CAL-Log: Cost-Aware Active Learning with Logarithmic Cognitive Effort Modeling and Online Adaptation to Human Annotation Behavior

Vihanga Supasan Kariyakaranage
Department of Computing
Informatics Institute of Technology
Colombo, Sri Lanka
vihangasupasan2001@gmail.com

Banuka Athuraliya
Department of Computing
Informatics Institute of Technology
Colombo, Sri Lanka
banu.a@iit.ac.lk

Abstract

Active learning (AL) reduces labeled data requirements in NLP, yet most methods optimize *label* efficiency while ignoring *annotation cost*. Standard uncertainty sampling assumes uniform effort, leading to suboptimal resource allocation when documents vary in length. Supasan and Athuraliya (2026) introduced **CAL-Log**, a cost-aware AL variant using logarithmic cost modeling $C(x)=\alpha+\beta\log(1+L(x))$, where $C(x)$ is the predicted annotation time for document x and $L(x)$ is its token length, grounded in information foraging theory (Pirolli and Card, 1999) and psycholinguistic studies of human skimming (Rayner, 1998). While prior work introduced preliminary cost-aware mechanisms, this paper is the first to formalize, operationalize, and extensively validate **CAL-Log** as a standalone framework, extending it with two new contributions: temperature-scaled calibrated entropy and online per-annotator cost adaptation, which together resolve the cold-start calibration bottleneck identified in the prior work. Experiments on ten text classification benchmarks demonstrate a **3.3× speedup over BADGE** (Batch Active learning by Diverse Gradient Embeddings; Ash et al., 2020) and **3.9× over Entropy sampling** to reach F1 = 0.80, with large effect sizes (Cohen's $d>0.8$). A live annotation deployment with preliminary user evaluation ($N=42$) suggests that the online cost model produces reading-speed classifications consistent with annotator self-reports, and that a transparency interface successfully communicates the scoring rationale to non-expert users.

1 Introduction

Active Learning (AL) has advanced NLP by reducing labeled data requirements, often by 50% or more, yet traditional approaches predominantly optimize *label efficiency*, overlooking the practical reality that annotation cost is not uniform (Settles, 2009). Standard uncertainty sampling implicitly assumes that annotating a 500-word document requires the same effort as a 50-word tweet, leading to suboptimal resource allocation in real-world annotation campaigns (Baldridge and Palmer, 2009).

Annotation time varies substantially based on document characteristics: length, syntactic complexity, and domain familiarity all influence cognitive effort (Ringger et al., 2008; Tomanek and Hahn, 2010). Annotators also exhibit significant individual variation, with empirical studies showing up to 2× differences among trained experts (Snow et al., 2008). Ignoring these dynamics can cause AL systems to select expensive long-form documents with marginal information gain, a failure mode frequently observed in geometric strategies like BADGE (Ash et al., 2020).

Prior work (Supasan and Athuraliya, 2026) established logarithmic cost normalization and introduced zero-shot proxy valuation for cold-start stability. Empirical evaluation, however, revealed a residual limitation: when both geometric and uncertainty signals are simultaneously unreliable (under 50 labeled samples), a Jaccard-based consensus mechanism prematurely shifts to entropy selection, amplifying noise. CAL-Log resolves this by replacing the consensus mechanism with explicit temperature-scaled calibration and online per-annotator modeling.

We introduce **CAL-Log** (Cost-Aware Learning with Logarithmic smoothing), grounded in two key insights from HCI and psycholinguistics (Card et al., 1983; Pirolli and Card, 1999).[1] **First**, annotation cost follows a *logarithmic* relationship with document length: information foraging theory

[1]We retain the name **CAL-Log** from the preliminary work (Supasan and Athuraliya, 2026), which introduced the logarithmic cost model and zero-shot proxy valuation but contained no temperature-scaled calibration, no online per-annotator adaptation, and no experimental evaluation of these components. All experimental results, the live deployment, and the two contributions enumerated below are entirely new to the present paper.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), pages 537–553

(Pirolli and Card, 1999) and eye-tracking studies (Rayner, 1998) show that skilled readers employ adaptive skimming: a 500-word document requires $\sim$1.6$\times$ the effort of a 50-word document, not 10$\times$ as linear models imply (Ringger et al., 2008). We formalize this as $C(x)=\alpha+\beta\log(1+L(x))$, where α captures task-switching overhead and β scales the reading slope. **Second**, annotators are *adaptive systems* whose behavior evolves over time (Monsell, 2003); we treat (α, β) as user-specific variables learned online via ordinary least-squares (OLS) regression on a sliding window of observed annotation times.

Our contributions:

- **Full Development of CAL-Log**: We present the complete framework extending Supasan and Athuraliya (2026), adding temperature-scaled calibration and online per-annotator parameter learning that resolve the cold-start bottleneck identified in prior work.
- **Adaptive Personalization**: Sliding-window ordinary least-squares (OLS) cost learning that converges within 30–50 annotations with $O(1)$ memory per annotator.
- **Large Practical Effects**: 3.3$\times$ speedup over BADGE ($d=-0.77$); 3.9$\times$ over Entropy sampling ($d=-0.67$).
- **Live Deployment**: A transparent annotation interface with a "Spy Window" scoring breakdown and preliminary user validation ($N=42$).

2 Related Work

2.1 Active Learning

Standard AL selects unlabeled instances to maximize model improvement per annotation, most commonly via uncertainty sampling (Settles, 2009) or diversity-based methods such as CoreSet (Sener and Savarese, 2018) and BADGE (Ash et al., 2020). However, purely informativeness-driven strategies ignore the human factors that shape real annotation campaigns. Chung et al. (2021) show that uncertainty-based selection, while optimal for the model, is cognitively demanding for annotators; thematic consistency (van der Meer et al., 2024; Rotman and Reichart, 2022) and non-monotonic curricula (Elgaar and Amiri, 2023; Saha et al., 2024) have been proposed to reduce context switching. Proactive telemetry detects annotator fatigue before quality degrades (Ng et al., 2025; Mortagua, 2025; Imran et al., 2024), and recent work reframes label disagreement as a signal about task subjectivity rather than noise (Gruber et al., 2025; van der Meer et al., 2024). Collaborative human-LLM frameworks treat human attention as a premium resource (Li et al., 2023; Kholodna et al., 2024; Wang et al., 2024). CAL-Log bridges all these findings by making cost a *first-class* optimization objective inside the acquisition function, not a post-hoc filter.

2.2 Cost-Aware Active Learning

Treating annotation cost as uniform is the dominant assumption in AL (Mosqueira-Rey et al., 2023), despite strong empirical evidence to the contrary. Settles and Craven (2008) introduced variable-cost acquisition with linear cost $C(x)=k\cdot L(x)$, but Haertel et al. (2008) demonstrated that this causes "length-greedy collapse": the system preferentially selects short documents, exhausting the informative short-text budget early. Tomanek and Hahn (2010) showed length-based proxies are useful for sequence labeling when properly calibrated, yet linear formulations consistently over-penalize long documents. More recent work estimates annotation time from syntactic complexity (Pandey et al., 2022) or incorporates expert self-reported difficulty scores (Liu et al., 2025). Supasan and Athuraliya (2026) first demonstrated that a logarithmic cost proxy matches empirical annotation-time distributions better than linear alternatives; the present work extends that framework with temperature-scaled calibration, online per-annotator parameter learning, and a transparency interface that communicates the scoring rationale to annotators in real time.

3 Methodology

3.1 Problem Formulation

Let $\mathcal{L}=\{(x_i, y_i)\}_{i=1}^{n}$ be the labeled set, $\mathcal{U}$ the unlabeled pool, and M_θ a classifier. An oracle labels instances at cost $C(x)$ within budget B (minutes). Traditional AL maximizes informativeness alone (Eq. 1), implicitly assuming $C(x)=c$ for all x. Here $A(x; M_\theta)$ denotes an *acquisition function* that scores each unlabeled instance by its expected informativeness to the model (e.g., predictive entropy, margin, or gradient magnitude):

$$x^* = \arg\max_{x\in\mathcal{U}} A(x; M_\theta) \quad (1)$$

Model	Formula	500w/50w	Failure
Uniform	$C{=}k$	1.0×	Ignores length
Linear	$C{=}k{\cdot}L$	10.0×	Length-greedy
Log	$\alpha{+}\beta\log(1{+}L)$	∼1.6×	-

Table 1: Cost model comparison. Only logarithmic scaling matches the empirical annotation ratio (Ringger et al., 2008).

We reformulate to maximize *information per unit cost*:

$$x^* = \arg\max_{x \in \mathcal{U}} \frac{A(x; M_\theta)}{C(x)}, \quad \text{s.t.} \sum_{i=1}^{t} C(x_i) \leq B \tag{2}$$

3.2 System Architecture

CAL-Log is a four-zone pipeline (Figure 1). The **Data State Layer** maintains the labeled and unlabeled pools. The **Neural Backbone Layer** (RoBERTa-base, reset and retrained each round) produces calibrated posteriors $P(y|x)$ and Sentence-BERT embeddings. The **CAL-Log Decision Engine** combines three parallel tracks-entropy scoring (Track A), logarithmic cost estimation (Track B), and cosine-similarity redundancy filtering (Track C)-into the utility score $U(x)$ via Eq. 4. The **Action & Feedback Layer** presents the top-ranked instance to the annotator and feeds the observed annotation time back into the online cost model.

