IN-CONTEXT CASCADE: FEW-SHOT ADAPTATION FOR BUDGET-CONSTRAINED LABELING AT SCALE

Anonymous authors
Paper under double-blind review

ABSTRACT

Labeling large pools of unlabeled text items with respect to a natural language predicate is a common bottleneck in data annotation, content moderation, and enterprise AI pipelines. While large language models can automate this task, exhaustive labeling by a costly oracle such as a frontier model or a human annotator is infeasible for pools of thousands to millions of items. We propose the *In-Context Cascade* (ICC), a budget-constrained pool labeling framework that combines in-context learning (ICL) adaptation of a cheap model with targeted oracle escalation. ICC allocates a small *harvest budget* B_h from total budget B to oracle-label items near the decision boundary, uses these as ICL demonstrations to adapt the cheap model across the remaining pool, and reserves an *escalation budget* $B_e = B - B_h$ to resolve residual uncertainty. We provide theoretical guarantees on coverage bounds, harvest strategy comparison, optimal budget allocation, and escalation thresholds, leveraging these for the algorithm. Experiments on four datasets with oracle budget B from 1% to 20% of pool size N show that ICC consistently outperforms uncertainty-based cascade baselines, with the largest gains often at the tightest budgets where ICL adaptation is most valuable.

1 INTRODUCTION AND RELATED WORK

Large-scale text annotation is a fundamental bottleneck in modern AI pipelines: training data curation, content moderation, regulatory compliance, and enterprise knowledge management all require accurate binary labels for thousands to millions of items with respect to a natural language predicate. Large language models (LLMs) have emerged as powerful annotators, capable of labeling items with accuracy approaching that of human experts (He et al., 2024; Gilardi et al., 2023). Yet exhaustive annotation by a high-quality oracle such as a frontier LLM or human expert remains prohibitively expensive at scale, while out-of-the-box cheap models often fall short of the accuracy required for downstream use. This tension between annotation quality and cost motivates the need for principled methods that allocate a limited oracle budget intelligently across a large unlabeled pool.

Prior work addresses parts of the pool labeling problem but not their combination. LLM cascade methods route queries between models of different capability: FrugalGPT (Chen et al., 2023) jointly optimizes cascade composition and acceptance thresholds under a cost constraint; TREACLE (Zhang et al., 2024) applies cascades directly to pool labeling; and a growing body of work studies routing signals, online adaptation, and multi-tier designs (Zhang et al., 2023; Ding et al., 2024; Gupta et al., 2024; Nie et al., 2024; Fanconi & van der Schaar, 2025; Ong et al., 2025; Valkanas et al., 2026). Online Cascade Learning (Nie et al., 2024) is closest in spirit, adapting the small model online via imitation learning, but requires gradient updates and operates over streams rather than fixed pools with a global budget. Semantic operator frameworks such as LOTUS (Patel et al., 2025), Stretto (Sanmartino et al., 2026), and BARGAIN (Zeighami et al., 2025) treat the proxy-oracle pattern as a batch processing primitive with accuracy guarantees, while co-annotation frameworks (Li et al., 2023) route uncertain items to human annotators. In all of these approaches, the cheap model is held fixed — oracle calls serve exclusively as direct label assignments or routing signals, and their value as ICL demonstrations for task-conditioned adaptation is left unexploited. Recent work on in-context distillation (Sarukkai et al., 2025; Wu et al., 2025) similarly uses strong model outputs as demonstrations for weaker models, but targets agentic and generative settings with unbounded query streams rather than a fixed unlabeled pool under a hard oracle budget constraint.

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.

We propose the *In-Context Cascade* (ICC): a budget-constrained pool labeling framework that combines in-context learning (ICL) adaptation of a cheap model with targeted oracle escalation. ICC allocates a small *harvest budget* to oracle-label informative items near the decision boundary, uses these as ICL demonstrations (Brown et al., 2020) to adapt the cheap model across the remaining pool, and reserves an *escalation budget* to resolve residual uncertainty. ICL effectiveness is known to depend on demonstration selection (Liu et al., 2022) and label distribution (Zhao et al., 2021), motivating the targeted harvest strategies we develop. The harvest phase draws on ideas from active learning (Settles, 2009), which selects informative items for labeling, though active learning operates under fundamentally different assumptions: it targets model training via gradient updates rather than full-pool annotation, requires iterative retraining, and does not exploit labeled examples as ICL demonstrations. Reliable uncertainty estimates underpin both harvest targeting and oracle escalation, making pool labeling a practical application of uncertainty quantification in LLMs (Guo et al., 2017; Desai & Durrett, 2020; Kadavath et al., 2022) where accurate confidence scores determine what to harvest, which predictions to escalate, and where to set optimal thresholds. We leverage temperature scaling (Guo et al., 2017) for post-hoc calibration and token logit-based confidences (Kadavath et al., 2022; Tian et al., 2023) throughout the pipeline.

The central insight of ICC is that oracle calls serve a dual purpose: beyond directly labeling individual items, a small set of oracle-labeled examples can substantially improve the cheap model's accuracy across the *entire* remaining pool via in-context learning. Existing cascade methods treat oracle calls purely as direct label assignments, discarding the demonstrative value of labeled examples. ICC explicitly exploits this value by decomposing the oracle budget into a component that oracle-labels a targeted set of informative items to serve as ICL demonstrations, and one that oracle-labels items after ICL adaptation. The optimal split between these two uses of the oracle budget is governed by the rate at which ICL adaptation reduces the cheap model's uncertainty across the pool, which we characterize theoretically and estimate from data. This dual-purpose view of oracle calls is absent from all prior work on cascades, active learning, and demonstration selection, and motivates a fundamentally different framework with new theoretical results not derivable from any prior combination. Our main **contributions** are as follows:

- **The ICC framework.** A five-stage pipeline for budget-constrained pool labeling that combines ICL adaptation of a cheap model with targeted oracle escalation, strictly generalizing existing cascade approaches as special cases (see Fig.1).
- **Theoretical analysis.** Coverage bounds for harvesting items (Theorem 1), an optimal budget allocation result under parametric decay assumptions (Theorem 2), a harvest strategy comparison showing uncertainty sampling dominates random sampling with advantage growing for easier predicates (Theorem 3), and optimal escalation thresholds under calibration (Theorem 4).
- **Practical allocation.** A data-driven procedure for estimating the optimal harvest fraction β^* from a small ICL pilot, connecting theoretical parameters to practice using observable quantities.
- **Empirical validation.** Experiments on four real-world datasets demonstrate that ICC consistently outperforms uncertainty-based cascades by up to 5-10 percentage points on accuracy across budget levels $B \in \{1\%, 2\%, 5\%, 10\%, 20\%\}$ of the pool size N, with the largest gains often observed at the smallest budgets.

2 Problem Formulation

Definition 1 (**Budgeted Pool Labeling Problem**). Let $\mathcal{D} = \{x_1, \ldots, x_N\}$ be an unlabeled pool of N items, $\mathcal{T}$ a natural language predicate, and $\mathcal{D}_{cal} = \{(x_i, y_i)\}_{i=1}^{n_{cal}}$ a small and separate caibration set of n_{cal} items with ground truth labels. An expensive oracle $f^+ : \mathcal{D} \to \{0, 1\}$ assigns accurate binary labels at cost one per call. A cheap model $f^- : \mathcal{D} \to \{0, 1\}$ labels items at negligible cost but with lower accuracy $\mathrm{err}(f^-) > \mathrm{err}(f^+) \approx \varepsilon^+$. Given an oracle budget $B \ll N$, the goal is to produce predicted labels $\hat{y}_i \in \{0, 1\}$ for all N items using at most B oracle calls, minimizing the overall error rate $\frac{1}{N}\sum_{i=1}^{N} \mathbf{1}[\hat{y}_i \neq f^+(x_i)]$.

This setting differs from active learning, where labels serve as training data via gradient updates rather than as the primary annotation output, and where fine-tuning at the tight budgets motivating ICC (as low as 1% to 5% of pool size) is infeasible. It also differs from standard LLM cascades (Chen et al., 2024) in that f^- is adapted to $\mathcal{T}$ via in-context learning before routing decisions are made, exploiting oracle labels as demonstrations rather than routing signals alone.

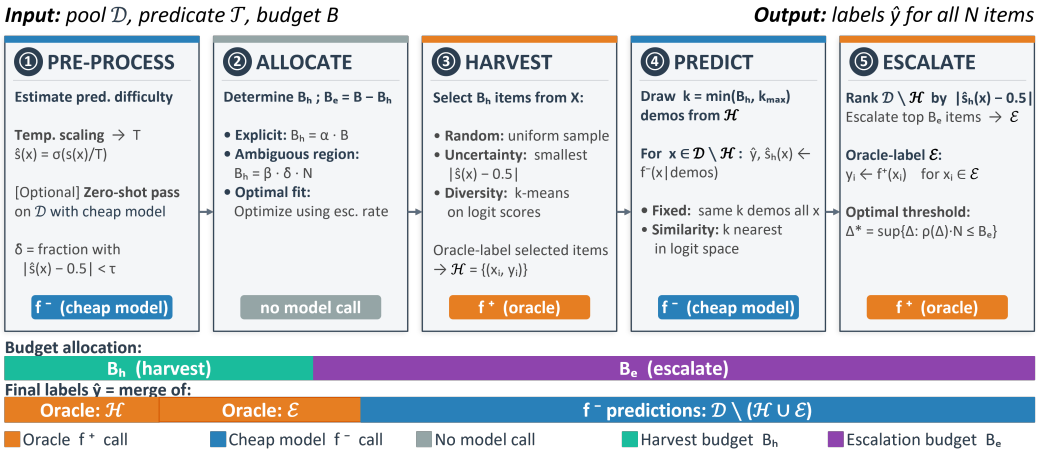

Figure 1: The ICC pipeline. Given pool $\mathcal{D}$, predicate $\mathcal{T}$, and oracle budget B: **(1) Pre-process**: calibrate f^- and estimate predicate difficulty $\hat{\delta}$; **(2) Allocate**: split B into harvest budget B_h and escalation budget B_e; **(3) Harvest**: oracle-label B_h targeted items to form demonstration set $\mathcal{H}$; **(4) Predict**: run ICL-adapted f^- on $\mathcal{D} \setminus \mathcal{H}$; **(5) Escalate**: oracle-label the B_e most uncertain items $\mathcal{E}$. Orange: oracle f^+; blue: cheap model f^-.

