

# HyPAC: Cost-Efficient LLMs–Human Hybrid Annotation with PAC Error Guarantees

Hao Zeng<sup>\*1</sup> Huipeng Huang<sup>\*1</sup> Xinhao Qu<sup>2</sup> Jianguo Huang<sup>3</sup> Bingyi Jing<sup>4,1</sup> Hongxin Wei<sup>1</sup>

## Abstract

Data annotation often involves multiple sources with different cost-quality trade-offs, such as fast large language models (LLMs), slow reasoning models, and human experts. In this work, we study the problem of routing inputs to the most cost-efficient annotation source while controlling the labeling error on test instances. We propose **HyPAC**, a method that adaptively labels inputs to the most cost-efficient annotation source while providing distribution-free guarantees on annotation error. HyPAC calibrates two decision thresholds using importance sampling and upper confidence bounds, partitioning inputs into three regions based on uncertainty and routing each to the appropriate annotation source. We prove that HyPAC achieves the minimum expected cost with a probably approximately correct (PAC) guarantee on the annotation error, free of data distribution and pre-trained models. Experiments on common benchmarks demonstrate the effectiveness of our method, reducing the annotation cost by 78.51% while tightly controlling the annotation error.

## 1. Introduction

Data annotation is an important step in building machine learning systems (Snow et al., 2008; Gligorić et al., 2024). In practice, an annotation system generally involves multiple sources (Qin et al., 2025; Wang et al., 2024) with different cost-quality trade-offs (Yue et al., 2023; Dekoninck et al., 2025), such as fast but less accurate large language models (LLMs), slow but more accurate reasoning models with extended inference (e.g., chain-of-thought prompting (Chen et al., 2025; Xu et al., 2025) or test-time compute (Snell

<sup>\*</sup>Equal contribution <sup>1</sup>Department of Statistics and Data Science, Southern University of Science and Technology, China

<sup>2</sup>Department of Statistics, University of California at Riverside, USA

<sup>3</sup>College of Computing and Data Science, Nanyang Technological University, Singapore

<sup>4</sup>School of Artificial Intelligence, The Chinese University of Hong Kong, Shenzhen, China. Correspondence to: Hongxin Wei <weihx@sustech.edu.cn>.

Preprint. February 4, 2026.

et al., 2024; Zhou et al., 2025)), and human annotators who provide the most accurate labels but at the highest cost (Li et al., 2023). Selecting the appropriate annotation source for each input is therefore crucial for achieving cost-efficient annotations while maintaining quality.

In the literature, there are many heuristic methods designed for *hybrid annotation* (Yue et al., 2023; Dekoninck et al., 2025) with the core idea: Route easy examples to fast, cheap models and send difficult examples to more accurate but expensive sources, such as stronger models or human annotators. However, existing hybrid labeling approaches often rely on heuristics to select annotation sources, without formal guarantees on the error rate (Li et al., 2023). This limitation raises a fundamental issue:

*How to design hybrid annotation that balances cost and quality with provable error control?*

In this paper, we introduce **HyPAC** (**H**ybrid annotation with **PAC** guarantees), a novel method for cost-efficient multi-source annotation with formal error guarantees. We formulate hybrid annotation as a constrained cost optimization problem: minimize annotation cost while controlling the error rate below a user-specified tolerance. HyPAC uses importance sampling and upper confidence bounds to set decision thresholds for routing examples to different labeling sources (weak models, strong models, or humans), ensuring error control with high probability while achieving the minimum expected annotation cost. This provides practitioners with a principled way to optimize the cost-quality trade-off across the full spectrum of available annotators.

Theoretically, we provide formal guarantees for HyPAC in terms of both error control and cost efficiency. First, we establish that HyPAC satisfies the  $(\epsilon, \alpha)$ -PAC annotation guarantee (Definition 3.1, Theorem 4.3): with probability at least  $1 - \alpha$ , the selected thresholds control the error rate below the user-specified tolerance  $\epsilon$ , and this guarantee is distribution-free. Second, we show that HyPAC achieves the minimum cost among all threshold-based methods that satisfy the same PAC constraint (Theorem 4.4). Both results rely on the monotonicity of risk and cost functions in the threshold parameters (Lemmas 4.1 and 4.2), which enables fixed-sequence testing without multiple testing correction.

We evaluate the effectiveness of HyPAC across diverse benchmark datasets including general tasks (MMLU-Redux (Gema et al., 2025)), math reasoning tasks (MATH-500 (Lightman et al., 2023), MATH-L5 (Hendrycks et al., 2021), and Zebra-Logic (Lin et al., 2025)), and coding tasks (HumanEval<sup>+</sup> (Liu et al., 2023)). The results demonstrate that HyPAC significantly reduces the annotation cost compared to expert-only annotation while achieving the target error control. For example, on MATH-500 with Qwen3-4B-Instruct-2507 (Yang et al., 2025) as the non-thinking model and Qwen3-4B-Thinking-2507 as the thinking model, HyPAC reduces the annotation cost by 78.51%, while controlling the annotation error below 5%. Through ablation studies, we show that HyPAC outperforms baseline methods in cost and labeling error, the logits-based uncertainty score is more reliable than the verbalized uncertainty score, and our method is effective under various cost functions and UCBs. We provide the code for reproducing our main experiments in this anonymous repository: <https://anonymous.4open.science/r/HyPAC-B5AD>.