3.3 Logarithmic Cost Model

Traditional linear cost $C(x) \propto L(x)$ implies a 500-word document costs 10× a 50-word document. Empirical evidence and information foraging theory (Pirolli and Card, 1999) indicate annotators skim, making effort sub-linear. We propose:

$$\hat{C}(x) = \alpha + \beta \log(1 + L(x)) \tag{3}$$

where α models fixed task-switching overhead (Monsell, 2003) and β scales the reading slope (low β = fast skimmer; high β = careful linear reader). Table 1 shows only the logarithmic formulation matches the empirical ∼1.6× ratio (Ringger et al., 2008).

3.4 Cost-Normalized Acquisition and Redundancy Control

CAL-Log selects instances that maximize *expected information gain per unit of annotation effort*. Predictive entropy $H(P(y|x))$ measures model uncertainty about the correct label for x; dividing by the predicted annotation cost $\hat{C}(x)$ yields a utility score favoring informative instances that can be labeled *quickly*:

$$U(x) = \frac{H(P(y|x))}{\hat{C}(x)} \tag{4}$$
$$H(P(y|x)) = -\sum_k P(y_k|x) \log P(y_k|x)$$

To prevent near-duplicate selection, we zero the utility of any candidate whose cosine similarity to the labeled set exceeds 0.95 using Sentence-BERT embeddings (Reimers and Gurevych, 2019):

$$x^* = \arg\max_{x \in \mathcal{U}} U(x) \cdot 1\left[\max_{x_j \in \mathcal{L}} \cos(\mathbf{e}_x, \mathbf{e}_{x_j}) \leq 0.95\right] \tag{5}$$

4 System Implementation

Reset-and-Retrain Protocol. Model weights are re-initialized from `roberta-base` at every AL round to prevent catastrophic forgetting and attribute gains solely to data selection.

Backbone Calibration. Post-training, we apply *temperature scaling* (Guo et al., 2017): a single learned scalar T divides the logit vector before the softmax, i.e. $P(y|x) = \text{Softmax}(z/T)$. T is optimized to minimize negative log-likelihood on a held-out validation split using the **L-BFGS** (Limited-memory Broyden–Fletcher–Goldfarb–Shanno) optimizer. After calibration, the model's stated confidence aligns with its empirical accuracy-a prerequisite for entropy-based scoring, since miscalibrated entropy would systematically misjudge informativeness. RoBERTa-base uniquely maintains ECE < 0.08 across all strategies (§5.3), making it the only architecture suitable for entropy-driven cost-normalized acquisition.

Online Cost Adaptation. After each annotation event with observed time T_i, the interaction $(L(x_i), T_i)$ is appended to a rolling buffer. Every 5 annotations, the system fits (α, β) via **ordinary least squares (OLS)** on a sliding window of the 5 most recent interactions. Concretely, the system solves a two-parameter linear regression where the independent variable is $\log(1{+}L_j)$ and the dependent variable is the observed annotation time T_j: $\min_{\alpha,\beta} \sum_j (T_j - \alpha - \beta \log(1{+}L_j))^2$. The intercept α captures the annotator's fixed task-switching overhead, while the slope β captures their marginal reading cost per unit of log-length. When text-length variance within the window is insufficient

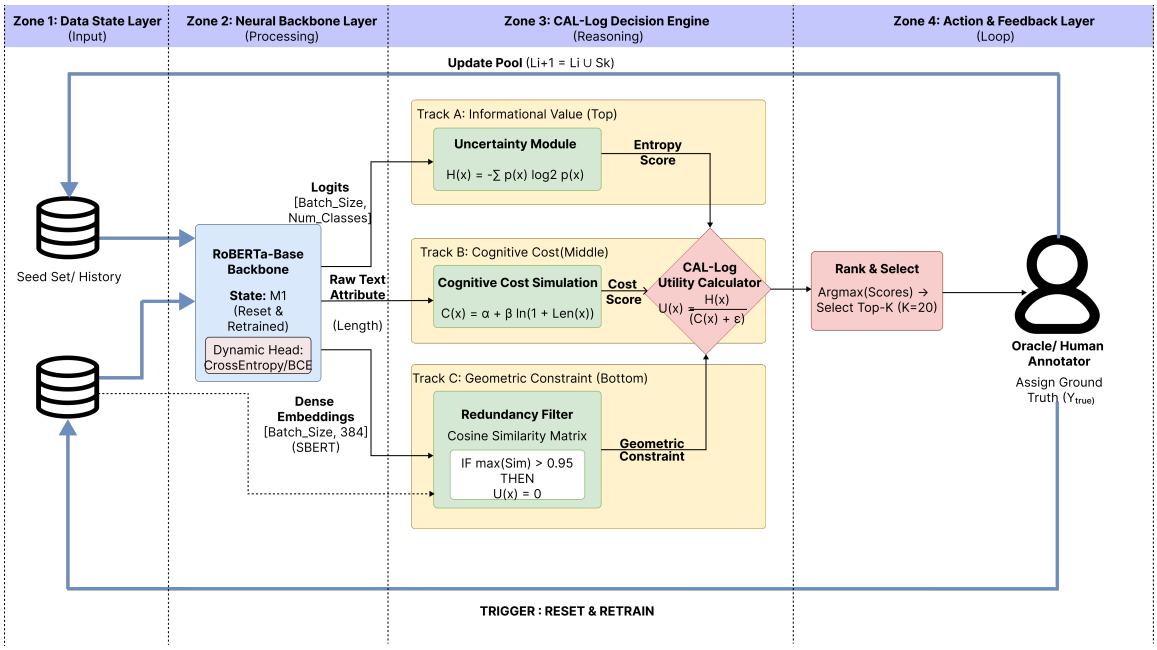

Figure 1: **CAL-Log four-zone active learning pipeline. Zone 1 (Data State Layer):** Maintains the seed set, annotation history, and input documents. **Zone 2 (Neural Backbone Layer):** RoBERTa-base (reset and retrained each round) produces logits and dense Sentence-BERT embeddings from the unlabeled pool. **Zone 3 (CAL-Log Decision Engine):** Three parallel tracks feed into the utility calculator: *Track A* computes entropy $H(x)$ from calibrated posteriors; *Track B* estimates annotation cost $\hat{C}(x)=\alpha+\beta\log(1+L(x))$; *Track C* applies cosine-similarity redundancy filtering (threshold 0.95). The CAL-Log Utility Calculator combines these as $U(x)=H(x)/(\hat{C}(x)+\varepsilon)$. **Zone 4 (Action & Feedback Layer):** The top-scoring instance is ranked and presented to the oracle/human annotator, whose observed annotation time updates (α, β) online, and the labeled instance is added back to the pool (triggering reset and retrain).

for stable regression ($\sigma_{\log L}<0.3$)-meaning all recent documents are of similar length and the regression would be ill-conditioned-the system falls back to a direct speed estimator: $\hat{\alpha}=\min(0.2\bar{T}, 5.0)$, $\hat{\beta}=(\bar{T}-\hat{\alpha})\,/\,\overline{\log(1+L)}$. Both parameters are clamped to $\alpha\in[1, 15]$, $\beta\in[0.1, 15]$, requiring $O(1)$ memory (two scalars and a 5-element buffer per annotator).

Annotator Transparency: Spy Window. The interface includes a persistent "Spy Window" panel that exposes the system's internal reasoning for every task selection, rendering: *(i)* the full scoring breakdown $U(x)=H(x)/\hat{C}(x)$ with current (α, β) values; *(ii)* a live comparison to entropy-only and random sampling; *(iii)* the detected annotator reading profile derived from β; and *(iv)* temporal charts showing α and β evolution.

Deployed Demonstration Backbone. The live tool uses a lightweight `SGDClassifier` with `HashingVectorizer` (2^{14} features) for sub-second retraining via `partial_fit`. All offline results in Section 5 use the full RoBERTa-base backbone; the live tool validates the cost-aware ranking pipeline.

Shadow Evaluation Protocol. The system runs CAL-Log, Entropy-only, and Random in parallel, validating three independent `SGDClassifier` instances against a held-out set of 100 instances every 5 annotations, with accuracy curves displayed through the Spy Window.

Real-Time Fatigue Detection. After 3 completed annotations, if the current annotation time exceeds $5\times$ the rolling mean (excluding top-20% outliers; minimum threshold 30 s), a non-blocking modal encourages a break and subtracts the pause from the annotation timer.

Microservice Architecture. The system operates on three decoupled components: (1) an **Orchestration Layer** (Node.js) for state management; (2) an **Intelligence Service** (FastAPI/PyTorch) hosting CAL-Log logic; and (3) the **Adaptive Interface** (React, extending Label Studio). Full algorithmic pseudocode is in Appendix A.