In this work, we consider the problem where the oracle is perfect ($f^+(x_i) = y_i$, $\varepsilon^+ = 0$), which is the standard assumption in prior cascade work including FrugalGPT and TREACLE, and enables a clean theoretical treatment of the budget allocation problem.

3 In-Context Cascade: Methodology and Analysis

We outline our proposed solution for the budgeted pool labeling problem: the **In-Context Cascade (ICC)** is a five-step pipeline that combines in-context learning (ICL) adaptation of the cheap model with targeted oracle escalation. The key insight is that a small *harvest set* $\mathcal{H}$ of oracle-labeled examples can substantially improve the cheap model's accuracy on the remaining items via ICL, reducing the number of uncertain predictions that require direct oracle resolution on the *escalation set* $\mathcal{E}$. ICC therefore decomposes the oracle budget B into two complementary uses: a harvest budget B_h for ICL adaptation and an escalation budget $B_e = B - B_h$ for direct oracle resolution of residual uncertainty. Fig. 1 illustrates the pipeline with the following steps:

1. **Pre-process.** Fit a calibrator on a small held-out set and optionally run the cheap model zero-shot on all pool items to obtain calibrated confidence scores $\hat{p}(x_i)$.
2. **Allocate.** Determine B_h and B_e using the calibrated confidence distribution.
3. **Harvest.** Select B_h items from the pool and oracle-label them, forming the harvest set $\mathcal{H}$.
4. **Predict few-shot.** Run the ICL-adapted cheap model on all non-harvested items $\mathcal{D} \setminus \mathcal{H}$ using demonstrations drawn from $\mathcal{H}$.
5. **Escalate.** Oracle-label B_e non-harvested items, thereby forming the escalation set $\mathcal{E}$, and replace cheap model predictions with oracle predictions.

ICC generalizes several existing approaches as special cases: $B_h = 0$ reduces to uncertainty-based cascade labeling (Chen et al., 2024), and $B_e = 0$ reduces to harvest-only ICL adaptation. The full ICC pipeline with $B_h > 0$ and $B_e > 0$ dominates both when the cheap model benefits from ICL demonstrations and when residual uncertainty exists after adaptation. The central design decisions are: (i) how to split the budget B between harvesting (B_h) and escalation (B_e), (ii) which items to harvest, i.e. how to construct $\mathcal{H}$ (of size B_h) to maximize cheap model accuracy on $\mathcal{D} \setminus \mathcal{H}$, and (iii) which items to escalate, i.e. how to construct $\mathcal{E}$ (of size B_e). We address each in the following subsections, with algorithm components motivated by our theoretical analyses.

3.1 THE PRE-PROCESS PHASE: CALIBRATION AND ZERO-SHOT PREDICTION

The pre-processing step prepares two inputs used by all subsequent stages: calibrated confidence scores for the calibration set, and optionally for the full pool.

Calibration. The cheap model f^- assigns a raw confidence score $s(x) \in \mathbb{R}$ to each item x — typically a log-probability or logit derived from the model's token distribution over `Yes`/`No` outputs. Raw scores are often poorly calibrated (Guo et al., 2017), meaning they do not reliably reflect the true probability that the predicate holds. We fit a temperature scaler (Guo et al., 2017) on a calibration set $\mathcal{D}_{cal}$. Temperature scaling learns a single scalar $T > 0$ that maps raw scores to calibrated probabilities: $\hat{p}(x) = \sigma\left(\frac{s(x)}{T}\right)$, where σ denotes the sigmoid function. Temperature T is fitted by minimizing the negative log-likelihood on $\mathcal{D}_{cal}$. Well-calibrated $\hat{p}(x)$ values have a natural interpretation as the cheap model's uncertainty: items with $\hat{p}(x)$ close to 0.5 are genuinely uncertain, while items near 0 or 1 are confidently classified. Calibration quality depends on n_{cal}; empirically, 20–50 items suffice for reliable temperature estimates across the datasets we consider.

Zero-shot prediction (optional). Running f^- zero-shot on all N pool items obtains calibrated confidence scores $\{\hat{p}(x_i)\}_{i=1}^N$, enabling data-driven budget allocation (Section 3.2) and uncertainty-based harvest item selection (Section 3.3). This is a one-time cost amortized across all subsequent experiments. When unavailable, all downstream quantities are estimated from $\mathcal{D}_{cal}$ alone.

Definition 2 (Ambiguous Region and Predicate Difficulty). We define the *ambiguous region* as the set of pool items for which the cheap model is genuinely uncertain: $\mathcal{A}_\tau = \{x \in \mathcal{D} : |\hat{p}(x) - 0.5| < \tau\}$, where $\tau \in (0, 0.5]$ is an ambiguity threshold over the calibrated scores. The *predicate difficulty* $\delta = |\mathcal{A}_\tau|/N$ is the fraction of pool items in the ambiguous region.

When the full-pool pass is available, δ is estimated directly as: $\hat{\delta} = \frac{1}{N}\sum_{i=1}^N \mathbf{1}[|\hat{p}(x_i) - 0.5| < \tau]$, otherwise, it is estimated from $\mathcal{D}_{cal}$. A large $\hat{\delta}$ indicates a hard predicate where many items require careful disambiguation; a small $\hat{\delta}$ indicates that most items are confidently classified zero-shot. As we show next, $\hat{\delta}$ plays a central role in determining the optimal budget allocation.

3.2 THE ALLOCATE PHASE

Given the calibrated confidence scores from pre-processing, the allocation step determines the harvest budget B_h and escalation budget $B_e = B - B_h$. We first introduce the escalation rate, then present the theoretical results that guide allocation, and finally describe three practical allocation approaches.

Definition 3 (Escalation Rate). Given a harvest set H of size B_h with oracle labels, let f_H^- denote the ICL-adapted cheap model using demonstrations drawn from H. The *escalation rate* is the fraction of non-harvested items whose adapted predictions remain uncertain. It is estimated from the calibration set when the full-pool pass is unavailable and directly from pool predictions otherwise, in which case $\rho(B_h) = \frac{1}{N-B_h}\sum_{x \notin H} \mathbf{1}[|\hat{p}_H(x) - 0.5| < \tau]$, where $\hat{p}_H(x)$ is the calibrated confidence of f_H^- on item x. The escalation rate measures residual uncertainty after ICL adaptation, so items with $\rho(B_h)$ close to 0.5 still require oracle resolution despite adaptation.

Theoretical guidance. We establish two results that guide allocation. The first one only requires a sampling assumption as follows:

(A1) Independent harvest sampling. Each item in $\mathcal{H}$ is drawn independently from $\mathcal{D}$, so $\mathbf{1}[x \in \mathcal{A}_\tau]$ are i.i.d. Bernoulli(δ) random variables and $|\mathcal{H} \cap \mathcal{A}_\tau| \sim \text{Binomial}(B_h, \delta)$.

Theorem 1 (Ambiguous Demonstration Coverage). *Let assumption **(A1)** hold. For a target number of ambiguous demonstrations $k^* > 0$ and confidence level $1 - \alpha \in (0, 1)$, it suffices to harvest $B_h \geq \frac{k^*}{\delta}\log\frac{1}{\alpha}$ items to ensure $|\mathcal{H} \cap \mathcal{A}_\tau| \geq k^*$ with probability at least $1 - \alpha$.*

Implications. This formalizes the intuition that the required harvest budget scales inversely with δ: harder predicates (large δ) need fewer items per demonstration since ambiguous examples are abundant, while easy predicates (small δ) require more harvesting to accumulate k^* useful demonstrations. The result is independent of pool size N.

The result above does not consider constraints while determining B_h. Subsequent results require understanding the error from allocations. Let $\mathrm{err}(f_\emptyset^-) = \Pr_{x\sim\mathcal{D}}(f_\emptyset^-(x) \neq f^+(x))$ and $\mathrm{err}(f_\mathcal{H}^-) = \Pr_{x\sim\mathcal{D}\setminus\mathcal{H}}(f_\mathcal{H}^-(x) \neq f^+(x))$ denote error rates of the unadapted and adapted cheap models, respectively, both measured against the oracle f^+ as ground truth. We make the following additional assumptions, subsets of which will be used for subsequent results:

- **(A2) Error decomposition.** The error rate of the cheap model decomposes as $\mathrm{err}(f_\mathcal{H}^-) = \delta \cdot \mathrm{err}_{\mathrm{amb}}(f_\mathcal{H}^-) + (1-\delta) \cdot \mathrm{err}_{\mathrm{unamb}}(f_\mathcal{H}^-)$, where $\mathrm{err}_{\mathrm{amb}}$ and $\mathrm{err}_{\mathrm{unamb}}$ denote error rates on items inside and outside $\mathcal{A}_\tau$, respectively.
- **(A3) ICL adaptation concentrates on ambiguous items.** ICL adaptation does not significantly affect the error rate on unambiguous items: $\mathrm{err}_{\mathrm{unamb}}(f_\mathcal{H}^-) \approx \mathrm{err}_{\mathrm{unamb}}(f_\emptyset^-)$. Error improvement concentrates on $\mathcal{A}_\tau$.
- **(A4) Monotone improvement.** The error rate $\mathrm{err}_{\mathrm{amb}}(f_\mathcal{H}^-)$ decreases monotonically in $|\mathcal{H} \cap \mathcal{A}_\tau|$, and each item in $\mathcal{H} \cap \mathcal{A}_\tau$ provides an i.i.d. observation about the oracle's labeling behavior in the ambiguous region.
- **(A5) Monotone escalation rate decay.** The escalation rate $\rho(B_h)$ is continuously differentiable and strictly decreasing in B_h, with $\rho(0) = \rho_0 > 0$ and $\lim_{B_h\to\infty} \rho(B_h) = 0$, and convex in B_h. Therefore $\rho'(B_h) < 0$ and $\rho''(B_h) \geq 0$.

