QUORUM: QUality-Optimized Routing Using Multiple annotators

Anonymous ACL submission

Abstract

Data annotation remains a central bottleneck in natural language processing, requiring human effort to obtain high-quality labels at scale. While Large Language Models (LLMs) offer a fast and cost-effective alternative, their reliability is highly instance-dependent: they perform well on simple inputs but often fail on examples requiring nuanced reasoning or contextual understanding. In this work, we address this challenge with **QUORUM** (**QU**ality-**O**ptimized **R**outing **U**sing **M**ultiple annotators), a budget-aware routing framework that dynamically assigns each instance to human or LLM annotators under a fixed annotation budget. Unlike prior approaches relying on model confidence or uncertainty estimates, QUORUM leverages feature-based signals to estimate instance difficulty and supports multiple annotations per instance, combining them through agreement-based rewards to improve reliability. We evaluate QUORUM across diverse closed- and open-ended annotation tasks in English and multilingual settings, and QUORUM improves annotation quality by up to 34.4% while reducing costs by 8.8% over competing methods. Code can be found at https://anonymous.4open.science/r/QUORUM-D63F.

1 Introduction

Human annotation is costly and difficult to scale across domains and languages[1]. LLMs offer a faster and cheaper alternative, although annotation quality can vary substantially (Shimizu et al., 2025; Tan et al., 2024). In practice, some instances can be reliably annotated by LLMs, whereas others require nuanced reasoning, domain expertise, or cultural understanding (Hendrycks et al., 2021). The challenge is therefore no longer whether humans or LLMs should annotate data, but *how to allocate annotation effort between them under limited budgets*. A natural solution is to dynamically route instances between annotators. Existing approaches largely rely on uncertainty estimation, using model confidence to decide when human supervision is required (Li et al., 2023; Gligorić et al., 2025). While effective, these methods often depend on calibrated confidence scores, additional inference steps, or model-specific uncertainty signals, increasing computational overhead and limiting applicability across annotators (Cecere et al., 2025).

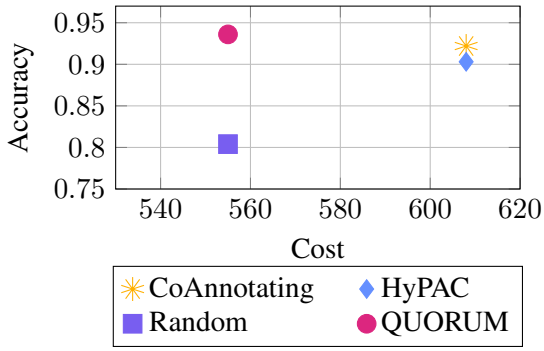

Figure 1: Pareto frontier of cost versus performance on AG's News dataset with 5,320 human annotations. The standalone LLM achieves 0.834 average accuracy. Costs denote cumulative annotation expenses.

In this work, we propose **QUORUM** (**QU**ality-**O**ptimized **R**outing **U**sing **M**ultiple annotators), a budget-aware routing framework that dynamically assigns instances to human or LLM annotators to jointly optimize annotation quality and cost. Unlike prior approaches, QUORUM does not rely on uncertainty estimation. Instead, it learns routing policies from contextual and budget signals, deciding not only *who* should annotate an instance, but also *whether additional annotations are worth their cost*. This naturally enables multi-annotator strategies, where multiple LLMs or humans can be combined when agreement is insufficient. In fact, the name QUORUM is inspired by collective decision-making, where a quorum represents the minimum agreement required to reach a reliable

[1] sourcebae.com/blog/what-is-data-annotation

outcome [2]. Similarly, QUORUM aggregates annotations from humans and LLMs, dynamically requesting additional supervision only when needed.

Figure 1 previews this trade-off on the AG's News dataset: QUORUM achieves the highest annotation quality while matching the cost of the cheapest competing methods, placing it on the Pareto frontier. Our main contributions are:

- We propose QUORUM, a budget-aware routing framework for allocating annotations between human and LLM annotators, supporting multi-annotator aggregation when needed.
- We provide a theoretically grounded formulation of budget-aware routing, with guarantees on posterior concentration and asymptotic convergence toward oracle routing.
- QUORUM achieves up to 34.4% relative improvement in annotation quality while reducing cost by up to 8.8% across English and multilingual open- and closed-ended tasks.

2 Related Work

Recent work on human–LLM collaboration for annotation has increasingly focused on balancing annotation quality and cost through adaptive routing strategies (Tan et al., 2024). A common starting point is the use of uncertainty or confidence signals to decide whether a sample should be handled by a model or escalated to a human annotator.

Uncertainty-based routing. Early approaches, such as those proposed by Wang et al. (2021) and extended by Wang et al. (2024), rely on model-provided confidence scores (*e.g.*, logits from LLM APIs) to determine whether an LLM prediction is sufficiently reliable. Samples with low confidence are routed to human annotators. While intuitive, this strategy has two main limitations: (i) it depends on the availability of calibrated confidence scores, which are not consistently provided across models, and (ii) it requires an additional model inference step before routing decisions can be made, introducing non-negligible computational overhead. To overcome these limitations, subsequent work has proposed alternative uncertainty estimation mechanisms. Li et al. (2023) introduce a human–LLM co-annotation framework that models uncertainty via self-evaluated confidence scores and entropy over multiple model outputs, formulating the problem as a multi-objective optimization task over annotation quality and cost. Similarly, Gligorić et al. (2025) propose Confidence Driven Inference (CDI), where models explicitly verbalize confidence estimates that are then used to decide when human supervision is required. Finally, Zeng et al. (2026) propose HyPAC, a framework with PAC-style guarantees that partitions the input space using uncertainty and applies importance sampling with upper confidence bounds to optimize annotation allocation under budget constraints.

Alternative routing strategies. Beyond confidence-based methods, other approaches incorporate external knowledge or learned error signals to improve routing decisions. Huang et al. (2024b) propose ARAIDA, which augments model predictions with retrieval from nearest labeled neighbors and adaptively weights model and retrieval signals based on estimated error. In a related direction, Huang et al. (2024a) introduce SANT, which integrates active learning with error-aware triage mechanisms to prioritize samples for human annotation under limited budgets. More recent work explores learning-based and theoretically grounded routing policies. Miranda et al. (2025) learn a performance prediction model that selects the optimal combination of human and LLM annotations based on a preference dataset, effectively learning routing decisions from data.

Overall, QUORUM avoids dependence on model-derived uncertainty and additional inference steps, relying on feature-based difficulty signals to enable a more general and cost-efficient routing strategy under budget constraints.

3 Method

We consider a dataset $\mathcal{D}$ of T samples:

$$\mathcal{D} = \{d_t\}_{t=1}^{T} \quad d_t \in \mathbb{R}^d$$

At each timestep t, the router selects annotators from an arm set, where an arm denotes an available action whose selection yields a reward and an associated cost (Lattimore and Szepesvári, 2020):

$$\mathcal{A} = \{1, \ldots, K, H\}$$

Where K denotes the number of LLM annotators and H denotes the human annotator. Each arm $a \in \mathcal{A}$ is associated with a cost $c_a > 0$. The objective is to maximize annotation quality under a budget constraint B. Figure 2 provides an overview of

[2] dictionary.cambridge.org/dictionary/english/quorum

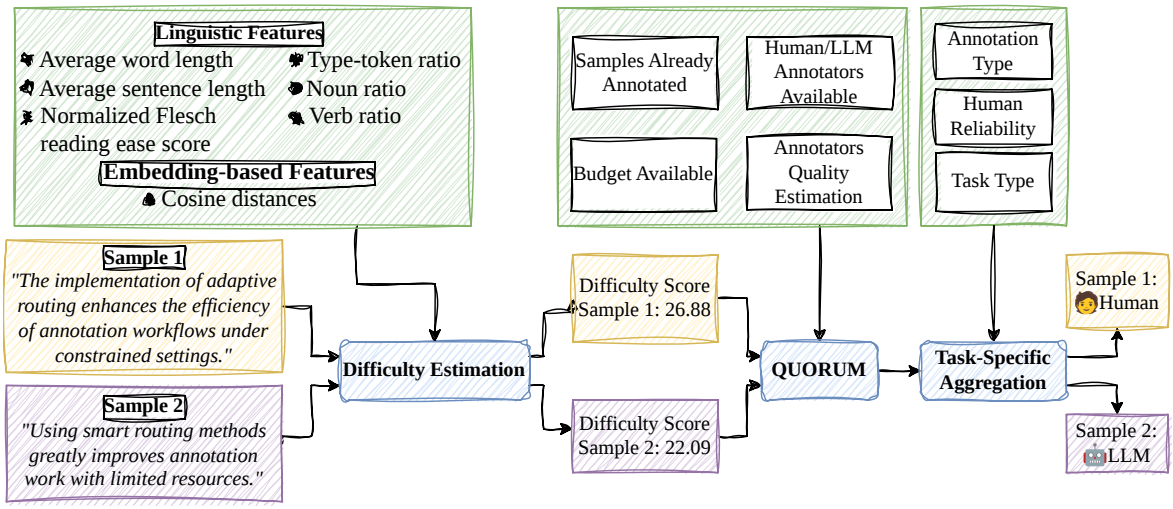

Figure 2: QUORUM, a budget-aware annotation router estimating instance difficulty using linguistic and semantic features, assigning samples to human or LLM annotators to maximize annotation quality while minimizing cost.

the proposed QUORUM framework, including difficulty estimation, adaptive routing, budget-aware adaptation, and task-specific aggregation.

3.1 Feature Definition

Our approach relies on feature-based signals to estimate the complexity of each input instance. These features serve as a proxy for instance difficulty and replace the use of model-derived uncertainty employed in prior work (Zeng et al., 2026).

For each sample d_t, we construct a feature representation $\phi(d_t) \in \mathbb{R}^m$ based on two components.

Linguistic features We extract a set of scalar features designed to capture lexical and syntactic complexity (Ziller et al., 2026; Yang et al., 2019). These include: (i) average word length, (ii) type-token ratio, (iii) average sentence length, (iv) normalized Flesch reading ease score (Flesch, 1948), (v) noun ratio, and (vi) verb ratio. Together, these features characterize properties such as vocabulary sophistication, structural complexity, and information density, which are known to correlate with annotation difficulty.

Embedding-based features To capture semantic properties, each instance is encoded into a dense vector representation using a pretrained text multilingual encoder[3]. We then compute the cosine distances (1− cosine similarities) between d_t and its k nearest neighbors in the embedding space. The resulting distance vector reflects the degree of semantic isolation of the instance, with higher distances indicating atypical or out-of-distribution samples potentially more difficult to annotate (Hassan and Alikhani, 2023).

The final feature vector $\phi(d_t)$ is obtained by concatenating the linguistic and embedding-based features and applying normalization. This representation is used as input to the routing policy. Additional feature details are reported in Appendix A.1.

Context Construction For each sample, the router first builds a contextual representation:

$$x_t = [\phi(d_t), b_t, p_t, m_t]$$

Where $\phi(d_t)$ denotes normalized input features, b_t is the remaining budget fraction, $p_t = t/T$ represents progress through the dataset, and m_t is a binary mask indicating which LLMs have already been queried for that sample. Consequently, each routing decision depends not only on the sample itself, but also on previously consumed resources and earlier annotations. This contextual vector becomes the input to all subsequent estimation stages.