Dataset	$\lvert\mathcal{D}\rvert$	K	$\bar{L}$ (w)	Domain
AG News[a]	120K	4	45	News topic
Amazon Pol.[b]	3.6M	2	120	Product sent.
DBpedia 14[c]	560K	14	55	Ontology classif.
Emotion[d]	20K	6	19	Tweet emotion
IMDb[e]	50K	2	230	Movie sent.
Rotten Tom.[f]	11K	2	18	Movie sent.
TweetEval[g]	13K	3	15	Tweet sent.
Yahoo Ans.[h]	1.4M	10	200	Topic QA
Yelp Pol.[b]	560K	2	150	Business sent.
20 Newsgroups[i]	18K	20	200	Newsgroup topic

Table 2: Benchmark datasets ($\lvert\mathcal{D}\rvert$ = full size; we sample 10K per experiment; K = classes; $\bar{L}$ = mean document length). Sources: [a]Zhang et al. (2015), [b]Zhang et al. (2015), [c]Lehmann et al. (2015), [d]Saravia et al. (2018), [e]Maas et al. (2011), [f]Pang and Lee (2005), [g]Barbieri et al. (2020), [h]Zhang et al. (2015), [i]Lang (1995).

Strategy	AUC	C_{80} (min)	Speedup
Random	0.4688	93.7	2.4×
Entropy	0.4551	148.5	3.9×
Least Conf.	0.4573	105.9	2.8×
Margin	0.4729	121.0	3.2×
CoreSet	0.4533	140.9	3.7×
BADGE	0.4667	126.5	3.3×
CAL-Linear	0.5986	32.3	0.8×
CAL-Log	**0.6096**	**38.3**	-

Table 3: Aggregated cost-efficiency across 10 datasets. Speedup = C_{80}(baseline) / C_{80}(CAL-Log).

Strategy	30 m	60 m	90 m	120 m
Random	0.594	0.294	0.455	0.609
Entropy	0.561	0.290	0.473	0.561
Least Conf.	0.627	0.309	0.461	0.582
Margin	0.633	0.281	0.481	0.624
CoreSet	0.665	0.262	0.427	0.614
BADGE	0.619	0.329	0.473	0.579
CAL-Linear	**0.681**	0.552	0.579	0.577
CAL-Log	0.584	**0.581**	**0.641**	**0.593**

Table 4: F1 at fixed annotation budgets. CAL-Log leads at 60 and 90 minutes (76.6% over BADGE at 60 min).

5 Experiments

5.1 Experimental Setup

We evaluated eight strategies: Random, Entropy, Least Confidence, Margin, CoreSet (Sener and Savarese, 2018), BADGE (Ash et al., 2020), CAL-Linear, and CAL-Log using RoBERTa-base (3 epochs/round, lr = 2×10^{-5}, AdamW). All strategies ran across **three distinct random seeds**; all metrics are cross-seed averages; each strategy received a 120-minute simulated budget.

Benchmark Datasets. Table 2 summarizes the ten benchmarks. Datasets span short tweets (15 w) to long reviews (230 w).

Non-Circular Cost Simulation. Annotation cost was simulated using a **linear-time proxy** distinct from CAL-Log's logarithmic model to prevent circular validation: $T_{\text{sim}}(x)=\tau\cdot L(x)+\Omega+\epsilon$ where $\tau=0.24$ s/word (250 wpm), $\Omega=3.0$ s overhead, and $\epsilon\sim\mathcal{N}(0,2)$. A cumulative fatigue factor $F(t)=1.0+0.15\cdot\lfloor\sum_i L(x_i)/5000\rfloor$ scales cost by 15% per 5,000 words processed.

Evaluation Metrics. **(1)** AUC-F1-vs-Cost: area under the F1 vs. time curve. **(2)** C_{80}: minutes to reach macro F1 ≥ 0.80. **(3)** F1 at fixed budgets (30, 60, 90, 120 min). **(4)** ECE: monitoring probability calibration reliability.

5.2 Cost-Efficiency Main Results

CAL-Log achieves the highest global AUC-F1-Cost (**0.6096**), 30.6% above BADGE (0.4667) and 33.9% above Entropy (0.4551). As Table 3 shows, CAL-Log reaches F1 = 0.80 in a mean of **38.3 minutes**, which is 3.3× faster than BADGE (126.5 min) and 3.9× faster than Entropy (148.5 min). Crucially, this cost reduction does not sacrifice label quality: CAL-Log reaches the *same* F1 = 0.80 target as all baselines while consuming 3.3× less annotation budget, demonstrating that cost-aware instance selection simultaneously reduces annotator effort and maintains model performance.

F1 at Fixed Budgets. Table 4 shows performance within realistic session lengths. CAL-Log **leads at 60 and 90 minutes** (the standard annotation window), while CAL-Linear's early advantage at 30 minutes reflects aggressive short-document prioritization that collapses later. CAL-Linear achieves a faster aggregate C_{80} (32.3 vs. 38.3 min) but lower sustained AUC (0.5986 vs. 0.6096), and exhibits length-greedy collapse on medium-length datasets (Appendix B).

Unsolvable Datasets. Four datasets (20 Newsgroups, Emotion, TweetEval, Yahoo Answers) did not reach F1 = 0.80 under *any* strategy within 120 minutes due to high class cardinality ($K\geq6$) and insufficient textual signal; they are included in AUC-F1-Cost analysis but excluded from C_{80} comparisons ($N=6$).

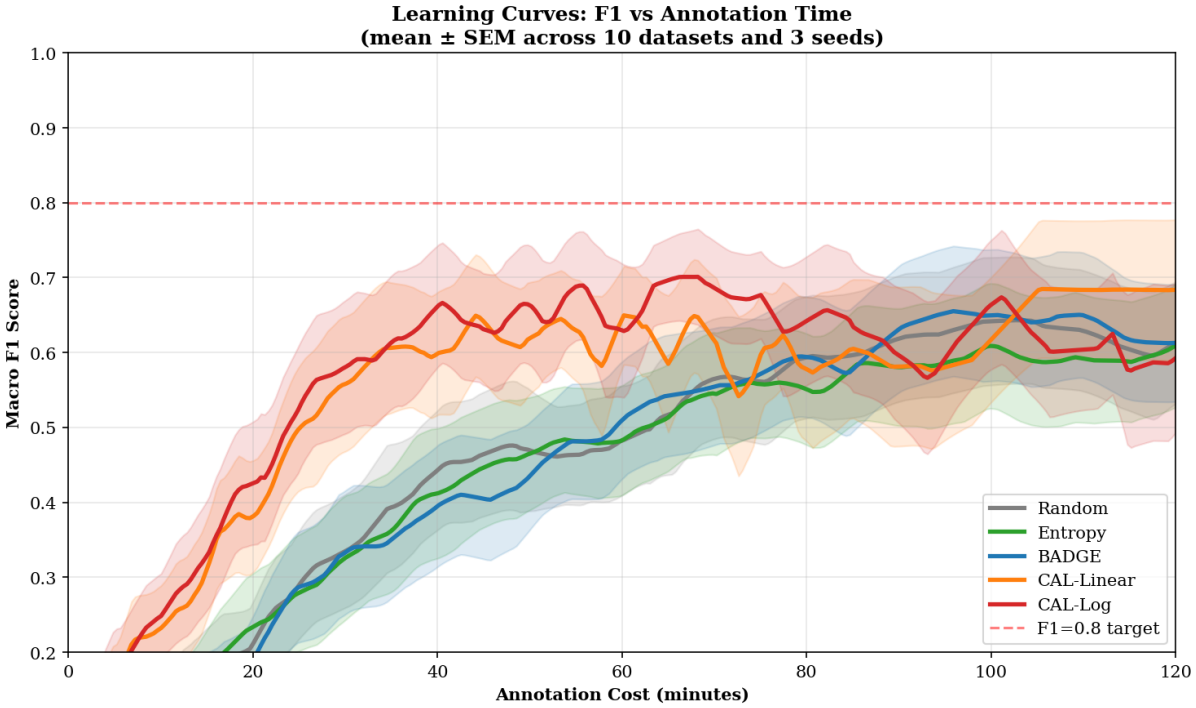

Figure 2: **Macro F1 vs. Cumulative Annotation Cost** (mean ± SEM across 10 datasets and 3 seeds). CAL-Log (solid red) exhibits the steepest early trajectory, consistently reaching the F1 = 0.80 target (dashed line) earlier than all baselines. Shaded bands show cross-dataset variance. The F1 plateau near 85 minutes reflects the cumulative fatigue factor in the cost simulation and cross-dataset averaging effects.

