

Certified Triage for Suggestion Detection in Customer Reviews

Aniket Upadhyay[†] and Aakash Trivedi[†] and Dhruv Kumar and Pratik Narang

Department of Computer Science & Information Systems,

Birla Institute of Technology and Science, Pilani, India

p20241007@pilani.bits-pilani.ac.in, f20191076P@alumni.bits-pilani.ac.in

dhruv.kumar@pilani.bits-pilani.ac.in, pratik.narang@pilani.bits-pilani.ac.in

Abstract

Detecting *suggestions* in reviews forces a trade-off: an LLM on every review is slow and costly; a cheap classifier is fast but its error rate is *uncontrolled*, measurable after the fact, not setttable in advance. We present a **certified triage layer**: a calibration step computes exactly where a classifier may act alone so its error there stays below an operator-chosen target α ; the rest is escalated. On a 14,995-review gold set over 68 industry categories, natural-weighted to a $\sim 9.5\%$ deployment positive rate, it auto-handles 74–93% of the stream at *certified* error for $\alpha \in \{2, 4, 5, 10\}\%$. Heuristic thresholding violates the $\alpha=5\%$ target on 36–49% of draws and an in-expectation conformal gate on 25–28%; only a *betting* certificate bounds the violation rate by δ , making a strict 2% SLA rigorous at 74% order-free coverage. Cost-sensitive training, not threshold tuning, lowers the certifiable error floor to 0.39%, and ~ 200 target labels restore it under cross-industry shift. Of 13 open-weight LLMs (20B–675B), 11 are worse detectors than our fine-tuned 435M encoder, *supervision, not scale, is why* (within-model control, §5.6), and the certificate rides on either backbone. Escalation targets a *human*: on the deferred slice the best LLM oracle still errs at 21.5%.

1 Introduction

Customer feedback arrives faster than anyone can read it, and *suggestions* operationally useful slice (“add online booking,” “open earlier on weekends”) are rare and scattered among evaluative text. Two obvious approaches fail in production: an LLM on every review is the default reflex, but its cost and latency are infeasible at scale (our classifier sustains ~ 40 reviews/s on one GPU versus ~ 1.3 /s for a batched-LLM pipeline, §5.3) and as we find in §5.10 it is not even more accurate here; while a

small classifier is cheap but its error rate is *uncontrolled*. The distinction matters: anyone can *measure* a classifier’s test error, but a fixed threshold cannot give an operations team a bound that holds at a level *they* choose, on the items the system handles alone, on a new draw. Deployment decisions are made against service-level agreements (SLAs) of exactly that form: a team automates a step only if it can say, and defend, “among the reviews this decides on its own, at most $X\%$ are wrong”, with X set before deployment and auditable after it. Nothing in a fixed-threshold pipeline provides that dial.

We provide that missing capability: a detection/triage layer with a provable error guarantee at minimum cost *under a linear, symmetric, fixed-oracle-cost model*. An operator sets α ; a small classifier labels each review and emits an uncertainty score; a conformal deferral gate, calibrated on a small gold set, decides where the classifier may act alone so that the error among auto-handled reviews is provably $\leq \alpha$; the uncertain remainder is escalated. The system maximizes the auto-handled fraction (*coverage*) subject to that guarantee.

Contributions. (1) A certified triage formulation with two guarantee strengths, an in-expectation conformal gate and a finite-sample high-probability *betting* certificate valid under the importance weighting needed to target a realistic base rate (§3). (2) A cost–coverage frontier with confidence intervals and an explicit cost model on 68 industry categories (§5.2, §5.3). (3) A validity demonstration: heuristic thresholding violates the SLA on 36–49% of draws; conformal control holds the *mean* but not the violation rate; only betting holds both (§5.4). (4) Cost-sensitive training that makes a strict 2% SLA certifiable at high coverage, and a domain-shift study with a quantified cheap-recalibration recipe (§5.5, §5.7). (5) A drift stress test with monitoring (§5.8), a SemEval cross-check (§5.9), and a 13-model evaluation of open-weight LLMs from 20B to 675B as detectors and escalation oracles,

[†]Equal contribution.

with a within-model control that isolates *supervision* from architecture and scale (§5.10, §5.6); escalation targets a human because no LLM tier is accurate enough on the escalated slice for end-to-end error to inherit the guarantee.

2 Related Work

Suggestion mining and LLM feedback systems. The closest applied work is Trivedi et al. (2026), a hybrid pipeline pairing a high-recall classifier with an instruction-tuned LLM for suggestion extraction and clustering; it optimizes extraction accuracy, not a quality guarantee. SemEval-2019 Task 9 (Negi et al., 2019) standardized the task; a wave of 2025–2026 LLM feedback systems optimize quality or actionability. None provides a statistical guarantee on its automated decisions.

Selective prediction and learning to defer. Selective classification provides high-probability risk certificates on the accepted region under i.i.d. calibration (Geifman and El-Yaniv, 2017), and learning-to-defer systems optimize a joint deferral loss without such a certificate (Madras et al., 2018; Mozannar and Sontag, 2020). Our setting breaks the i.i.d. assumption both works rely on: calibration is balanced while deployment runs at a 9.5% positive rate, so the risk must be estimated under per-class importance weights. Heavy weighting invalidates the binomial and normal-approximation bounds those certificates use, which is why we certify with a betting martingale.

LLM cascades and routing. FrugalGPT (Chen et al., 2023) and RouteLLM (Ong et al., 2025) cascade/route between models to cut cost, using heuristic or learned routers without an error SLA. We differ in two ways: the routing threshold carries a finite-sample error certificate, and contrary to the usual cascade assumption that a larger model is automatically the quality oracle, we find that an *unsupervised* LLM is a worse detector at this strict low-base-rate boundary regardless of size, while a *supervised* one is better than our encoder but still errs far above α on the escalated slice. Either way the cheap model carries the certificate and escalation targets a human.

Conformal risk control. We build on conformal risk control (Angelopoulos et al., 2024) and risk-controlling prediction (Bates et al., 2021). For high-probability control under importance weighting we use betting confidence sequences (Waudby-

Smith and Ramdas, 2024), whose time-uniform foundation is Howard et al. (2021) and whose near-optimality for bounded means (Shekhar and Ramdas, 2023) explains why betting dominates Hoeffding–Bentkus and normal-approximation intervals at the strict- α , small-sample end, exactly the regime of §5.5. For shift we draw on weighted/adaptive conformal inference (Tibshirani et al., 2019; Gibbs and Candes, 2021) and Mondrian conformal prediction (Vovk et al., 2003).

3 Method

Triage gate. A small model assigns each review a probability p of containing a suggestion and a confidence $c = \max(p, 1 - p)$. The gate auto-handles reviews with $c \geq \tau$ (predicting the confident class) and escalates the rest. τ is chosen on a held-out gold calibration set so the error among auto-handled reviews is $\leq \alpha$.

Target distribution and weights. The deployment stream is far more skewed than a balanced annotation sample. We estimate the deployment positive rate at $\hat{\pi} = 9.5\%$ (§4) and target the natural distribution with per-class importance weights w_i (positives 0.19, negatives 1.81), so the controlled risk is $R(\tau) = \frac{\sum_{i \in A} w_i \mathbb{I}[\text{wrong}_i]}{\sum_{i \in A} w_i}$ over the accepted set A . Heavy weighting reduces the effective sample size $\text{ESS} = (\sum w)^2 / \sum w^2$ (Appendix Table 10), which governs how tightly the guarantee can be set. Validity is robust to misspecifying $\hat{\pi}$; coverage is not: re-weighting to 5% or 15% (vs. 9.5%) shifts coverage by ≤ 2 points at $\alpha=5\%$, but at $\alpha=2\%$ a 15% assumption costs 21 coverage points (Appendix Table 21). $\hat{\pi}$ is therefore an economic input, not a safety-critical one.

In-expectation guarantee. Following conformal risk control (Angelopoulos et al., 2024), we pick the largest-coverage τ with $\frac{n}{n+1} \hat{R}(\tau) + \frac{1}{n+1} \leq \alpha$ (with $n=\text{ESS}$). This controls $\mathbb{E}[R]$ under exchangeability but, with max-coverage selection, can slightly overshoot on a given draw—motivating a high-probability certificate. Plugging ESS into the CRC bound is a *heuristic*; we label it as such wherever it appears.

Weighted betting certificate. Certifying $R(\tau) \leq \alpha$ is equivalent to $\mathbb{E}[w(\mathbb{I}[\text{wrong}] - \alpha)] \leq 0$. Affine-mapping the bounded variable onto $Y \in [0, 1]$ (using the maximum weight as scale) turns this into $\mathbb{E}[Y] \leq \alpha$, which we test with a betting martingale $K_t = \prod_{i \leq t} (1 - \lambda_i(Y_i - \alpha))$

with predictable λ_i (Waudby-Smith and Ramdas, 2024); by Ville’s inequality (Ville, 1939), $K_t \geq 1/\delta$ certifies $R < \alpha$ with error $\leq \delta$. We select τ over a fixed grid with a Bonferroni union bound (so the reported coverage is valid for the data-chosen threshold). Unlike a Wilson normal approximation, this is rigorous under the heavy weighting; synthetic validation confirms the false-certify rate is $\leq \delta$ even at $\alpha = 2\%$ (Appendix Table 12).