Definition 4 (Total Residual Error). Under assumption **(A2)**, the *total residual error* is the expected number of incorrectly labeled non-harvested items not corrected by escalation: $\mathcal{E}(B_h) = \mathrm{err}(f_{B_h}^-) \cdot \max[\rho(B_h) \cdot N - B_e,\ 0]$, where the unambiguous error term is omitted as it is constant in B_h under assumption **(A3)**.

Theorem 2 (Optimal Budget Allocation). *Let assumptions **(A1)** and **(A5)** hold. Approximating* $\mathrm{err}(f_{B_h}^-)$ *as constant near the optimum, the optimal harvest size B_h^* minimizing total residual error $\mathcal{E}(B_h)$ subject to $0 \leq B_h \leq B$ satisfies $\rho'(B_h^*) \cdot N = -1$, i.e. the marginal reduction in the number of uncertain items equals one.*

There may be closed form solutions under parametric assumptions on ρ. For instance, when $\rho(B_h) = \rho_0(1+B_h)^{-\gamma}$ (power law), $B_h^* = (\gamma\rho_0 N)^{1/(1+\gamma)} - 1$, and when $\rho(B_h) = \rho_0 e^{-\lambda B_h}$, $B_h^* = \frac{1}{\lambda}\log(\lambda\rho_0 N)$. In both cases B_h^* is independent of the total budget B and is clipped to $[0, B]$ in practice. The condition $\rho'(B_h^*)N = -1$ has a unique solution under assumption **(A5)** since ρ' is monotone increasing from $\rho'(0) < -1/N$ to $\rho'(\infty) = 0$. See Appendix A for detailed derivations.

Practical allocation. Theorems 1 and 2 motivate the following practical approaches to setting B_h. All methods set $B_h = \lfloor \beta \cdot B \rfloor$ but differ in how they infer harvest fraction $\beta \in [0, 1]$:

1. **Explicit.** The practitioner directly specifies $\beta = B_h/B$, bypassing estimation entirely. This is the most robust option when prior knowledge of the predicate difficulty is available.
2. **Target demonstrations.** Set $\beta = \min\left(1,\ \dfrac{k^* \log(1/\alpha)}{\hat{\delta} \cdot B}\right)$, where $k^* \leq k_{\max}$ is a target number of ambiguous demonstrations and $1-\alpha$ is a desired confidence level. By Theorem 1, this guarantees $|\mathcal{H} \cap \mathcal{A}_\tau| \geq k^*$ with probability at least $1-\alpha$, using only $\hat{\delta}$ from pre-processing with no ICL pilot calls required. We recommend $k^* = \lfloor k_{\max}/4 \rfloor$ and $\alpha = 0.1$ as practical defaults.
3. **Ambiguous region.** Set $\beta = \hat{\delta}$, allocating a harvest fraction equaling estimated predicate difficulty. This is a data-driven heuristic that does not need a target k^*: the expected number of ambiguous demonstrations scales as $\delta^2 \cdot B$ according to Theorem 1, increasing with predicate difficulty, and $\hat{\delta}$ is directly available from the zero-shot confidence distribution with no ICL pilot calls required.
4. **Optimal (fitted).** Estimate decay parameter γ or λ from a small ICL pilot experiment on the calibration set by running f^- with a few demonstration sizes $B_h \in \{8, 16, 32\}$ and fitting the observed escalation rate decay and corresponding optimal β. This directly instantiates Theorem 2 and is most appropriate when the decay parameter is expected to vary substantially across datasets.

When assumption **(A5)** is violated, i.e. $\rho(B_h)$ is non-monotone, as can occur when ICL demonstration imbalance shifts the cheap model's confidence distribution (Section 3.4), the fitted approach is unreliable and the other allocation approaches are recommended instead.

3.3 THE HARVEST PHASE

The harvest phase selects B_h items from the pool, oracle-labels them, and returns the labeled set $\mathcal{H}$ for use as ICL demonstrations in the predict phase. The quality of $\mathcal{H}$ depends on how well it covers the ambiguous region $\mathcal{A}_\tau$.

Definition 5 (**Ambiguous Coverage**). The *ambiguous coverage* of a harvest set $\mathcal{H}$ is the fraction of harvest examples in the ambiguous region: $\phi(\mathcal{H}) = |\mathcal{H} \cap \mathcal{A}_\tau|/|\mathcal{H}|$. A strategy with higher expected coverage $\phi_k = \mathbb{E}[\phi(\mathcal{H}_k)]$ concentrates more oracle calls near the decision boundary and therefore requires fewer harvest examples to achieve the same error reduction.

We consider three harvest strategies, each selecting B_h items from the pool $\mathcal{D}$:

- **Random.** Select B_h items uniformly at random. Simple but covers $\mathcal{A}_\tau$ only at rate $\phi_{\text{rand}} = \delta$.
- **Uncertainty.** Rank all items by $|\hat{p}(x) - 0.5|$ ascending and select the B_h most uncertain. Targets $\mathcal{A}_\tau$, achieving $\phi_{\text{unc}} \to 1$ as calibration improves. Requires the zero-shot pass from pre-processing.
- **Diversity.** Cluster items by their calibrated probabilities using k-means and select one representative per cluster, filling remaining slots randomly. Ensures coverage across the full confidence distribution, including ambiguous items, with $\phi_{\text{div}} \geq \phi_{\text{rand}}$.

Theorem 3 (**Harvest Strategy Comparison**). *Under assumptions **(A1)–(A4)**, for a target number of ambiguous demonstrations $k^* > 0$ and confidence level $1 - \alpha$, the expected harvest budget required by strategy k to ensure $|\mathcal{H}_k \cap \mathcal{A}_\tau| \geq k^*$ satisfies*

$$B_h^{(k)} \geq \frac{k^*}{\phi_k \cdot \delta} \log \frac{1}{\alpha},$$

where $\phi_k = \mathbb{E}[\phi(\mathcal{H}_k)]$ is the expected ambiguous coverage of strategy k (Definition 5). The coverages satisfy $\phi_{\text{rand}} \leq \phi_{\text{div}} \leq \phi_{\text{unc}}$, with $\phi_{\text{rand}} = \delta$ and $\phi_{\text{unc}} \to 1$ as calibration improves. Consequently the required harvest sizes satisfy $B_h^{(\text{unc})} \leq B_h^{(\text{div})} \leq B_h^{(\text{rand})}$, with ratio $B_h^{(\text{rand})}/B_h^{(\text{unc})} = \phi_{\text{unc}}/\delta \geq 1$, which grows as $\delta \to 0$: the advantage of uncertainty harvest over random sampling is largest for easy predicates with few ambiguous items.

Implications. The coverage ordering reflects the targeting mechanism of each strategy. Random sampling covers $\mathcal{A}_\tau$ at rate δ by construction, requiring $B_h^{(\text{rand})} \propto k^*/\delta^2$ harvest items to accumulate k^* ambiguous demonstrations. Diversity sampling improves on this by spreading harvest items across the confidence distribution, while uncertainty sampling achieves the highest coverage by directly targeting items closest to the decision boundary, requiring only $B_h^{(\text{unc})} \propto k^*/\delta$ items. The advantage grows as $\delta \to 0$: for easy predicates, random sampling rarely finds informative demonstrations while uncertainty sampling targets them directly. For hard predicates ($\delta \to 1$), all strategies converge and random sampling is nearly as effective as targeted approaches.

3.4 THE PREDICT PHASE

Given the labeled harvest set $\mathcal{H}$, the predict phase runs the ICL-adapted cheap model $f_{\mathcal{H}}^-$ on all non-harvested items $\mathcal{D} \setminus \mathcal{H}$, using demonstrations drawn from $\mathcal{H}$. We select $k = \min(|\mathcal{H}|, k_{\max})$ demonstrations from $\mathcal{H}$, where $k_{\max}$ is a cap that prevents context length issues and avoids label bias amplification from large imbalanced demonstration sets (Zhao et al., 2021). Three demonstration selection strategies are supported: **fixed** selection draws the same k demonstrations for all items (one batch call, most efficient); **random** selection samples a fresh set of k demonstrations per item; and **similarity** selection retrieves the k harvest items with logit confidence scores closest to each target item, matching demonstrations to items by difficulty level. The predict phase returns calibrated confidence scores $\hat{p}_{\mathcal{H}}(x_i)$ for all $x_i \in \mathcal{D} \setminus \mathcal{H}$, which serve as inputs to the escalate phase.

3.5 THE ESCALATE PHASE

The escalate phase selects $B_e = B - B_h$ items from the non-harvested pool $\mathcal{D} \setminus \mathcal{H}$ and oracle-labels them, replacing the cheap model's predictions with ground truth for the most uncertain items.