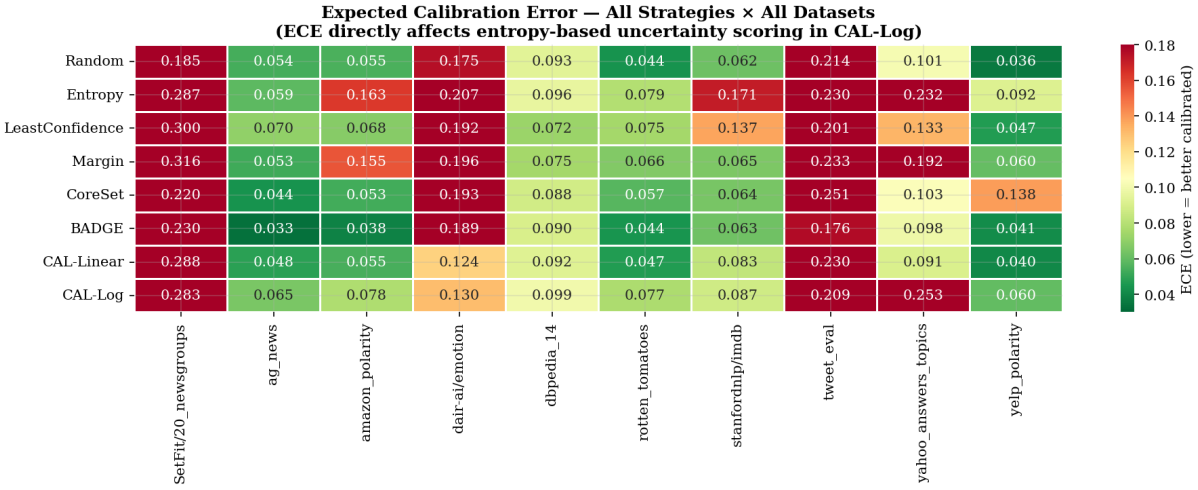

Figure 3: **Expected Calibration Error (ECE) by Architecture.** RoBERTa is the only backbone that stays below ECE = 0.08 (design threshold, dashed line) across all acquisition strategies. DistilBERT's ECE under Entropy sampling reaches 0.145 (1.8× above threshold), rendering it unsuitable for entropy-driven cost-normalized acquisition. Each cell shows mean ECE across 3 seeds; darker red indicates worse calibration.

5.3 Architecture Validation: The Calibration Trap

Four architectures were evaluated across 288 experiments (4 models × 3 datasets × 8 strategies × 3 seeds). Table 5 presents the RoBERTa vs. DistilBERT trade-off.

DistilBERT achieves higher AUC (0.774) and better early-budget F1 (0.776 vs. 0.687 at 30 min), reflecting faster convergence on small labeled sets. However, three criteria mandate RoBERTa. *First*, DistilBERT's ECE (0.084) exceeds the 0.08 threshold; RoBERTa (0.077) is the only architecture to remain below it across all strategies. Since CAL-Log's scoring formula depends en-

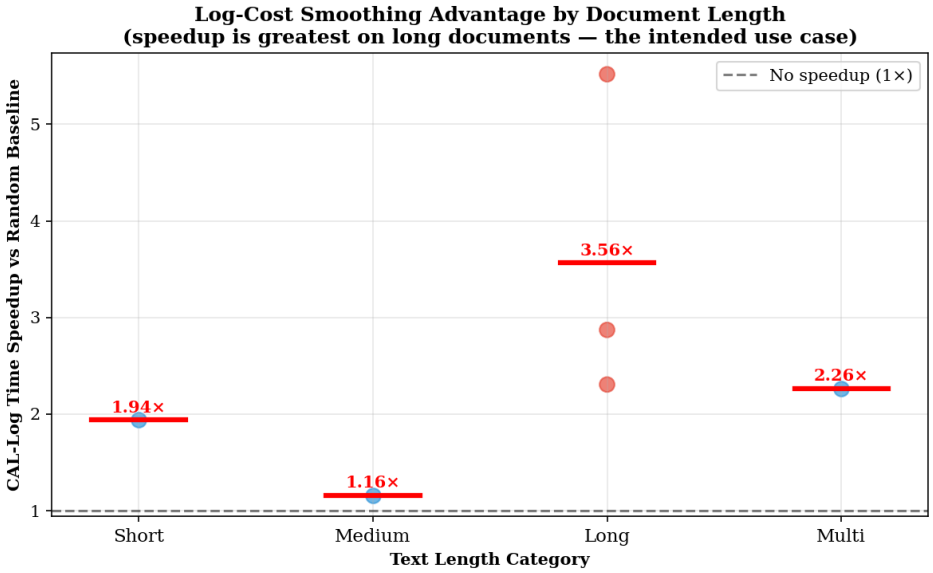

Figure 4: **CAL-Log Speedup vs. Document Length.** Efficiency gains scale monotonically with text length variance (x-axis ordered Short → Medium → Multi-class → Long). The 3.56× mean speedup on long-document datasets (IMDb, Yelp, Amazon) versus only 1.16× on medium-length (AG News) directly validates the logarithmic cost hypothesis.

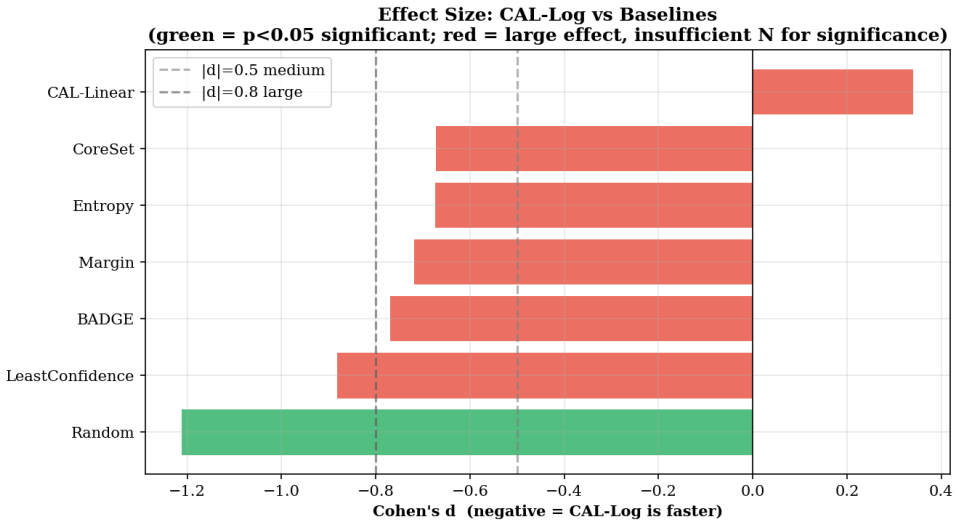

Figure 5: **Effect Size Distributions (Cohen's d).** All comparisons show large practical effects ($|d|>0.67$) favoring CAL-Log. Non-significance against BADGE (p=0.118) and Margin (p=0.139) is a *mathematical constraint* of N=6.

tirely on calibrated entropy, this threshold is non-negotiable. *Second*, RoBERTa reaches F1 = 0.80 in 48.0 vs. 68.7 minutes, a 30% annotation cost saving at deployment. *Third*, DistilBERT's calibration degrades catastrophically under uncertainty-based strategies (ECE = 0.145 for Entropy), while RoBERTa remains stable (0.079). BERT-base (ECE = 0.118) and DeBERTa-v3 (mean F1 = 0.471) are excluded.

5.4 Ablation and Document Length Analysis

Table 6 quantifies each component's contribution.

Removing the **diversity filter** is the most harmful component-level removal (ΔAUC -0.006), with the sharpest penalty on IMDb where redundant long reviews dominate. High α=10.0 is the most severe overall degradation (-0.031): over-smoothing collapses the denominator to pure entropy. Extended ablation plots are in Appendix C.

Criterion	RoBERTa	DistilBERT
ECE (CAL-Log)†	**0.077**	0.084
AUC-F1-Cost	0.763	**0.774**
Cost to F1 = 0.80	**48.0** min	68.7 min
F1 @ 30 min	0.687	**0.776**
F1 @ 60 min	0.760	**0.801**
F1 @ 90 min	**0.815**	0.803
Final F1	**0.830**	0.821
ECE under Entropy	**0.079**	0.145

Table 5: RoBERTa vs. DistilBERT under CAL-Log. †Design threshold = 0.08. DistilBERT exceeds the ECE threshold and collapses catastrophically under Entropy (0.145 vs. 0.079), invalidating entropy-based acquisition despite competitive AUC.

Variant	AUC	C_{80}	ΔAUC
CAL-Log (full)	0.7507	32.9	-
CAL-Linear	0.7473	37.4	−0.003
No diversity filter	0.7448	38.9	−0.006
No temperature scaling	0.7490	32.8	−0.002
No fatigue model	0.7550	30.2	+0.004
α=10.0 (over-smooth)	0.7199	36.7	−0.031
α=2.0 (low base)	0.7469	44.9	−0.004
β=6.0 (high slope)	0.7566	36.2	+0.006
β=1.0 (low slope)	0.7486	37.9	−0.002

Table 6: Ablation study (4 datasets: AG News, Amazon, Rotten Tomatoes, IMDb). α=10.0 is the most damaging variant (ΔAUC −0.031).

Interpretation of α and β. α captures *fixed task-switching overhead*: a large α (e.g., α>8) indicates an annotator who pauses substantially between items regardless of length, reflecting careful orientation before labeling. β captures *marginal reading cost per unit of log-length*: high β (β>5) indicates near-linear reading behavior scaling steeply with length, while low β (β<1.5) reflects aggressive skimming. High α=10.0 collapses the cost denominator toward a constant, reducing CAL-Log to plain entropy sampling; high β=6.0 improves performance on long-document datasets by penalizing verbose instances more.