Cost-sensitive training. The error among the *most confident* predictions is dominated by heavily weighted false positives when training on a balanced sample but deploying at 9.5%. Class-weighting the loss to the natural prior removes this over-confidence and lowers the certifiable error floor (§5.5).

Order-free certification. The betting martingale scans the calibration set sequentially, so its wealth and hence the certified coverage depends on the data *order* (78.8% vs. 73.9% at $\alpha=2\%$). Choosing an order after seeing the data would void the guarantee. We therefore certify with the *permutation-averaged e-value* $\bar{K}_n = \frac{1}{P} \sum_p K_n^{(p)}$ over P random permutations, rejecting when $\bar{K}_n \geq 1/\delta$: each terminal wealth satisfies $\mathbb{E}[K_n] \leq 1$ under H_0 , so $\bar{K}_n$ does too, and Markov’s inequality gives a valid level- δ test. (We average *terminal* wealth, not the running maximum, since $\sup_t K_t$ is not an e-value.) This is deterministic given the data and reproducible by an auditor. Hoeffding–Bentkus and Wilson bounds are permutation-invariant and need no such treatment. **A deployment audit should quote the order-free number** (Table 4).

Cost model. With per-decision costs $c_{\text{clf}} \approx 0$, c_{oracle} for the escalation target, and escalation fraction p_{esc} , the expected cost is $N[c_{\text{clf}} + p_{\text{esc}}c_{\text{oracle}}]$. Under this linear model with an approximately-free classifier, a fixed per-escalation oracle cost, and symmetric FP/FN costs, coverage-maximization is cost-minimization and the optimization reduces to a one-dimensional threshold search; we claim no stronger optimality, and the contribution is the certified *constraint*, not the objective (asymmetric costs reduce to a per-class threshold). We report \$/1k reviews, throughput, and end-to-end error as a function of oracle accuracy.

4 Experimental Setup

Data and provenance. We start from 15,000 hand-labeled English reviews (balanced 7,500/7,500) drawn from the Yelp Open Dataset (Yelp, 2023) and the Amazon Reviews corpus (Hou et al., 2026) released by the McAuley Lab. Five rows with conflicting annotator labels on identical text are dropped, leaving 14,995 (7,497 positive / 7,498 negative). Reviews span 68 industry *categories* assigned post-hoc from source metadata, grouped into four macro-domains (Product, Service, Digital/Media, Hybrid retail). Each review has a binary suggestion label and an extracted suggestion span; positives are further tagged with one of five surface forms (optative, conditional, modal/deontic, imperative, interrogative). The deployment positive rate $\hat{\pi} = 9.5\%$ is estimated from a random sample of 10,000 unfiltered reviews from the same corpora (Wald 95% CI [8.9%, 10.1%]); per-class natural weights map the balanced eval sample to that rate, and **all errors are reported natural-weighted**. Source review ids are reused across per-industry source files, so we split *group-by-normalized-text* (not by id) into 60/20/20 train/calibration/test (8,997/2,998/3,000), stratified by label $\times$ macro-domain, with zero text overlap; calibration is reserved untouched for the gate and test is sealed.

Annotation. Each of the 14,995 reviews was labeled by one of three annotators against a written guideline defining a suggestion (a proposed or desired change, recommendation, or request; purely evaluative or descriptive text is excluded). A 2,000-review subset was independently labeled by all three. Pre-adjudication inter-annotator agreement on that subset is Fleiss’ $\kappa = 0.73$ (substantial on the Landis–Koch scale; pairwise Cohen $\kappa \in [0.731, 0.732]$). Model-assisted flagging surfaced borderline items (10.4% of the full gold); these and all disagreements were adjudicated jointly with written rationales. The released gold labels are the adjudicated labels: on the 2,000-row subset they coincide with majority consensus on every row. After adjudication, residual disagreement versus gold is 6.67% of annotator votes (400/6,000) and 20% of items still have a single dissenter. The certified guarantee is relative to these labels. A homogeneous independent-flip reading of $\kappa = 0.73$ would imply a $\sim 7\%$ noise floor, but that model is inconsistent with the 0.39% measured error on the auto-handled region; we therefore bound noise on

the region the certificate actually covers (§5.11) rather than inverting κ corpus-wide.

Models. We benchmark six encoders (Appendix Table 9) and deploy DeBERTa-v3-large (He et al., 2021) with cost-sensitive (natural-weighted) training as the headline score model. Open-weight LLMs are evaluated as detectors and candidate escalation oracles: 13 models from 20B to 675B across eight families, including gpt-oss:120b and deepseek-v3.1:671b (temperature 0, served via Ollama; cloud-hosted models at the endpoint’s default precision, which we did not control). We make no claim about proprietary frontier models.

5 Results

5.1 Classifier

Appendix Table 9 compares six encoders on the sealed test set. Balanced F1 barely moves with model size (0.81 $\rightarrow$ 0.88), but the ranking quality that drives triage coverage—natural-weighted AUPRC—rises sharply (0.51 $\rightarrow$ 0.75). At the natural base rate the raw classifier at threshold 0.5 has precision only 0.43 (recall 0.87) yet ROC-AUC 0.95: the ranking is strong but the fixed operating point is poor—exactly the gap the gate exploits. The deployed score model is the cost-sensitive deberta-v3-large-natural (balanced F1 0.863; natural F1 0.592).

5.2 Cost–coverage frontier with confidence intervals (RQ1)

Selecting τ on the reserved calibration set and measuring realized error on the sealed test, the in-expectation gate (cost-sensitive DeBERTa-v3-large) auto-handles 80–98% of the stream with realized error tracking α (Table 1; CIs over 300 calibration resamples). A classifier-only system has 100% coverage but 11.0% uncontrolled natural error at $\tau=0.5$ —failing every SLA; even a natural-error-tuned threshold reaches only 6.2% and still fails a 5% SLA at 95% confidence (Appendix Table 20). As anticipated by max-coverage selection, the in-expectation mean can sit at or marginally above α (e.g. 5.3% at $\alpha=5\%$); the betting certificate (§5.5) provides the strict high-probability guarantee at slightly lower coverage. **Table 1 is heuristic, not certified**; a deployment should quote Table 4 / Table 29.

α	coverage [95% CI]	realized err [95% CI]
2%	80% [78, 82]	2.1% [1.6, 2.6]
4%	86% [83, 88]	4.0% [3.2, 5.1]
5%	89% [86, 92]	5.3% [3.9, 6.6]
10%	98% [96, 100]	10.0% [9.0, 11.0]

Table 1: **Heuristic operating points, not the certified ones.** The in-expectation conformal gate plugs the effective sample size into the CRC bound ($n=ESS$): it controls $\mathbb{E}[R]$ but not the per-draw violation rate, and its realized mean sits at or marginally above α at the tightest settings. Sealed test, 95% CIs over 300 calibration resamples.

α	p_{esc}	\$/1k (LLM)	\$/1k (human)
2%	21%	\$0.047	\$51.6
5%	12%	\$0.026	\$29.0
10%	2%	\$0.004	\$ 4.4
all-oracle: \$0.226/1k (LLM), \$250/1k (human)			

Table 2: Cost model. Symmetric FP/FN, fixed oracle cost; LLM \$/call from measured tokens (~ 723 in/196 out), human at \$0.25/review.

5.3 Cost and throughput

Table 2 gives the explicit cost. The classifier is effectively free; the escalated fraction p_{esc} drives cost. With a human oracle at \$0.25/review, triage cuts review cost from \$250/1k (all-human) to \$51.6/1k at $\alpha=2\%$ and \$4.4/1k at $\alpha=10\%$. The classifier sustains ~ 40 reviews/s on one GPU versus ~ 1.3 /s for a 12-worker LLM pipeline at 9.1s/call— $30\times$ faster—so all-LLM is infeasible at stream scale. End-to-end error depends on the oracle (Appendix Table 13); even a 90%-accurate human keeps end-to-end error near α . An operations A/B in human-hours and dollars is in Appendix Table 24.