Our contributions are summarized:

1. We formulate hybrid labeling as a **constrained cost optimization problem** with multiple labeling sources, and provide a provable framework for controlling the labeling error rate.
2. We propose **HyPAC**, a novel practical method that calibrates decision thresholds using importance sampling and upper confidence bounds. HyPAC is agnostic to the score function, supports **input-dependent costs**, and could **extend to multiple labeling sources**.
3. We provide **distribution-free guarantees** on both error control and cost optimality: the error rate stays below a user-specified tolerance with high probability, and HyPAC achieves the minimum cost among all threshold-based methods with the same constraint.

**Notation.** Let  $\mathcal{X}$  denote the input space and  $\mathcal{Y}$  denote the label space. We denote by  $\mathcal{D}$  the distribution of  $(x, y)$ . We use  $\tilde{f}_1 : \mathcal{X} \rightarrow \mathcal{Y}$  for the non-thinking model,  $\tilde{f}_2 : \mathcal{X} \rightarrow \mathcal{Y}$  for the thinking model, and  $y$  for the human annotation (ground truth). The cost functions of these three sources are denoted by  $c_1(x)$ ,  $c_2(x)$ , and  $c_h(x)$ , respectively, where each cost may depend on the input  $x$ . The score function  $U : \mathcal{X} \rightarrow [0, 1]$  measures the uncertainty of the annotation system for input  $x$ . The calibration set is denoted by  $\mathcal{S}_{\text{cal}} = \{(x_i, y_i)\}_{i=1}^m$ . When  $\mathcal{S}_{\text{cal}}$  appears under an expectation or probability operator, it refers to the randomness induced by the calibration dataset.

## 2. Hybrid annotation

In this section, we present the problem of hybrid annotation for large language models and introduce the key concepts required for our method. While our framework applies to general multi-source settings, we focus on a three-source setup in this paper for clarity: a non-thinking model, a thinking model, and human annotation. The framework can be easily extended to more sources (see Appendix C).

We consider a *hybrid annotation* task where input samples  $x \in \mathcal{X}$  need to be labeled with ground truth label  $y \in \mathcal{Y}$ , i.e., human annotation. In this paper, we use non-thinking models as examples of fast models and thinking models as examples of slow models:

- **Non-thinking model**  $\tilde{f}_1 : \mathcal{X} \rightarrow \mathcal{Y}$ : a fast LLM using standard inference without extended reasoning, which is cheap but less accurate. For example, Qwen3-4B-Instruct generates responses quickly with lower computational cost.
- **Thinking model**  $\tilde{f}_2 : \mathcal{X} \rightarrow \mathcal{Y}$ : a slow but more powerful LLM using extended reasoning (e.g., chain-of-thought), which is more accurate but more expensive. For example, Qwen3-4B-Thinking produces longer reasoning chains and requires more tokens per response.
- **Human annotation**  $y$ : labels from human experts. In most cases, we treat human labels as ground truth.

The goal of hybrid annotation is to design a decision rule  $\hat{f}$  that selects the annotation source for each input  $x$ , such that the overall annotation error rate is controlled while minimizing the expected annotation cost. Formally, we define the annotation error as

$$R(\hat{f}) = \mathbb{E}_{(x,y) \sim \mathcal{D}} [\ell(y, \hat{f}(x))], \quad (1)$$

where  $\ell : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}_+$  is a loss function (e.g., 0-1 loss for classification, semantic loss for text generation). Given a cost function  $C(\hat{f}, x)$ , the goal is to solve:

$$\min_{\hat{f}} \mathbb{E}_{x \sim \mathcal{D}_x} C(\hat{f}, x) \text{ s.t. } R(\hat{f}) \leq \epsilon, \quad (2)$$

where  $\epsilon$  is the target error threshold. We assume the quality of annotation sources follows: human annotation has the highest quality, followed by the thinking model, then the non-thinking model.

However, existing hybrid labeling methods often rely on heuristics for routing decisions (Li et al., 2023; Wang et al., 2024; Li, 2024; Yuan et al., 2025), without formal guarantees on the resulting error rate, which may lead to unpredictable performance in downstream tasks. To address this challenge, we formalize the hybrid annotation problem with strict error rate control, providing formal guarantees on annotation quality while minimizing cost and enabling practitioners to meet pre-specified error thresholds reliably.

### 3. HyPAC

In this section, we present **Hybrid** annotation with **PAC** guarantees (HyPAC), our method for cost-efficient hybrid annotation with rigorous labeling quality control guarantees. We first define the PAC annotation problem and show how to express it as a constrained optimization over threshold parameters. We then present our threshold calibration method, which uses importance sampling and upper confidence bounds to identify valid threshold pairs that satisfy the PAC constraint while minimizing annotation cost.

#### 3.1. PAC annotation

First, we consider a score  $U : \mathcal{X} \rightarrow [0, 1]$  that measures the uncertainty of the annotation system for input  $x$ . Given two thresholds  $(u_1, u_2)$  with  $0 \leq u_1 \leq u_2 \leq 1$ , we define the decision rule  $\hat{f} = T_{u_1, u_2} : \mathcal{X} \rightarrow \mathcal{Y}$  as:

$$T_{u_1, u_2}(x) = \begin{cases} \tilde{f}_1(x) & \text{if } U(x) \leq u_1 \\ \tilde{f}_2(x) & \text{if } u_1 < U(x) \leq u_2 \\ y & \text{if } U(x) > u_2 \end{cases}$$

Figure 1 illustrates this three-tier routing mechanism.