Document Length Analysis. CAL-Log's efficiency scales *monotonically* with length variance (Figure 4). On **long** datasets (IMDb, Yelp, Amazon; 120-230 w), the mean speedup is **3.56**× (Yelp: 5.52×, Amazon: 2.87×, IMDb: 2.31×). On **medium** datasets (AG News; ∼45 w), the gain drops to 1.16×. This monotonic scaling (Medium 1.16× → Short 1.94× → Multi 2.26× → Long 3.56×) directly validates the logarithmic cost hypothesis.

5.5 Statistical Significance and Effect Sizes

Wilcoxon Signed-Rank tests confirm: Random (p=0.031, d=−1.21, 59.2% faster); BADGE (p=0.118, d=−0.77, 69.8% faster); Entropy (p=0.160, d=−0.67, 74.2% faster).

Non-significance against BADGE and stronger baselines is a **mathematical constraint**: with N=6 paired observations (four datasets unsolvable by all strategies), the minimum achievable two-tailed Wilcoxon p-value is 0.031. Following Wasserstein and Lazar (2016), we emphasize practical effect sizes over p-value thresholds: all comparisons show $|d|$>0.67 with perfectly consistent direction across solvable datasets.

Post-Hoc Power Analysis. Achieving α=0.05 with 80% power for the BADGE comparison (d=−0.77) requires $N\geq$15 datasets - 2.5× what is available - motivating effect-size reporting (Cohen's d) as the primary significance measure when dataset counts are inherently limited.

5.6 Preliminary User Evaluation

A live annotation tool was deployed for structured user evaluation. Evaluators (N=42) completed IMDb sentiment annotation sessions using the full CAL-Log pipeline and Spy Window interface, then completed a structured feedback instrument: *(1)* self-reported reading style; *(2)* agreement with system's automatic profile classification; and *(3)* five Likert-scale items (1-5) assessing demonstration effectiveness, mathematical transparency, visible adaptation, Spy Window utility, and overall simulation believability.

Findings. 30 of 42 evaluators (71%) reported that the system's automatic reading profile matched or partially matched their self-reported style. Mean Likert ratings ranged from 3.8 (overall believability) to 4.3 (mathematical transparency), with Spy Window utility at 4.2, providing *preliminary evidence* that the online cost model and transparency interface perform as intended. **Limitation**: This is an uncontrolled convenience sample (N=42); controlled experiments with a pre-registered protocol are planned.

6 Conclusion

CAL-Log demonstrates that simple, interpretable cost-aware selection can outperform complex gradient-based methods. By modeling annotation cost with $C(x)=\alpha+\beta\log(1+L(x))$, grounded in

information foraging theory and extending the logarithmic cost principle established by Supasan and Athuraliya (2026), CAL-Log avoids cost-greedy collapse without gradient calculations or clustering. Experiments on ten benchmarks show **3.3× speedup over BADGE** and **3.9× over Entropy** to reach F1 = 0.80, with large effect sizes ($|d|>0.67$) that scale monotonically with document length variance. A preliminary live evaluation ($N=42$) suggests both the accuracy of reading-profile detection and the effectiveness of the transparency interface.

Future work includes extension to sequence labeling and span annotation, crowdsourced annotator validation at scale, and richer cost proxies incorporating syntactic complexity and domain difficulty.

Limitations

Offline Simulation Reliance. Primary results use a linear-time cost proxy, standard in AL research (Settles, 2009), but this cannot fully capture fatigue, distraction, or task difficulty beyond document length. **Cold-Start Latency.** The sliding-window OLS requires 30–50 annotations to converge, limiting suitability for micro-tasking crowd work. **Modality and Task Scope.** The logarithmic cost model is optimized for text classification and may not generalize to NLI tasks or fine-grained annotation (NER, relation extraction). **Threats to Validity.** The cost simulation and acquisition both assume document length as the primary cost driver, all benchmarks are English-only, and simulated annotation time cannot capture difficulty beyond length (e.g., sarcasm in short tweets vs. straightforward sentiment in long reviews).

Backbone Generalizability. All primary experiments use RoBERTa-base as the backbone, selected because it is the only tested architecture that maintains ECE < 0.08 across all acquisition strategies (§5.3). The cost-aware ranking mechanism itself is model-agnostic: any backbone that produces well-calibrated posterior probabilities can be substituted without modifying the acquisition function. However, we have not yet verified that larger models (e.g., DeBERTa-v3-large, encoder–decoder architectures) or parameter-efficient fine-tuning methods (e.g., LoRA) maintain sufficient calibration for entropy-based scoring. Testing additional backbones is a priority for future work.

Cost-Aware Baselines. Our cost-aware comparisons are limited to CAL-Linear, the linear-cost variant of our own framework. While alternative cost-aware methods exist (Pandey et al., 2022; Liu et al., 2025), they rely on cost signals unavailable at selection time in our setting (syntactic parse features, oracle difficulty scores) or target different annotation tasks (sequence labeling), precluding direct comparison without substantial reimplementation under different assumptions. Future work should benchmark against reimplementations of alternative cost-aware acquisition functions under matched conditions.

Ethics Statement

CAL-Log collects annotation timing data to personalize the cost model. Timing data should be anonymized in production and used solely to improve the annotation experience, not to increase surveillance or impose speed targets. The system adapts to the annotator's natural pace; organizations should monitor annotation quality concurrently to ensure efficiency gains do not compromise label reliability or annotator well-being.

Acknowledgments

We thank the anonymous reviewers for their constructive feedback, which substantially improved the clarity of the methodology and experimental presentation. We also thank the 42 evaluators who participated in the live annotation study. Generative AI tools (ChatGPT, Claude) were used solely for language refinement and clarity of the manuscript text; they did not contribute to the core scientific methodology, experimental design, implementation, or analysis.

References

Jordan T. Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. 2020. Deep batch active learning by diverse, uncertain gradient lower bounds. In *International Conference on Learning Representations*.

Jason Baldridge and Alexis Palmer. 2009. How well does active learning actually work? Time-based evaluation of cost-reduction strategies for language documentation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 296–305, Singapore. Association for Computational Linguistics.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa-Anke, and Leonardo Neves. 2020. TweetEval: Unified benchmark and comparative evaluation

for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650. Association for Computational Linguistics.

Stuart K. Card, Thomas P. Moran, and Allen Newell. 1983. *The Psychology of Human-Computer Interaction*. CRC Press, Boca Raton.

Chaeyeon Chung, Jungsoo Lee, Kyungmin Park, Junsoo Lee, Minjae Kim, Mookyung Song, Yeonwoo Kim, Jaegul Choo, and Sungsoo Ray Hong. 2021. Understanding human-side impact of sampling image batches in subjective attribute labeling. In *Proceedings of the ACM on Human-Computer Interaction*, volume 5, pages 1–26.

Mohamed Elgaar and Hadi Amiri. 2023. HuCurl: Human-induced curriculum discovery. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1862–1877, Toronto, Canada. Association for Computational Linguistics.

Cornelia Gruber, Helen Alber, Bernd Bischl, Göran Kauermann, Barbara Plank, and Matthias Aßenmacher. 2025. Revisiting active learning under (human) label variation. In *Proceedings of the 4th Workshop on Perspectivist Approaches to NLP*, pages 75–86, Suzhou, China. Association for Computational Linguistics.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR.

Robbie Haertel, Eric Ringger, Kevin Seppi, James Carroll, and Peter McClanahan. 2008. Assessing the costs of sampling methods in active learning for annotation. In *Proceedings of ACL-08: HLT, Short Papers*, pages 65–68, Columbus, Ohio. Association for Computational Linguistics.

Md Abdullah Al Imran, Farnad Nasirzadeh, and Chandan Karmakar. 2024. Designing a practical fatigue detection system: A review on recent developments and challenges. *Journal of Safety Research*, 90:389–407.

Nataliia Kholodna, Sahib Julka, Mohammad Khodadadi, Muhammed Nurullah Gumus, and Michael Granitzer. 2024. LLMs in the loop: Leveraging large language model annotations for active learning in low-resource languages. In *Machine Learning and Knowledge Discovery in Databases. Applied Data Science Track: European Conference, ECML PKDD 2024, Proceedings, Part X*, pages 397–412, Vilnius, Lithuania. Springer.

Ken Lang. 1995. NewsWeeder: Learning to filter netnews. *Machine Learning*, 27:331–339.

Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Christian Bizer. 2015. DBpedia – a large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web*, 6(2):167–195.

Minzhi Li, Taiwei Shi, Caleb Ziems, Min-Yen Kan, Nancy Chen, Zhengyuan Liu, and Diyi Yang. 2023. 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*, pages 1487–1505, Singapore. Association for Computational Linguistics.

Shang Liu, Zhongze Cai, Hanzhao Wang, Zhongyao Ma, and Xiaocheng Li. 2025. Incentivizing high-quality human annotations with golden questions. *ArXiv:2505.19134*.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Stephen Monsell. 2003. Task switching. *Trends in Cognitive Sciences*, 7(3):134–140.

Diana Mortagua. 2025. Improving annotator selection in active learning using a mood and fatigue-aware recommender system. *Computing Research Repository. ArXiv:2507.23756*.

Eduardo Mosqueira-Rey, Elena Hernández-Pereira, David Alonso-Ríos, José Bobes-Bascarán, and Ángel Fernández-Leal. 2023. Human-in-the-loop machine learning: A state of the art. *Artificial Intelligence Review*, 56(4):3005–3054.