5.4 Validity vs. heuristic thresholding (RQ2)

We draw 200 random calibration sets of size n , pick τ three ways, and measure realized error on the fixed sealed test (Table 3). The heuristic threshold—targeting α empirically, the prevailing practice—violates the 5% SLA on 36–49% of draws at every n , with mean error exceeding α at small n . Conformal holds the *mean* $\leq \alpha$ but still violates on 25–28% of draws. Only the betting certificate holds the violation rate $\leq \delta$ (Appendix Table 11, Appendix Figure 2). The coverage cost is largest at small n , where the weighted ESS is small (Appendix Table 10: $n=100 \Rightarrow ESS \approx 46$ accepted). Re-running the $\alpha=5\%$ betting comparison at $B=1,000$ gives 4/1,000 violations (Clopper–

n	method	mean err	viol.	cov.
100	heuristic	5.4%	0.49	0.80
	conformal	4.2%	0.25	0.65
	certified	3.4%	0.08	0.36
2000	heuristic	4.7%	0.36	0.82
	conformal	4.6%	0.28	0.82
	certified	3.9%	0.00	0.80

Table 3: RQ2, $\alpha=5\%$, 200 draws. Heuristic violates the SLA on 36–49% of draws at every n . Conformal controls the mean, not the violation rate. Betting controls both.

method ($\alpha=2\%$)	A	B	C
Weighted RCPS-HB	60.8	60.8 [‡]	56.8 [0,74]
Weighted RCPS-Wilson	68.7	68.7 [‡]	66.1 [0,74]
Betting, as-calibrated	78.8	74.2 [74,79]	75.0 [74,79]
Betting, order-averaged e-value	73.9	73.9 [‡]	—

Table 4: Certified coverage (%) under each reporting procedure. **A**: deterministic operating point on the full calibration set in a pre-registered (id-sorted) order—the number a deployment audit should quote. **B**: same data, re-ordered. **C**: bootstrap over calibration draws—a variability diagnostic, *not* an operating point. [‡]order-invariant by construction. We headline the order-free 73.9%.

Pearson 95% [0.0011, 0.0102]), an order of magnitude inside $\delta=0.10$.

5.5 Strict- α rigor and cost-sensitive training

A Wilson normal-approximation upper bound is anti-conservative under heavy weighting: at $\alpha=2\%$ it violates the intended δ on 24–49% of real-data draws, whereas the betting certificate’s false-certify rate stays $\leq \delta$ in synthetic validation, tight at the boundary (Appendix Table 12, Appendix Figure 3). With a balanced-trained classifier the betting certificate *defers* at 2% because the natural error floor sits at $\sim 1.8\%$. **Cost-sensitive training** drops the floor to 0.39% and makes $\alpha=2\%$ certifiable. The as-calibrated operating point is 79% coverage with 0/200 violations; because that figure is order-dependent we headline the order-free e-value coverage of **73.9%** (Table 4), still +5 points over corrected-Wilson RCPS (68.7%) and +13 over Hoeffding–Bentkus (60.8%; Table 29). Balanced F1 is essentially unchanged (0.876 $\rightarrow$ 0.863); natural F1 *improves* (0.580 $\rightarrow$ 0.592). The strict-SLA fix is a better objective, not threshold tuning.

5.6 Supervision vs. architecture and scale

A fine-tuned encoder versus few-shot decoders cannot by itself attribute the difference to supervision rather than to architecture or scale. We add the missing within-model control: the same Llama-3-8B (Grattafiori et al., 2024) evaluated 0-shot, 8-shot, and LoRA-fine-tuned (Hu et al., 2021) on the *identical* 8,997 training rows, with the same prompt, class weights, token budget, seed, and scoring rule (log-odds of “Yes” vs. “No”). The rows differ in exactly one factor: whether the adapter has been trained (Table 5).

Supervision dominates by 15–21 $\times$. Against the best no-supervision configuration (0-shot, 50.1%) supervision is worth 41.9 points; against the paper’s own 8-shot prompt, 59.8. Varying architecture *and* scale at matched supervision—a 435M encoder against an 8B decoder—is worth 2.8 points (11.0% $\rightarrow$ 8.2%). Two details: 1,000 labels on the 8B model (13.1%) already match deepseek-v3.1:671b at 8-shot (13.7%); and 8-shot is *worse* than 0-shot in natural error (67.9% vs. 50.1%) despite better ranking, because the exemplars raise recall and at a 9.5% base rate each false positive carries 1.81 $\times$ weight.

The fine-tuned decoder is better, not worse. At 8.2% natural error it beats our encoder’s 11.0%. We therefore make no claim that encoders outperform LLMs as detectors; the claim is that *unsupervised* LLMs are poor at a strict low-base-rate boundary irrespective of size (11 of 13 significantly worse; Table 7), and that supervision transfers to decoders too. The encoder is the deployment choice on **cost**—20 $\times$ fewer parameters, $\sim 30\times$ lower per-review latency—not on capability.

The certificate is architecture-agnostic. Running the same weighted betting certificate on the fine-tuned decoder’s scores holds the SLA in all eight cells (Table 6). The decoder’s better ranking buys coverage at loose α ; the encoder’s ultra-confident region is cleaner at strict α . The triage layer, not the backbone, is the contribution.

5.7 Robustness under domain shift (RQ3)

A single pooled certificate is only a *marginal* guarantee. Table 8 gives the macro-domain transfer matrix at $\alpha=5\%$: calibrating on one domain and deploying on another inflates error up to 15.3% (Product $\rightarrow$ Service), while a Mondrian (per-domain) gate (Vovk et al., 2003) restores per-domain validity. No on-diagonal domain signif-

model	supervision	nat err	bal acc	ROC
<i>one model, one prompt, one scoring rule—only supervision varies</i>				
Llama-3-8B	0-shot	50.1 [47.8,52.4]	0.563	0.581
Llama-3-8B	8-shot ICL	67.9 [65.8,69.9]	0.578	0.644
Llama-3-8B	FT, 300 labels	23.1 [21.2,25.0]	0.605	0.687
Llama-3-8B	FT, 1,000 labels	13.1 [11.4,14.8]	0.864	0.941
Llama-3-8B	FT, 8,997 labels	8.2 [7.0,9.3]	0.875	0.952
<i>matched supervision, 20× parameter difference</i>				
DeBERTa-v3-large (435M)	FT, 8,997	11.0 [9.6,12.5]	0.867	0.945
Llama-3-8B (8B)	FT, 8,997	8.2 [7.0,9.3]	0.875	0.952

Table 5: Within-model control; natural-weighted error (%) with 95% CIs in brackets. Supervision accounts for $\sim 95\%$ of the effect; architecture and scale for $\sim 5\%$. Against the best no-supervision row (0-shot) supervision is worth 41.9 points; against the paper’s own 8-shot prompt, 59.8. Architecture *and* scale at matched supervision are worth 2.8 (11.0% $\rightarrow$ 8.2%). 1,000 labels on the 8B model already match a 671B model at 8-shot (13.7%). The fine-tuned decoder is *better* than the encoder: the encoder is the deployment choice on cost, not capability.

backbone	$\alpha=2\%$	$\alpha=4\%$	$\alpha=5\%$	$\alpha=10\%$
encoder (DeBERTa-v3-large)	78.8 (1.65)	81.3 (2.52)	82.8 (3.01)	93.2 (7.53)
decoder (Llama-3-8B, LoRA)	74.0 (1.53)	81.9 (2.51)	85.1 (3.11)	93.0 (5.60)

Table 6: Certified coverage (%) with realized error (%) in parentheses, for the 435M encoder and the 8B decoder. The same weighted betting certificate on either backbone: the SLA holds in all eight cells. These are as-calibrated operating points (procedure A); the order-free headline at $\alpha=2\%$ is 73.9% (Table 4).

icantly exceeds α once CIs and ESS are taken into account (Appendix Table 22; Hybrid ESS 162). Recalibration is cheap (Appendix Table 14): $k \approx 200$ target-domain labels restores Product $\rightarrow$ Service from 15.3% to 5.9%; $k=50$ is too small to certify (ESS ~ 25).

5.8 Drift stress test

As an offline proxy—*not* shadow traffic or a live A/B—we stream the sealed test in 30 batches of 100 with a fixed gate ($\alpha=5\%$). In-distribution, per-batch error fluctuates (40% of batches exceed α) but the running mean is $4.2\% \leq \alpha$. An injected covariate shift (batches tilt 80% toward harder Service/Hybrid content) raises the mean to 5.6%, breaking the SLA; a monitor on a rolling labeled-audit window trips at batch 21, and recalibrating on the recent labeled slice returns the mean to 2.0% (Appendix Figure 4). The certificate is not fire-and-forget: paired with cheap auditing and periodic recalibration it survives this synthetic drift. Live/shadow-traffic and A/B validation remain future work.

detector (strict prompt)	n	nat err	gap vs. clf
classifier (FT, 435M)	800	10.5	—
deepseek-v3.1:671b [†]	400	13.7 [9.9,18.0]	+0.9 n.s.
nemotron-3-ultra	396	17.4 [13.1,22.2]	+4.5 n.s.
qwen3.5:397b	799	19.0 [15.6,22.4]	+8.4
gpt-oss:120b [†]	800	21.9 [18.2,25.7]	+11.3
glm-5.2	799	23.6 [19.8,27.3]	+13.1
deepseek-v4-pro	798	23.5 [19.7,27.2]	+12.9
gpt-oss:120b (re-run)	800	24.4 [20.6,28.3]	+13.8
nemotron-3-super	382	25.7 [19.9,31.7]	+12.3
nemotron-3-nano:30b	800	27.1 [22.9,31.2]	+16.5
minimax-m3	800	27.7 [23.7,31.9]	+17.2
gemma4:31b	800	31.2 [26.9,35.3]	+20.7
mistral-large-3:675b	800	31.2 [27.1,35.5]	+20.7
kimi-k2.6 [‡]	790	31.7 [27.5,36.0]	+21.1
gpt-oss:20b	774	36.6 [32.3,41.1]	+26.3

Table 7: Open-weight LLMs as detectors, same sealed-test reviews and same strict prompt; natural-weighted error (%) with 95% CIs in brackets. Gaps are *paired* bootstrap differences against the classifier on identical items. n varies because rows lost to HTTP-429 rate limiting are excluded as infrastructure failures and differ per model; every gap is a paired bootstrap against the classifier on that model’s own surviving items. 11 of 13 distinct models are significantly worse than a fine-tuned 435M encoder (survives Holm and Bonferroni; Appendix Table 28). The two exceptions are *not shown to differ*, not shown to be better. [†]cached from the original evaluation; deepseek-v3.1:671b has since been retired. An independent re-run of gpt-oss:120b eight months later gives 24.4% against 21.9%. [‡]kimi-k2.6 needed a text fallback on 59.5% of items.

cal $\downarrow$ /test $\rightarrow$	Prod	Serv	Dig	Hyb
Product	4.9%	15.3%	10.6%	10.2%
Service	1.6%	5.3%	2.1%	3.4%
Digital/Media	2.3%	6.4%	4.4%	4.7%
Hybrid	3.1%	9.7%	5.8%	6.2%

Table 8: Cross-domain transfer (calibrate row, deploy column), realized error at $\alpha=5\%$. Off-diagonal shift breaks the marginal guarantee.