Theorem 4 (**Optimal Escalation Thresholds**). *Given calibrated confidence scores, a fixed escalation budget B_e, and threshold gap Δ such that $t^{\text{lo}} = \frac{1}{2} - \frac{\Delta}{2}$ and $t^{\text{hi}} = \frac{1}{2} + \frac{\Delta}{2}$, the optimal threshold gap*

Δ^* minimizing total residual error $\mathcal{E}(B_h)$ subject to the budget constraint satisfies

$$\Delta^* = \sup\{\Delta \in [0,1] : \rho(\Delta) \cdot (N - B_h) \leq B_e\},$$

where $\rho(\Delta) = \Pr_{x \sim \mathcal{D} \setminus \mathcal{H}}(\hat{p}_{\mathcal{H}}(x) \in [t^{\text{lo}}, t^{\text{hi}}])$ is the fraction of non-harvested items within the threshold interval. The minimum residual error satisfies $\mathcal{E}^* \leq \mathcal{E}(B_h) - B_e \cdot (\frac{1}{2} - \frac{\Delta^*}{2})$, confirming that larger B_e reduces residual error.

Implications. Theorem 4 applies well-established statistical principles regarding optimal rejection thresholds based on calibrated probabilities (Chow, 1970) and confirms that the optimal escalation policy is top-k by uncertainty: sort non-harvested items by $|\hat{p}_{\mathcal{H}}(x_i) - 0.5|$ ascending and escalate until B_e is exhausted, with Δ^* set at the confidence gap of the last escalated item. The result relies on well-calibrated confidence scores as miscalibrated scores can distort the uncertainty ranking.

The final labels are assembled by merging oracle labels from sets $\mathcal{H}$ and $\mathcal{E}$ with cheap model predictions for all remaining items $\mathcal{D} \setminus (\mathcal{H} \cup \mathcal{E})$. The complete ICC pipeline is summarized in Fig. 1 and as pseudocode in Appendix B, which includes all hyperparameters and implementation details.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate ICC on four binary classification datasets spanning legal, financial, and social media domains. **Overruling** (Zheng et al., 2021) contains 2,400 legal sentences labeled for whether they overrule a prior legal decision. We subsample 1,000 items for the remaining three datasets to keep evaluation tractable. **Headlines** (Sinha & Khandait, 2021) is drawn from a commodity market sentiment dataset and contains financial news headlines labeled for positive price direction. **ToxiGen** (Hartvigsen et al., 2022) contains statements labeled for toxic or hateful content using human annotations. **Movie Reviews** (Lao et al., 2025) are labeled for sentiment polarity, drawn from the SemBench benchmark. These are standard benchmarks for annotation evaluation; larger-scale evaluation is supported by the theoretical results, which scale to arbitrary N (Theorems 1–4).

Models. We use ground truth annotations as perfect oracle f^+, simulating a human annotator with zero error rate. This isolates the contribution of ICC's pipeline from oracle quality. We use the popular **Llama-3.1-8B-Instruct** (Grattafiori et al., 2024) as cheap model f^- in the main paper but also demonstrate robustness on three other open-source models: **Granite-3.3-8B-Instruct** (Granite Team, IBM, 2024), **Granite-4.0-H-Tiny**, and **Llama-3.3-70B-Instruct** (see Appendix C.1).

Baselines. We compare ICC against three baselines. **RandomOracle+Majority** randomly allocates B oracle calls and defaults to majority class for the remainder. **RandomOracle+Cheap** randomly allocates B oracle calls and uses f^- zero-shot for the remainder. **FixedCascade** escalates the B most uncertain items to the oracle and uses f^- zero-shot for the remainder (Chen et al., 2024); this is the strongest non-ICL baseline and the primary comparison point for ICC. Parameter-efficient fine-tuning methods such as LoRA (Hu et al., 2022) operate under fundamentally different assumptions – requiring gradient access, multiple training epochs, and longer adaptation time – and target a different deployment regime than ICC's inference-only pipeline; we therefore treat them as out-of-scope.

Evaluation. We report mean accuracy over $n = 5$ runs for stochastic methods (random and diversity harvest) and single runs for deterministic methods (uncertainty harvest). Fractional oracle budgets $B/N \in \{1\%, 2\%, 5\%, 10\%, 20\%\}$ are considered. The harvest budget B_h is set using one of four approaches: ambiguous region, target demonstrations, power law, and exponential fit (Section 3.2), with parameters estimated from a small ICL pilot on the calibration set. We use ambiguity threshold parameter $\tau = 0.2$, $k_{\max} = 32$ demonstrations, fixed demonstration selection, and top-k escalation unless otherwise specified. In our main experiments, confidence scores are calibrated using temperature scaling (Guo et al., 2017) on a held-out calibration set of 50 items (drawn stratified by label), but results are robust to this choice (see Appendix C.5).

4.2 Main Results

Figure 2 shows cost-accuracy curves for three ICC variants and three baselines across all four datasets. We consider **ICC-UF-λ**, which sets harvest fraction β^* from the fitted exponential; **ICC-UF-δ**,

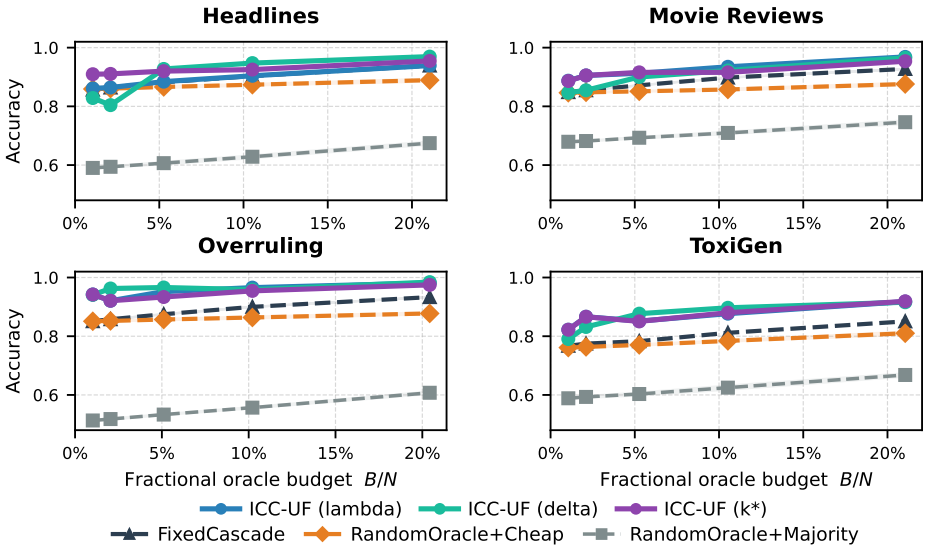

Figure 2: Cost-accuracy curves across four datasets using Llama-3.1-8B-Instruct as the cheap model. ICC variants (solid lines) consistently outperform baselines (dashed lines) across all budget levels $B/N \in \{1\%, 2\%, 5\%, 10\%, 20\%\}$ with the largest gains often at the smallest budgets.

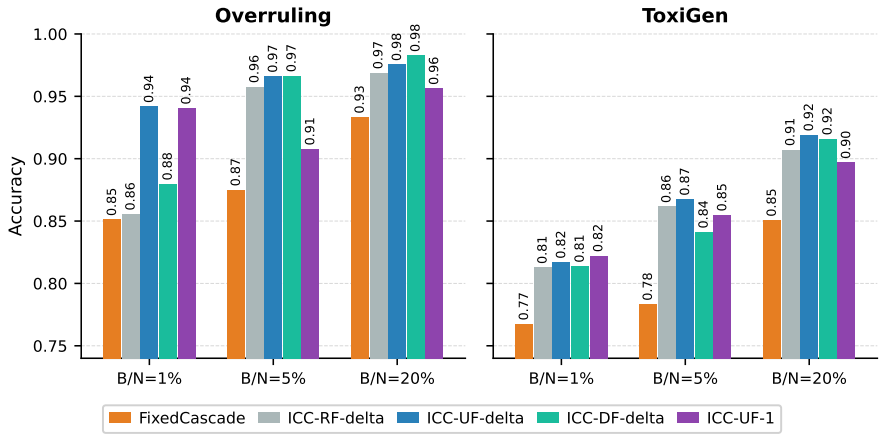

Figure 3: Harvest strategy comparison across budget levels on Overruling and ToxiGen for Llama-3.1-8B-Instruct as cheap model. Each group shows five methods that progressively add ICC components, isolating the contribution of ICL adaptation, harvest strategy, and the harvest-escalation split.

where β^* captures the empirical ambiguous region; and **ICC-UF-k^***, which sets β^* based on target demonstrations. All variants use an uncertainty-based harvest strategy (denoted 'U') and fixed demonstration selection (denoted 'F') with $k_{\max} = 32$. We see that ICC variants consistently outperform all baselines across every budget level, particularly on Overruling and ToxiGen, with the largest gains at the tightest budgets: they achieve almost +10 percentage points over FixedCascade at $B/N = 1\%$ on Overruling, where the entire oracle budget is too small for escalation alone to be effective. On Headlines, gains are modest and the advantage narrows at larger budgets. The three ICC variants show broadly similar performance across the datasets, confirming that the algorithmic allocation approaches can provide reliable data-driven alternatives to manual hyperparameter selection.

Table 1: Accuracy at $B/N = 5\%$ for six allocation strategies across four datasets. $\beta = 0$ recovers FixedCascade (no ICL); $\beta = 1$ allocates all the budget to harvesting. Bold indicates best per dataset.

