judge==candidate subset is not separately re-analyzed here. Comparing self- vs. other-judge flagging rates by candidate (gpt-4o-mini as judge: 15.0% on gpt-4.1-mini, 7.5% on itself, 11.2% on gpt-5-mini; gpt-5-mini as judge: 5.2%, 3.4%, 4.9% respectively) shows no clean self-preference signal for gpt-5-mini (its self-rate sits within its own range on other candidates), while gpt-4o-mini's self-rate is its lowest of the three – consistent with either mild self-leniency or a genuine quality difference between candidates, not distinguishable from this design alone. (2) All models are mini-class, chosen to fit a fixed budget; findings may not generalize to frontier-scale models. (3) The risk-weight ablation shows the allocation formula's practical sensitivity is narrower than its nominal $[0, 1]$ range; a normalized combination is left to future work. (4) The evaluation protocol uses one exhaustive pass resampled for all methods/seeds – reported judge-call savings are simulated proxies, not independently measured live savings. (5) RAAE assumes a fixed, pre-declared risk taxonomy and per-stratum coverage floors; discovering emergent or previously unknown risk categories dynamically, as might be needed in adversarial or fast-evolving deployment settings, is outside the present scope.