5.9 External cross-check: SemEval-2019

SemEval-2019 Task 9 Subtask A (Negi et al., 2019) is sentence-level suggestion detection on software-forum posts, with an independent annotation team. The test split has 833 sentences at a 10.4% positive rate, close to our 9.5%. Our product-review classifier transfers *zero-shot* at ROC-AUC 0.93 (F1 0.65 untuned; 8.5% error), and a fresh DeBERTa-v3-base fine-tuned on SemEval reaches F1 0.772—competitive with the Subtask A leaderboard (OleNet 0.781, Yimmon 0.763). Few-shot LLMs on the same split, including mistral-large-3:675b, sit at F1 0.66–0.69 and er-

ror 8.9–10.3%, by the same precision-at-low-base-rate mechanism (Appendix Table 27). We bound the claim to a fixed, narrowly defined boundary, a low base rate, and $\sim 10^3$ – 10^4 in-domain labels; we make no claim for PII or safety classification.

5.10 LLMs as detectors and escalation oracles

We evaluate 13 open-weight models (20B–675B, eight families) as detectors on the same sealed-test reviews with the same strict prompt (Table 7), scored per model on its own surviving items rather than on a common subset. A three-step prompt ladder on gpt-oss:120b lowers natural error from 41.8% (zero-shot) to 27.4% (few-shot) to 21.9% (strict); deepseek-v3.1:671b reaches 13.7% (Appendix Table 15). With an adequate generation budget, genuine format-failure rates are 0% for every model except kimi-k2.6, which needed a text fallback on 59.5% of items; the residual is excluded rather than scored as an error, a convention that is therefore inert for 12 of the 13 models.

The gap is a precision-at-low-base-rate effect, not low competence. On *balanced* accuracy the strongest LLMs are competitive (0.83/0.77 vs. the classifier’s 0.87); the natural-weighted gap is precision-driven. Scale helps *within* a family (gpt-oss 20B→120B: 36.6% → 24.4%) but saturates far above α , and *across* families parameter count predicts little (675B at 31.2% vs. 671B at 13.7%). Combinations do not rescue the few-shot setting: unanimity across ten models reaches 9.2% natural error but halves recall to 0.489 (Appendix Table 26).

Paired tests, not a 3.2-point gap. Comparing each LLM against the classifier on *exactly the same* reviews (Appendix Table 16), gpt-oss:120b is significantly worse (+11.3, 95% CI [7.5, 15.1]), but deepseek-v3.1:671b is statistically indistinguishable (+0.9, [−4.6, 6.2], $p=0.37$): the original 3.2-point figure compared deepseek against the classifier on a different subset. 11 of 13 models remain significantly worse after Holm and Bonferroni (Appendix Table 28). We say “not shown to differ” for the two exceptions; neither interval excludes a 5-point gap.

Implication for triage. On the slice the gate actually escalates, an LLM *is* substantially better than the classifier it replaces (Appendix Table 17: 21.5% vs. 56.2% at $\alpha=5\%$), so an LLM is a legitimate cost-reducing *middle tier*. A three-tier cascade (classifier → LLM → human on disagreement) holds the 5% target end-to-end at 59% of the

human cost (Appendix Table 25); at $\alpha=2\%$ it does not. Because $21.5\% \gg \alpha$, end-to-end error under an LLM oracle cannot inherit the $\leq \alpha$ guarantee. Escalation targets a human *whenever the SLA must hold end to end*.

5.11 Label noise on the auto-handled region

A paper that certifies 2% owes a quantitative noise bound, not an inversion of κ . Under symmetric flip noise, $\mathbb{E}[R_{\text{obs}}] = \eta + R_{\text{true}}(1 - 2\eta) \geq \eta$, so any measured error upper-bounds the noise rate of that region with no annotator model. The measured natural-weighted floor is 0.39% (calibration) / 0.49% (test); **effective noise on the auto-handled region is therefore $\leq 0.39\%$** , an order of magnitude below $\alpha=2\%$. The inequality requires noise independent of the model’s prediction; the 17.4% error on accepted-borderline items (Appendix Table 18) is evidence against the classifier having absorbed annotator idiosyncrasies.

The gate itself filters ambiguity: borderline mass is 3.4% of the auto-handled slice against 11.3% of the escalated slice (3.3 $\times$). A label-corruption stress test (Appendix Table 23) finds **zero SLA violations across 22 conditions**; as uniform noise approaches α , the certificate refuses to certify rather than issuing a false promise. Label noise makes an observed-label certificate *conservative* for true-label risk, not anti-conservative.

5.12 What the gate escalates

Appendix Table 19 profiles the escalated slice at $\alpha=5\%$. Escalation is not uniform: interrogative suggestions are deferred at 56.7%, borderline items at 46.2% vs. 24.1% for the rest, and Service at 38.2% vs. Product at 15.4%. That is the qualitative content behind the quantitative guarantee: the gate concentrates on the forms and domains where the decision boundary is contested.

6 Conclusion

We presented a certified triage layer for suggestion detection: a cheap classifier with a conformal deferral gate that holds an operator-chosen error SLA, down to a rigorously certified, order-free 2% at 74% coverage. It beats heuristic thresholding on validity, is robust across industry categories with cheap recalibration, and is benchmark-competitive. Across 13 open-weight LLMs, *supervision, not scale*, fixes this strict low-base-rate boundary, and the certificate rides on either backbone: the layer, not the model, is the contribution.

Limitations

We have no production evidence. Every number here is offline, on held-out labels; §5.8 is a *drift stress test* with a synthetically injected covariate shift, not shadow traffic, and there is no live A/B. For a paper built around deployment SLAs that is the central remaining gap, and we state it first.

The conformal guarantee is on the auto-handled error and, in its in-expectation form, is marginal and controls only $\mathbb{E}[R]$ —it violates α on 25–28% of individual draws (Table 3); only the betting certificate controls the violation rate. Validity assumes the calibration sample is exchangeable with deployment, stress-tested in §5.7–§5.8. The certificate is relative to the gold labels; §5.11 bounds effective label noise on the auto-handled region at $\leq 0.39\%$, but an α below a region’s noise floor is not certifiable and the certificate then correctly *defers*. **Coverage, though not validity, is sensitive to the base-rate estimate at strict α .** End-to-end error depends on the escalation oracle: an LLM oracle is markedly better than the classifier on the escalated slice (21.5% vs. 56.2%) but far above α , so only a human oracle lets end-to-end error inherit the guarantee. Our LLM evidence covers *open-weight* models only, at a serving precision we did not control for the cloud-hosted ones; we make no claim about proprietary frontier models. The data is 68 industry *categories* assigned post-hoc across two public corpora, not a production stream from one platform. Remaining work: live/shadow-traffic A/B evidence, asymmetric-cost extensions, and richer oracle models.

Ethics Statement

Annotation. The three annotators were near-native English speakers based in India, members of the research group but not involved in this project, and were paid according to institute norms; no external crowdworkers were engaged and no speed or label-value bonuses applied. Guidelines, borderline criteria, and adjudication rationales are released with the data.

Labor impact. A layer that auto-handles 74–93% of a review stream at certified error displaces triage work, and—by construction—the residual queue is the *hardest* slice, a more sustained cognitive workload for the reviewers who remain. Operators should read α as a workload dial as well as a quality dial, and staff the escalation queue

accordingly.

Data. Reviews come from the Yelp Open Dataset and the Amazon Reviews corpus (McAuley Lab) under their respective licenses; we redistribute derived labels, spans, and taxonomy—not raw review text—and no additional user metadata.

Acknowledgments

This work was supported by BirdEye Inc., Palo Alto, California, USA. The authors gratefully acknowledge BirdEye Inc. for its support of this research.