Method	Overruling	ToxiGen	Headlines	Movie Reviews
FixedCascade ($\beta = 0$; no harvest)	0.875	0.783	0.882	0.872
ICC-UF ($\beta = 1$; no escalate)	0.914	0.851	0.920	**0.916**
ICC-UF-δ (ambiguous region)	**0.966**	**0.877**	**0.927**	0.900
ICC-UF-k^* (target demonstrations)	0.934	0.851	0.920	**0.916**
ICC-UF-γ (power law)	0.929	0.872	0.826	0.907
ICC-UF-λ (exponential)	0.952	0.852	0.884	0.913

4.3 Ablation Study

ICC Component Decomposition. Figure 3 isolates the contribution of each ICC component by comparing five methods across two datasets and three budget levels. Moving left to right within each group, the methods progressively add components: **FixedCascade** uses oracle escalation with no ICL; **ICC-RF-delta** adds ICL with random harvest; **ICC-UF-delta** replaces random with uncertainty-based harvest; **ICC-DF-delta** uses diversity harvest; and **ICC-UF-1** allocates the full budget to harvest with no escalation ($\beta = 1$). This design directly tests the value of ICL adaptation (FixedCascade vs. ICC-RF-delta), the effect of harvest strategy (ICC-RF vs. ICC-UF vs. ICC-DF), and the importance of the harvest-escalation split (ICC-UF-delta vs. ICC-UF-1). All ICC variants outperform FixedCascade on both datasets, confirming that ICL adaptation provides consistent gains beyond oracle escalation alone. Uncertainty harvest outperforms random harvest at small budgets where targeted selection is critical, with the gap narrowing at larger budgets as random harvest covers the ambiguous region through volume; this is consistent with Theorem 3. Full-budget harvest with no escalation (ICC-UF-1) consistently underperforms, confirming the importance of the harvest-escalation split.

Allocation Method Comparison. Table 1 compares six allocation strategies for **ICC-UF** at a fixed budget $B/N = 5\%$ across all datasets, with settings as used for the main results (Section 4.2). The boundary cases $\beta = 0$ (pure escalation, equivalent to FixedCascade) and $\beta = 1$ (pure harvesting) span the space of possible allocations. The four principled allocation approaches (ambiguous region, target demonstrations, power law, exponential) each set β from data without manual tuning using method-specific parameters. We observe that pure harvest ($\beta = 1$) consistently outperforms FixedCascade ($\beta = 0$) by 3–7 percentage points, confirming that ICL adaptation adds value beyond direct escalation alone. Among the four proposed allocation approaches, ICC-UF-δ (requiring only τ and no ICL pilot calls) is the most reliable allocation strategy, with ICC-UF-k^* often competitive. The fitted approaches ICC-UF-γ and ICC-UF-λ do well on most datasets but fall short on Headlines, consistent with the non-monotone escalation rate decay observed during allocation estimation on that dataset.

Appendix C includes additional experiments on cheap model robustness, demonstration selection strategy, as well as effect of ambiguity threshold, maximum demonstrations, and calibration set size.

5 Conclusion

We have presented the In-Context Cascade (ICC), a principled framework for budget-constrained pool labeling that unifies in-context learning adaptation with targeted oracle escalation. Unlike prior cascade methods that treat oracle calls purely as direct label assignments, ICC exploits the demonstrative value of oracle labels to improve cheap model accuracy across the entire pool, with theoretical guarantees characterizing stages of the pipeline. For applications where oracle access is expensive, such as annotation, content moderation, and enterprise AI pipelines, ICC can consistently achieve accuracy close to full oracle labeling at a fraction of the cost.

ICC is designed for API-accessible annotation pipelines operating under a fixed oracle budget, and is evaluated on binary classification predicates with a perfect oracle. These scope choices are deliberate and instructive: they isolate the contribution of ICL adaptation and targeted escalation in a theoretically tractable setting. Natural directions for **future work** include extension to multi-class and structured prediction settings, evaluation under noisy oracles such as weaker LLMs, embedding-based demonstration selection for a richer suite of similarity metrics, and a principled approach to calibration in the ICL setting.

AI USE STATEMENT

In this work, we used generative AI tools for editing and proposing formatting ideas about presenting some of the material. Additionally, we used generative AI tools for to assist with the Python implementation of the proposed methods and visualization plots. We have reviewed all AI-assisted code by running unit tests for smaller and contained functionalities. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

ICC is designed to reduce annotation cost in large-scale text labeling pipelines, with intended applications in content moderation, regulatory compliance, and enterprise data curation. As with any annotation automation framework, there is a risk of misuse. Deploying it with a poorly specified predicate, a biased cheap model, or an insufficiently accurate oracle could propagate systematic labeling errors at scale, with potential harms in high-stakes decision-making contexts such as content moderation or legal document classification. Practitioners should validate ICC outputs against ground truth before deployment, ensure that the oracle used for harvesting and escalation meets the accuracy requirements of the target application, and audit for biases in cheap model predictions before applying the processed outputs to model training. The datasets used in this work are publicly available benchmarks; ToxiGen in particular contains synthetic toxic and hateful content generated for research purposes, and we use only human-annotated labels from this dataset.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we have included the following materials. All theoretical assumptions (**A1**–**A5**) are stated explicitly in Section 3.2, and complete proofs of all four theoretical results (Theorems 1–4) are provided in Appendix A. The complete ICC pipeline is presented as pseudocode in Algorithm 1 (Appendix B), covering all five stages with branching logic for all configurable options, including four allocation approaches, three harvest strategies, two demonstration selection strategies, and the demonstration balance flag. Experimental hyperparameters are fully specified in Section 4.1: ambiguity threshold $\tau = 0.2$, $k_{\max} = 32$ demonstrations, fixed demonstration selection, top-k escalation, and a held-out calibration set of 50 items drawn stratified by label. All four datasets used in the main experiments are publicly available benchmarks (Overruling, Headlines, ToxiGen, Movie Reviews) with citations provided in Section 4.1. Additional experimental results on cheap model robustness, demonstration selection strategy, ambiguity threshold sensitivity, maximum demonstrations, and calibration budget fraction are reported in Appendix C, providing sufficient detail to reproduce the ablation studies. Cache interactions are noted throughout Algorithm 1: zero-shot cheap model calls and oracle calls are served from cache after pre-population, so only the ICL predict calls in Stage 4 constitute live API calls during experiments.

REFERENCES

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020.

Lingjiao Chen, Matei Zaharia, and James Zou. Frugalgpt: How to use large language models while reducing cost and improving performance. *arXiv preprint arXiv:2305.05176*, 2023.

Lingjiao Chen, Matei Zaharia, and James Zou. FrugalGPT: How to use large language models while reducing cost and improving performance. *Transactions on Machine Learning Research*, 2024.

C. K. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, 1970.

Shrey Desai and Greg Durrett. Calibration of pre-trained transformers. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 295–302, 2020.

Dujian Ding, Ankur Mallick, Chi Wang, Robert Sim, Subhabrata Mukherjee, Victor Rühle, Laks V. S. Lakshmanan, and Ahmed Hassan Awadallah. Hybrid LLM: Cost-efficient and quality-aware query routing. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

Claudio Fanconi and Mihaela van der Schaar. Cascaded language models for cost-effective human-AI decision-making. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120, 2023. doi: 10.1073/pnas.2305016120.

Granite Team, IBM. Granite 3.0 language models, 2024. URL `https://github.com/ibm-granite/granite-3.0-language-models`.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, et al. The Llama 3 herd of models, 2024.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pp. 1321–1330. PMLR, 2017.

Neha Gupta, Harikrishna Narasimhan, Wittawat Jitkrittum, Ankit Singh Rawat, Aditya Krishna Menon, and Sanjiv Kumar. Language model cascades: Token-level uncertainty and beyond. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. ToxiGen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3309–3326. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.acl-long.234.

Xingwei He, Zhenghao Lin, Yeyun Gong, A-Long Jin, Hang Zhang, Chen Lin, Jian Jiao, Siu Ming Yiu, Nan Duan, and Weizhu Chen. AnnoLLM: Making large language models to be better crowdsourced annotators. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, pp. 165–190, Mexico City, Mexico, 2024. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield Doddsand Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *preprint arXiv: 2207.05221 [cs.CL]*, 2022.

Jiale Lao, Andreas Zimmerer, Olga Ovcharenko, Tianji Cong, Matthew Russo, Gerardo Vitagliano, Michael Cochez, Fatma Özcan, Gautam Gupta, Thibaud Hottelier, H. V. Jagadish, Kris Kissel, Sebastian Schelter, Andreas Kipf, and Immanuel Trummer. SemBench: A benchmark for semantic query processing engines, 2025.

Minzhi Li, Taiwei Shi, Caleb Ziems, Min-Yen Kan, Nancy F. Chen, Zhengyuan Liu, and Diyi Yang. CoAnnotating: Uncertainty-guided work allocation between human and large language models for data annotation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 1487–1505, Singapore, 2023. Association for Computational Linguistics.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pp. 100–114, Dublin, Ireland and Online, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.deelio-1.10.

Lunyiu Nie, Zhimin Ding, Erdong Hu, Christopher Jermaine, and Swarat Chaudhuri. Online cascade learning for efficient inference over streams. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.

Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E. Gonzalez, M. Waleed Kadous, and Ion Stoica. RouteLLM: Learning to route LLMs with preference data. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025.

Liana Patel, Siddharth Jha, Melissa Pan, Harshit Gupta, Parth Asawa, Carlos Guestrin, and Matei Zaharia. Semantic operators: A declarative model for rich, ai-based data processing, 2025. URL https://arxiv.org/abs/2407.11418.

Gabriele Sanmartino, Matthias Urban, Paolo Papotti, and Carsten Binnig. The Stretto execution engine for LLM-augmented data systems, 2026.

Vishnu Sarukkai, Asanshay Gupta, James Hong, Michaël Gharbi, and Kayvon Fatahalian. In-context distillation with self-consistency cascades: A simple, training-free way to reduce LLM agent costs, 2025.

Burr Settles. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009. URL https://burrsettles.com/pub/settles.activelearning.pdf.

Ankur Sinha and Tanmay Khandait. Impact of news on the commodity market: Dataset and results. In Kohei Arai (ed.), *Advances in Information and Communication. FICC 2021*, volume 1364 of *Advances in Intelligent Systems and Computing*, pp. 589–601. Springer, Cham, 2021.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5433–5442, 2023.