Then we rewrite its annotation error as

$$R(\hat{f}) = R(u_1, u_2) = \mathbb{E}_{(x, y) \sim \mathcal{D}} [\ell(y, T_{u_1, u_2}(x))]$$

For cost, let  $c_1(x)$ ,  $c_2(x)$ , and  $c_h(x)$  be the cost functions for using the non-thinking model, thinking model, and human annotation on input  $x$ , respectively. In practice, these costs often depend on the input, for example, the number of tokens generated by the model. Under costs of sources  $c_1(x)$ ,  $c_2(x)$ , and  $c_h(x)$ , the cost function is:

$$\begin{aligned} C(u_1, u_2, x) = & c_1(x) \cdot \mathbf{1}\{U(x) \leq u_1\} \\ & + c_2(x) \cdot \mathbf{1}\{u_1 < U(x) \leq u_2\} \\ & + c_h(x) \cdot \mathbf{1}\{U(x) > u_2\} \end{aligned}$$

We assume that the costs satisfy  $c_1(x) \leq c_2(x) \leq c_h(x)$  for almost all  $x$ , reflecting the cost hierarchy of the three sources. Then the goal Eq. (2) becomes,

$$\begin{aligned} \min_{u_1, u_2} & \mathbb{E}_{x \sim \mathcal{D}_x} [C(u_1, u_2, x)] \\ \text{subject to} & R(u_1, u_2) \leq \epsilon, \quad 0 \leq u_1 \leq u_2 \leq 1 \end{aligned} \tag{3}$$

Directly solving this optimization Eq. (3) is impossible because we do not know the true risk  $R(u_1, u_2)$  without access to the ground truth labels  $y$  for all inputs. Thus, we consider a relaxed form where the constraint holds in a PAC style. Given a calibration set  $\mathcal{S}_{\text{cal}}$ , we seek thresholds  $(\hat{u}_1, \hat{u}_2)$  that minimize the expected cost  $\mathbb{E}_{x \sim \mathcal{D}_x} [C(\hat{u}_1, \hat{u}_2, x)]$  while satisfying the  $(\epsilon, \alpha)$ -PAC annotation as follows.

**Definition 3.1**  $((\epsilon, \alpha)\text{-PAC}$  annotation). An annotation procedure using thresholds  $(u_1, u_2)$  is  $(\epsilon, \alpha)$ -PAC if

$$\mathbb{P}_{\mathcal{S}_{\text{cal}}} (R(u_1, u_2) \leq \epsilon) \geq 1 - \alpha, \tag{4}$$

where the risk  $R(u_1, u_2)$  is evaluated on the test distribution  $\mathcal{D}$ , and the probability is over the calibration set  $\mathcal{S}_{\text{cal}}$ .

Since multiple threshold pairs may satisfy this condition, we call the set of all such pairs the  $(\epsilon, \alpha)\text{-PAC feasible region}$ . It is inspired by the PAC efficient (Zeng et al., 2025), which considers a two-level reasoning system with fast and slow models. We extend it to multiple annotators with a human annotator. We aim to find thresholds that satisfy the PAC guarantee while minimizing the annotation cost:

$$\begin{aligned} \min_{u_1, u_2} & \mathbb{E}_{x \sim \mathcal{D}_x} [C(u_1, u_2, x)] \\ \text{s.t.} & \mathbb{P}_{\mathcal{S}_{\text{cal}}} (R(u_1, u_2) \leq \epsilon) \geq 1 - \alpha, \\ & 0 \leq u_1 \leq u_2 \leq 1. \end{aligned} \tag{5}$$

This framework naturally extends to more than three annotation sources by adding more thresholds to partition the confidence space into more regions, where each region is routed to a different annotation source. We discuss this extension in detail in Appendix C.

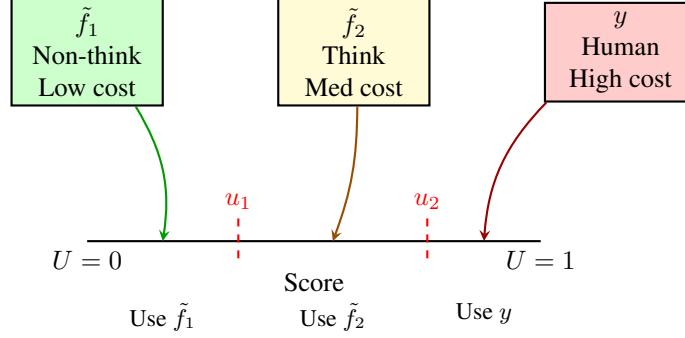
#### 3.2. Threshold calibration

The main challenge in solving this problem (5) is that we cannot directly check whether a threshold pair  $(u_1, u_2)$  satisfies the PAC condition, because the true risk  $R(u_1, u_2)$  is unknown without access to ground truth labels for all inputs. To address this, we use a valid upper confidence bound (UCB) on the risk: a bound that holds with high probability can serve as a substitute for the true risk to identify valid thresholds. In light of this, our method follows a two-step approach: first, we identify all threshold pairs that lie in the  $(\epsilon, \alpha)$ -PAC feasible region using the UCB; then, among valid thresholds, we select the one with the lowest cost.