References

- Anastasios Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. 2024. Conformal risk control. In *International conference on learning representations*, volume 2024, pages 55198–55218.
- Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael Jordan. 2021. Distribution-free, risk-controlling prediction sets. *Journal of the ACM (JACM)*, 68(6):1–34.
- Lingjiao Chen, Matei Zaharia, and James Zou. 2023. Frugalpt: How to use large language models while reducing cost and improving performance. *arXiv preprint arXiv:2305.05176*.
- Yonatan Geifman and Ran El-Yaniv. 2017. Selective classification for deep neural networks. *Advances in neural information processing systems*, 30.
- Isaac Gibbs and Emmanuel Candes. 2021. Adaptive conformal inference under distribution shift. In *Advances in Neural Information Processing Systems*, volume 34, pages 1660–1672. Curran Associates, Inc.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.
- Yupeng Hou, Jiacheng Li, Xiangjun Fu, Zhankui He, An Yan, Xiusi Chen, and Julian McAuley. 2026. Bridging language and items for retrieval and recommendation: Benchmarking llms as semantic encoders. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3251–3265.

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

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

David Madras, Toni Pitassi, and Richard Zemel. 2018. Predict responsibly: improving fairness and accuracy by learning to defer. *Advances in neural information processing systems*, 31.

Hussein Mozannar and David Sontag. 2020. Consistent estimators for learning to defer to an expert. In *International conference on machine learning*, pages 7076–7087. PMLR.

Sapna Negi, Tobias Daudert, and Paul Buitelaar. 2019. Semeval-2019 task 9: Suggestion mining from online reviews and forums. In *Proceedings of the 13th international workshop on semantic evaluation*, pages 877–887.

Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E Gonzalez, Mohammed Kadous, and Ion Stoica. 2025. Routellm: Learning to route llms from preference data. In *International Conference on Learning Representations*, volume 2025, pages 34433–34448.

Shubhanshu Shekhar and Aaditya Ramdas. 2023. On the near-optimality of betting confidence sets for bounded means. *arXiv preprint arXiv:2310.01547*.

Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candès, and Aaditya Ramdas. 2019. Conformal prediction under covariate shift. *Advances in neural information processing systems*, 32.

Aakash Trivedi, Aniket Upadhyay, Pratik Narang, Dhruv Kumar, and Praveen Kumar. 2026. A hybrid supervised-llm pipeline for actionable suggestion mining in unstructured customer reviews. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 823–836.

Jean Ville. 1939. *Etude critique de la notion de collectif*, volume 3. Gauthier-Villars Paris.

Vladimir Vovk, David Lindsay, Ilia Nouretdinov, and Alex Gammernan. 2003. Mondrian confidence machine. Workingpaper. On-line Compression Modelling project, <http://vovk.net/kp>, Working Paper 4.

Ian Waudby-Smith and Aaditya Ramdas. 2024. Authors’ reply to the discussion of ‘estimating means of bounded random variables by betting’. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(1):53–61.

Yelp. 2023. Yelp open dataset. <https://www.yelp.com/dataset>.

A Supplementary figures

The first three plots below show data already tabulated in the main body, for readers who prefer the curve to the table. Figure 1 is the cost–coverage frontier of Table 1; Figure 2 is the validity comparison of Table 3; and Figure 3 is the strict- α rigor comparison of Table 12. Figure 4 is the drift stress test of §5.8.

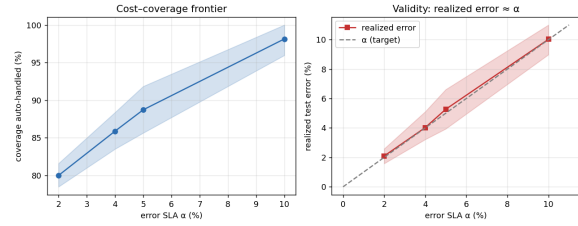


Figure 1: Cost–coverage frontier (cost-sensitive DeBERTa-v3-large, in-expectation gate; 95% CI band over 300 calibration resamples). Left: coverage auto-handled vs. the SLA α . Right: realized test error tracks the α target (dashed). Matches Table 1.

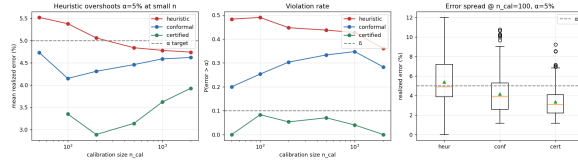


Figure 2: Validity (RQ2, $\alpha=5\%$): mean realized error and violation rate vs. calibration size n for heuristic, conformal, and certified selection, and the error spread at $n=100$ (cf. Table 3).

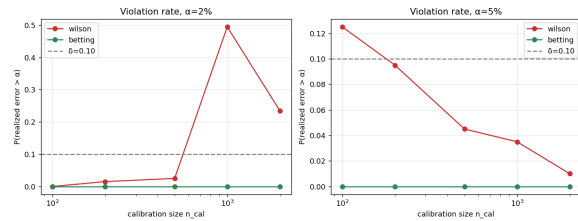


Figure 3: Strict- α rigor: violation rate $P(\text{realized error} > \alpha)$ vs. calibration size for the Wilson normal-approximation bound vs. the weighted betting certificate (target $\leq \delta=0.10$, dashed). Under heavy importance weighting Wilson reaches a violation rate of 0.49 at $\alpha=2\%$ —above δ —while betting holds 0.00 at every n .

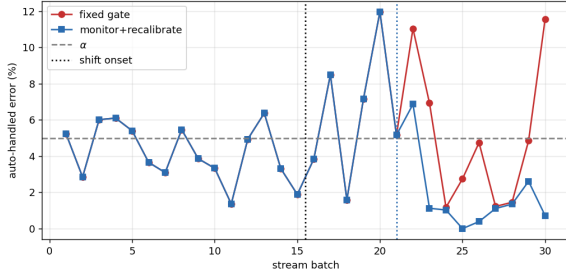


Figure 4: Drift stress test: per-batch auto-handled error under a fixed gate vs. a monitor-and-recalibrate policy. Shift onset at batch 16; the monitor trips and recalibrates, restoring the SLA. This is an injected covariate shift on the sealed test, not production evidence.

B Encoder sweep and calibration effective sample size

Two inputs to the deployed configuration. §5.1 reports that ranking quality, not balanced F1, is what sets triage coverage; Table 9 is the six-encoder sweep behind that statement and behind the choice of DeBERTa-v3-large as the score model. §3 notes that importance weighting to the 9.5% deployment base rate shrinks the calibration set’s effective sample size, which governs how tightly the certificate can be run; Table 10 gives the shrinkage at four nominal sizes.

Encoder	F1 _{bal}	AUPRC _{nat}	ROC _{nat}	p95 ms
DistilBERT	0.809	0.514	0.878	2.7
RoBERTa-base	0.845	0.647	0.924	5.1
DeBERTa-v3-small	0.853	0.611	0.927	6.4
DeBERTa-v3-base	0.870	0.650	0.934	12.4
RoBERTa-large	0.872	0.709	0.942	16.0
DeBERTa-v3-large	0.876	0.750	0.945	35.5

Table 9: Encoder sweep (sealed test, balanced training). Bigger models help *ranking* (AUPRC/ROC), which sets triage coverage, more than balanced F1. The deployed gate uses the cost-sensitive DeBERTa-v3-large (balanced F1 0.863).

nominal n	ESS (full)	ESS (accepted)
100	62	46
500	305	275
2000	1208	1064
2998	1809	1570

Table 10: Effective sample size of the weighted calibration sets ($\alpha=5\%$ accepted subset). Heavy weighting shrinks ESS well below the row count.

C What each guarantee controls

The three threshold-selection rules compared in §5.4 differ in which guarantee they actually deliver, not just in how much coverage they buy. Table 11 states that side by side. Table 12 then measures the quantity that matters for a deployment promise—the rate at which a method certifies an SLA that does not hold—on synthetic data where the true risk is known and on the real calibration set.

method	guarantee		measured ($\alpha=5\%$)		
	mean	tail	mean err	viol.	cov.
heuristic (empirical τ)	×	×	4.7–5.4%	0.36–0.49	0.82
in-expectation conformal	✓	×	4.2–4.6%	0.25–0.28	0.82
weighted betting (ours)	✓	✓	3.4–3.9%	0.00–0.08	0.80

Table 11: What each method actually controls, over 200 draws at $\alpha=5\%$. “mean” is the in-expectation guarantee $\mathbb{E}[R] \leq \alpha$; “tail” is the high-probability guarantee $P(R > \alpha) \leq \delta$. Conformal risk control bounds the *mean* realized error, not the per-draw violation rate; only the betting certificate bounds the latter.

setting	bound	false-certify
$\mu=\alpha=0.05$ (synthetic)	betting	0.10
$\mu=\alpha=0.02$ (synthetic)	betting	0.10
$R=\alpha=0.02$, weighted (syn)	betting	0.07
$\alpha=0.02$, real data	Wilson	0.24–0.49
$\alpha=0.02$, real data	betting	0.00

Table 12: False-certify rate against the target $\delta=0.10$. The betting certificate is rigorous (false-certify $\leq \delta$, tight at the boundary on synthetic data; 0/200 violations on real data) where the uncorrected Wilson approximation is not. $\delta=0.10$ throughout.