3.2 Human-LLM Routing with Calibration

Feature-Space Uncertainty Before routing begins, the model allocates approximately 20% of the affordable human budget to a calibration stage. This step mitigates uncertainty in early routing decisions by establishing initial estimates of LLM reliability. The objective is to estimate the reliability of each LLM before online decisions are made. Calibration samples are selected according to feature-space uncertainty:

$$u_t = \sqrt{(\phi(d_t) - \mu)^T \Sigma^{-1} (\phi(d_t) - \mu)}$$

[3] huggingface.co/NovaSearch/stella_en_1.5B_v5

where μ and Σ are statistics over contextual features. Samples with highest uncertainty receive human labels. Human supervision is then used to estimate correctness of each LLM prediction:

$$r_{a,t} = \mathbb{1}(\hat{y}_t^{(a)} = y_t)$$

The resulting calibration rewards initialize annotator-specific posteriors.

Annotator Quality Estimation Each arm maintains Gaussian posterior parameters:

$$\theta_a \sim \mathcal{N}(\mu_a, \Sigma_a)$$

Updated using Bayesian linear regression:

$$\Sigma_a' = \left(\Sigma_a^{-1} + \frac{xx^T}{\sigma^2}\right)^{-1}$$

$$\mu_a' = \Sigma_a' \left(\Sigma_a^{-1}\mu_a + \frac{r}{\sigma^2}x\right)$$

Posterior means encode expected annotator quality, whereas posterior covariance captures uncertainty. Human annotators are initialized with optimistic priors (higher mean and lower variance), reflecting greater expected reliability. This follows an assumption in annotation research that human labels generally provide stronger supervision signals than automated annotators (Gligorić et al., 2025).

The learned posterior parameters subsequently determine routing choices.

Routing Mechanism Given context x_t, expected quality for each available LLM is estimated as $q_a = \mu_a^T x_t$. To avoid premature exploitation, routing includes decaying exploration:

$$\epsilon_t = \frac{\epsilon_0}{1 + \gamma t}$$

Where ϵ_0 denotes the initial exploration probability and γ controls the rate at which exploration decreases over time. Therefore, early decisions emphasize exploration, whereas later stages increasingly exploit calibrated annotator estimates. At each timestep, with probability ϵ_t, the router explores by randomly selecting an available LLM:

$$a_t \sim \text{Uniform}(\mathcal{A}_{\text{LLM}})$$

Otherwise, routing exploits current quality estimates and selects the highest-scoring annotator:

$$a_t = \arg\max_a q_a$$

When predicted quality falls below a threshold τ_H, the router escalates to human annotation:

$$a_t = \begin{cases} H, & q_{a^*} < \tau_H \\ a^*, & \text{otherwise} \end{cases}$$

3.3 Iterative Annotation Selection

Budget-Aware Adaptive Refinement Every routing decision updates consumed cost. Human budgets are partitioned into a working component and a reserved component $B = B_{\text{working}} + B_{\text{reserved}}$. The reserved part prevents exhaustion of expert annotations before difficult late-stage samples appear.

After initial routing, samples are ranked for possible additional annotations. Difficulty is estimated from posterior uncertainty:

$$D_t = \frac{1}{K}\sum_a x_t^T \Sigma_a x_t$$

We further define the annotation disagreement among collected labels:

$$\Delta_t = 1 - \frac{\max_y \text{count}(y)}{|\mathcal{Y}_t|}$$

where $\mathcal{Y}_t$ is the set of annotations for sample t. For each sample, a priority score combines difficulty, disagreement and missing human supervision:

$$P_t = D_t + \Delta_t + \mathbb{1}(\text{no human})$$

Samples with highest priority receive extra annotations until budgets are exhausted. Hence annotation becomes an iterative refinement process rather than a single routing decision.

Task-specific Aggregation The final stage aggregates collected annotations differently depending on task type. For classification, agreement is computed from label frequency with human weight $w_H = 10$ and LLM weight $w_{\text{LLM}} = 1$.

Consensus is achieved when weighted agreement exceeds a threshold. Final outputs are determined through weighted majority voting. For summarization, exact agreement is inappropriate. Instead, summaries are represented through embeddings and semantic agreement is measured via average cosine similarity:

$$r_t = 2\left(\frac{\sum_{i \neq j} \cos(e_i, e_j)}{|\mathcal{Y}_t|(|\mathcal{Y}_t| - 1)}\right) - 1$$

Where $e_i, e_j \in \mathbb{R}^d$ denote the embeddings. Consensus is reached when semantic similarity exceeds

a threshold. The final summary is selected as the annotation closest to the embedding centroid of all summaries. Therefore, in summarization we are rewarding, due to absence of labels, semantic consistency among annotators. The output of this aggregation stage constitutes the final prediction. The pseudocode of QUORUM is in Appendix A.4.

3.4 Theoretical Properties

Proposition 3.1 (Posterior concentration)**.** *Assuming bounded contextual representations $\|x_t\|_2 \leq L$ and bounded true annotator parameters $\|\theta_a^*\|_2 \leq S$, with Gaussian noise:*

$$r_{a,t} = x_t^\top \theta_a^* + \eta_t, \qquad \eta_t \sim \mathcal{N}(0, \sigma^2)$$

For any annotator a, with probability at least $1-\delta$:

$$\left| x_t^\top (\mu_{a,t} - \theta_a^*) \right| \leq \beta_t \sqrt{x_t^\top \Sigma_{a,t} x_t}$$

With:

$$\beta_t = \sigma \sqrt{d \log \left(\frac{1 + tL^2/\lambda}{\delta} \right)} + \sqrt{\lambda} S.$$

We have that:

$$x_t^\top \Sigma_{a,t} x_t \to 0$$

As $t \to \infty$, implying that uncertainty estimates become increasingly accurate over time.

Annotator quality estimates concentrate around the true parameters as observations accumulate. Therefore, the uncertainty-driven difficulty score used by QUORUM becomes progressively more reliable. A detailed proof is provided in Appendix A.2.

Proposition 3.2 (Vanishing average regret)**.** *Under the same assumptions of Proposition 3.1, let:*

$$a_t^* \in \arg\max_{a \in \mathcal{A}} x_t^\top \theta_a^*$$

be the context-dependent oracle annotator, and define:

$$R(T) = \sum_{t=1}^{T} \left(x_t^\top \theta_{a_t^*}^* - x_t^\top \theta_{\hat{a}_t}^* \right)$$

If the core routing rule uses decaying exploration $\epsilon_t = \epsilon_0/(1+\gamma t)$, the oracle annotator is uniformly separated from the second-best annotator, and exploration provides sufficiently diverse samples for each feasible annotator, then:

$$\lim_{T \to \infty} \frac{R(T)}{T} = 0$$

Implying asymptotic convergence toward oracle routing.

Proposition 3.2 implies that the average loss relative to an optimal routing policy vanishes over time. Therefore, QUORUM asymptotically approaches the performance of an oracle with knowledge of annotator quality. Proof is provided in Appendix A.3.

4 Experiments

We evaluate QUORUM across multiple tasks, budget regimes, and annotation settings. Additional implementation details, prompts, and hyperparameters are reported in Appendix B.

Baselines We compare against representative annotation routing baselines: ARAIDA (Huang et al., 2024b) (retrieval-augmented error-aware routing), CoAnnotating (Li et al., 2023) and CDI (Gligorić et al., 2025) (confidence-based human–LLM collaboration), HyPAC (Zeng et al., 2026) (PAC-guaranteed budget-aware routing) and SANT (Huang et al., 2024a) (active learning with selective triage). We also insert the Random baseline, while we avoid inserting Wang et al. (2021), which was improved by Li et al. (2023) and Miranda et al. (2025) due to reproducibility problems. Since SANT and ARAIDA rely on label-based confidence estimation, they cannot be extended to summarization and are omitted in those settings.

Datasets We evaluate our approach on 8 datasets spanning classification (AG's News, SST2, IMDB and PubMed), summarization (CNN and XLSum), and question-answering (Global-MMLU and MMLU-Redux) tasks, as shown in Table 5. We select AG's News, IMDB, SST2 and MMLU-Redux to compare with related work (Zeng et al., 2026). Global-MMLU and XLSum are included to assess performance in multilingual settings.

Experimental Settings We study two regimes:

- AUDITOR STYLE: routing under a fixed budget of human annotations, deciding between human and LLM supervision.
- DOLLARS: routing under monetary constraints with heterogeneous annotator costs, allowing selection among multiple annotators.

We use Claude Sonnet 4.5, Nova Pro v1.0, and Qwen3-32B as annotators. We assume heterogeneous annotation, reflecting the relative pricing

Table 1: Mean ± std in the AUDITOR STYLE setting. For each dataset, we report the number of samples, Accuracy or ROUGE-1 depending on the task (↑), Machine Cumulative Accuracy (↑), and total annotation cost (↓) for varying percentages of human-labeled data. **Bold** indicates the best-performing method, while underlined the second best. Value in parentheses represents the average performance obtained by LLM annotators on that dataset.