**Importance sampling for unbiased risk estimation.** We use importance sampling (Owen & Zhou Associate, 2000; Tokdar & Kass, 2010) to obtain unbiased risk estimates. For each example  $x_i$  in the calibration set, we query the ground truth label  $y_i$  with probability  $p_i$ . The importance-weighted risk estimator is:

$$\hat{R}_{\text{IS}}(u_1, u_2) = \frac{1}{m} \sum_{i=1}^m \frac{Z_i}{p_i} \cdot \ell(y_i, T_{u_1, u_2}(x_i)) \tag{6}$$

where  $Z_i \sim \text{Bernoulli}(p_i)$  indicates whether we queried the ground truth for example  $i$ . This estimator is unbiased:  $\mathbb{E}[\hat{R}_{\text{IS}}(u_1, u_2)] = R(u_1, u_2)$ . In practice, we often use uniform sampling with  $p_i = p$  for all  $i$ . The choice of  $p$  involves



**Figure 1. Hybrid annotation with three sources and two thresholds.** The routing rule uses score  $U(x)$  to select the appropriate annotation source: when  $U(x) \leq u_1$ , use the non-thinking model  $\tilde{f}_1$ ; when  $u_1 < U(x) \leq u_2$ , use the thinking model  $\tilde{f}_2$ ; when  $U(x) > u_2$ , use human annotation.

a trade-off: a larger  $p$  gives more accurate estimates but requires more ground truth queries, while a smaller  $p$  reduces calibration cost but increases variance.

**UCB construction.** Given the importance-weighted risk estimates, we construct an upper confidence bound (UCB) for each threshold pair  $(u_1, u_2)$ . Define the importance-weighted loss for each sample as:

$$W_i(u_1, u_2) = \frac{Z_i}{p_i} \cdot \ell(y_i, T_{u_1, u_2}(x_i))$$

so that  $\hat{R}_{\text{IS}}(u_1, u_2) = \frac{1}{m} \sum_{i=1}^m W_i(u_1, u_2)$ . Since  $\{W_i\}_{i=1}^m$  are i.i.d. random variables with  $\mathbb{E}[W_i] = R(u_1, u_2)$ , by the Central Limit Theorem (CLT), the sample mean is asymptotically normal:

$$\frac{\hat{R}_{\text{IS}}(u_1, u_2) - R(u_1, u_2)}{\hat{\sigma}_W(u_1, u_2)/\sqrt{m}} \xrightarrow{d} \mathcal{N}(0, 1)$$

where  $\hat{\sigma}_W(u_1, u_2)$  is the sample standard deviation of  $\{W_i\}$ . This yields the CLT-based UCB:

$$\hat{L}_{u_1, u_2}(\alpha) = \hat{R}_{\text{IS}}(u_1, u_2) + z_{1-\alpha} \frac{\hat{\sigma}_W(u_1, u_2)}{\sqrt{m}} \quad (7)$$

where  $z_{1-\alpha}$  is the  $(1 - \alpha)$ -quantile of the standard normal distribution. Other valid choices (e.g., Hoeffding-type or empirical Bernstein bounds) can be used to achieve distribution-free finite-sample guarantees.

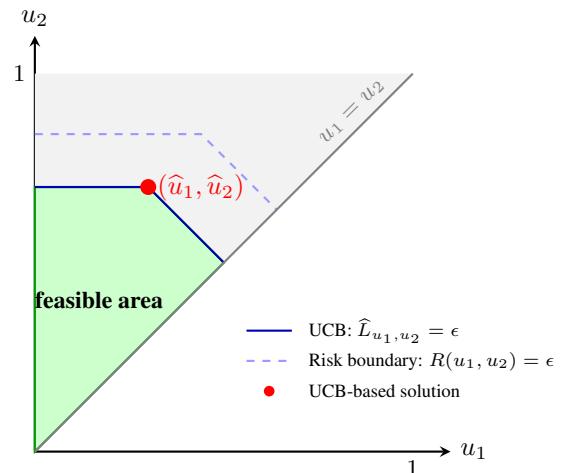
**Threshold selection.** Given the UCBs for all threshold pairs, we select the pair that minimizes cost while ensuring the error rate is controlled. We estimate the expected cost using the empirical cost on the calibration set:

$$\hat{C}(u_1, u_2) = \frac{1}{m} \sum_{i=1}^m C(u_1, u_2, x_i)$$

The optimization problem then becomes:

$$\begin{aligned} (\hat{u}_1, \hat{u}_2) &= \arg \min_{u_1, u_2} \hat{C}(u_1, u_2) \\ \text{s.t. } &\hat{L}_{u_1, u_2}(\alpha) \leq \epsilon \\ &0 \leq u_1 \leq u_2 \leq 1 \end{aligned} \quad (8)$$

We then solve the optimization problem in Eq. (8) by searching over a given threshold grid. This calibration procedure ensures that the selected thresholds control the error rate on future test data with high probability, while minimizing the annotation cost. We illustrate it in Figure 2.