D End-to-end error and cheap recalibration

The certificate bounds error on the auto-handled stream only. What a deployment experiences end to end also depends on how good the escalation oracle is, and on whether the gate still holds after a domain change. Table 13 sweeps oracle accuracy a and, in its caption, separates four numbers at $\alpha=2\%$ that are easy to conflate. Table 14 is the recalibration curve behind the ~ 200 -label claim of §5.7.

α	auto err	$a=1.0$	$a=0.95$	$a=0.90$
2%	1.9%	1.5%	2.5%	3.5%
5%	5.1%	4.5%	5.1%	5.7%
10%	10.0%	9.8%	9.9%	10.0%

Table 13: End-to-end error vs. oracle accuracy a on escalated cases. Four related numbers at $\alpha=2\%$ appear in this paper and should not be conflated: 2.1% is the heuristic-ESS mean over resamples (Table 1); 1.9% is the auto-handled error used in this end-to-end table; 1.65% is the as-calibrated certified operating point (Table 6); 0.39% is the classifier’s error floor.

k	src-only	50	100	200	400
err (Prod→Serv)	15.3%	—	11.2%	5.9%	5.6%

Table 14: Cheap recalibration: k target-domain labels; ~ 200 restore the 5% SLA. “—”: $k=50$ cannot certify (ESS too small).

E LLM detectors: prompt ladder, paired tests, escalated slice

Three supporting views of the LLM evaluation in §5.10. Table 15 is the prompt ladder on the two models of the original evaluation, which is where the zero-shot / few-shot / strict progression is measured. Table 16 is the paired comparison against the classifier on identical reviews, which is what corrects the earlier unpaired 3.2-point figure. Table 17 restricts attention to the slice the gate actually defers, which is the only slice on which an escalation oracle is ever used.

detector (prompt)	bal. acc	nat err	prec.
classifier (fine-tuned)	0.873	10.5%	—
deepseek-v3.1:671b (strict)	0.767	13.7%	0.85
gpt-oss:120b (strict)	0.830	21.9%	0.79
(few-shot)	0.824	27.4%	0.76
(zero-shot)	0.747	41.8%	0.67

Table 15: Prompt ladder on the original two LLMs, same 800/400 sealed-test reviews. The large natural-weighted gap is a precision-at-9.5%-base-rate effect. See Table 7 for the 13-model comparison and Table 16 for paired tests.

detector (strict)	n	LLM err	clf, <i>same</i>	paired gap
gpt-oss:120b	800	21.9 [18.1,25.8]	10.5	+11.3 [7.5,15.1]
deepseek-v3.1:671b	400	13.7 [9.8,18.0]	12.8	+0.9 [−4.6,6.2]

Table 16: Paired comparison on identical reviews; natural-weighted error (%) with 95% CIs in brackets. The 3.2-point classifier-vs-deepseek gap in the original evaluation compared deepseek against the classifier on a *different* subset. On deepseek’s own 400 reviews the classifier scores 12.8%; the paired gap is not distinguishable from zero ($p=0.37$).

escalated slice ($\alpha=5\%$)	n	LLM err	clf err
deepseek-v3.1:671b (strict)	113	21.5%	56.2%
gpt-oss:120b (strict)	217	40.1%	49.6%

Table 17: LLM quality on the slice the gate actually escalates. An LLM *is* a better oracle than the classifier it replaces, but $21.5\% \gg \alpha$, so an LLM oracle cannot make end-to-end error inherit the $\leq \alpha$ guarantee.

F Escalation profile and label-noise routing

What the gate defers, described rather than counted. Table 18 shows where annotator-flagged ambiguity ends up: borderline mass is $3.3\times$ denser on the escalated slice than on the auto-handled one, which is the routing behaviour the label-noise bound of §5.11 depends on. Table 19 profiles the escalated slice by suggestion form, annotator flag and macro-domain.

slice ($\alpha=2\%$)	borderline (nat. wt)	nat err
auto-handled	3.4%	1.65%
escalated	11.3%	45.7%
auto-handled, non-borderline	0%	1.09%
auto-handled, borderline	100%	17.4%

Table 18: The gate routes annotator-flagged ambiguity to the human ($3.3\times$ the auto-handled rate). Combined with $\mathbb{E}[R_{\text{obs}}] \geq \eta$ under symmetric flip noise, the measured 0.39% error floor bounds effective label noise on the auto-handled region at $\leq 0.39\%$.

facet	value	% escalated
suggestion form	Interrogative	56.7
	Direct imperative	46.9
	Conditional	39.9
	Modal/deontic	35.7
	Optative	34.5
annotator flag	borderline	46.2
	not borderline	24.1
macro-domain	Service	38.2
	Product	15.4

Table 19: Qualitative profile of the escalated slice ($\alpha=5\%$). The gate concentrates on indirect forms, annotator-flagged borderline items, and the Service domain. Of residual auto-handled error, 26 false positives carry 63% of the natural-weighted mass and 145 false negatives 37% (FP weight 1.81 vs. FN 0.19).

G Tuned classifier-only baseline

The classifier-only comparison in §5.2 is quoted at the default $\tau=0.5$ operating point, which invites the objection that the baseline is being under-tuned. Table 20 removes it: we re-select τ on the calibration split under four different criteria and score each on the sealed test. The best achievable natural-weighted error is 6.2% rather than 11%, but it is bought by pushing τ to 0.945, which doubles balanced error, and it still fails a 5% SLA at 95% confidence. Tuning τ on calibration is precisely the heuristic whose per-draw violation rate Table 3 measures.

classifier-only policy	τ	nat err	bal err	F1
0.5 (original)	0.500	11.0 [9.6,12.5]	13.3%	0.863
Youden’s J (cal)	0.445	11.2 [9.8,12.6]	13.3%	0.864
F1-optimal (cal)	0.130	13.5 [11.9,15.1]	12.8%	0.873
nat-error-optimal (cal)	0.945	6.2 [5.5,6.9]	25.0%	0.675

Table 20: A *tuned* classifier-only baseline (thresholds selected on calibration, evaluated on sealed test); natural-weighted error (%) with 95% CIs in brackets. The best natural error is 6.2%, not 11%—but it still fails a 5% SLA at 95% confidence, it is near-degenerate (balanced error doubles), and it carries no guarantee: tuning τ on calibration is the heuristic of Table 3.

H Base-rate misspecification at strict α

The gate targets the deployment base rate $\hat{\pi}$ through per-class importance weights, so getting $\hat{\pi}$ wrong is a live deployment risk rather than a modelling detail. Table 21 re-runs the certificate with π assumed at 5%, at our estimated 9.5%, and at 15%. Validity survives every assumption; coverage does not. At the loose $\alpha=5\%$ setting the penalty is under

two points, but at $\alpha=2\%$ assuming 15% costs 21 coverage points. $\hat{\pi}$ is therefore an economic input worth estimating well, not a safety-critical one.

assumed π	$\alpha=2\%$		$\alpha=5\%$	
	cov.	realized err	cov.	realized err
5%	80.6%	1.16%	83.8%	2.54%
9.5% (ours)	78.8%	1.65%	82.7%	3.01%
15%	57.4%	1.03%	81.5%	3.59%

Table 21: Validity is robust; coverage is not. The SLA holds at every assumed π , but at $\alpha=2\%$ a $\pi=15\%$ assumption costs 21 coverage points. $\hat{\pi}$ is an economic input, not a safety-critical one.

I Mondrian per-domain ESS

§5.7 reports that a single pooled certificate gives only a *marginal* guarantee, and that a Mondrian (per-domain) gate restores per-domain validity. Table 22 gives the numbers behind that claim: per-domain calibration and test sizes, the effective sample size left after weighting, and the realized error with its interval. No domain significantly exceeds α . The Hybrid retail interval is much the widest because its ESS is the smallest at 162, and per-domain calibration at that scale should be expected to move by ± 3 points from draw to draw—which is the practical argument for more calibration labels per domain rather than a tighter nominal α .

domain	n_{cal}	ESS	n_{test}	err	95% CI
Product	1370	956	1373	4.9%	[3.7, 6.2]
Service	884	444	885	5.6%	[3.2, 8.4]
Digital/media	417	250	416	4.4%	[2.3, 7.0]
Hybrid retail	327	162	326	5.4%	[2.4, 9.2]

Table 22: Mondrian gate at $\alpha=5\%$. No domain significantly exceeds α ; the Hybrid interval is the widest because ESS is smallest. Mondrian calibration at $\text{ESS} \approx 160$ should be expected to wobble by ± 3 points.

J Label-noise stress test

§5.11 bounds effective label noise on the auto-handled region from the measured error, with no annotator model. Table 23 attacks the same question from the other side: corrupt the calibration labels at a known rate η , re-select τ on the corrupted labels, and score the result against the *clean* sealed test. We do this both for uniform noise and for noise concentrated on annotator-flagged borderline items, the more realistic failure mode. Across all 22 cells no draw that certified ever exceeded α on clean labels. What degrades as η approaches α

is not validity but willingness to certify at all: under 2% uniform noise the certificate refuses outright, which is the correct behaviour once the labels can no longer support the promise.