Dataset	Method	10%			30%			50%			70%		
		Accuracy/ROUGE	MCA	Cost	Accuracy/ROUGE	MCA	Cost	Accuracy/ROUGE	MCA	Cost	Accuracy/ROUGE	MCA	Cost
AG's News (LLM: 0.834) 7,600 samples	Random	0.729 ± 0.001	0.163 ± 0.012	144	0.805 ± 0.002	0.146 ± 0.005	281	0.794 ± 0.002	0.145 ± 0.004	418	0.804 ± 0.001	0.154 ± 0.002	555
	ARAIDA	0.787 ± 0.010	$\mathbf{0.255 \pm 0.050}$	152	0.834 ± 0.013	0.218 ± 0.012	304	0.861 ± 0.035	0.217 ± 0.014	456	0.886 ± 0.055	0.215 ± 0.010	608
	CDI	0.800 ± 0.001	0.229 ± 0.012	152	0.841 ± 0.001	0.215 ± 0.005	304	0.887 ± 0.003	0.222 ± 0.005	456	0.921 ± 0.002	0.220 ± 0.002	608
	CoAnnotating	0.800 ± 0.001	0.233 ± 0.001	152	0.844 ± 0.001	0.223 ± 0.001	304	0.889 ± 0.001	0.224 ± 0.001	456	0.922 ± 0.001	0.221 ± 0.001	608
	HyPAC	0.794 ± 0.007	0.220 ± 0.016	152	0.840 ± 0.013	$\mathbf{0.226 \pm 0.007}$	304	0.875 ± 0.033	0.225 ± 0.009	456	0.903 ± 0.052	0.224 ± 0.008	608
	SANT	0.798 ± 0.007	0.236 ± 0.026	152	0.838 ± 0.009	0.216 ± 0.005	304	0.867 ± 0.026	0.216 ± 0.006	456	0.896 ± 0.041	0.217 ± 0.006	608
	QUORUM	$\mathbf{0.800 \pm 0.001}$	0.224 ± 0.010	**144**	$\mathbf{0.845 \pm 0.002}$	0.225 ± 0.005	**281**	$\mathbf{0.890 \pm 0.001}$	$\mathbf{0.227 \pm 0.003}$	**418**	$\mathbf{0.936 \pm 0.001}$	$\mathbf{0.226 \pm 0.001}$	**555**
SST2 (LLM: 0.944) 870 samples	Random	0.902 ± 0.003	0.067 ± 0.027	16	0.921 ± 0.003	0.062 ± 0.009	32	0.934 ± 0.002	0.062 ± 0.004	48	0.943 ± 0.003	0.061 ± 0.004	64
	ARAIDA	0.939 ± 0.002	0.047 ± 0.020	17	0.946 ± 0.003	0.043 ± 0.005	35	0.948 ± 0.006	0.044 ± 0.007	52	0.949 ± 0.009	0.045 ± 0.008	70
	CDI	0.941 ± 0.003	0.055 ± 0.029	17	0.952 ± 0.003	0.054 ± 0.010	35	0.970 ± 0.004	0.069 ± 0.009	52	0.985 ± 0.002	0.070 ± 0.004	70
	CoAnnotating	0.948 ± 0.001	0.126 ± 0.001	17	0.959 ± 0.001	0.077 ± 0.001	35	0.964 ± 0.002	0.057 ± 0.005	52	0.978 ± 0.002	0.060 ± 0.003	70
	HyPAC	0.966 ± 0.007	0.173 ± 0.035	17	0.950 ± 0.003	0.048 ± 0.011	35	0.969 ± 0.003	$\mathbf{0.100 \pm 0.034}$	52	0.969 ± 0.003	0.088 ± 0.028	70
	SANT	0.940 ± 0.002	0.045 ± 0.018	17	0.951 ± 0.003	0.051 ± 0.008	35	0.963 ± 0.007	0.056 ± 0.007	52	0.975 ± 0.011	0.058 ± 0.008	70
	QUORUM	$\mathbf{0.968 \pm 0.001}$	$\mathbf{0.188 \pm 0.006}$	**16**	$\mathbf{0.969 \pm 0.003}$	$\mathbf{0.124 \pm 0.032}$	**32**	$\mathbf{0.971 \pm 0.001}$	0.071 ± 0.001	**48**	$\mathbf{0.990 \pm 0.001}$	$\mathbf{0.097 \pm 0.001}$	**64**
MMLU-Redux (LLM: 0.633) 3,000 samples	Random	0.706 ± 0.002	0.204 ± 0.021	57	0.691 ± 0.003	0.216 ± 0.010	111	0.767 ± 0.004	0.196 ± 0.007	165	0.901 ± 0.004	0.224 ± 0.006	219
	ARAIDA	0.749 ± 0.011	0.278 ± 0.018	60	0.794 ± 0.013	0.265 ± 0.008	120	0.811 ± 0.036	0.266 ± 0.006	180	0.826 ± 0.059	0.266 ± 0.006	240
	CoAnnotating	0.772 ± 0.001	0.430 ± 0.001	60	$\mathbf{0.836 \pm 0.001}$	$\mathbf{0.357 \pm 0.001}$	120	0.889 ± 0.001	0.321 ± 0.001	180	0.939 ± 0.001	0.300 ± 0.001	240
	HyPAC	0.764 ± 0.003	0.355 ± 0.026	60	0.832 ± 0.003	0.344 ± 0.009	120	0.884 ± 0.003	0.311 ± 0.006	180	0.936 ± 0.003	0.297 ± 0.002	240
	CDI	0.759 ± 0.003	0.307 ± 0.029	60	0.808 ± 0.003	0.270 ± 0.009	120	0.865 ± 0.004	0.275 ± 0.008	180	0.923 ± 0.005	0.279 ± 0.006	240
	SANT	0.754 ± 0.008	0.278 ± 0.018	60	0.807 ± 0.009	0.271 ± 0.006	120	0.854 ± 0.027	0.275 ± 0.007	180	0.899 ± 0.046	0.274 ± 0.006	240
	QUORUM	$\mathbf{0.778 \pm 0.002}$	$\mathbf{0.497 \pm 0.023}$	**57**	0.815 ± 0.002	0.287 ± 0.007	**111**	$\mathbf{0.891 \pm 0.002}$	$\mathbf{0.373 \pm 0.003}$	**165**	$\mathbf{0.954 \pm 0.002}$	$\mathbf{0.364 \pm 0.003}$	**219**
IMDB (LLM: 0.954) 25,000 samples	Random	0.901 ± 0.001	0.016 ± 0.006	475	0.912 ± 0.001	0.015 ± 0.002	925	0.923 ± 0.001	0.025 ± 0.001	1375	0.944 ± 0.001	0.022 ± 0.001	1825
	ARAIDA	0.948 ± 0.002	0.080 ± 0.024	500	0.960 ± 0.003	0.057 ± 0.001	1000	0.965 ± 0.008	0.056 ± 0.002	1500	0.970 ± 0.013	0.056 ± 0.002	2000
	CDI	0.946 ± 0.001	0.100 ± 0.004	500	0.961 ± 0.001	0.052 ± 0.001	1000	0.969 ± 0.001	0.046 ± 0.001	1500	0.985 ± 0.001	0.056 ± 0.001	2000
	CoAnnotating	0.946 ± 0.001	0.006 ± 0.001	500	0.960 ± 0.001	0.049 ± 0.001	1000	$\mathbf{0.981 \pm 0.001}$	$\mathbf{0.071 \pm 0.001}$	1500	0.990 ± 0.001	0.064 ± 0.001	2000
	HyPAC	0.947 ± 0.001	0.058 ± 0.012	500	0.950 ± 0.001	0.055 ± 0.005	1000	0.954 ± 0.001	0.055 ± 0.004	1500	0.957 ± 0.001	0.054 ± 0.002	2000
	SANT	0.946 ± 0.001	0.100 ± 0.001	500	0.957 ± 0.001	0.057 ± 0.001	1000	0.957 ± 0.001	0.057 ± 0.001	1500	0.957 ± 0.001	0.057 ± 0.001	2000
	QUORUM	$\mathbf{0.951 \pm 0.001}$	$\mathbf{0.056 \pm 0.002}$	**475**	$\mathbf{0.962 \pm 0.001}$	$\mathbf{0.058 \pm 0.001}$	**925**	0.973 ± 0.001	0.056 ± 0.001	**1375**	$\mathbf{0.993 \pm 0.001}$	$\mathbf{0.064 \pm 0.001}$	**1825**
PubMed (LLM: 0.701) 30,000 samples	Random	0.616 ± 0.001	0.317 ± 0.008	562	0.708 ± 0.001	0.279 ± 0.004	1094	0.731 ± 0.002	0.250 ± 0.004	1627	0.752 ± 0.001	0.259 ± 0.002	2159
	ARAIDA	0.666 ± 0.019	0.410 ± 0.017	591	0.728 ± 0.018	0.368 ± 0.005	1183	0.748 ± 0.049	0.364 ± 0.002	1775	0.762 ± 0.072	0.362 ± 0.006	2366
	CDI	0.677 ± 0.001	0.362 ± 0.008	591	0.749 ± 0.001	0.362 ± 0.003	1183	0.820 ± 0.001	0.360 ± 0.002	1775	0.892 ± 0.001	0.360 ± 0.002	2366
	CoAnnotating	0.678 ± 0.001	0.375 ± 0.001	591	0.750 ± 0.001	0.364 ± 0.001	1183	0.821 ± 0.001	0.361 ± 0.001	1775	0.890 ± 0.001	0.356 ± 0.001	2366
	HyPAC	0.678 ± 0.001	0.370 ± 0.005	591	0.749 ± 0.001	0.361 ± 0.003	1183	0.820 ± 0.001	0.359 ± 0.002	1775	0.891 ± 0.001	0.358 ± 0.002	2366
	SANT	0.673 ± 0.011	0.366 ± 0.036	591	0.714 ± 0.001	0.365 ± 0.001	1183	0.714 ± 0.001	0.364 ± 0.001	1775	0.714 ± 0.001	0.365 ± 0.001	2366
	QUORUM	$\mathbf{0.683 \pm 0.001}$	$\mathbf{0.424 \pm 0.011}$	**562**	$\mathbf{0.762 \pm 0.001}$	$\mathbf{0.404 \pm 0.003}$	**1094**	$\mathbf{0.835 \pm 0.001}$	$\mathbf{0.389 \pm 0.001}$	**1627**	$\mathbf{0.905 \pm 0.001}$	$\mathbf{0.377 \pm 0.001}$	**2159**
CNN (LLM: 0.363) 11,490 Samples	Random	0.402 ± 0.001	-	10455	0.521 ± 0.001	-	31137	0.619 ± 0.001	-	51819	0.781 ± 0.001	-	72501
	CDI	0.423 ± 0.001	-	11604	0.556 ± 0.001	-	34584	$\mathbf{0.682 \pm 0.001}$	-	57564	0.807 ± 0.004	-	80544
	CoAnnotating	0.423 ± 0.001	-	11604	0.550 ± 0.001	-	34584	0.677 ± 0.001	-	57564	0.810 ± 0.001	-	80544
	HyPAC	0.423 ± 0.001	-	11604	0.550 ± 0.001	-	34584	0.677 ± 0.001	-	57564	0.806 ± 0.001	-	80544
	QUORUM	$\mathbf{0.425 \pm 0.001}$	-	**10455**	$\mathbf{0.560 \pm 0.001}$	-	**31137**	0.678 ± 0.001	-	**51819**	$\mathbf{0.816 \pm 0.001}$	-	**72501**

structure of Amazon Bedrock[4] models as of May 2026. Specifically, we assign a cost of \$0.05 per annotation for Claude, \$0.03 for Nova Pro, \$0.01 for Qwen, and \$0.10[5] for human annotations.

Unlike several baselines, QUORUM does not rely on uncertainty estimation, which requires additional LLM inference and therefore incurs extra annotation cost. For uncertainty-based methods, we reproduce prior settings using self-verbalized confidence (Gligorić et al., 2025). Prompt templates and details are reported in Appendix B.

Metrics Following prior work (Huang et al., 2024b), we report Accuracy for classification and ROUGE-1 for summarization. We additionally evaluate Machine Cumulative Accuracy (MCA) (Huang et al., 2024b), measuring how often instances routed to humans correspond to cases where LLM predictions would fail:

$$\text{MCA} = \frac{\sum_{i \in H} \mathbb{1}[\hat{y}_i \neq y_i]}{|H|}$$

Higher MCA indicates more effective allocation of humans. MCA is computed using Claude as the reference LLM, being the most expensive annotator. MCA is omitted for summarization tasks. We also report total annotation cost.

Implementation Details We run all experiments using 10 random seeds for classification and 3 for summarization and report the mean and standard deviation across runs. Hyperparameters are selected based on sensitivity analysis (reported in Appendix C.2), using the configurations that achieved the best overall performance. We set the human escalation threshold to $\tau_H = 0.5$ and the parameters $\epsilon_0 = 0.3$ and $\gamma = 0.005$. QUORUM is initialized with $\mu_a = \mathbf{0}$ and $\Sigma_a = I$ for LLM annotators, while human annotators are assigned a prior $\mu_H = 0.3 \cdot \mathbf{1}$ and $\Sigma_H = 0.5I$. The maximum number of annotations per sample is set to 3, with a minimum of 1. The agreement threshold for consensus-based stopping is 0.66. Following previous research (Gligorić et al., 2025), we use the human-provided labels from the original datasets as gold-standard annotations for evaluation purposes.

5 Results

We investigate the following research questions:

RQ1. Does QUORUM improve annotation quality

[4] aws.amazon.com/bedrock/pricing

[5] https://aitaggers.com.au/pricing

Table 2: Mean $\pm$ std in the DOLLARS setting. For each dataset, we report the full human annotation cost, Accuracy or ROUGE-1 depending on the task (↑), and Machine Cumulative Accuracy (↑) under varying budget constraints. **Bold** indicates the best-performing method, underlined the second best. Row values indicate the available annotation budget, expressed as a percentage of the cost required to annotate the entire dataset using human annotators.