**Figure 2. Illustration of UCB-based threshold selection in the  $(u_1, u_2)$  space.** The green region shows the  $(\epsilon, \alpha)$ -PAC feasible region where  $\hat{L}_{u_1, u_2}(\alpha) \leq \epsilon$  and  $u_1 \leq u_2$ . The solid blue line is the UCB boundary, and the dashed blue line is the true risk boundary. Since the UCB upper bounds the true risk, the UCB boundary lies inside the true risk boundary. The red point marks the optimal solution selected by HyPAC, which maximizes thresholds (i.e., minimizing cost) within the feasible region.

**Deployment.** Once we have calibrated the thresholds  $(\hat{u}_1, \hat{u}_2)$ , the deployment phase is straightforward. For each

**Algorithm 1** HyPAC

---

**Input:** Calibration set  $\mathcal{S}_{\text{cal}} = \{x_i\}_{i=1}^m$ , test example  $x$ , models  $\tilde{f}_1, \tilde{f}_2$ , score function  $U$ , error tolerance  $\epsilon$ , significance level  $\alpha$ , sampling probability  $p$ , threshold grid  $\mathcal{G}$

**Output:** Label  $\hat{y}$  for test example  $x$

```

1: // Calibration phase
2: for  $i = 1$  to  $m$  do
3:   Compute score  $U(x_i)$  and predictions  $\tilde{f}_1(x_i), \tilde{f}_2(x_i)$ 
4:   Sample  $Z_i \sim \text{Bernoulli}(p)$ ; if  $Z_i = 1$ , query ground truth  $y_i$ 
5: end for
6: for each  $(u_1, u_2) \in \mathcal{G}$  with  $u_1 \leq u_2$  do
7:   Compute importance-weighted risk estimate  $\hat{R}_{\text{IS}}(u_1, u_2)$  via Eq. (6)
8:   Compute UCB  $\hat{L}_{u_1, u_2}(\alpha)$  via Eq. (7)
9: end for
10: Compute  $(\hat{u}_1, \hat{u}_2)$  via Eq. (8)
11: // Deployment phase
12: Compute score  $U(x)$ 
13: Return  $T_{\hat{u}_1, \hat{u}_2}(x)$ 

```

---

test example  $x$ , we compute the score  $U(x)$  and route the example to the appropriate annotation source based on the calibrated thresholds.

The complete procedure, including both calibration and deployment phases, is summarized in Algorithm 1. Next, we provide a theoretical analysis showing that this method achieves PAC annotation guarantees and cost optimality.

## 4. Theoretical analysis

In this section, we provide a theoretical analysis of the HyPAC method. We first show that both the risk and cost functions are monotone in the threshold parameters (Section 4.1), which is key to establishing the PAC guarantee and cost optimality. We then present our main results (Section 4.2): a PAC guarantee showing that the selected thresholds control the error rate (Theorem 4.3), i.e., PAC annotation, and a cost optimality result showing that HyPAC achieves the minimum cost among all PAC-valid solutions (Theorem 4.4).

### 4.1. Monotonicity

We now prove that the risk function  $R(u_1, u_2)$  is monotone in both threshold parameters. This result is crucial: it implies that to minimize cost while controlling risk, we should select the largest thresholds within the PAC feasible region.

**Lemma 4.1** (Risk monotonicity). *The risk function*

$$R(u_1, u_2) = \mathbb{E}_{(x,y) \sim \mathcal{D}} [\ell(y, T_{u_1, u_2}(x))]$$

*is monotone non-decreasing in both  $u_1$  and  $u_2$ . Specifically,*

*for any fixed  $u_2$  and  $u_1 < u'_1$  with  $u'_1 \leq u_2$ , we have  $R(u_1, u_2) \leq R(u'_1, u_2)$ . Similarly, for any fixed  $u_1$  and  $u_2 < u'_2$ , we have  $R(u_1, u_2) \leq R(u_1, u'_2)$ .*

The complete proof is given in Appendix A.1. The monotonicity property ensures that the feasible region  $\{(u_1, u_2) : R(u_1, u_2) \leq \epsilon, 0 \leq u_1 \leq u_2 \leq 1\}$  has a nested structure: If  $(u_1, u_2)$  is feasible, then any  $(u'_1, u'_2)$  with  $u'_1 \leq u_1$  and  $u'_2 \leq u_2$  is also feasible. This nested structure enables fixed-sequence testing without multiple testing correction, as we show in the proof of Theorem 4.3.

**Lemma 4.2** (Cost monotonicity). *Assuming the costs satisfy  $c_1(x) \leq c_2(x) \leq c_h(x)$  (cost ordering), we have: Both the expected cost  $\mathbb{E}_{x \sim \mathcal{D}_x}[C(u_1, u_2, x)]$  and the empirical cost  $\hat{C}(u_1, u_2) = \frac{1}{m} \sum_{i=1}^m C(u_1, u_2, x_i)$  (for any finite sample  $\{x_i\}_{i=1}^m$ ) are monotone non-increasing in both  $u_1$  and  $u_2$ . Furthermore, if  $c_1(x) < c_2(x) < c_h(x)$  strictly for almost all  $x$ , both the expected cost and empirical cost are strictly decreasing in both  $u_1$  and  $u_2$ .*

The proof is given in Appendix A.2. These two monotonicity results have important practical implications. Risk monotonicity (Lemma 4.1) tells us that larger thresholds lead to higher risk (more errors). Cost monotonicity (Lemma 4.2) tells us that larger thresholds lead to lower cost (more examples routed to cheaper sources).