η	uniform noise			borderline-concentrated		
	cert.	cov.	clean err	cert.	cov.	clean err
0	1.00	78.8	1.65	1.00	78.8	1.65
0.5%	0.36	68.1	0.85	1.00	75.2	1.21
1%	0.03	69.1	0.88	0.99	73.2	1.05
2%	0.00	—	—	0.98	69.7	0.90
3%	0.00	—	—	0.95	65.5	0.76
5%	0.00	—	—	0.73	61.7	0.64

Table 23: Calibration labels corrupted at rate η , τ re-selected, error scored against the *clean* sealed test (200 draws/cell, $\alpha=2\%$). No draw that certified ever exceeded α in any of the 22 cells. Where it cannot hold the SLA the certificate refuses to certify rather than promising falsely.

K Operations A/B and three-tier cascade

Two operational readings of the same gate. Table 24 converts the $\alpha=5\%$ operating point into human-hours and dollars over a 100k-review stream and sets it against three alternative policies. The comparison that matters is C versus D. The heuristic threshold buys 5.4 coverage points and \$13 per 1k, and it does hold the SLA *on this draw*—but it is the same rule that violates on 36–49% of draws, so what the certificate actually costs is ~ 45 human-hours per 100k reviews for a promise that survives resampling.

Table 25 then asks whether an LLM can absorb the escalated slice as a middle tier, with a human auditing only the cases where the LLM and the classifier disagree. At $\alpha=5\%$ it works: the three-tier policy holds the target end to end at 59% of the all-human cost and cuts human touches from 17.9% to 10.6%. At $\alpha=2\%$ it does not, because the LLM’s error on the deferred slice (Table 17) is an order of magnitude above the target.

policy ($\alpha=5\%$)	cov.	auto err	SLA	hrs/100k	\$/1k
A: all-human	0%	—	✓	833	250.00
B: classifier-only @0.5	100%	11.0%	×	0	0.00
C: heuristic threshold	88.1%	5.0%	✓ [†]	99	29.70
D: certified gate	82.7%	3.0%	✓	144	43.13

Table 24: Operations A/B over a 100k-review stream (human oracle \$0.25 and 30 s per review). C vs. D is the relevant comparison: the heuristic buys 5.4 coverage points and \$13/1k and holds the SLA *on this draw*—[†]but violates it on 36–49% of draws. The certificate costs ~ 45 human-hours per 100k.

policy ($\alpha=5\%$, deepseek oracle)	e2e err	human	\$/1k
2-tier, no oracle	12.8%	0%	\$0.00
2-tier, LLM oracle	6.6%	0%	\$0.04
2-tier, human oracle	2.8%	17.9%	\$44.82
3-tier: LLM + human on disagr.	4.4%	10.6%	\$26.65

Table 25: Three-tier cascade at $\alpha=5\%$: the LLM handles the escalated slice and a human audits only LLM-vs-classifier disagreements. “e2e err” is end-to-end error and “human” the share of reviews a human touches. The 3-tier policy holds the 5% target end-to-end at 59% of the human cost. At $\alpha=2\%$ it does not (end-to-end 4.0%).

L LLM combinations

§5.10 reports that ensembling does not rescue the few-shot setting. Table 26 gives the combinations behind that claim, over the ten models with near-complete coverage of the common item set. The union is worst, because at a 9.5% base rate every extra false positive carries $1.81\times$ weight. Only unanimity—requiring all ten to say yes—reaches the classifier’s natural-weighted error, and even then the paired gap of -1.25 points is not significant. It gets there by halving recall to 0.489, which for a suggestion detector means discarding half the signal the system exists to find.

combination ($n=761$ common items)	nat err	bal acc	prec.	rec.
LLM union (any says yes)	54.1 [49.5,58.6]	0.692	—	—
LLM majority (10 models)	26.7 [22.8,30.8]	0.788	—	—
classifier + LLMs, majority	23.3 [19.5,27.3]	0.800	—	—
LLM intersection (all 10 yes)	9.2 [7.2,11.4]	0.720	0.906	0.489
classifier alone, same items	10.4 [7.7,13.2]	0.869	0.892	0.837

Table 26: Combinations over the ten models with near-complete coverage; natural-weighted error (%) with 95% CIs in brackets. Unanimity reaches the classifier’s natural-weighted error (paired gap -1.25 , 95% CI $[-4.6, +2.0]$, n.s.) by halving recall—fatal for a suggestion detector.

M SemEval-2019 Task 9A replication

§5.9 cross-checks the detector on SemEval-2019 Task 9 Subtask A: an independently annotated benchmark, in a different domain (software-forum posts), at sentence rather than review granularity, and at a 10.4% positive rate close to our 9.5%. Table 27 gives the full comparison on that split. The supervision finding replicates—a fine-tuned encoder leads, and the few-shot LLMs sit 3.5–4.9 F1 points behind—and it replicates by the same mechanism as on our data: LLM precision of 0.50–0.54 against the encoder’s 0.69, at recall above 0.93. The margin is smaller here than on the re-

view corpus, which is what we would expect on a benchmark whose positive class is more lexically marked.

detector on SemEval-2019 9A	F1	err	prec.	rec.
DeBERTa-v3-base, FT on SemEval-A	0.772	5.4%	0.691	0.874
deepseek-v4-pro (few-shot)	0.686	8.9%	0.544	0.931
gpt-oss:120b (few-shot)	0.669	10.1%	0.509	0.977
mistral-large-3:675b (few-shot)	0.664	10.3%	0.503	0.977
<i>our review classifier, zero-shot</i>	<i>0.647</i>	<i>8.5%</i>	<i>0.570</i>	<i>0.747</i>

Table 27: The supervision finding replicates on an independent task (different annotators, software-forum domain, sentence granularity, 10.4% base rate). The mechanism is the same: LLM precision 0.50–0.54 vs. encoder 0.69. The margin is smaller (3.5–4.9 points).

N Multiplicity correction

The breadth comparison in Table 7 is 14 simultaneous paired tests against the same classifier, so its significance claims need a multiplicity correction before they can be quoted. Table 28 re-tests all of them with a paired sign-flip permutation test at 20,000 permutations: uncorrected, then under Holm–Bonferroni, then under plain Bonferroni. The same 12 of 14 comparisons survive all three, which is 11 of the 13 distinct models once the gpt-oss:120b re-run is counted with its original. The two that do not survive—deepseek-v3.1:671b and nemotron-3-ultra—are *not shown to differ* from the classifier rather than shown to beat it; neither interval excludes a 5-point gap in either direction.

correction over $m=14$ paired tests	significant	threshold
uncorrected	12/14	$p < 0.05$
Holm–Bonferroni	12/14	FWER 0.05
plain Bonferroni	12/14	$p < 0.05/14$

Table 28: 13 simultaneous detector comparisons, re-tested with a paired sign-flip permutation test (20,000 permutations). Eleven of thirteen distinct models remain significantly worse under Holm and Bonferroni. The two exceptions are not shown to differ; neither interval excludes a 5-point gap.

O RCPS comparison

Betting is one of several ways to turn a calibration set into a high-probability bound, so the comparison worth making holds everything else fixed. Table 29 runs unweighted RCPS with a Hoeffding–Bentkus bound, weighted RCPS-HB, weighted RCPS-Wilson and the weighted betting certificate on identical scores, the identical τ grid, the same $\delta=0.10$ and the same Bonferroni correction; the

only thing that differs between rows is the upper-confidence bound itself. Unweighted RCPS targets balanced error and so defers entirely at strict α here. All three weight-aware methods control the SLA, and betting does so at the highest coverage, with the margin widest exactly where a deployment feels it—at $\alpha=2\%$. By $\alpha=10\%$ the three have converged.

Table 30 isolates that strict setting and separates the two things a practitioner needs from a bound: safety, measured as the false-certify rate, and power, measured as the coverage it can still certify. The two baselines that are easiest to reach for—unweighted RCPS-HB and an uncorrected ESS-Wilson interval—are unsafe here, at false-certify rates of 0.99 and 0.30–0.49 against a δ of 0.10. Among the safe methods betting is both the most powerful and the only one that still certifies below the full calibration set, holding 73% coverage on all 200 draws at $n=2000$ where HB and Wilson defer.