Dataset	Method	10%		30%		50%		70%	
		Accuracy/ROUGE	MCA	Accuracy/ROUGE	MCA	Accuracy/ROUGE	MCA	Accuracy/ROUGE	MCA
AG's News (LLM: 0.834) Human Cost: 760	Random	0.678 ± 0.001	0.059 ± 0.070	0.679 ± 0.001	0.064 ± 0.050	0.681 ± 0.001	0.050 ± 0.028	0.688 ± 0.001	0.064 ± 0.013
	ARAIDA	0.778 ± 0.001	$\mathbf{0.120 \pm 0.001}$	0.783 ± 0.012	0.060 ± 0.001	0.786 ± 0.018	$\mathbf{0.110 \pm 0.001}$	0.792 ± 0.030	$\mathbf{0.095 \pm 0.001}$
	CDI	0.778 ± 0.001	0.120 ± 0.001	0.779 ± 0.001	0.060 ± 0.001	0.780 ± 0.001	0.110 ± 0.001	0.783 ± 0.001	0.095 ± 0.001
	CoAnnotating	0.778 ± 0.001	0.069 ± 0.001	0.778 ± 0.001	0.069 ± 0.001	0.778 ± 0.001	0.069 ± 0.001	0.778 ± 0.001	0.069 ± 0.001
	HyPAC	0.783 ± 0.003	0.100 ± 0.017	0.785 ± 0.005	0.094 ± 0.016	0.790 ± 0.008	0.097 ± 0.015	0.797 ± 0.013	0.094 ± 0.010
	SANT	0.781 ± 0.009	0.120 ± 0.001	0.783 ± 0.012	0.060 ± 0.001	0.786 ± 0.018	0.110 ± 0.001	0.792 ± 0.030	0.095 ± 0.001
	QUORUM	$\mathbf{0.794 \pm 0.001}$	0.084 ± 0.001	$\mathbf{0.809 \pm 0.005}$	$\mathbf{0.099 \pm 0.001}$	$\mathbf{0.829 \pm 0.010}$	0.094 ± 0.001	$\mathbf{0.887 \pm 0.028}$	0.085 ± 0.003
SST2 (LLM: 0.944) Human Cost: 87	Random	0.906 ± 0.002	0.015 ± 0.091	0.916 ± 0.001	0.020 ± 0.089	0.928 ± 0.001	0.038 ± 0.061	0.940 ± 0.002	0.040 ± 0.027
	ARAIDA	0.936 ± 0.002	0.000 ± 0.002	0.937 ± 0.002	0.000 ± 0.002	0.936 ± 0.002	0.000 ± 0.002	0.937 ± 0.002	0.045 ± 0.002
	CDI	0.936 ± 0.002	0.000 ± 0.002	0.936 ± 0.002	0.000 ± 0.003	0.936 ± 0.003	0.000 ± 0.003	0.937 ± 0.003	0.045 ± 0.003
	CoAnnotating	0.937 ± 0.003	0.000 ± 0.003	0.937 ± 0.003	0.000 ± 0.003	0.937 ± 0.003	0.000 ± 0.003	0.937 ± 0.003	0.000 ± 0.003
	HyPAC	0.938 ± 0.004	0.016 ± 0.024	0.937 ± 0.001	0.020 ± 0.037	0.940 ± 0.003	0.051 ± 0.030	0.941 ± 0.003	0.036 ± 0.016
	SANT	0.936 ± 0.002	0.000 ± 0.003	0.936 ± 0.001	0.000 ± 0.001	0.937 ± 0.003	0.000 ± 0.001	0.939 ± 0.005	0.045 ± 0.001
	QUORUM	$\mathbf{0.944 \pm 0.001}$	$\mathbf{0.021 \pm 0.001}$	$\mathbf{0.950 \pm 0.001}$	$\mathbf{0.022 \pm 0.001}$	$\mathbf{0.953 \pm 0.001}$	$\mathbf{0.054 \pm 0.001}$	$\mathbf{0.967 \pm 0.001}$	$\mathbf{0.048 \pm 0.001}$
MMLU-Redux (LLM: 0.633) Human Cost: 300	Random	0.630 ± 0.001	0.370 ± 0.142	0.631 ± 0.001	0.379 ± 0.072	0.634 ± 0.001	0.405 ± 0.051	0.671 ± 0.002	0.383 ± 0.032
	ARAIDA	0.730 ± 0.001	0.111 ± 0.001	0.731 ± 0.001	0.421 ± 0.002	0.732 ± 0.002	0.410 ± 0.002	0.734 ± 0.002	0.329 ± 0.002
	CDI	0.730 ± 0.002	0.111 ± 0.002	0.731 ± 0.002	0.421 ± 0.002	0.732 ± 0.002	0.410 ± 0.002	0.734 ± 0.002	0.329 ± 0.002
	CoAnnotating	0.595 ± 0.002	0.336 ± 0.002	0.626 ± 0.002	$\mathbf{0.482 \pm 0.002}$	0.686 ± 0.002	$\mathbf{0.542 \pm 0.002}$	$\mathbf{0.795 \pm 0.002}$	$\mathbf{0.543 \pm 0.002}$
	HyPAC	0.735 ± 0.001	$\mathbf{0.401 \pm 0.048}$	0.737 ± 0.001	0.428 ± 0.043	0.741 ± 0.001	0.427 ± 0.044	0.749 ± 0.002	0.417 ± 0.037
	SANT	0.730 ± 0.002	0.111 ± 0.002	0.731 ± 0.002	0.421 ± 0.002	0.732 ± 0.001	0.410 ± 0.001	0.734 ± 0.001	0.329 ± 0.001
	QUORUM	$\mathbf{0.735 \pm 0.001}$	0.399 ± 0.012	$\mathbf{0.747 \pm 0.001}$	0.389 ± 0.181	$\mathbf{0.759 \pm 0.001}$	0.492 ± 0.058	0.780 ± 0.001	0.439 ± 0.015
IMDB (LLM: 0.954) Human Cost: 2,500	Random	0.916 ± 0.003	0.053 ± 0.008	0.927 ± 0.003	0.056 ± 0.011	0.937 ± 0.003	0.053 ± 0.008	0.928 ± 0.003	0.037 ± 0.005
	ARAIDA	0.946 ± 0.001	0.110 ± 0.001	0.946 ± 0.001	0.097 ± 0.001	0.947 ± 0.001	0.097 ± 0.001	0.947 ± 0.001	0.071 ± 0.001
	CDI	0.946 ± 0.001	0.110 ± 0.001	0.946 ± 0.001	0.097 ± 0.001	0.947 ± 0.001	0.097 ± 0.001	0.947 ± 0.001	0.071 ± 0.001
	CoAnnotating	0.947 ± 0.001	0.026 ± 0.001	0.947 ± 0.001	0.014 ± 0.001	0.947 ± 0.001	0.007 ± 0.001	0.947 ± 0.003	0.007 ± 0.003
	HyPAC	0.947 ± 0.003	0.040 ± 0.006	0.947 ± 0.003	0.035 ± 0.004	0.948 ± 0.001	0.038 ± 0.008	0.950 ± 0.001	0.039 ± 0.005
	SANT	0.946 ± 0.003	0.110 ± 0.003	0.946 ± 0.003	0.097 ± 0.003	0.947 ± 0.003	0.097 ± 0.003	0.947 ± 0.002	0.071 ± 0.002
	QUORUM	$\mathbf{0.949 \pm 0.002}$	$\mathbf{0.118 \pm 0.040}$	$\mathbf{0.952 \pm 0.001}$	$\mathbf{0.124 \pm 0.003}$	$\mathbf{0.960 \pm 0.003}$	$\mathbf{0.160 \pm 0.002}$	$\mathbf{0.968 \pm 0.006}$	$\mathbf{0.153 \pm 0.001}$
PubMed (LLM: 0.701) Human Cost: 3,000	Random	0.342 ± 0.001	0.174 ± 0.036	0.614 ± 0.001	0.097 ± 0.025	0.627 ± 0.001	0.113 ± 0.015	0.638 ± 0.001	0.172 ± 0.010
	ARAIDA	0.647 ± 0.016	0.206 ± 0.002	0.650 ± 0.021	0.175 ± 0.002	0.646 ± 0.002	0.192 ± 0.002	0.651 ± 0.001	0.217 ± 0.002
	CDI	0.642 ± 0.002	0.206 ± 0.002	0.643 ± 0.002	0.175 ± 0.002	0.646 ± 0.002	0.192 ± 0.002	0.651 ± 0.001	0.217 ± 0.001
	CoAnnotating	0.655 ± 0.003	0.182 ± 0.001	0.655 ± 0.003	0.182 ± 0.001	0.655 ± 0.002	0.182 ± 0.001	0.656 ± 0.001	0.182 ± 0.001
	HyPAC	0.650 ± 0.006	0.192 ± 0.015	0.654 ± 0.007	0.197 ± 0.013	0.660 ± 0.011	0.188 ± 0.015	0.673 ± 0.018	0.188 ± 0.008
	SANT	0.647 ± 0.016	0.206 ± 0.001	0.650 ± 0.021	0.175 ± 0.001	0.648 ± 0.006	$\mathbf{0.192 \pm 0.001}$	0.664 ± 0.039	0.217 ± 0.001
	QUORUM	$\mathbf{0.660 \pm 0.004}$	$\mathbf{0.278 \pm 0.002}$	$\mathbf{0.678 \pm 0.009}$	$\mathbf{0.272 \pm 0.001}$	$\mathbf{0.714 \pm 0.019}$	0.176 ± 0.001	$\mathbf{0.787 \pm 0.039}$	$\mathbf{0.242 \pm 0.002}$
CNN (LLM: 0.363) Human Cost: 1,149	Random	0.342 ± 0.001	-	0.295 ± 0.001	-	0.341 ± 0.001	-	0.352 ± 0.001	-
	CDI	0.361 ± 0.001	-	0.364 ± 0.001	-	0.368 ± 0.001	-	0.378 ± 0.001	-
	CoAnnotating	0.391 ± 0.001	-	0.422 ± 0.001	-	0.486 ± 0.001	-	0.613 ± 0.001	-
	HyPAC	0.373 ± 0.001	-	0.378 ± 0.001	-	0.388 ± 0.001	-	0.407 ± 0.001	-
	QUORUM	$\mathbf{0.476 \pm 0.001}$	-	$\mathbf{0.491 \pm 0.001}$	-	$\mathbf{0.501 \pm 0.002}$	-	$\mathbf{0.638 \pm 0.002}$	-

under limited annotation budgets?

RQ2. Can feature-based difficulty estimation replace uncertainty without sacrificing performance?

RQ3. Does allocating multiple annotations improve final prediction quality?

RQ4. Does QUORUM transfer across languages?

RQ1. Budget-Aware Routing Evaluation

We evaluate QUORUM under two routing settings: AUDITOR STYLE and DOLLARS.

AUDITOR STYLE Table 1 reports performance under varying human annotation budgets. Overall, QUORUM achieves the strongest cost–performance trade-off, matching or improving performance while reducing annotation cost across datasets. The total cost for each sample is computed as the sum of per-call costs incurred for each annotator queried. Improvements are most pronounced in low-budget regimes (10%–30%), where efficient allocation of human labels is critical. Gains are especially evident on challenging benchmarks such as PubMed and MMLU-Redux. For summarization tasks, QUORUM obtains the highest ROUGE-1 scores while maintaining the lowest annotation cost, showing generalization beyond classification settings. In a few cases at 10% budget, performance falls below the average standalone LLM score (in parentheses). This is expected, as the LLM baseline reflects average annotator performance, whereas under strict budgets the system relies on the cheapest model.

DOLLARS Table 2 evaluates routing under fixed monetary constraints with heterogeneous annotator costs. In this setting, QUORUM benefits from increased budgets, while several baselines plateau early, indicating limited adaptability as additional resources become available. This effect is evident on a challenging benchmark such as PubMed, where QUORUM shows larger gains between low- and high-budget regimes. These results highlight more efficient budget allocation through selective rather than uniform use of expensive annotators. For summarization, QUORUM achieves the strongest ROUGE-1 scores, confirming that adap-

Table 3: Mean ± 95% confidence intervals accuracy (↑) with uncertainty-based routing, routing based on random features and our feature-based difficulty estimation. Budget is expressed as a percentage of dataset size.