Antonios Valkanas, Soumyasundar Pal, Pavel Rumiantsev, Yingxue Zhang, and Mark Coates. C3PO: Optimized large language model cascades with probabilistic cost constraints for reasoning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=e4IlBqhbTO.

Yu Wu, Shuo Wu, Ye Tao, Yansong Li, and Anand D. Sarwate. From deferral to learning: Online in-context knowledge distillation for LLM cascades, 2025.

Sepanta Zeighami, Shreya Shankar, and Aditya Parameswaran. Cut costs, not accuracy: Llm-powered data processing with guarantees, 2025. URL https://arxiv.org/abs/2509.02896.

Jieyu Zhang, Ranjay Krishna, Ahmed H. Awadallah, and Chi Wang. EcoAssistant: Using LLM assistant more affordably and accurately, 2023.

Xuechen Zhang, Zijian Huang, Ege Onur Taga, Carlee Joe-Wong, Samet Oymak, and Jiasi Chen. Efficient contextual LLM cascades through budget-constrained policy learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Tony Z. Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12697–12706. PMLR, 2021.

Lucia Zheng, Neel Guha, Brandon R. Anderson, Peter Henderson, and Daniel E. Ho. When does pretraining help? Assessing self-supervised learning for law and the CaseHOLD dataset. In *Proceedings of the 18th International Conference on Artificial Intelligence and Law*, pp. 159–168. Association for Computing Machinery, 2021.

A THEOREM PROOFS

PROOF OF THEOREM 1

Proof. By assumption **(A1)**, each harvested item falls in $\mathcal{A}_\tau$ independently with probability δ, so the number of ambiguous demonstrations $m = |\mathcal{H} \cap \mathcal{A}_\tau|$ follows a $\mathrm{Binomial}(B_h, \delta)$ distribution with mean $\mathbb{E}[m] = B_h\delta$.

We require $m \geq k^*$ with probability at least $1 - \alpha$. Applying Hoeffding's inequality to the sum of B_h independent Bernoulli(δ) random variables:

$$\Pr(m \leq B_h\delta - t) \leq \exp\left(-\frac{2t^2}{B_h}\right)$$

for any $t > 0$. Setting $t = B_h\delta - k^*$, which requires $k^* \leq B_h\delta$, the event $\{m \leq k^* - 1\}$ is contained in $\{m \leq B_h\delta - t\}$, so:

$$\Pr(m < k^*) \leq \exp\left(-\frac{2(B_h\delta - k^*)^2}{B_h}\right).$$

Requiring this to be at most α:

$$\frac{2(B_h\delta - k^*)^2}{B_h} \geq \log\frac{1}{\alpha}.$$

Let $u = B_h\delta - k^* \geq 0$. Then $B_h = (u + k^*)/\delta$ and the condition becomes:

$$\frac{2u^2\delta}{u + k^*} \geq \log\frac{1}{\alpha}.$$

A sufficient condition is $u \geq k^*$, i.e. $B_h\delta \geq 2k^*$, which gives $B_h \geq 2k^*/\delta$. Under this condition $u \geq k^*$ and:

$$\frac{2u^2\delta}{u + k^*} \geq \frac{2u^2\delta}{2u} = u\delta = (B_h\delta - k^*)\delta \geq \frac{B_h\delta^2}{2}$$

where the last inequality uses $B_h\delta \geq 2k^*$ so $k^* \leq B_h\delta/2$. Requiring $B_h\delta^2/2 \geq \log(1/\alpha)$ gives $B_h \geq 2\log(1/\alpha)/\delta^2$. Combining with the requirement $B_h \geq 2k^*/\delta$ and setting $C = 2\max(k^*, \log(1/\alpha)/\delta)$:

$$B_h \geq \frac{k^*}{\delta}\log\frac{1}{\alpha}$$

is a sufficient condition that satisfies both requirements simultaneously for $k^* \geq 1$ and $\delta \in (0, 1)$, completing the proof. □

PROOF OF THEOREM 2

Proof. **Setup.** Under the constant-error approximation $\mathrm{err}(f^-_{B_h}) \approx \bar{\varepsilon}$, minimizing $\mathcal{E}(B_h) = \bar{\varepsilon} \cdot \max[\rho(B_h)N - (B - B_h), 0]$ is equivalent to minimizing

$$F(B_h) = \rho(B_h) \cdot N + B_h - B,$$

over $B_h \in [0, B]$, where we consider the interior region where $\rho(B_h)N > B - B_h$ so the max is not binding.

First-order condition. Differentiating $F(B_h)$ with respect to B_h:

$$F'(B_h) = \rho'(B_h) \cdot N + 1.$$

Setting $F'(B_h^*) = 0$ gives the FOC:

$$\rho'(B_h^*) \cdot N = -1.$$

Since $\rho'(B_h) < 0$ by assumption **(A2)**, the left side is negative and equals $-1 < 0$ — consistent. Since $\rho''(B_h) \geq 0$ by assumption **(A2)** (convexity), $F''(B_h) = \rho''(B_h) \cdot N \geq 0$, so F is convex and any solution to the FOC is a global minimum.

Existence and uniqueness. By assumption **(A2)**, ρ' is continuous, monotone increasing (from convexity), with $\lim_{B_h \to 0} \rho'(B_h) < 0$ and $\lim_{B_h \to \infty} \rho'(B_h) = 0$. The FOC $\rho'(B_h^*)N = -1$ therefore has a unique solution provided $\rho'(0) < -1/N$, i.e. the initial marginal decay rate is sufficiently large. When this condition fails, $F'(B_h) > 0$ for all $B_h \geq 0$ and the minimum is at $B_h^* = 0$ (pure escalation). The solution is clipped to $[0, B]$ if the unconstrained optimum falls outside this interval.

Power law closed form. Let $\rho(B_h) = \rho_0(1 + B_h)^{-\gamma}$ with $\rho_0, \gamma > 0$. Then:

$$\rho'(B_h) = -\gamma\rho_0(1 + B_h)^{-\gamma-1}.$$

Substituting into the FOC:

$$-\gamma\rho_0(1 + B_h^*)^{-\gamma-1} \cdot N = -1 \quad \Longrightarrow \quad (1 + B_h^*)^{\gamma+1} = \gamma\rho_0 N.$$

Solving:

$$B_h^* = (\gamma\rho_0 N)^{1/(1+\gamma)} - 1.$$

This is valid provided $\gamma\rho_0 N > 1$, i.e. $\rho'(0)N = -\gamma\rho_0 N < -1$, which is the existence condition above. The result is independent of B.

Exponential closed form. Let $\rho(B_h) = \rho_0 e^{-\lambda B_h}$ with $\rho_0, \lambda > 0$. Then:

$$\rho'(B_h) = -\lambda\rho_0 e^{-\lambda B_h}.$$

Substituting into the FOC:

$$-\lambda\rho_0 e^{-\lambda B_h^*} \cdot N = -1 \quad \Longrightarrow \quad e^{\lambda B_h^*} = \lambda\rho_0 N.$$

Solving:

$$B_h^* = \frac{1}{\lambda} \log(\lambda\rho_0 N).$$

This is valid provided $\lambda\rho_0 N > 1$, i.e. $\rho'(0)N = -\lambda\rho_0 N < -1$, again matching the existence condition. The result is independent of B.

Interpretation. In both cases B_h^* is determined solely by the decay parameters (γ or λ), the initial escalation rate ρ_0, and pool size N — not by the total budget B. The optimal escalation budget is then $B_e^* = B - \min(B_h^*, B)$, with $B_h^* = B$ (all budget to harvest) when the unconstrained optimum exceeds B, and $B_h^* = 0$ (pure escalation) when the existence condition $\rho'(0)N < -1$ fails. □

PROOF OF THEOREM 3

Proof. **Sample complexity bound.** Under strategy k, each harvested item falls in $\mathcal{A}_\tau$ with probability $\phi_k \cdot \delta$ — the joint probability of being selected by strategy k (probability ϕ_k) and lying in $\mathcal{A}_\tau$ (probability δ). By assumption **(A1)**, selections are independent, so $m_k = |\mathcal{H}_k \cap \mathcal{A}_\tau| \sim \text{Binomial}(B_h^{(k)}, \phi_k\delta)$. Applying the proof of Theorem 1 with δ replaced by $\phi_k\delta$ gives:

$$B_h^{(k)} \geq \frac{k^*}{\phi_k \cdot \delta} \log \frac{1}{\alpha}.$$

Coverage ordering. *Random:* Items are drawn uniformly from $\mathcal{D}$ independently of $\mathcal{A}_\tau$, so $\phi_{\text{rand}} = \Pr(x \in \mathcal{A}_\tau) = \delta$ by definition.

Uncertainty: Strategy unc selects items by ascending $|\hat{p}(x) - 0.5|$, directly targeting $\mathcal{A}_\tau$. Under perfect calibration and $B_h \leq |\mathcal{A}_\tau|$, all selected items lie in $\mathcal{A}_\tau$, giving $\phi_{\text{unc}} = 1$. Under imperfect calibration, $\phi_{\text{unc}} \leq 1$ with $\phi_{\text{unc}} \to 1$ as calibration improves.

Diversity: Strategy div spreads selections across the confidence distribution, ensuring representation from all confidence regions including $\mathcal{A}_\tau$, giving $\phi_{\text{div}} \geq \phi_{\text{rand}} = \delta$. Since diversity does not exclusively target $\mathcal{A}_\tau$, $\phi_{\text{div}} \leq \phi_{\text{unc}}$.