α	method	cov. [95% CI]	nat. err	viol.	f-cert.
2%	Unweighted RCPS-HB	— (defers)	—	0.00	0.00
	Weighted RCPS-HB	57 [0.74]	0.8%	0.00	0.00
	Weighted RCPS-Wilson	67 [61.74]	0.8%	0.00	0.00
	Weighted betting (ours)	75 [74.79]	1.2%	0.00	0.00
4%	Unweighted RCPS-HB	— (defers)	—	0.00	0.00
	Weighted RCPS-HB	79 [79.81]	1.8%	0.00	0.00
	Weighted RCPS-Wilson	80 [79.83]	1.9%	0.00	0.00
	Weighted betting (ours)	81 [79.84]	2.4%	0.00	0.00
5%	Unweighted RCPS-HB	— (defers)	—	0.00	0.00
	Weighted RCPS-HB	82 [79.85]	2.7%	0.00	0.00
	Weighted RCPS-Wilson	82 [79.85]	2.9%	0.00	0.00
	Weighted betting (ours)	83 [79.86]	3.0%	0.01	0.01
10%	Unweighted RCPS-HB	90 [85.94]	5.8%	0.00	0.00
	Weighted RCPS-HB	93 [90.95]	7.5%	0.00	0.00
	Weighted RCPS-Wilson	94 [92.96]	7.8%	0.00	0.00
	Weighted betting (ours)	93 [89.95]	7.2%	0.00	0.00

Table 29: RCPS baselines vs. the weighted betting certificate (cost-sensitive DeBERTa-v3-large, 200 bootstrap calibration draws, sealed-test natural-weighted error). Identical scores, τ grid, $\delta=0.10$ and Bonferroni correction throughout; methods differ *only* in the UCB/test. $\text{viol.} = P(\text{certify} \wedge \text{nat. err} > \alpha)$; f-cert. is that rate conditional on certifying; both must stay $\leq \delta$. Unweighted RCPS targets the balanced error and so defers at strict α here; all three weight-aware methods control the SLA, and betting does so at the highest coverage (largest margin at strict α ; the three converge at $\alpha=10\%$). Coverage here is the bootstrap mean (procedure C). The order-free audit number at $\alpha=2\%$ is 73.9% (Table 4); the as-calibrated point is 79%.

method ($\alpha=2\%$)	f-cert.	cov. (full n)	cert. at $n \leq$
Unweighted RCPS-HB [†]	0.99	—	(unsafe)
ESS-Wilson, uncorr. [†]	0.30–0.49	—	(unsafe)
Weighted RCPS-HB	0.00	61%	2998
Weighted RCPS-Wilson	0.00	69%	2998
Weighted betting (ours)	0.00	79%	2000

Table 30: Strict $\alpha=2\%$. Top: the baselines practitioners reach for false-certify the SLA ([†]measured on the balanced classifier, whose $\sim 1.8\%$ natural-error floor sits at the SLA edge; the uncorrected ESS-Wilson is the Table 12 variant, here 0.49 at $n=1000$, 0.30 at $n=2000$). Bottom: all three weight-aware methods control the SLA on the cost-sensitive classifier; betting certifies it at the highest coverage and is the only rigorous method still feasible below the full calibration set ($n=2000$: 73% coverage on all 200 draws, vs. deferral for HB and Wilson).

P Hyperparameters and reproducibility

Code and data. Code and derived data are released at <https://huggingface.co/datasets/au-hf/certifiedTriageSM>: the weighted betting certificate, the RCPS baselines and the ESS-Wilson diagnostic, gate calibration and the τ search, the cost and throughput model, the LLM scoring harness, and the analysis and plotting code for every table and figure. Derived labels, spans, and the industry taxonomy ship under CC-BY-4.0; code is MIT; raw review text remains under the Yelp and Amazon licenses. The README lists environment, seeds, and the command sequence per table and figure.

Data and splits. The gold set is 15,000 reviews (balanced 7,500/7,500) over 68 industry categories and 4 macro-domains, from the Yelp Open Dataset and Amazon Reviews (McAuley Lab); the unit of analysis is the whole review. A 2,000-row subset was triple-annotated (pre-adjudication Fleiss $\kappa=0.73$); after adjudication, residual disagreement versus gold is 6.67% of votes. We drop texts whose duplicates carry conflicting labels (14,995 rows remain) and split **60/20/20** into train/calibration/test, grouping by normalized text (not review_id, which is reused across industries) so no text straddles two splits (leakage = 0), stratified by (label $\times$ macro-domain), seed 42. This yields train 8,997 (4,497 pos.), calibration 2,998 (1,501 pos.), and sealed test 3,000 (1,499 pos.). All headline metrics are *natural-weighted* to the deployment base rate ($\approx 9.5\%$ positive) via per-row weights 0.1904 (positive) and 1.8096 (negative); the maximum weight $w_{\max} = 1.8096$ is used by

the weighted certificates. Calibration is reserved untouched for the conformal phase.

Encoder classifiers. All encoders are fine-tuned with AdamW, a linear schedule, warmup ratio 0.06, weight decay 0.01, max sequence length 512, evaluation batch size 64, bf16, and seed 42; we select the checkpoint with best dev F1 (10% of train held out, stratified). Cost-sensitive (“-natural”) variants weight the cross-entropy loss by the deployment class priors ([neg 1.81, pos 0.19]). Per-model settings are in Table 31. The headline classifier is deberta-v3-large-natural.

model	params	ep.	batch ($\times$ accum)	lr	cls. wt.
DistilBERT	67M	3	16×1	2e-5	—
DeBERTa-v3-small	142M	3	16×1	2e-5	—
RoBERTa-base	125M	3	16×1	2e-5	—
DeBERTa-v3-base	184M	3	16×1	2e-5	—
RoBERTa-large	355M	3	8×2	1e-5	—
DeBERTa-v3-large	435M	3	8×2	1e-5	—
DeBERTa-v3-base-nat.	184M	3	16×1	2e-5	nat.
DeBERTa-v3-large-nat.	435M	2	8×2	1e-5	nat.

Table 31: Encoder fine-tuning hyperparameters. Effective batch size is 16 throughout; large models use gradient checkpointing. “nat.” = natural (deployment) class-weighted loss; the last row is the headline classifier.

Conformal deferral gate. The gate auto-handles a review when classifier confidence $c = \max(p, 1-p) \geq \tau$ and escalates otherwise. We sweep SLAs $\alpha \in \{2, 4, 5, 10\}\%$ and select the maximum-coverage τ on the 400-point grid $\text{linspace}(0.50, 0.999, 400)$ over the calibration split, using three rules: the raw empirical heuristic ($\hat{e} \leq \alpha$), the CRC in-expectation correction ($\frac{n}{n+1}\hat{e} + \frac{1}{n+1} \leq \alpha$), and a certified one-sided Wilson upper bound at $\delta = 0.10$ (90%) with effective sample size $n_{\text{eff}} = (\sum w)^2 / \sum w^2$. The τ grid is 400 points on $[0.50, 0.999]$. Temperature scaling (single scalar, fit on dev by natural-weighted NLL, 400 iters, lr 0.05) is reported as a diagnostic only and not applied (the gate is invariant to monotone rescaling); ECE uses 15 bins. Test-error confidence intervals use 2,000 bootstrap resamples.

Distribution-free certificates. The weighted betting confidence sequence (Waudby-Smith and Ramdas, 2024) uses the predictable plug-in (running mean/variance, priors 0.5/0.25), $\lambda = \sqrt{2 \ln(1/\delta) / (\hat{v} t)}$ capped at $0.75 / (1 - \alpha)$, and the Ville threshold $K_t \geq 1/\delta$. Threshold selection scans the grid $\text{linspace}(0.70, 0.999, 45)$ with a Bonferroni union bound $\delta/45$, $\delta = 0.10$,

minimum accepted count 5. The RCPS comparison (unweighted RCPS-HB, weighted RCPS-HB, weighted RCPS-Wilson, weighted betting; plus an uncorrected ESS-Wilson diagnostic at $z=1.2816$) uses the identical grid, δ , Bonferroni correction, and minimum-count, with $\alpha \in \{2, 4, 5, 10\}\%$, seed 20260616, and 200 calibration draws (bootstrap at full n ; subsample sweep at $n \in \{500, 1000, 2000, 2998\}$). The small- n study (RQ2) uses 300 without-replacement draws per $n_{\text{cal}} \in \{50, 100, 200, 500, 1000, 2000\}$, $\alpha \in \{2, 5, 10\}\%$, $\delta = 0.10$.

LLM arm. We query open-weight models via Ollama at temperature 0. The original evaluation used gpt-oss:120b ($n=800$) and deepseek-v3.1:671b ($n=400$); the revision expands to 13 models (20B–675B). The strict prompt uses 8 boundary few-shot examples and a span-first JSON output. With an adequate generation budget, genuine format-failure rates are 0% for every model except kimi-k2.6, which needed a text fallback on 59.5% of items; the residual is excluded rather than scored as an error, a convention that is therefore inert for 12 of the 13 models. Dollar costs use a reference price of \$0.15/\$0.60 per 1M input/output tokens; the end-to-end cost model assumes \$0.226 per 1k escalated reviews (LLM) or \$250 per 1k (human expert audit). The Llama-3-8B supervision control is LoRA on the causal LM, supervised on the answer token of the same prompt, scored by log-odds of “Yes” vs. “No”.

SemEval-2019 Task 9A cross-check. Zero-shot transfer applies the product-review deberta-v3-large with max length 128. The fine-tuned check trains deberta-v3-base for 3 epochs, batch 32, lr $2e-5$, warmup 0.06, weight decay 0.01, max length 128, seed 42, bf16.

Environment. PyTorch 2.5.1+cu121, Transformers 4.46.3, on a single NVIDIA RTX A4500 (CUDA 12.2, bf16).