Budget	Dataset	LLM-Uncertainty Features	Random Features	Ours
10%	AG's News	0.800 ± 0.002	0.697 ± 0.009	$\mathbf{0.800 \pm 0.001}$
	SST2	0.948 ± 0.001	0.901 ± 0.012	$\mathbf{0.968 \pm 0.001}$
	Global-MMLU (es)	0.741 ± 0.001	0.518 ± 0.012	$\mathbf{0.749 \pm 0.002}$
70%	AG's News	0.932 ± 0.002	0.712 ± 0.007	$\mathbf{0.936 \pm 0.001}$
	SST2	0.978 ± 0.002	0.921 ± 0.010	$\mathbf{0.990 \pm 0.001}$
	Global-MMLU (es)	0.910 ± 0.001	0.692 ± 0.009	$\mathbf{0.912 \pm 0.001}$

Table 4: Effect of multiple annotations per instance on QUORUM's accuracy/ROUGE-1 across datasets.

#Annotations	AG's News	PubMed	MMLU-Redux	CNN
1	0.852	0.740	0.721	0.387
2	0.873	0.756	0.754	0.398
3	**0.891**	**0.791**	**0.766**	**0.410**

tive routing remains effective even when supervision quality is estimated via semantic similarity. An analysis of feature usage within the bandit is deferred to Appendix C.1 due to space constraints.

RQ2. Features vs Uncertainty

We compare our feature-based difficulty estimation with uncertainty-driven routing strategies. As shown in Table 3, our approach matches or outperforms uncertainty-based methods across all datasets and budget regimes. Crucially, uncertainty estimation requires additional LLM forward passes, introducing extra cost and computational overhead, whereas our features rely on lightweight, deterministic computations over the input text with no model inference. We compare against a random feature baseline, which performs worse, confirming that the improvement stems from the structured linguistic and semantic signals in our representation rather than feature dimensionality alone. Runtime comparisons in Appendix C.3 further show that our approach is not only more economical in terms of annotation cost, but also substantially faster than confidence estimation or model retraining procedures, improving scalability on larger datasets.

RQ3. Impact of Multiple Annotations

Table 4 evaluates the impact of increasing the number of annotations per instance on downstream performance in the DOLLARS setting. Across all datasets, we observe an improvement as the number of annotations increases from one to three, confirming that additional supervision signals provide complementary information that improves label quality. The gains are particularly pronounced on the CNN dataset, where performance increases from 0.387 to 0.410, suggesting that subjective or ambiguous instances benefit more from redundancy in annotations. These results support the design choice of QUORUM to dynamically allocate multiple annotations when uncertainty or disagreement is high, as additional labels consistently translate into higher-quality final predictions.

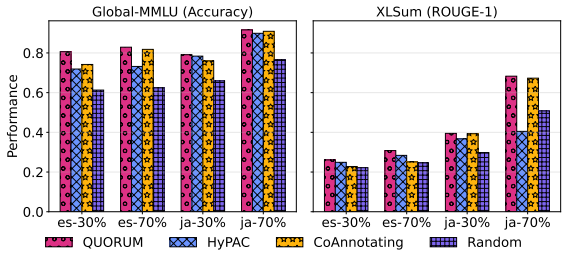

Figure 3: Performance under multilingual settings across Spanish (es) and Japanese (ja) benchmarks for classification (Global-MMLU) and summarization (XLSum) at different annotation budgets (30% and 70%).

RQ4. Multilingual Generalization

To assess whether routing decisions generalize beyond English, we evaluate QUORUM on Spanish and Japanese variants of classification (Global-MMLU) and summarization (XLSum) tasks in the DOLLARS setting. As shown in Fig. 3, QUORUM matches or outperforms competing methods across languages and annotation budgets. Improvements are visible under higher-budget regimes, where additional annotations are allocated more effectively. These gains are achieved while maintaining lower annotation costs than competitors, suggesting that the feature-based routing transfers well across languages without requiring language-specific uncertainty estimation. Due to space constraints, complete multilingual results can be found in Appendix C.4.

6 Conclusion

We introduced QUORUM, a budget-aware annotation routing framework dynamically allocating samples between human annotators and LLMs. Unlike prior approaches relying on uncertainty or confidence estimation, QUORUM learns instance difficulty from lightweight linguistic and semantic features, enabling cost-efficient and adaptive annotation strategies. Across a set of classification and summarization benchmarks, our method consistently achieves strong performance–cost trade-offs.

Limitations While QUORUM demonstrates strong performance across multiple benchmarks, we highlight a few directions for future work. First, our evaluation is conducted primarily on English-language datasets, and extending the framework to additional multilingual or cross-lingual settings would further validate its generality. Second, although we consider a diverse set of tasks, evaluating QUORUM on more specialized domains could provide additional insights into its behavior under different annotation regimes.

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. 2011. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24.

Nicola Cecere, Andrea Bacciu, Ignacio Fernández-Tobías, and Amin Mantrach. 2025. Monte carlo temperature: a robust sampling strategy for llm's uncertainty quantification methods. In *Proceedings of the 5th Workshop on Trustworthy NLP (TrustNLP 2025)*, pages 305–320.

Franck Dernoncourt and Ji-Young Lee. 2017. Pubmed 200k rct: a dataset for sequential sentence classification in medical abstracts. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 308–313.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology*, 32(3):221.

Aryo Pradipta Gema, Joshua Ong Jun Leang, Giwon Hong, Alessio Devoto, Alberto Carlo Maria Mancino, Rohit Saxena, Xuanli He, Yu Zhao, Xiaotang Du, Mohammad Reza Ghasemi Madani, and 1 others. 2025. Are we done with mmlu? In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5069–5096.

Kristina Gligorić, Tijana Zrnic, Cinoo Lee, Emmanuel Candes, and Dan Jurafsky. 2025. Can unconfident llm annotations be used for confident conclusions? In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3514–3533.

Peter Hall and Christopher C Heyde. 2014. *Martingale limit theory and its application*. Academic press.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703, Online. Association for Computational Linguistics.

Sabit Hassan and Malihe Alikhani. 2023. D-calm: A dynamic clustering-based active learning approach for mitigating bias. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5540–5553.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Chen Huang, Yang Deng, Wenqiang Lei, Jiancheng Lv, and Ido Dagan. 2024a. Selective annotation via data allocation: These data should be triaged to experts for annotation rather than the model. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 301–320.

Chen Huang, Yiping Jin, Ilija Ilievski, Wenqiang Lei, and Jiancheng Lv. 2024b. Araida: Analogical reasoning-augmented interactive data annotation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10660–10675.

Tor Lattimore and Csaba Szepesvári. 2020. *Bandit algorithms*. Cambridge University Press.

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.

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.

Lester James Validad Miranda, Yizhong Wang, Yanai Elazar, Sachin Kumar, Valentina Pyatkin, Faeze Brahman, Noah A Smith, Hannaneh Hajishirzi, and Pradeep Dasigi. 2025. Hybrid preferences: Learning to route instances for human vs. ai feedback. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7162–7200.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Seiji Shimizu, Hisada Shohei, Yutaka Uno, Shuntaro Yada, Shoko Wakamiya, and Eiji Aramaki. 2025. Exploring llm annotation for adaptation of clinical information extraction models under data-sharing restrictions. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 14678–14694.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, and 1 others. 2025. Global mmlu: Understanding and addressing cultural and linguistic biases in multilingual evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18761–18799.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large language models for data annotation and synthesis: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 930–957, Miami, Florida, USA. Association for Computational Linguistics.

Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2021. Want to reduce labeling cost? gpt-3 can help. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4195–4205.

Xinru Wang, Hannah Kim, Sajjadur Rahman, Kushan Mitra, and Zhengjie Miao. 2024. Human-llm collaborative annotation through effective verification of llm labels. In *Proceedings of the 2024 CHI conference on human factors in computing systems*, pages 1–21.

Yinfei Yang, Oshin Agarwal, Chris Tar, Byron C Wallace, and Ani Nenkova. 2019. Predicting annotation difficulty to improve task routing and model performance for biomedical information extraction. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1471–1480.

Hao Zeng, Huipeng Huang, Xinhao Qu, Jianguo Huang, Bingyi Jing, and Hongxin Wei. 2026. Hypac: Cost-efficient llms-human hybrid annotation with pac error guarantees. *arXiv preprint arXiv:2602.02550*.

Thomas Ziller, Shashikant Ilager, Alessandro Tundo, Ezio Bartocci, Leonardo Mariani, and Ivona Brandic. 2026. Greenserv: Energy-efficient context-aware dynamic routing for multi-model llm inference. *arXiv preprint arXiv:2601.17551*.

A QUORUM Description

A.1 Feature Description

Our approach relies on the following features:

- Linguistic features (Ziller et al., 2026): Six numerical values that capture information like sentence complexity, ambiguity, or other text characteristics that make a sample harder to label:
  - Average Word Length: Mean number of characters per word. Longer words often indicate more technical or complex vocabulary.
  - Type-Token Ratio: Ratio of unique words to total words. Higher values indicate greater lexical diversity, suggesting more varied or sophisticated language use.
  - Average Sentence Length: Mean number of words per sentence. Longer sentences may indicate syntactic complexity or convoluted structure. For languages without explicit word boundaries (*e.g.*, Japanese), word-length–based features were not employed due to the lack of a directly comparable segmentation scheme.
  - Flesch Reading Ease: Inverted and normalized (0-1) Flesch score (Flesch, 1948), where higher values indicate more difficult text. Combines syllable count and sentence length into a standard readability metric.
  - Noun Ratio: Proportion of nouns among POS-tagged content words. Higher noun density may indicate more entity-rich or descriptive text.
  - Verb Ratio: Proportion of verbs among POS-tagged content words. Higher verb density may indicate more action-oriented or procedural text.
- Embedding features (Hassan and Alikhani, 2023): Each sample is encoded using an embedder (Stella[6] for our experiments). For each sample, we compute the cosine distance to its 5 nearest neighbors in the embedding space. This produces a distance vector that serves as a proxy for sample atypicality or semantic isolation .

[6] https://huggingface.co/NovaSearch/stella_en_1.5B_v5

A.2 Proof of Proposition 3.1

For each annotator a, rewards follow:

$$r_{a,t} = x_t^\top \theta_a^* + \eta_t$$

With Gaussian noise $\eta_t \sim \mathcal{N}(0, \sigma^2)$. Bayesian linear regression updates yield posterior parameters:

$$\Sigma_t^{-1} = \Sigma_0^{-1} + \frac{1}{\sigma^2} \sum_{s<t} x_s x_s^\top$$

$$\mu_t = \Sigma_t \left(\Sigma_0^{-1} \mu_0 + \frac{1}{\sigma^2} \sum_{s<t} r_s x_s \right).$$

Applying the concentration bounds for Bayesian linear regression proved in Abbasi-Yadkori et al. (2011) (precisely Theorem 2 in Abbasi-Yadkori et al. (2011)), with probability $1 - \delta$:

$$\|\mu_t - \theta^*\|_{\Sigma_t^{-1}} \leq \beta_t.$$

Now:

$$\left| x_t^\top (\mu_{a,t} - \theta_a^*) \right| = \left| x_t^\top \Sigma_{a,t}^{1/2} \Sigma_{a,t}^{-1/2} \top (\mu_{a,t} - \theta_a^*) \right|$$
$$= \left| \left(\Sigma_{a,t}^{1/2} x_t \right)^\top \left(\Sigma_{a,t}^{-1/2} (\mu_{a,t} - \theta_a^*) \right) \right|$$

Using Cauchy-Schwarz we have:

$$\left| x_t^\top (\mu_{a,t} - \theta_a^*) \right| \leq \left\| \Sigma_{a,t}^{1/2} x_t \right\|_2 \|\mu_{a,t} - \theta_a^*\|_{\Sigma_{a,t}^{-1}}$$