### 4.2. PAC guarantee and cost optimal

We now state our main results. We first introduce the key assumption on the UCB construction.

**Assumption 1** (UCB validity). For each threshold pair  $(u_1, u_2)$  with  $0 \leq u_1 \leq u_2 \leq 1$  and any significance level  $\alpha \in (0, 1)$ , the upper confidence bound  $\hat{L}_{u_1, u_2}(\alpha)$  satisfies:

$$\mathbb{P}_{\mathcal{S}_{\text{cal}}} \left( R(u_1, u_2) \leq \hat{L}_{u_1, u_2}(\alpha) \right) \geq 1 - \alpha$$

where the probability is taken over the random draw of the calibration set  $\mathcal{S}_{\text{cal}}$ .

This assumption requires UCB to provide valid coverage for the true risk. In Section 3.2, we construct such a UCB. The validity of this UCB can be established using various methods, such as the central limit theorem (CLT), Hoeffding's inequality, or Bernstein bounds; see Appendix B. Under the assumption, we could have the validity of HyPAC:

**Theorem 4.3** (PAC guarantee). *Let  $(\hat{u}_1, \hat{u}_2)$  be the thresholds selected by Algorithm 1 with error tolerance  $\epsilon$  and significance level  $\alpha$ . Under Assumption 1, and assuming independent calibration/test sets and fixed pre-trained models  $\tilde{f}_1, \tilde{f}_2$ , we have:*

$$\mathbb{P}_{\mathcal{S}_{\text{cal}}} (R(\hat{u}_1, \hat{u}_2) \leq \epsilon) \geq 1 - \alpha$$

where the probability is taken over the random draw of the calibration set  $\mathcal{S}_{\text{cal}}$ .

**Table 1.** HyPAC achieves higher cost savings than both PAC labeling and CSE baselines while maintaining controlled annotation error. Results for annotation error and cost savings are reported at  $\epsilon = 0.05$  and  $\alpha = 0.05$  using the token-based cost function. The non-thinking model  $\tilde{f}_1$  is Qwen3-4B-Instruct-2507, and the thinking model  $\tilde{f}_2$  is Qwen3-4B-Thinking-2507. HyPAC is compared against three baselines: (1) PAC labeling with non-thinking models and experts ( $\tilde{f}_1 + \text{Human}$ ), (2) PAC labeling with thinking models and experts ( $\tilde{f}_2 + \text{Human}$ ), and (3) CSE. **Bold** numbers indicate superior results.

Datasets	Metric	Logits-based Score				Verbalized Score			
		$\tilde{f}_1 + \tilde{f}_2 + \text{Human}$	$\tilde{f}_1 + \text{Human}$	$\tilde{f}_2 + \text{Human}$	CSE	$\tilde{f}_1 + \tilde{f}_2 + \text{Human}$	$\tilde{f}_1 + \text{Human}$	$\tilde{f}_2 + \text{Human}$	CSE
MMLU-Redux	Error (%)	$4.09 \pm 0.94$	$4.00 \pm 1.20$	$3.97 \pm 1.34$	$3.96 \pm 4.70$	$2.97 \pm 0.47$	$2.02 \pm 0.14$	$0.60 \pm 0.08$	$3.37 \pm 2.58$
	Save (%)	<b><math>49.40 \pm 5.11</math></b>	$48.97 \pm 7.36$	$40.70 \pm 7.72$	$28.01 \pm 32.84$	<b><math>39.27 \pm 0.38</math></b>	$37.64 \pm 0.71$	$-6.03 \pm 0.22$	$36.06 \pm 14.34$
MATH-500	Error (%)	$4.00 \pm 1.71$	$3.39 \pm 1.77$	$3.09 \pm 1.54$	$4.14 \pm 4.45$	$4.22 \pm 2.99$	$3.41 \pm 1.70$	$1.75 \pm 2.20$	$4.40 \pm 1.62$
	Save (%)	<b><math>78.51 \pm 4.69</math></b>	$70.85 \pm 12.04$	$67.20 \pm 16.70$	$43.38 \pm 51.14$	<b><math>76.94 \pm 9.34</math></b>	$66.22 \pm 21.17$	$16.62 \pm 44.97$	$74.51 \pm 1.98$
MATH-L5	Error (%)	$4.17 \pm 1.34$	$3.46 \pm 1.74$	$3.39 \pm 1.56$	$4.20 \pm 3.96$	$4.35 \pm 8.30$	$2.88 \pm 1.41$	$1.78 \pm 1.66$	$4.32 \pm 4.89$
	Save (%)	<b><math>72.55 \pm 4.35</math></b>	$62.98 \pm 15.22$	$43.49 \pm 25.19$	$37.50 \pm 46.17$	<b><math>70.94 \pm 15.63</math></b>	$67.64 \pm 22.32$	$21.83 \pm 24.00$	$70.66 \pm 32.48$
Zebra-Logic	Error (%)	$4.48 \pm 1.30$	$3.86 \pm 1.65$	$3.63 \pm 1.34$	$4.58 \pm 2.62$	$4.47 \pm 6.27$	$1.34 \pm 2.22$	$0.14 \pm 0.15$	$4.34 \pm 1.69$
	Save (%)	<b><math>46.10 \pm 3.75</math></b>	$43.76 \pm 6.34$	$45.39 \pm 8.15$	$36.83 \pm 22.26$	<b><math>36.03 \pm 8.34</math></b>	$1.11 \pm 19.50$	$-22.17 \pm 0.44$	$5.56 \pm 15.42$
HumanEval <sup>+</sup>	Error (%)	$3.25 \pm 1.65$	$3.39 \pm 3.09$	$2.42 \pm 1.69$	$3.44 \pm 3.34$	$3.36 \pm 1.74$	$2.28 \pm 3.24$	$1.69 \pm 1.97$	$3.28 \pm 1.77$
	Save (%)	<b><math>86.48 \pm 5.35</math></b>	$62.13 \pm 23.17$	$75.26 \pm 24.46$	$67.51 \pm 21.39$	$70.55 \pm 34.50$	$30.49 \pm 42.60$	$49.97 \pm 43.01$	<b><math>87.46 \pm 1.63</math></b>