Lynnette Hui Xian Ng, Kokil Jaidka, Kaiyuan Tay, Hansin Ahuja, and Niyati Chhaya. 2025. Improving user behavior prediction: Leveraging annotator metadata in supervised machine learning models. *ArXiv:2503.21000*.

Rahul Pandey, Hemant Purohit, Carlos Castillo, and Valerie L. Shalin. 2022. Modeling and mitigating human annotation errors to design efficient stream processing systems with human-in-the-loop machine learning. *International Journal of Human-Computer Studies*.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 115–124. Association for Computational Linguistics.

Peter Pirolli and Stuart Card. 1999. Information foraging. *Psychological Review*, 106(4):643–675.

Keith Rayner. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124(3):372–422.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Eric Ringger, Marc Carmen, Robbie Haertel, Kevin Seppi, Deryle Lonsdale, Peter McClanahan, James Carroll, and Noel Ellison. 2008. Assessing the costs of machine-assisted corpus annotation through a user study. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Guy Rotman and Roi Reichart. 2022. Multi-task active learning for pre-trained transformer-based models. *Transactions of the Association for Computational Linguistics*, 10:1209–1228.

Rohan Saha, Abrar Fahim, Alona Fyshe, and Alex Murphy. 2024. Exploring curriculum learning for vision-language tasks: A study on small-scale multimodal training. In *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 65–81, Miami, FL, USA. Association for Computational Linguistics.

Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. 2018. CARER: Contextualized affect representations for emotion recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3687–3697, Brussels, Belgium. Association for Computational Linguistics.

Ozan Sener and Silvio Savarese. 2018. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*.

Burr Settles. 2009. Active learning literature survey. Technical Report 1648, University of Wisconsin–Madison Department of Computer Sciences.

Burr Settles and Mark Craven. 2008. An analysis of active learning strategies for sequence labeling tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 1070–1079, Honolulu, Hawaii. Association for Computational Linguistics.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Ng. 2008. Cheap and fast—but is it good? Evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 254–263, Honolulu, Hawaii. Association for Computational Linguistics.

Vihanga Supasan and Banuka Athuraliya. 2026. ALX0: Cost-aware active learning for cloud-scale NLP via zero-shot proxy valuation. In *Proceedings of the International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*.

Katrin Tomanek and Udo Hahn. 2010. A comparison of models for cost-sensitive active learning. In *Coling 2010: Posters*, pages 1247–1255, Beijing, China. Coling 2010 Organizing Committee.

Michiel van der Meer, Neele Falk, Pradeep K. Murukannaiah, and Enrico Liscio. 2024. Annotator-centric active learning for subjective NLP tasks. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18537–18555, Miami, Florida, USA. Association for Computational Linguistics.

Yifan Wang, David Stevens, Pranay Shah, Wenwen Jiang, Miao Liu, Xu Chen, Robert Kuo, Na Li, Boying Gong, Daniel Lee, Jiabo Hu, Ning Zhang, and Bob Kamma. 2024. Model-in-the-loop (MILO): Accelerating multimodal AI data annotation with LLMs. *ArXiv:2409.10702*.

Ronald L. Wasserstein and Nicole A. Lazar. 2016. The ASA statement on p-values: Context, process, and purpose. *The American Statistician*, 70(2):129–133.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, volume 28.

Appendices

A Algorithm Pseudocode

Algorithm 1 CAL-Log Active Learning Loop

Require: Unlabeled pool $\mathcal{U}$, labeled set $\mathcal{L}$, backbone M, budget B (minutes)
Ensure: Augmented labeled set $\mathcal{L}'$

Initialise: $\alpha \leftarrow 5.0;\ \beta \leftarrow 3.0;\ \texttt{buf} \leftarrow [\,]$
while $\sum_{i=1}^{t} C_i < B$ **do**
  // Step 1: Reset-and-Retrain
  $M \leftarrow \textsc{FineTune}\big(\texttt{roberta-base},\ \mathcal{L}\big)$
  $T_s \leftarrow \textsc{CalibrateTemp}\big(M,\ \mathcal{L}_{\text{val}}\big)$ {L-BFGS temperature scaling}
  // Step 2: Calibrated Uncertainty
  for all $x \in \mathcal{U}$ **do**
    $P(y \mid x) \leftarrow \text{Softmax}\big(M(x) \,/\, T_s\big)$
    $H(x) \leftarrow -\sum_{k=1}^{K} P(y_k \mid x)\ \log\ P(y_k \mid x)$
  end for
  // Step 3: Logarithmic Cost Prediction
  for all $x \in \mathcal{U}$ **do**
    $\hat{C}(x) \leftarrow \alpha + \beta\ \log\big(1 + L(x)\big)$
  end for
  // Step 4: Cost-Normalised Scoring
  for all $x \in \mathcal{U}$ **do**
    if $\max_{x_j \in \mathcal{L}} \cos(\mathbf{e}_x,\ \mathbf{e}_{x_j}) \le 0.95$ **then**
      $U(x) \leftarrow H(x)\ /\ \hat{C}(x)$
    else
      $U(x) \leftarrow 0$ {Redundancy penalty}
    end if
  end for
  // Step 5: Select, Annotate, Update
  $x^* \leftarrow \arg\max_{x \in \mathcal{U}}\ U(x)$
  $y^* \leftarrow \textsc{Annotator}(x^*)$
  $T_{\text{obs}} \leftarrow \textsc{MeasureTime}(x^*)$
  $\texttt{buf}.\text{append}\big(L(x^*),\ T_{\text{obs}}\big);$ retain last 5 entries
  if $|\texttt{buf}| \ge 5$ **then**
    if $\sigma_{\log L}(\texttt{buf}) \ge 0.3$ **then**
      $(\alpha,\ \beta) \leftarrow \text{OLS}(\texttt{buf})$ {Sliding-window least squares}
    else
      $\hat{\alpha} \leftarrow \min(0.2\,\bar{T},\ 5.0); \quad \hat{\beta} \leftarrow \big(\bar{T} - \hat{\alpha}\big)\ /\ \overline{\log(1+L)}$ {Low-variance fallback}
    end if
    Clamp: $\alpha \in [1,\ 15],\ \beta \in [0.1,\ 15]$
  end if
  $\mathcal{L} \leftarrow \mathcal{L} \cup \{(x^*,\ y^*)\}; \quad \mathcal{U} \leftarrow \mathcal{U} \setminus \{x^*\}$
end while
return $\mathcal{L}$

B Detailed Empirical Results

Dataset	Random	Entropy	LeastConf	Margin	CoreSet	BADGE	CAL-Lin.	CAL-Log
AG News	34.7	39.3	39.6	48.6	42.4	43.4	33.2	**29.9**
Amazon Polarity	83.3	69.9	74.3	64.4	85.3	81.6	**25.5**	29.0
DBpedia 14	71.4	72.4	81.7	68.3	65.0	66.9	33.7	**31.5**
Rotten Tomatoes	40.5	47.7	36.4	52.8	44.5	53.1	24.6	**20.9**
IMDb	229.5	537.5	317.2	415.8	509.7	418.0	**58.2**	99.5
Yelp Polarity	102.8	124.4	86.0	76.2	98.3	96.0	**18.5**	18.6
Mean	93.7	148.5	105.9	121.0	140.9	126.5	**32.3**	38.3

Table 7: **Cost to reach F1 = 0.80** (minutes) per dataset. Four datasets (20 Newsgroups, Emotion, TweetEval, Yahoo Answers) were unsolvable within the 120-minute budget by *all* strategies and are excluded. Best result per dataset in **bold**.

Dataset	Random	Entropy	LeastConf	Margin	CoreSet	BADGE	CAL-Lin.	CAL-Log
AG News	0.054	0.059	0.070	0.053	0.044	0.033	0.048	0.065
Amazon Polarity	0.055	0.163	0.068	0.155	0.053	0.038	0.055	0.078
Emotion	0.175	0.207	0.192	0.196	0.193	0.189	0.124	0.130
IMDb	0.062	0.171	0.137	0.065	0.064	0.063	0.083	0.087
TweetEval	0.214	0.230	0.201	0.233	0.251	0.176	0.230	0.209

Table 8: **Expected Calibration Error (ECE)** for a representative subset of datasets. ECE > 0.15 on Emotion and TweetEval indicates that uncertainty sampling becomes unreliable without temperature scaling.