Namely:

$$|x_t^\top (\mu_t - \theta^*)| \leq \beta_t \sqrt{x_t^\top \Sigma_t x_t}$$

If the annotator is sampled infinitely often on sufficiently diverse contexts, so that:

$$\lambda_{\min}(\Sigma_t^{-1}) \to \infty$$

Then for bounded contexts $\|x_t\|_2 \leq L$:

$$x_t^\top \Sigma_t x_t \leq \frac{L^2}{\lambda_{\min}(\Sigma_t^{-1})} \to 0$$

Thus posterior uncertainty shrinks asymptotically and quality estimates converge toward the true annotator parameters. □

A.3 Proof of Proposition 3.2

Proposition 3.2 (formal). *Assume the linear contextual reward model:*

$$r_{a,t} = x_t^\top \theta_a^* + \eta_{a,t}$$

with bounded contexts $\|x_t\|_2 \leq L$, bounded parameters $\|\theta_a^*\|_2 \leq S$, and conditionally sub-Gaussian noise. For each context x_t, define the oracle annotator:

$$a_t^* \in \arg\max_{a \in \mathcal{A}} x_t^\top \theta_a^*$$

Assume that the oracle annotator is uniformly separated from the second-best annotator, i.e.:

$$\Delta_{\min} = \inf_{t \geq 1} \left(x_t^\top \theta_{a_t^*}^* - \max_{a \neq a_t^*} x_t^\top \theta_a^* \right) > 0$$

Consider the core routing rule that explores with probability:

$$\epsilon_t = \frac{\epsilon_0}{1 + \gamma t},$$

and otherwise exploits according to:

$$\hat{a}_t \in \arg\max_{a \in \mathcal{A}} x_t^\top \mu_{a,t}$$

Assume that the exploration mechanism samples each feasible annotator infinitely often on sufficiently diverse contexts, so that, for every annotator a:

$$\frac{\lambda_{\min}(\Sigma_{a,t}^{-1})}{\beta_{a,t}^2} \to \infty$$

Then, on the posterior concentration event of Proposition 3.1, the average regret vanishes:

$$\lim_{T \to \infty} \frac{R(T)}{T} = 0$$

Where:

$$R(T) = \sum_{t=1}^{T} \left(x_t^\top \theta_{a_t^*}^* - x_t^\top \theta_{\hat{a}_t}^* \right)$$

Proof. Let:

$$\Delta_t = x_t^\top \theta_{a_t^*}^* - x_t^\top \theta_{\hat{a}_t}^*$$

denote the instantaneous regret.

We work on the high-probability event from Proposition 3.1, on which, for every annotator a:

$$\left| x_t^\top (\mu_{a,t} - \theta_a^*) \right| \leq \beta_{a,t} \sqrt{x_t^\top \Sigma_{a,t} x_t}$$

First, we show that the prediction error vanishes. Since $\|x_t\|_2 \leq L$:

$$x_t^\top \Sigma_{a,t} x_t \leq \lambda_{\max}(\Sigma_{a,t}) \|x_t\|_2^2$$

Moreover:

$$\lambda_{\max}(\Sigma_{a,t}) = \frac{1}{\lambda_{\min}(\Sigma_{a,t}^{-1})}$$

Therefore:

$$x_t^\top \Sigma_{a,t} x_t \leq \frac{L^2}{\lambda_{\min}(\Sigma_{a,t}^{-1})}$$

Substituting this into the concentration bound gives:

$$\left| x_t^\top (\mu_{a,t} - \theta_a^*) \right| \leq \frac{\beta_{a,t} L}{\sqrt{\lambda_{\min}(\Sigma_{a,t}^{-1})}}$$

By assumption:

$$\frac{\lambda_{\min}(\Sigma_{a,t}^{-1})}{\beta_{a,t}^2} \to \infty$$

And hence:

$$\frac{\beta_{a,t} L}{\sqrt{\lambda_{\min}(\Sigma_{a,t}^{-1})}} \to 0$$

Thus, for every annotator a:

$$\left| x_t^\top (\mu_{a,t} - \theta_a^*) \right| \to 0$$

Since $\mathcal{A}$ is finite, this convergence holds uniformly over annotators:

$$\max_{a \in \mathcal{A}} \left| x_t^\top (\mu_{a,t} - \theta_a^*) \right| \to 0$$

Now use the margin assumption. Since the maximum prediction error over annotators goes to zero, there exists a finite time T_0 such that, for all $t \geq T_0$ and all $a \in \mathcal{A}$:

$$\left| x_t^\top (\mu_{a,t} - \theta_a^*) \right| \leq \frac{\Delta_{\min}}{4}.$$

Fix any $t \geq T_0$, and suppose the algorithm is in an exploitation round. We show that the greedy rule selects the oracle annotator.

For the oracle annotator a_t^*, we have:

$$x_t^\top \mu_{a_t^*,t} \geq x_t^\top \theta_{a_t^*}^* - \frac{\Delta_{\min}}{4}$$

For any non-oracle annotator $a \neq a_t^*$, the margin assumption gives:

$$x_t^\top \theta_a^* \leq x_t^\top \theta_{a_t^*}^* - \Delta_{\min}$$

Using again the prediction-error bound:

$$x_t^\top \mu_{a,t} \leq x_t^\top \theta_a^* + \frac{\Delta_{\min}}{4}$$

Combining the last two inequalities:

$$x_t^\top \mu_{a,t} \leq x_t^\top \theta_{a_t^*}^* - \Delta_{\min} + \frac{\Delta_{\min}}{4} = x_t^\top \theta_{a_t^*}^* - \frac{3\Delta_{\min}}{4}$$

On the other hand:

$$x_t^\top \mu_{a_t^*,t} \geq x_t^\top \theta_{a_t^*}^* - \frac{\Delta_{\min}}{4}$$

Therefore, for every $a \neq a_t^*$:

$$x_t^\top \mu_{a_t^*,t} > x_t^\top \mu_{a,t}$$

Hence the exploitation rule:

$$\hat{a}_t \in \arg\max_{a \in \mathcal{A}} x_t^\top \mu_{a,t}$$

selects $\hat{a}_t = a_t^*$. Thus every exploitation round after T_0 has zero regret.

For the exploration rounds, let $Z_t \in \{0, 1\}$ be the indicator that round t is an exploration round. By construction of the algorithm:

$$\mathbb{P}(Z_t = 1 \mid \mathcal{F}_{t-1}) = \epsilon_t$$

where $\mathcal{F}_{t-1}$ denotes the history before round t. Since Z_t is binary, this implies:

$$\mathbb{E}[Z_t \mid \mathcal{F}_{t-1}] = \epsilon_t$$

Defining:

$$Y_t = Z_t - \epsilon_t$$

Then:

$$\mathbb{E}[Y_t \mid \mathcal{F}_{t-1}] = \mathbb{E}[Z_t - \epsilon_t \mid \mathcal{F}_{t-1}] = \epsilon_t - \epsilon_t = 0$$

Defining M_T as:

$$M_T = \sum_{t=1}^{T} Y_t = \sum_{t=1}^{T} (Z_t - \epsilon_t)$$

We can show it is a martingale with bounded increments. Since:

$$\epsilon_t = \frac{\epsilon_0}{1 + \gamma t}$$

We have:

$$\sum_{t=1}^{T} \epsilon_t \leq \frac{\epsilon_0}{\gamma} \log(1 + \gamma T)$$

Therefore:

$$\frac{1}{T} \sum_{t=1}^{T} \epsilon_t \to 0$$

Since $Y_t = Z_t - \epsilon_t$ is a martingale difference sequence and $|Y_t| \leq 1$, we have:

$$\sum_{t=1}^{\infty} \frac{\mathbb{E}[Y_t^2]}{t^2} \leq \sum_{t=1}^{\infty} \frac{1}{t^2} < \infty$$

By the strong law for martingale differences (Hall and Heyde, 2014, Theorem 2.18), it follows that

$$\frac{1}{T} \sum_{t=1}^{T} (Z_t - \epsilon_t) \to 0 \quad \text{almost surely.}$$

Consequently:

$$\frac{1}{T} \sum_{t=1}^{T} Z_t = \frac{1}{T} \sum_{t=1}^{T} \epsilon_t + \frac{1}{T} \sum_{t=1}^{T} (Z_t - \epsilon_t) \to 0$$

Thus exploration rounds form a vanishing fraction of all rounds.

Finally, instantaneous regret is bounded. Indeed, since $\|x_t\|_2 \leq L$ and $\|\theta_a^*\|_2 \leq S$:

$$\left| x_t^\top \theta_a^* - x_t^\top \theta_b^* \right| \leq |x_t^\top \theta_a^*| + |x_t^\top \theta_b^*| \leq 2LS$$

for any a, b. Hence $\Delta_t \leq 2LS$.

After time T_0, exploitation rounds have zero regret, and only exploration rounds can contribute regret. Therefore:

$$R(T) \leq 2LS\, T_0 + 2LS \sum_{t=T_0}^{T} Z_t$$

Dividing by T, we obtain:

$$\frac{R(T)}{T} \leq \frac{2LS\, T_0}{T} + 2LS \frac{1}{T} \sum_{t=T_0}^{T} Z_t$$

The first term goes to zero because T_0 is finite. The second term goes to zero because exploration rounds have vanishing frequency. Therefore:

$$\frac{R(T)}{T} \to 0$$

□

A.4 QUORUM Pseudocode

Algorithm 1 summarizes QUORUM under the DOLLARS setting, where routing decisions are constrained by a monetary budget and annotators have heterogeneous costs. The procedure consists of an initial calibration phase, adaptive routing, optional budget-aware re-annotation of difficult samples, and task-specific aggregation. SELECT refers to the Routing Mechanism paragraph of Section 3.2. Algorithm 2 summarizes the AUDITOR STYLE setting, where a fixed human annotation budget is available and the router decides whether to allocate samples to humans or to the cheapest LLM annotator. Human supervision is concentrated on calibration examples and samples estimated to be close to previously observed model errors.

Algorithm 1 QUORUM — DOLLARS

Require: LLM predictions $\mathcal{P} = \{P_1, \ldots, P_K\}$, human labels $\mathbf{y}$, budget B
Ensure: Final outputs $\mathbf{o} = [o_1, \ldots, o_T]$
Initialize posteriors: $\boldsymbol{\mu}_a \leftarrow \mathbf{0}$, $\boldsymbol{\Sigma}_a \leftarrow \mathbf{I}$ for LLM arms
Phase 1: Calibration
$N_{\text{cal}} \leftarrow \lfloor 0.20 \cdot \lfloor B / c_{\text{human}} \rfloor \rfloor$
$\mathcal{C} \leftarrow$ top-N_{cal} samples by Mahalanobis uncertainty
for $t \in \mathcal{C}$ **do**
  Query human for y_t; for each LLM arm a: update posterior with reward r
end for
Phase 2: Routing
for each uncovered sample $t \notin \mathcal{C}$ **do**
  $a^* \leftarrow$ SELECT$(\mathbf{x}_t, \text{available arms}, t)$
  Record annotation from a^*; update budget
end for
Phase 3: Extra Annotations
while budget remains **do**
  Pick highest-priority sample by P_t
  Select arm; record annotation
end while
Phase 4: Aggregation
for $t = 1, \ldots, T$ **do**
  $o_t \leftarrow$ weighted majority vote
end for

B Experiments - Additional Details

Experimental Setup All experiments are conducted on a machine equipped with 96 CPU cores (AMD EPYC 7R32), 4 NVIDIA A10G GPUs, and 373 GB of system memory.