Budget ordering and ratio. Since $B_h^{(k)} \propto 1/\phi_k$ and $\phi_{\text{rand}} \leq \phi_{\text{div}} \leq \phi_{\text{unc}}$, the required budgets satisfy $B_h^{(\text{unc})} \leq B_h^{(\text{div})} \leq B_h^{(\text{rand})}$. The ratio follows directly:

$$\frac{B_h^{(\text{rand})}}{B_h^{(\text{unc})}} = \frac{\phi_{\text{unc}}}{\phi_{\text{rand}}} = \frac{\phi_{\text{unc}}}{\delta} \geq 1,$$

which grows as $\delta \to 0$ since $\phi_{\text{unc}} \to 1$ while $\delta \to 0$. □

PROOF OF THEOREM 4

Proof. **Setup.** After the predict phase, each non-harvested item $x \in \mathcal{D} \setminus \mathcal{H}$ has calibrated confidence $\hat{p}_{\mathcal{H}}(x)$ and pointwise error rate $\text{err}(f_{\mathcal{H}}^-, x) \approx \min(\hat{p}_{\mathcal{H}}(x),\, 1 - \hat{p}_{\mathcal{H}}(x))$ by the calibration assumption. The residual error on non-harvested items is

$$\mathcal{E}(B_h) = \sum_{x \in \mathcal{D} \setminus \mathcal{H}} \text{err}(f_{\mathcal{H}}^-, x) \approx \sum_{x \in \mathcal{D} \setminus \mathcal{H}} \min\big(\hat{p}_{\mathcal{H}}(x),\, 1 - \hat{p}_{\mathcal{H}}(x)\big).$$

Escalating item x replaces its cheap model error with oracle error $\varepsilon^+ \approx 0$, yielding a per-item benefit of $\min(\hat{p}_{\mathcal{H}}(x), 1 - \hat{p}_{\mathcal{H}}(x))$.

Optimality of threshold escalation. To minimize residual error subject to escalating at most B_e items, we should escalate those with the largest per-item benefit, i.e. those with $\min(\hat{p}(x), 1 - \hat{p}(x))$ largest, equivalently those with $|\hat{p}(x) - 0.5|$ smallest. This corresponds exactly to selecting items within a symmetric interval $[t^{\text{lo}}, t^{\text{hi}}]$ around 0.5. The budget constraint $\rho(\Delta) \cdot (N - B_h) \leq B_e$ limits how wide this interval can be over the $N - B_h$ non-harvested items, so the optimal policy exhausts the budget by choosing the widest feasible interval:

$$\Delta^* = \sup\left\{\Delta \in [0,1] : \rho(\Delta) \cdot (N - B_h) \leq B_e\right\}.$$

This is well-defined since $\rho(0) = 0 \leq B_e/(N - B_h)$ and $\rho(\Delta)$ is non-decreasing in Δ, so the feasible set is non-empty and closed, and the sup is attained.

Residual error bound. The total error reduction from escalating B_e items each with per-item error at least $\frac{1}{2} - \frac{\Delta^*}{2}$ before escalation is at least $B_e \cdot (\frac{1}{2} - \frac{\Delta^*}{2})$, giving:

$$\mathcal{E}^* \leq \mathcal{E}(B_h) - B_e \cdot \left(\tfrac{1}{2} - \tfrac{\Delta^*}{2}\right),$$

confirming that larger B_e reduces residual error. □

B ICC ALGORITHM

We present the complete ICC pipeline as pseudocode in Algorithm 1. The algorithm covers all five stages — pre-process, allocate, harvest, predict, and escalate — with branching logic for each configurable option: four allocation approaches, three harvest strategies, two demonstration selection strategies, and the balance flag for demonstration composition. Special cases are handled but not shown here explicitly, including $B_h = 0$ (pure escalation, equivalent to uncertainty-based cascade baselines) and $B_e = 0$ (pure harvest with ICL adaptation, no escalation). Cache interactions are noted throughout: zero-shot cheap model calls and oracle calls are served from cache after pre-population, so only the ICL predict calls in Stage 4 constitute live API calls during experiments.

C ADDITIONAL EXPERIMENTAL RESULTS

C.1 CHEAP MODEL ROBUSTNESS

Table 2 reports accuracy at $B/N = 5\%$ on Overruling for three methods across four cheap models, ranging from a 1B active-parameter hybrid MoE (Granite-4.0-H-Tiny) to a 70B parameter model (Llama-3.3-70B-Instruct). These experiments are conducted with a wider ambiguous region $\tau = 0.3$, since one of the models is larger and therefore also more confident, and with fixed demonstrations using maximum demonstration size of $k_{max} = 32$. ICC-UF-δ consistently outperforms both baselines across all cheap models, demonstrating that the method is robust to the choice of cheap model.

Table 2: Accuracy at $B/N = 5\%$ on Overruling for three methods across four cheap models. ICC-UF-δ uses uncertainty harvest with $\beta^* = \hat{\delta}$. Bold indicates best per model.

Cheap Model	RandomOracle+Cheap	FixedCascade	ICC-UF-δ
Granite-3.3-8B-Instruct	0.869	0.888	**0.962**
Granite-4.0-H-Tiny	0.879	0.889	**0.934**
Llama-3.1-8B-Instruct	0.857	0.875	**0.952**
Llama-3.3-70B-Instruct	0.919	0.933	**0.967**

C.2 Demonstration Selection Strategy

We compare three demonstration selection strategies within ICC on Overruling using Llama-3.1-8B-Instruct as the cheap model ($\tau = 0.2$, $k_{\max} = 32$, $\beta^* = \hat{\delta}$): **fixed** (ICC-UF) selection draws the same k demonstrations for all pool items in a single batch call; **random** (ICC-UR) samples a fresh set of k demonstrations per item; and **similarity** (ICC-US) retrieves the k harvest items with logit confidence scores closest to each target item. Figure 4 shows cost-accuracy curves for all three strategies alongside FixedCascade.

All three strategies substantially outperform FixedCascade across all budget levels, confirming that the choice of demonstration selection matters less than ICL adaptation itself. Fixed and random selection perform nearly identically throughout, validating fixed selection as the recommended default given its significantly lower computational cost (one batch call versus N individual calls). Similarity selection underperforms at small budgets, likely because the 1D logit-based proximity metric provides insufficient signal to differentiate demonstrations when the harvest set is small. However, it recovers at larger budgets (0.978 at $B/N = 20\%$). A similarity selection approach that uses full embeddings remains a potential direction for future work.

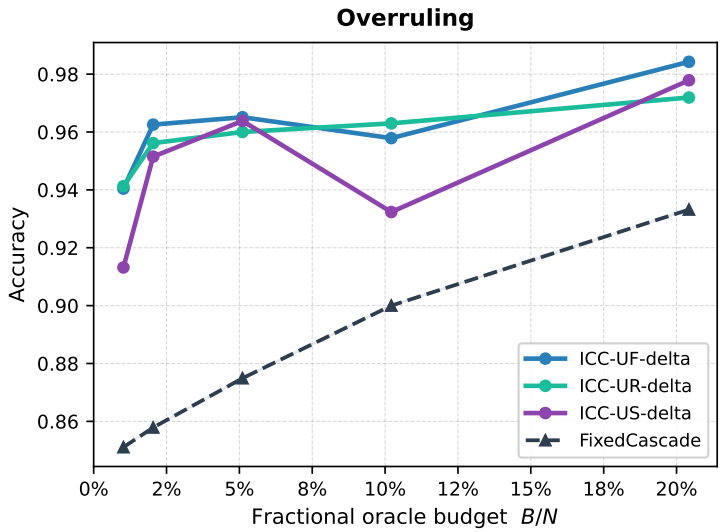

Figure 4: Demonstration selection strategy comparison on Overruling (Llama-3.1-8B-Instruct, τ=0.2, $k_{\max}$=32). Fixed and random selection perform comparably; similarity selection of demonstrations underperforms at small budgets.

C.3 Effect of Ambiguity Threshold τ

We evaluate the sensitivity of ICC-UF-δ to the ambiguity threshold τ, which controls both the definition of the ambiguous region $\mathcal{A}_\tau$ and the harvest fraction $\beta^* = \hat{\delta}(\tau)$. We sweep $\tau \in \{0.1, 0.2, 0.3, 0.4\}$ on Overruling using Llama-3.1-8B-Instruct as the cheap model with uncertainty harvest and all five budget levels.

Figure 5 shows the results. At large budgets ($B \geq 240$), accuracy is nearly flat across all τ values, varying by less than 0.01, indicating that ICC is robust to the choice of τ when the budget is sufficient. Sensitivity is higher at small budgets, particularly $B = 24$, where accuracy increases by nearly 8 percentage points from $\tau = 0.1$ to $\tau = 0.4$ as larger τ allocates more budget to harvest. Based on these results, $\tau = 0.2$ provides a reliable default, since it lies in the flat region for large budgets and near the inflection point for small budgets. This avoids both overly conservative and overly aggressive ambiguous region definitions.

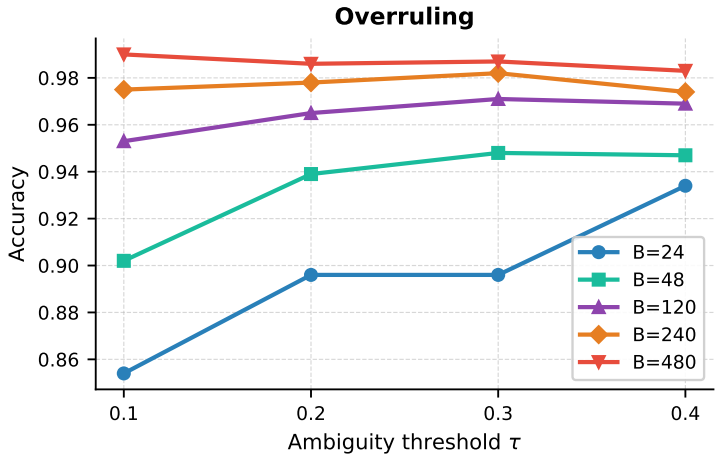

Figure 5: Accuracy vs. ambiguity threshold τ on Overruling (Llama-3.1-8B-Instruct, uncertainty harvest). Large budgets are insensitive to τ; small budgets benefit from larger τ values.