Dataset	Random	Entropy	LeastConf	Margin	CoreSet	BADGE	CAL-Lin.	CAL-Log
20 Newsgroups	0.013	0.012	0.005	0.007	0.019	0.003	0.202	**0.211**
AG News	0.712	0.697	0.702	0.695	0.697	0.700	**0.730**	0.722
Amazon Polarity	0.652	0.662	0.687	0.669	0.675	0.668	0.825	**0.831**
Emotion	0.357	0.364	0.334	0.333	0.290	0.315	**0.485**	0.272
DBpedia 14	0.579	0.525	0.502	0.538	0.493	0.567	0.787	**0.796**
Rotten Tomatoes	0.741	0.751	0.749	0.725	0.721	0.734	0.784	**0.794**
IMDb	0.286	0.286	0.312	0.328	0.270	0.326	0.647	**0.690**
TweetEval	0.521	0.476	0.451	0.493	0.501	**0.509**	0.455	0.494
Yahoo Answers	0.303	0.310	0.288	0.342	0.302	0.321	0.242	**0.432**
Yelp Polarity	0.524	0.467	0.545	0.598	0.566	0.525	0.829	**0.854**
Mean	0.469	0.455	0.457	0.473	0.453	0.467	0.599	**0.610**

Table 9: **AUC-F1-Cost** across all 10 datasets and 8 strategies. CAL-Log achieves the highest mean (0.610), 30.6% above BADGE (0.467). Best result per dataset in **bold**.

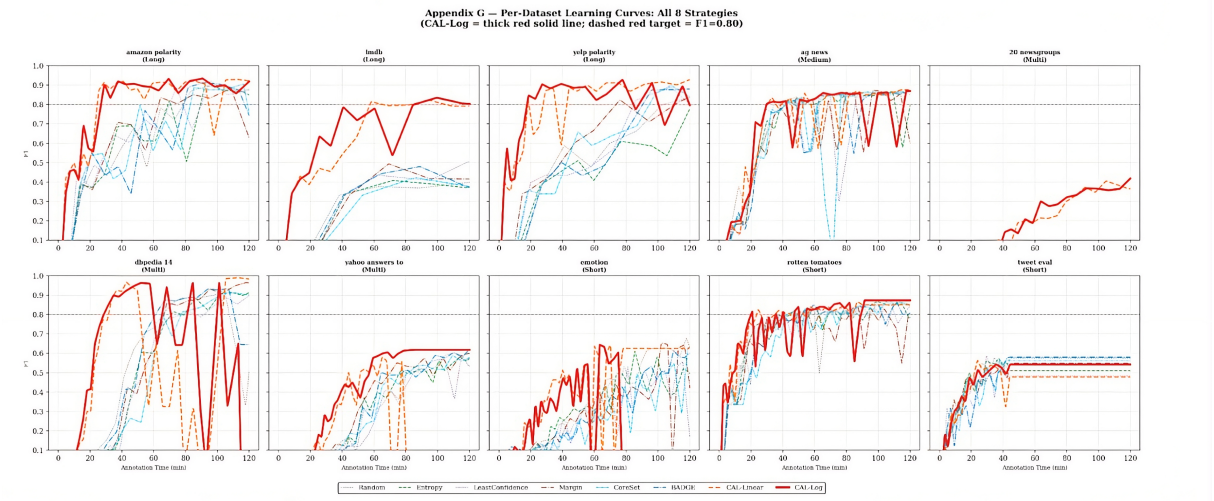

Figure 6: **Per-Dataset Learning Curves.** Macro F1 vs. annotation time across all 10 benchmarks (2×5 grid, ordered by increasing mean document length). CAL-Log (red) consistently reaches production-ready performance earlier on long-document datasets (IMDb, Yelp, Amazon), while on short-text datasets (Emotion, TweetEval) no strategy reaches F1 = 0.80 within the 120-minute budget.

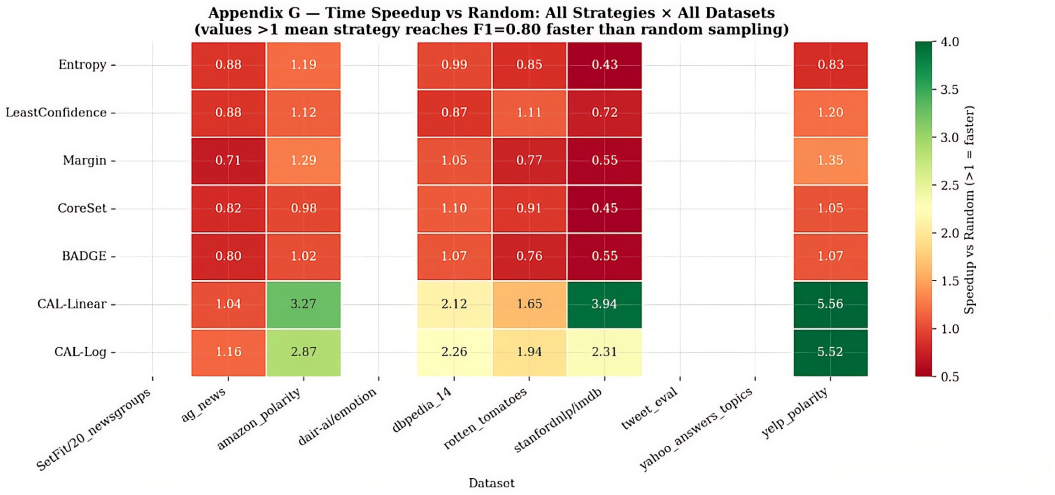

Figure 7: **Time Speedup vs. Random Sampling.** Heatmap showing convergence speed to F1 = 0.80 relative to Random baseline. Rows = strategies; columns = datasets (sorted by mean document length). Green = faster (>1.0×); gray = unsolvable within budget. CAL-Log and CAL-Linear achieve the largest speedups on long-document datasets (Yelp: 5.52×, IMDb: 2.31×).

C Ablation Study: Extended Visualizations

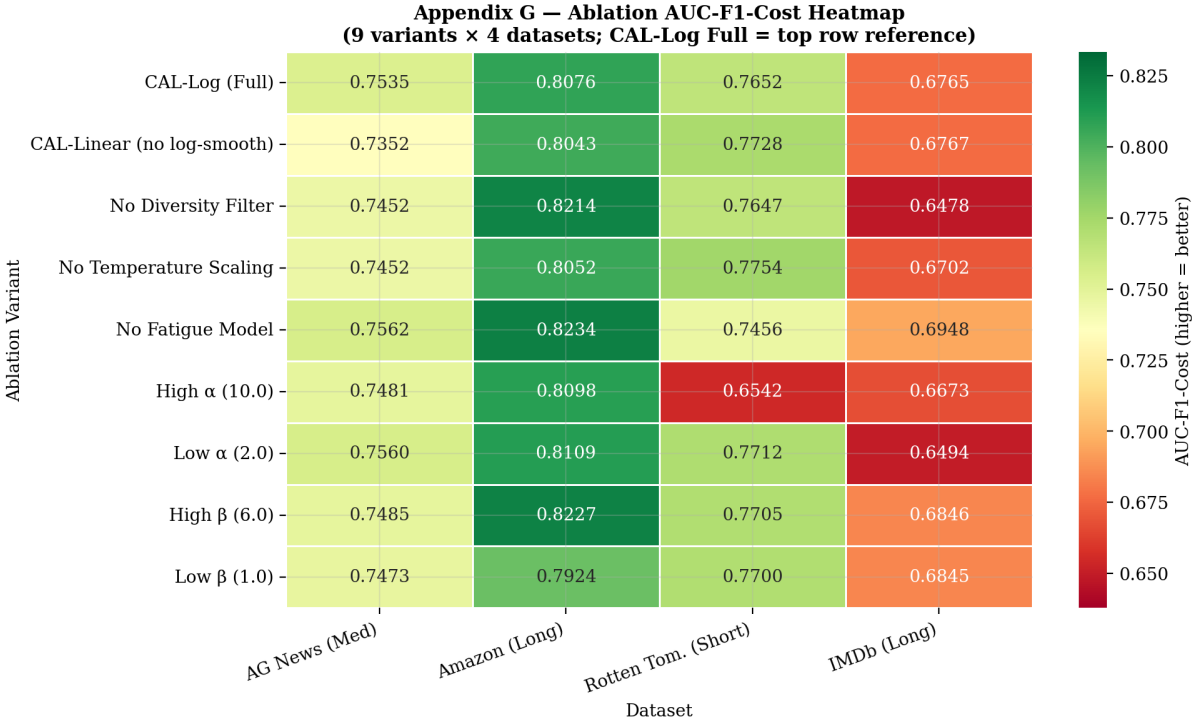

Figure 8: **Ablation Heatmap.** AUC-F1-Cost per ablation variant (rows) × dataset (columns). Cell annotations show ΔAUC from the full CAL-Log configuration. High α=10.0 (over-smoothing) and removing the diversity filter produce the most severe degradation across all datasets.