The complete proof is given in Appendix A.3. This theorem provides a rigorous guarantee that HyPAC controls the error rate with high probability. The guarantee is **distribution-free**: it holds for any data distribution  $\mathcal{D}$  and any pre-trained models  $\tilde{f}_1, \tilde{f}_2$ , without assumptions on the data structure or model quality. We also provide a finite-sample guarantee for the empirical risk; see Appendix A.5.

Beyond error control, HyPAC also achieves cost optimality among all methods that satisfy the PAC constraint.

**Theorem 4.4** (Cost optimality). *Let  $\mathcal{S}_{\text{cal}}$  be a calibration set and  $\mathcal{F}(\mathcal{S}_{\text{cal}}) = \{(u_1, u_2) : \hat{L}_{u_1, u_2}(\alpha) \leq \epsilon, 0 \leq u_1 \leq u_2 \leq 1\}$  be the PAC feasible region based on  $\mathcal{S}_{\text{cal}}$ . Assuming the cost ordering  $c_1(x) \leq c_2(x) \leq c_h(x)$ , the threshold pair  $(\hat{u}_1, \hat{u}_2)$  selected by Algorithm 1 satisfies:*

$$\mathbb{E}_{x \sim \mathcal{D}_x}[C(\hat{u}_1, \hat{u}_2, x)] = \min_{(u_1, u_2) \in \mathcal{F}(\mathcal{S}_{\text{cal}})} \mathbb{E}_{x \sim \mathcal{D}_x}[C(u_1, u_2, x)],$$

The proof is given in Appendix A.4. This result shows that HyPAC is not only a PAC annotation but also cost-optimal given the calibration data.

## 5. Experiments

In this section, we evaluate HyPAC on multiple benchmark datasets. Our experiments aim to answer two questions: (1) whether HyPAC achieves higher cost savings compared to existing provable methods with the same error control budgets; (2) whether HyPAC provides reliable error control compared to heuristic baselines.

### 5.1. Experimental setup

**Models.** We evaluate HyPAC on diverse open-source LLMs across different parameter scales. For the main experiments, we employ Qwen3-4B-Instruct-2507 (Yang et al., 2025) as the non-thinking model and Qwen3-4B-

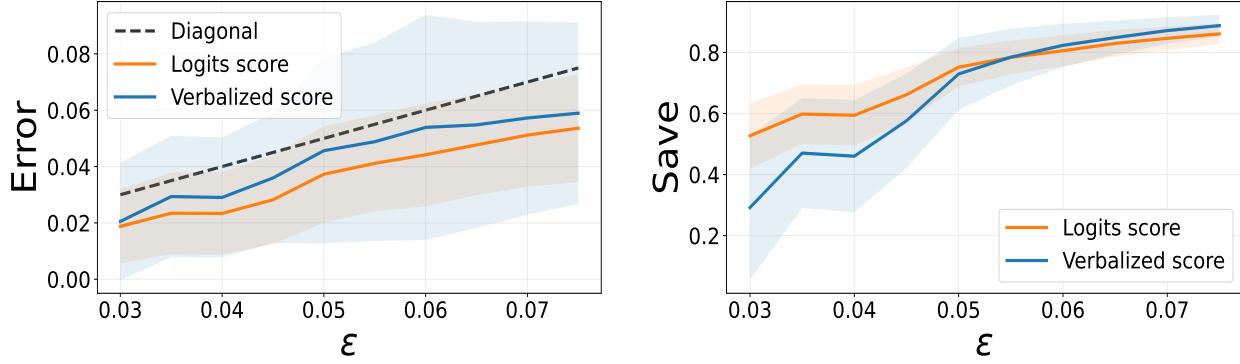
Thinking-2507 as the thinking model. In addition, we evaluate HyPAC using two alternative model pairs. Specifically, we use Llama-3.1-8B-Instruct (Grattafiori et al., 2024) as the non-thinking model with DeepSeek-R1-Distill-Llama-8B (DeepSeek-AI et al., 2025) as the thinking model, and Qwen2.5-32B-Instruct (Qwen et al., 2025) as the non-thinking model with DeepSeek-R1-Distill-Qwen-32B as the thinking model. The sampling temperature and other hyperparameters for LLMs are detailed in Appendix F.