Prompts The prompt we use to obtain LLM annotations for classification is the following:

Algorithm 2 QUORUM — AUDITOR STYLE

Require: LLM predictions $\mathcal{P}$, human labels $\mathbf{y}$, human budget B

Ensure: Final outputs $\mathbf{o}$

1: $\mathbf{p} \leftarrow$ predictions from cheapest LLM
2: $\mathcal{C} \leftarrow$ random $0.2B$ calibration samples
3: $\mathcal{E} \leftarrow \{t \in \mathcal{C} : p_t \neq y_t\}$
4: **if** $|\mathcal{E}| \geq 2$ **then**
5: Sort remaining samples by distance to error centroid (ascending)
6: **end if**
7: Assign human labels to $\mathcal{C} \cup$ top-$(B - |\mathcal{C}|)$ nearest samples; use LLM elsewhere

```
This is a text classification task. The
possible labels are [...] while the indices
of the labels are [...]. Output only 'Label:
the predicted index of the label that you
predict'. This is the sentence to classify:
[...].
```

The prompt we use to obtain LLM annotations for summarization is the following:

```
This is a text summarization task. Output
only the summarized text and nothing more.
Be concise and summarize the following text
[...]
```

The prompt we use to obtain LLM annotations for Q&A is the following:

```
This is a question-answering task. Select
the right answer from the available choices.
Output only 'Label: the predicted index of
the label that you predict'. This is the
question [...] and the possible answers
are: [...]
```

The prompt we use to obtain the confidence is:

```
How likely is it that the following text
is correctly summarized/classified? Output
the probability (a number between 0 and 1)
Text: [...] Prediction: [...]
```

We set `max_tokens=1000` and `temperature=1`.

Datasets We evaluate on a diverse set of benchmarks spanning several tasks classification with AG's News, SST2 (Socher et al., 2013), IMDB (Maas et al., 2011) and PubMed (Dernoncourt and Lee, 2017), multiple-choice QA with Global-MMLU (Singh et al., 2025) and MMLU-Redux (Gema et al., 2025) and abstractive summarization with CNN/DailyMail (See et al., 2017) and XLSum (Hasan et al., 2021). This selection covers varying difficulty levels, class distributions, and task formats.

Table 5: Number of classes, class distribution and number of samples of the selected datasets. NA indicates summarization datasets with no classes.

Name	#Classes	Class Distribution (%)	#Samples
AG's News	4	[25,25,25,25]	7,600
SST2	2	[49.9,51.1]	870
Global-MMLU*	4	[23, 25, 25, 27]	14,300
MMLU-Redux*	4	[22, 25, 25, 28]	3,000
IMDB*	2	[50,50]	25,000
PubMed*	5	[10.4,15.45, 33.42,7.89,32.84]	30,000
CNN*	NA	NA	11,490
XLSum*	NA	NA	4,763

C Additional Results

C.1 Feature Analysis of QUORUM

To understand how QUORUM allocates annotations, we inspect the learned bandit policy in Fig. 4. The arm selection proportions image (left) show that the router initially explores all annotators uniformly, then progressively converges toward lower-cost annotators for routine instances, while reserving higher-cost annotators (both human and LLM) for instances where predicted difficulty is highest.

The feature importance analysis (right) shows that b_t, the remaining budget fraction, is by far the dominant routing signal, followed by contextual features derived from m_t (the annotation mask) and linguistic components of $\phi(d_t)$ such as vocabulary richness. This confirms that routing decisions are primarily driven by budget state and annotation history rather than embedding-based features alone. In particular, the model learns to regulate both the overall annotation ratio and the human annotation ratio as explicit proxies for budget consumption and allocation pressure, allowing it to avoid premature depletion of high-cost annotations while still maintaining coverage on uncertain instances. The heatmap (center) further reveals arm-specific patterns: the human arm exhibits strong positive weights on budget and mask features, confirming that escalation to H is triggered when prior annotations are insufficient or conflicting.

C.2 Sensitivity analysis of QUORUM

To assess robustness to hyperparameter choices, we evaluate QUORUM under different values of the main routing and aggregation parameters (Table 6). Sensitivity experiments are performed on the AG's

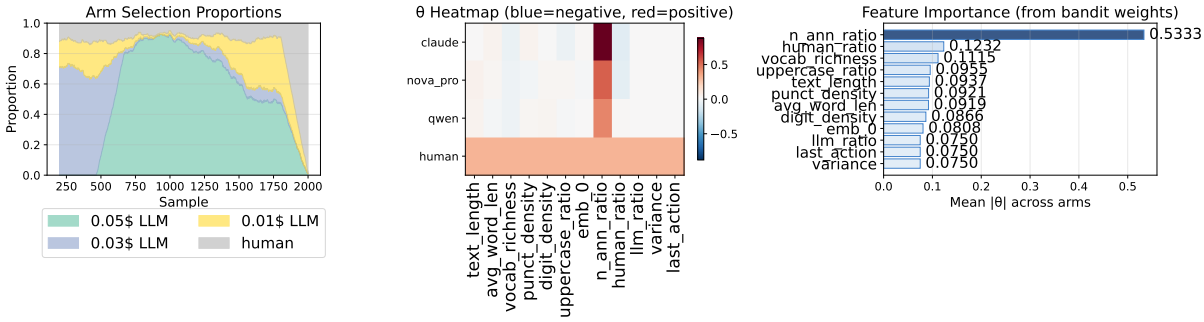

Figure 4: Analysis of the learned bandit routing policy on the AG's News dataset (window of 2000 samples) in the DOLLARS setting. **Left:** arm selection proportions over time, showing convergence from exploration to exploitation. **Center:** heatmap of posterior weight vectors across features and arms (blue=negative, red=positive). **Right:** feature importance computed as mean absolute weight across arms.

Table 6: Sensitivity analysis of QUORUM with respect to key routing and aggregation hyperparameters. We report downstream accuracy while varying the human escalation threshold (τ_H), human voting weight (w_H), initial exploration rate (ϵ_0), exploration decay factor (γ) and calibration size.

τ_H Value	Accuracy
0.1	0.780
0.3	0.781
0.5	**0.879**
0.7	0.875

(a) τ_H sensitivity.

w_H Value	Accuracy
1	0.790
5	0.831
10	**0.879**
15	0.871

(b) w_H sensitivity.

ϵ_0 Value	Accuracy
0.1	0.861
0.3	**0.879**
0.5	0.873
0.7	0.865

(c) ϵ_0 sensitivity.

γ Value	Accuracy
0.001	0.875
0.005	**0.879**
0.01	0.842
0.05	0.857

(d) γ sensitivity.

Calibration Size	Accuracy
0.2	**0.879**
0.4	0.874
0.6	0.871
0.8	0.873

(e) Calibration size sensitivity.

News dataset under the DOLLARS setting with a total budget of $450. Overall, performance remains relatively stable across configurations, suggesting that the framework is not highly sensitive to precise tuning. The human escalation threshold τ_H has the largest impact: moderate values ($\tau_H = 0.5$) achieve the best performance, whereas overly aggressive or overly conservative escalation leads to lower accuracy. This indicates that balancing reliance on LLMs and human supervision is critical. Similarly, increasing the human voting weight w_H improves performance up to an intermediate regime ($w_H = 10$), confirming that prioritizing human annotations is beneficial, but excessive weighting yields diminishing returns. Exploration parameters exhibit comparable behaviour. Moderate initial exploration ($\epsilon_0 = 0.3$) and slow decay ($\gamma = 0.005$) provide the strongest results, whereas excessive exploration or premature exploitation degrade performance. These findings support the design choice of combining adaptive exploration with budget-aware routing. Overall, the results suggest that QUORUM is robust to moderate hyperparameter variation and does not require extensive tuning to obtain competitive performance. The calibration-size analysis suggests that allocating a relatively small portion of the budget to the initial calibration stage is sufficient to obtain reliable annotator estimates. Performance peaks when 20% of the affordable budget is reserved for calibration, while larger calibration fractions gradually reduce accuracy. This indicates that excessive upfront calibration may limit the resources available for adaptive routing and iterative annotation refinement.

Table 7: Routing overhead (seconds) required before routing decisions across datasets (↓). ET denotes embedding extraction time, CT confidence estimation time via additional LLM prompting, and TT model training time.

Method	AG's News	PubMed	Global-MMLU
Random	0.007	0.035	0.014
ARAIDA	4.988+TT	19.879+TT	12.250+TT
CDI	0.503+CT	0.547+CT	0.661+CT
CoAnnotating	0.006+CT	0.038+CT	0.014+CT
HyPAC	0.046+CT	0.150+CT	0.075+CT
SANT	18.052+TT	161.014+TT	40.853+TT
QUORUM	0.494+ET	1.467+ET	0.593+ET

C.3 Time Analysis of QUORUM

Table 7 reports the routing overhead required by different routing strategies before annotation decisions can be made. Existing approaches relying on confidence estimation (CT) require an additional LLM call for every sample to obtain uncertainty signals, resulting in substantial computational costs. On datasets of roughly 2,000 instances, this process can require approximately two hours. Training-based methods (TT), such as ARAIDA and SANT, are also expensive due to supervised fitting procedures, typically requiring around 20 minutes for the same dataset size.

In contrast, QUORUM only performs embedding extraction (ET) and lightweight feature computation, reducing preprocessing to approximately four minutes on datasets of similar size. Across benchmarks, QUORUM consistently achieves lower computational overhead than confidence- or training-based alternatives while maintaining competitive annotation performance. These results suggest that replacing uncertainty estimation with feature-based difficulty signals improves not only routing effectiveness but also practical scalability.

C.4 Additional Multilingual Results

To assess the generalizability of our approach beyond English, we evaluate all methods on Japanese and Spanish language tasks using Global-MMLU (ja, es) for multiple-choice Question Answering and XLSum (ja, es) for abstractive summarization. These datasets allow us to test whether the annotation strategies remain effective when applied to a typologically distant language with different linguistic properties. Tables 8 and 9 report results under the AUDITOR STYLE and DOLLARS settings, respectively. Our method consistently achieves the best or second-best performance across both tasks and budget levels, demonstrating that its effectiveness is not limited to English and transfers well to multilingual scenarios. Notably, the performance gap with competing methods widens at higher budgets: at 70% human annotation, QUORUM reaches 0.946 accuracy on Global-MMLU (ja), outperforming the second-best method by nearly 4 percentage points. These results confirm that our method generalizes effectively to typologically distant languages, even when LLM annotators are less reliable.

Table 8: Performance in the AUDITOR STYLE setting under different human annotation budgets. We report Accuracy or ROUGE-1 depending on the task (↑), Machine Cumulative Accuracy (↑), and total annotation cost (↓) for varying percentages of human-labeled data. **Bold** indicates the best-performing method, while underlined the second best. Value in parentheses represents the average performance obtained by LLM annotators on that dataset.