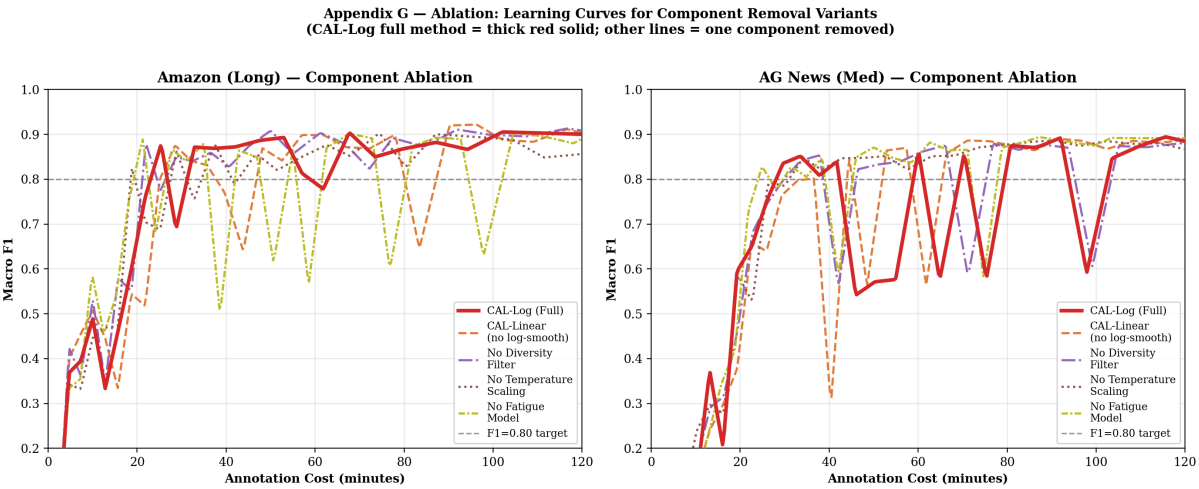

Figure 9: **Component Removal Learning Curves.** F1 vs. annotation time (0–120 min) comparing full CAL-Log (solid red) against four key ablation variants (dashed orange), with BADGE (solid blue) as reference. **(a)** No Diversity Filter: degradation concentrated on IMDb. **(b)** No Temperature Scaling: minimal impact. **(c)** High α=10.0: most severe degradation. **(d)** CAL-Linear: competitive early but diverges after 60 minutes.

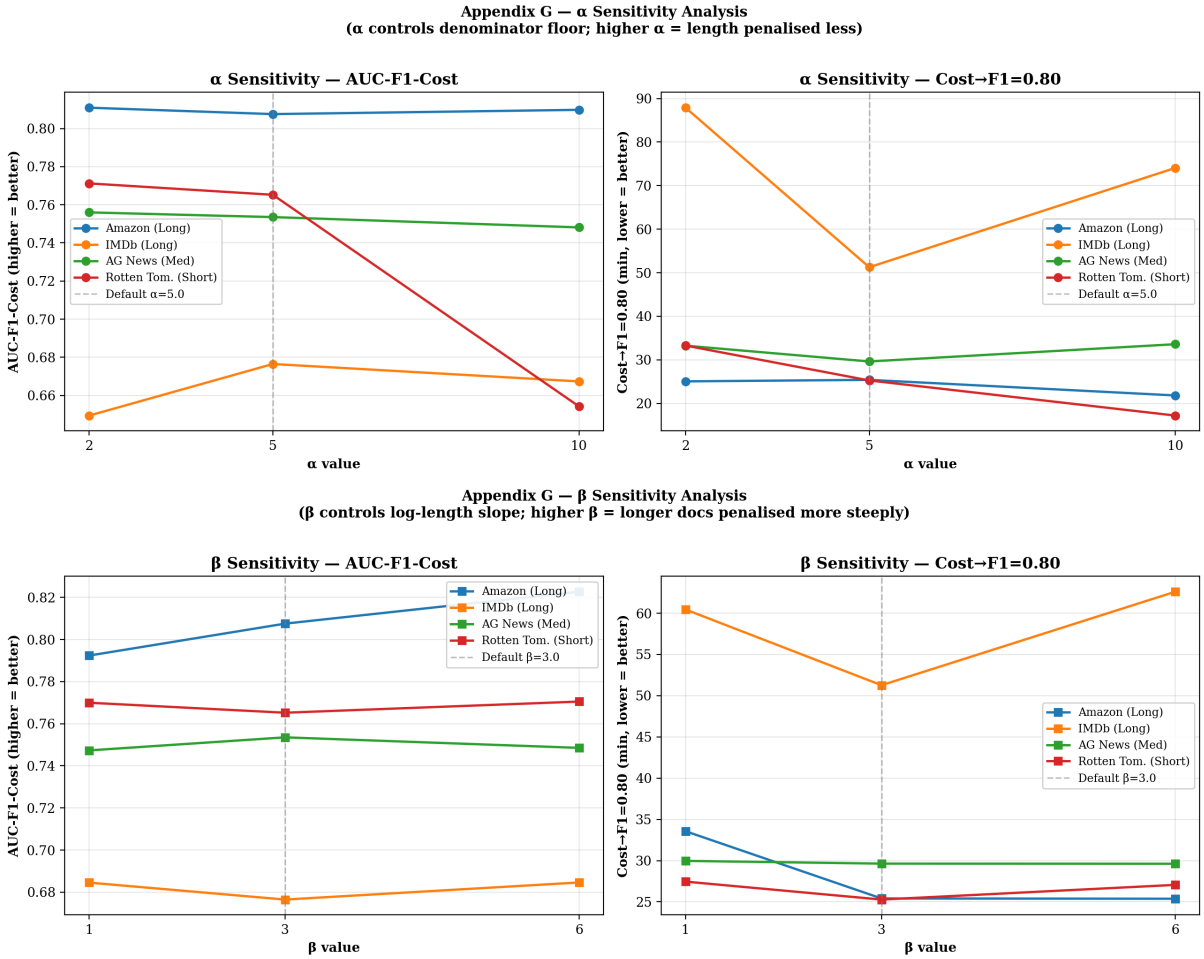

Figure 10: **Parameter Sensitivity: α and β. Left:** AUC-F1-Cost vs. α (2.0, 5.0, 10.0). Default α=5.0 provides optimal convergence; α=10.0 severely degrades performance. **Right:** AUC-F1-Cost vs. β (1.0, 3.0, 6.0). β=3.0 is robust across all length distributions. Thin lines = individual datasets; thick line = cross-dataset mean.

D Architecture Validation: Extended Tables

Strategy	RoBERTa	DistilBERT	BERT
CAL-Log	**0.077**	0.084†	0.118†
Entropy	**0.079**	0.145†	0.117†
BADGE	**0.057**	0.065	0.071

Table 10: **Expected Calibration Error (ECE) across backbones.** †Exceeds the design threshold of 0.08. RoBERTa is the only architecture that remains below threshold across *all* strategies. DistilBERT's ECE under Entropy (0.145) represents a 1.8× exceedance, invalidating its confidence outputs for entropy-based scoring in multi-class settings.

Strategy	RoBERTa	DistilBERT	BERT
CAL-Log	0.118	**0.058**	0.278
Entropy	**0.034**	0.274	0.083
Margin	**0.039**	0.218	0.280

Table 11: **F1 standard deviation across 3 random seeds** (lower = more reproducible). BERT exhibits extreme instability under cost-aware strategies (std = 0.278 for CAL-Log), driven by calibration failure on AG News varying by seed.

RoBERTa vs. (n=3)	p**-value**	**AUC diff**	**Cost saving**
DistilBERT	0.750	−0.011	20.7 min
BERT	0.750	+0.004	29.0 min
DeBERTa-v3	0.250	+0.314	never reached

Table 12: **Wilcoxon signed-rank tests: RoBERTa vs. each baseline** (n=3). Non-significance is a mathematical floor: minimum achievable p-value with n=3 is 0.25. Architecture selection rests on ECE threshold compliance and cost-to-target, not raw F1 significance.

Method	Complexity	Requires
Entropy	$O(n \cdot K)$	Forward pass
Margin	$O(n \cdot K)$	Forward pass
CoreSet	$O(n^2 \cdot d)$	Embeddings
BADGE	$O(n \cdot K \cdot d)$	Gradients
CAL-Log	$O(n \cdot K)$	Forward pass

Table 13: **Per-round computational complexity** (n=$|\mathcal{U}|$, K= classes, d= embedding dimension). CAL-Log adds only a scalar division to entropy computation. In deployment, CAL-Log re-ranks 10K+ instances in <1 s vs. BADGE's 8-12× longer gradient computation.

E Broader Impact

Cost-aware active learning has the potential to democratize NLP model development by reducing annotation budgets. This disproportionately benefits under-resourced research groups, non-profit organizations, and projects targeting low-resource languages where annotation expertise is scarce and expensive. By optimizing the allocation of human cognitive effort, CAL-Log can reduce the total hours required for dataset creation, potentially improving working conditions for professional annotators.

However, cost-aware systems also carry risks. Optimizing purely for speed may inadvertently pressure annotators to work faster, degrading label quality. We emphasise that CAL-Log is designed to *adapt to the annotator's natural pace*; it does not impose externally-defined speed targets. Organizations deploying cost-aware AL systems should:

- Monitor annotation quality metrics concurrently with efficiency gains.
- Ensure that cost-model outputs (e.g., reading-speed classification) are not used for performance evaluation of individual annotators.
- Verify that efficiency improvements do not come at the expense of annotator well-being or label reliability.

F Reproducibility Details

Backbone: RoBERTa-base, 3 epochs/round, lr = 2×10^{-5}, AdamW, temperature scaling via L-BFGS. **Cost model:** α_0=5.0, β_0=3.0, OLS on 5-event sliding window, $\alpha\in[1, 15]$, $\beta\in[0.1, 15]$. **Redundancy:** cosine similarity threshold 0.95. **Evaluation:** 3 random seeds per strategy per dataset; 120-minute simulated budget. Code will be released upon publication.