C.4 Effect of Maximum Demonstrations k_{max}

We evaluate the sensitivity of ICC to the maximum number of ICL demonstrations $k_{\max}$ on Overruling and ToxiGen using Llama-3.1-8B-Instruct ($\tau = 0.2$, uncertainty harvest, $\beta^* = \hat{\delta}$, fixed demonstration selection). We vary $k_{\max} \in \{8, 16, 32, 64\}$, keeping all other hyperparameters fixed.

Figure 6 shows accuracy at budget levels $B/N \in \{1\%, 5\%, 20\%\}$ for each value of $k_{\max}$. Accuracy is nearly identical across all values of $k_{\max}$ at small and medium budgets ($B/N \leq 5\%$), confirming that ICC is robust to the number of demonstrations when the harvest set is small. Differences emerge at $B/N = 20\%$, where $k_{\max} = 64$ slightly underperforms smaller values, likely due to label noise or context length effects from large demonstration sets. We recommend $k_{\max} = 32$ as a robust default: it performs at or near the top across all budget levels.

C.5 Effect of Calibration Budget Fraction B_{cal}/B

We evaluate the sensitivity of ICC to the calibration set size, as represented by the relative fraction when compared to the full budget, B_{cal}/B, on Overruling and ToxiGen using Llama-3.1-8B-Instruct ($\tau = 0.2$, uncertainty harvest, $\beta^* = \hat{\delta}$, fixed demonstration selection). We consider total budget fraction $B/N = 5\%$ and vary the effective calibration budget fraction $B_{cal}/B \in \{10, 20, 50\}\%$, keeping all other hyperparameters fixed. Note that a total budget fraction of 5% on a dataset of size $N = 1000$ corresponds to a total budget size of 50, and a calibration budget fraction of 10% implies that there are only 5 items in the calibration set. Similar to the setup in the main paper, items in the calibration set are held separately and not included for evaluation.

Table 3 shows minor differences in accuracy across the chosen calibration budget fractions. This indicates that ICC appears to be robust to the size of the initial calibration set and is reasonably functional even with very limited calibration.

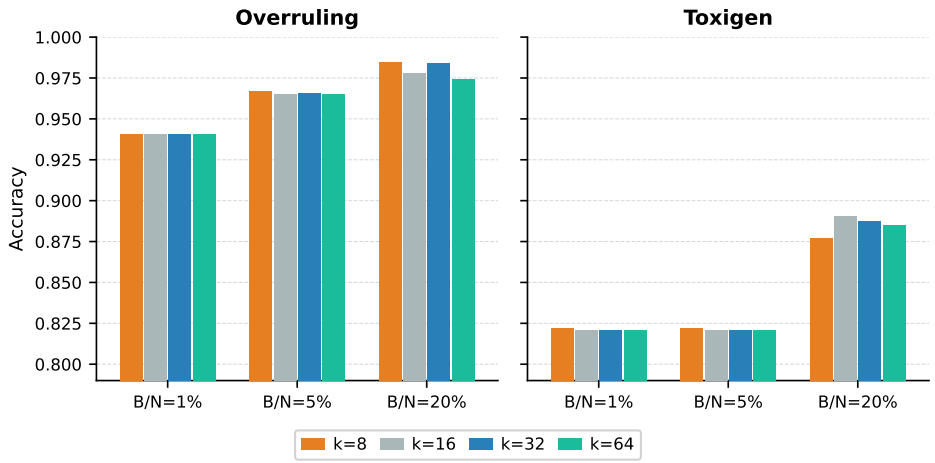

Figure 6: Effect of maximum demonstrations $k_{\max}$ on the Overruling and ToxiGen datasets (Llama-3.1-8B, τ=0.2, uncertainty harvest, fixed selection).

Table 3: Accuracy at $B/N = 5\%$ on Overruling and ToxiGen across three different effective calibration budget fractions B_{cal}/B. The ICC method deployed is ICC-UF-δ, which uses uncertainty harvest with $\beta^* = \hat{\delta}$. Bold indicates best per calibration budget fraction.

Dataset	B_{cal}/B = 10 %	B_{cal}/B = 20 %	B_{cal}/B = 50 %
Overruling	0.95	**0.96**	0.95
ToxiGen	0.82	0.83	**0.84**

Algorithm 1 In-Context Cascade (ICC) — Full Pipeline

Require: $\mathcal{D}$, $\mathcal{T}$, B, f^+, f^-, $\mathcal{D}_{\text{cal}}$, τ, [β or (k^* and α) or γ or λ], strategy, $k_{\max}$, n_c, balance (Boolean)
Ensure: Labels $\hat{y}_i \in \{0, 1\}$ for all $x_i \in \mathcal{D}$

 # —- Stage 1: Pre-process —-
Fit temperature $T \leftarrow \arg\min_T \text{NLL}(f^-, \mathcal{D}_{\text{cal}}, T)$
$s(x_i) \leftarrow f^-(x_i \mid \mathcal{T}, \text{demos}=\emptyset)$ for all $x_i \in \mathcal{D}$ ▷ zero-shot; served from cache if available
$\hat{p}(x_i) \leftarrow \sigma(s(x_i)/T)$ for all $x_i \in \mathcal{D}$
$\hat{\delta} \leftarrow \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}[|\hat{p}(x_i) - 0.5| < \tau]$ ▷ predicate difficulty

 # —- Stage 2: Allocate —-
if explicit **then**
    $\beta^* \leftarrow \beta$
else if ambiguous region or target demonstrations **then**
    $\beta^* \leftarrow \hat{\delta}$ (ambiguous region); $\beta^* \leftarrow \min\left(1, \frac{k^* \log(1/\alpha)}{\hat{\delta} \cdot B}\right)$ (target demonstrations)
else if optimal fit **then**
    Estimate β^* from γ or λ and other parameters
end if
$B_h \leftarrow \lfloor \beta^* \cdot B \rfloor, \quad B_e \leftarrow B - B_h$

 # —- Stage 3: Harvest —-
if random **then**
    $\mathcal{H}_{\text{idx}} \leftarrow \text{Sample}(\{1, \ldots, N\}, B_h)$ ▷ uniform without replacement
else if uncertainty **then**
    $\mathcal{H}_{\text{idx}} \leftarrow \text{ArgSort}_{B_h}(|\hat{p}(x_i) - 0.5|, \text{ascending})$ ▷ most uncertain items; $\hat{p}$ from Stage 1
else if diversity **then**
    Cluster $\{\hat{p}(x_i)\}$ into n_c groups via k-means
    $\mathcal{H}_{\text{idx}} \leftarrow$ one representative per cluster (nearest centroid), fill remainder randomly
end if
$y_i \leftarrow f^+(x_i \mid \mathcal{T})$ for all $i \in \mathcal{H}_{\text{idx}}$ ▷ oracle labels; served from cache
$\mathcal{H} \leftarrow \{(x_i, y_i) : i \in \mathcal{H}_{\text{idx}}\}$

 # —- Stage 4: Predict —-
$\mathcal{R} \leftarrow \mathcal{D} \setminus \{x_i : i \in \mathcal{H}_{\text{idx}}\}$ ▷ non-harvested items
$k \leftarrow \min(|\mathcal{H}|, k_{\max})$
if $k = 0$ **then**
    $\hat{y}(x), s(x) \leftarrow f^-(x \mid \mathcal{T}, \text{demos}=\emptyset)$ for all $x \in \mathcal{R}$ ▷ zero-shot fallback; served from cache
else
    if fixed **then**
        $\text{pos} \leftarrow \{(x, y) \in \mathcal{H} : y = 1\}, \quad \text{neg} \leftarrow \{(x, y) \in \mathcal{H} : y = 0\}$
        if balance **then**
            $\text{demos} \leftarrow \text{Sample}(\text{pos}, \lfloor k/2 \rfloor) \cup \text{Sample}(\text{neg}, k - \lfloor k/2 \rfloor)$
        else
            $\text{demos} \leftarrow \text{Sample}(\mathcal{H}, k)$
        end if
        $\hat{y}(x), s(x) \leftarrow f^-(x \mid \mathcal{T}, \text{demos})$ for all $x \in \mathcal{R}$ ▷ single batch call
    else if similarity **then**
        for each $x \in \mathcal{R}$ **do**
            $\text{demos} \leftarrow k$ items from $\mathcal{H}$ nearest to x by $|\hat{p}(\cdot) - \hat{p}(x)|$
            $\hat{y}(x), s(x) \leftarrow f^-(x \mid \mathcal{T}, \text{demos})$
        end for
    end if
    $\hat{p}_{\mathcal{H}}(x) \leftarrow \sigma(s(x)/T)$ for all $x \in \mathcal{R}$
end if

 # —- Stage 5: Escalate —-
$\mathcal{E}_{\text{idx}} \leftarrow \text{ArgSort}_{B_e}(|\hat{p}_{\mathcal{H}}(x_i) - 0.5|, \text{ascending})$ ▷ most uncertain non-harvested items
$y_i \leftarrow f^+(x_i \mid \mathcal{T})$ for all $i \in \mathcal{E}_{\text{idx}}$ ▷ oracle labels; served from cache

 # —- Assemble final labels —-
$\hat{y}_i \leftarrow y_i$ **if** $i \in \mathcal{H}_{\text{idx}} \cup \mathcal{E}_{\text{idx}}$ **else** $\hat{y}(x_i)$
return $\{\hat{y}_i\}_{i=1}^{N}$