Dataset	Method	10%			30%			50%			70%		
		Accuracy/ROUGE	MCA	Cost	Accuracy/ROUGE	MCA	Cost	Accuracy/ROUGE	MCA	Cost	Accuracy/ROUGE	MCA	Cost
Global-MMLU (ja) (LLM: 0.640) 14,040 samples	Random	0.641 ± 0.001	0.305 ± 0.013	267	0.729 ± 0.002	0.313 ± 0.007	519	0.729 ± 0.002	0.165 ± 0.003	772	0.796 ± 0.002	0.243 ± 0.002	1025
	ARAIDA	0.679 ± 0.014	0.383 ± 0.109	281	0.732 ± 0.018	0.312 ± 0.012	562	0.759 ± 0.052	0.315 ± 0.015	842	0.782 ± 0.082	0.318 ± 0.018	1123
	CDI	0.691 ± 0.001	0.355 ± 0.014	281	0.751 ± 0.002	0.318 ± 0.006	562	0.823 ± 0.002	0.335 ± 0.004	842	0.893 ± 0.002	0.338 ± 0.002	1123
	CoAnnotating	0.692 ± 0.001	0.358 ± 0.001	281	0.760 ± 0.001	0.348 ± 0.001	562	0.818 ± 0.001	0.323 ± 0.001	842	0.909 ± 0.001	0.361 ± 0.001	1123
	HyPAC	0.715 ± 0.001	0.349 ± 0.009	281	$\mathbf{0.784 \pm 0.001}$	$\mathbf{0.350 \pm 0.002}$	562	$\mathbf{0.838 \pm 0.001}$	0.331 ± 0.002	842	0.899 ± 0.010	0.342 ± 0.008	1123
	SANT	0.689 ± 0.011	0.389 ± 0.072	281	0.738 ± 0.019	0.307 ± 0.044	562	0.761 ± 0.038	0.335 ± 0.058	842	0.780 ± 0.064	0.293 ± 0.010	1123
	QUORUM	$\mathbf{0.721 \pm 0.001}$	$\mathbf{0.394 \pm 0.008}$	**267**	0.748 ± 0.001	0.335 ± 0.005	**519**	$\mathbf{0.857 \pm 0.001}$	$\mathbf{0.364 \pm 0.003}$	**772**	$\mathbf{0.946 \pm 0.001}$	$\mathbf{0.400 \pm 0.002}$	**1025**
Global-MMLU (es) (LLM: 0.676) 14,300 samples	Random	0.710 ± 0.001	0.289 ± 0.011	267	0.788 ± 0.001	0.289 ± 0.004	519	0.826 ± 0.001	0.260 ± 0.002	772	0.853 ± 0.001	0.229 ± 0.002	1025
	ARAIDA	0.727 ± 0.013	0.358 ± 0.112	281	0.777 ± 0.013	0.275 ± 0.004	562	0.800 ± 0.042	0.278 ± 0.004	842	0.822 ± 0.069	0.280 ± 0.005	1123
	CDI	0.740 ± 0.001	0.295 ± 0.009	281	0.795 ± 0.001	0.281 ± 0.003	562	0.854 ± 0.002	0.286 ± 0.004	842	0.912 ± 0.001	0.288 ± 0.001	1123
	CoAnnotating	0.741 ± 0.001	0.303 ± 0.001	281	$\mathbf{0.802 \pm 0.001}$	0.303 ± 0.001	562	0.852 ± 0.001	0.283 ± 0.001	842	0.910 ± 0.001	0.284 ± 0.001	1123
	HyPAC	0.742 ± 0.001	0.313 ± 0.007	281	0.802 ± 0.001	$\mathbf{0.305 \pm 0.003}$	562	0.855 ± 0.001	0.289 ± 0.003	842	$\mathbf{0.915 \pm 0.001}$	0.292 ± 0.001	1123
	SANT	0.738 ± 0.009	0.324 ± 0.072	281	0.798 ± 0.011	0.298 ± 0.008	562	0.851 ± 0.035	$\mathbf{0.310 \pm 0.013}$	842	0.900 ± 0.058	$\mathbf{0.308 \pm 0.011}$	1123
	QUORUM	$\mathbf{0.749 \pm 0.002}$	$\mathbf{0.379 \pm 0.016}$	**267**	0.798 ± 0.002	0.290 ± 0.006	**519**	$\mathbf{0.856 \pm 0.002}$	0.290 ± 0.004	**772**	0.912 ± 0.001	0.287 ± 0.001	**1025**
		ROUGE-1	MCA	Cost	ROUGE-1	MCA	Cost	ROUGE-1	MCA	Cost	ROUGE-1	MCA	Cost
XLSum (ja) (LLM: 0.166) 880 samples	Random	0.243 ± 0.003	-	17	0.388 ± 0.006	-	32	0.532 ± 0.005	-	48	0.678 ± 0.001	-	64
	CDI	0.247 ± 0.001	-	18	0.381 ± 0.004	-	34	0.529 ± 0.004	-	53	0.673 ± 0.003	-	71
	CoAnnotating	0.250 ± 0.001	-	18	0.395 ± 0.001	-	34	0.541 ± 0.001	-	53	0.682 ± 0.001	-	71
	HyPAC	0.255 ± 0.004	-	18	0.367 ± 0.003	-	34	0.385 ± 0.001	-	53	0.405 ± 0.003	-	71
	QUORUM	$\mathbf{0.258 \pm 0.002}$	-	**17**	$\mathbf{0.403 \pm 0.001}$	-	**32**	$\mathbf{0.543 \pm 0.002}$	-	**48**	$\mathbf{0.683 \pm 0.002}$	-	**64**
XLSum (es) (LLM: 0.229) 4,763 samples	Random	0.301 ± 0.001	-	90	0.454 ± 0.001	-	176	0.609 ± 0.001	-	261	0.763 ± 0.001	-	347
	CDI	0.303 ± 0.001	-	95	0.456 ± 0.002	-	190	0.612 ± 0.002	-	285	0.768 ± 0.001	-	380
	CoAnnotating	0.304 ± 0.001	-	95	0.459 ± 0.001	-	190	0.614 ± 0.001	-	285	0.768 ± 0.001	-	380
	HyPAC	0.304 ± 0.001	-	95	0.460 ± 0.001	-	190	0.614 ± 0.001	-	285	0.768 ± 0.001	-	380
	QUORUM	$\mathbf{0.307 \pm 0.001}$	-	**90**	$\mathbf{0.461 \pm 0.001}$	-	**176**	$\mathbf{0.616 \pm 0.001}$	-	**261**	$\mathbf{0.771 \pm 0.001}$	-	**347**

Table 9: Performance in the DOLLARS setting under a fixed monetary budget. We report Accuracy or ROUGE-1 depending on the task (↑) and Machine Cumulative Accuracy (↑) under varying budget constraints. **Bold** indicates the best-performing method, while underlined the second best. Row values indicate the available annotation budget, expressed as a percentage of the cost required to annotate the entire dataset using human annotators.

Dataset	Method	10%		30%		50%		70%	
		Accuracy/ROUGE	MCA	Accuracy/ROUGE	MCA	Accuracy/ROUGE	MCA	Accuracy/ROUGE	MCA
Global-MMLU (ja) (LLM: 0.640) Human Cost: 1,404	Random	0.665 ± 0.002	0.313 ± 0.008	0.659 ± 0.002	0.343 ± 0.007	0.729 ± 0.002	0.245 ± 0.003	0.766 ± 0.002	0.243 ± 0.002
	ARAIDA	0.696 ± 0.028	0.354 ± 0.109	0.732 ± 0.018	0.312 ± 0.012	0.759 ± 0.052	0.315 ± 0.015	0.782 ± 0.082	0.318 ± 0.018
	CDI	0.726 ± 0.002	0.350 ± 0.008	0.751 ± 0.002	0.318 ± 0.006	0.823 ± 0.002	0.335 ± 0.004	0.893 ± 0.002	0.338 ± 0.002
	CoAnnotating	0.727 ± 0.001	0.353 ± 0.001	0.760 ± 0.001	0.348 ± 0.001	0.818 ± 0.001	0.323 ± 0.001	0.909 ± 0.001	0.361 ± 0.001
	HyPAC	0.749 ± 0.001	0.351 ± 0.004	0.784 ± 0.001	0.350 ± 0.002	$\mathbf{0.838 \pm 0.001}$	0.331 ± 0.002	0.899 ± 0.010	0.342 ± 0.008
	SANT	0.719 ± 0.021	0.381 ± 0.075	0.738 ± 0.019	0.307 ± 0.044	0.761 ± 0.038	$\mathbf{0.335 \pm 0.058}$	0.780 ± 0.064	0.293 ± 0.010
	QUORUM	$\mathbf{0.758 \pm 0.001}$	$\mathbf{0.393 \pm 0.003}$	$\mathbf{0.791 \pm 0.001}$	$\mathbf{0.405 \pm 0.005}$	0.808 ± 0.001	0.331 ± 0.003	$\mathbf{0.916 \pm 0.001}$	$\mathbf{0.382 \pm 0.002}$
Global-MMLU (es) (LLM: 0.676) Human Cost: 143	Random	0.612 ± 0.001	0.305 ± 0.044	0.613 ± 0.001	0.270 ± 0.043	0.635 ± 0.001	0.185 ± 0.038	0.624 ± 0.001	0.206 ± 0.014
	ARAIDA	0.712 ± 0.002	$\mathbf{0.844 \pm 0.002}$	0.714 ± 0.002	$\mathbf{0.848 \pm 0.002}$	0.716 ± 0.002	$\mathbf{0.557 \pm 0.002}$	0.718 ± 0.002	0.362 ± 0.002
	CDI	0.712 ± 0.001	0.844 ± 0.001	0.714 ± 0.001	0.848 ± 0.001	0.716 ± 0.001	0.557 ± 0.001	0.718 ± 0.001	0.362 ± 0.001
	CoAnnotating	$\mathbf{0.727 \pm 0.001}$	0.376 ± 0.001	0.742 ± 0.001	0.425 ± 0.001	0.771 ± 0.001	0.450 ± 0.001	0.818 ± 0.001	0.320 ± 0.001
	HyPAC	0.717 ± 0.001	0.286 ± 0.025	0.719 ± 0.001	0.298 ± 0.021	0.724 ± 0.001	0.301 ± 0.019	0.732 ± 0.001	0.303 ± 0.009
	SANT	0.712 ± 0.001	0.844 ± 0.001	0.714 ± 0.001	0.848 ± 0.001	0.716 ± 0.001	0.557 ± 0.001	0.718 ± 0.001	0.362 ± 0.001
	QUORUM	0.725 ± 0.001	0.749 ± 0.001	$\mathbf{0.806 \pm 0.001}$	0.745 ± 0.001	$\mathbf{0.808 \pm 0.001}$	0.471 ± 0.001	$\mathbf{0.829 \pm 0.001}$	$\mathbf{0.386 \pm 0.002}$
		ROUGE-1	MCA	ROUGE-1	MCA	ROUGE-1	MCA	ROUGE-1	MCA
XLSum (ja) (LLM: 0.166) Human Cost: 88	Random	0.234 ± 0.004	-	0.298 ± 0.006	-	0.432 ± 0.005	-	0.508 ± 0.001	-
	CDI	0.323 ± 0.003	-	0.381 ± 0.004	-	0.529 ± 0.004	-	0.673 ± 0.003	-
	CoAnnotating	0.326 ± 0.001	-	0.393 ± 0.001	-	0.521 ± 0.001	-	0.673 ± 0.001	-
	HyPAC	0.319 ± 0.004	-	0.367 ± 0.003	-	0.385 ± 0.001	-	0.405 ± 0.003	-
	QUORUM	$\mathbf{0.330 \pm 0.005}$	-	$\mathbf{0.395 \pm 0.001}$	-	$\mathbf{0.543 \pm 0.002}$	-	$\mathbf{0.683 \pm 0.002}$	-
XLSum (es) (LLM: 0.229) Human Cost: 476	Random	0.219 ± 0.001	-	0.222 ± 0.001	-	0.231 ± 0.001	-	0.246 ± 0.002	
	CDI	0.228 ± 0.002	-	0.231 ± 0.002	-	0.236 ± 0.002	-	0.246 ± 0.002	-
	CoAnnotating	0.226 ± 0.002	-	0.228 ± 0.002	-	0.234 ± 0.002	-	0.251 ± 0.002	-
	HyPAC	0.243 ± 0.002	-	0.249 ± 0.002	-	0.261 ± 0.002	-	0.284 ± 0.001	-
	QUORUM	$\mathbf{0.251 \pm 0.001}$	-	$\mathbf{0.261 \pm 0.001}$	-	$\mathbf{0.282 \pm 0.001}$	-	$\mathbf{0.308 \pm 0.001}$	-