**Uncertainty scores.** We adopt two complementary score functions: a logits-based score for the white-box case and a verbalized score for the black-box case. In particular, let  $y = (y_1, y_2, \dots, y_l)$  be a generated sequence of length  $l$ . We define the logits-based score as:

$$U_{\text{logits}}(x) = 1 - \frac{1}{l} \sum_{j=1}^l \mathbb{P}(y_j \mid y_1, \dots, y_{j-1}, x)$$

where  $\mathbb{P}(y_j \mid y_1, \dots, y_{j-1}, x)$  is the conditional probability of token  $y_j$ . This score measures the average confidence of the model across all generated tokens. For the verbalized score, we let the model explicitly state its self-reported confidence. In this study, we report the average confidence over 10 trials, and the prompts are listed in Table 4.

**Datasets.** We conduct experiments on general tasks, math and text reasoning tasks, and coding tasks, using MMLU-Redux (Gema et al., 2025), MATH-500 (Lightman et al., 2023), MATH-L5 (Hendrycks et al., 2021), Zebra-Logic (Lin et al., 2025), and HumanEval<sup>+</sup> (Liu et al., 2023) datasets. Table 2 reports the accuracy of each LLM on these datasets. For each dataset, we partition the original test set into a calibration set and a test set randomly. Table 3 provides details on the specific splitting strategies.



**Figure 3. The logits-based uncertainty score achieves better cost-error tradeoff than the verbalized score.** Error and cost savings of HyPAC with different uncertainty score functions under a confidence level of  $\alpha = 0.05$ . Experiments are conducted on MATH-500. The shaded areas represent standard deviations.

**Cost functions.** We consider two cost functions: token-based and API-based. For the token-based cost function, the cost of annotating a sample is defined as the number of tokens generated by the LLM. The expert annotation cost is set to the maximum allowed new tokens, which is 32,768. For the API cost function, the annotation cost is computed:

$$c_{\text{API}}^{\tilde{f}} = p_{\text{in}}^{\tilde{f}} \cdot n_{\text{in}} + p_{\text{out}}^{\tilde{f}} \cdot n_{\text{out}},$$

where  $n_{\text{in}}$  and  $n_{\text{out}}$  denote the numbers of input and generated tokens, respectively, and  $p_{\text{in}}^{\tilde{f}}$  and  $p_{\text{out}}^{\tilde{f}}$  are the model-specific per-token prices. For expert annotation, we adopt the thinking model’s pricing, and set  $n_{\text{out}}$  to 32,768. The exact per-token prices are provided in Appendix F.

**Baselines.** We compare HyPAC against four baselines: (1) PAC labeling (Candès et al., 2025) with non-thinking models and experts, (2) PAC labeling with thinking models and experts, (3) cascaded selective evaluation (CSE) (Jung et al., 2024), (4) CoAnnotating (Li et al., 2023). We provide a detailed introduction to these baselines in Appendix E.

**Evaluation metrics.** We evaluate the annotation performance by measuring the following metrics: **annotation error** defined in Eq. (1), and **cost savings**, defined as the relative reduction in annotation cost achieved by the hybrid annotation methods compared to expert-only annotation:

$$\left(1 - \frac{\sum_{i=1}^{n_{\text{test}}} C(\hat{f}, x_i)}{\sum_{i=1}^{n_{\text{test}}} C(y, x_i)}\right) \times 100\%.$$

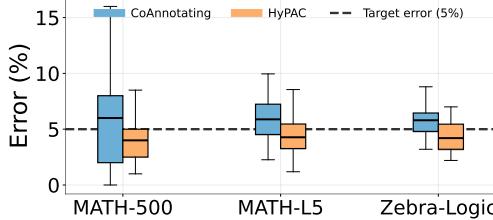
**Implementation Details** Across all experiments, we fix the confidence level at  $\alpha = 0.05$  and the sampling weight at  $p = 0.9$ , while varying  $\epsilon$ . Each experiment is repeated 100 times, and we report the mean annotation error and cost savings. Unless otherwise specified, we use the CLT-based UCB, the token-based cost function, and the binary 0–1 loss. More details of implementation are provided in Appendix F.

## 5.2. Results

**HyPAC tightly controls annotation error and reduces annotation costs compared with other provable methods.** Table 1 compares HyPAC with baseline methods on five annotation tasks. HyPAC keeps the error below  $\epsilon$  across all benchmarks while saving more costs. For example, on HumanEval<sup>+</sup> with  $\epsilon = 0.05$  and  $\alpha = 0.05$  using logits-based scores, HyPAC has an average error of 3.47% and cost savings of 86.69%. In contrast, PAC labeling baselines save 60.89% and 72.86% of costs, respectively. HyPAC increases cost savings by 13% over the best baseline, showing that using both models is better for saving costs. We see similar high savings on other datasets and score functions. HyPAC provides a balance between cost and accuracy in most cases. Appendix G.6 shows results for two other model pairs.

**HyPAC controls the annotation error with different uncertainty scores and is more stable in the logits-based score.** HyPAC works with any uncertainty score and ensures error control while reducing costs regardless of the score function used. Figure 3 compares logits-based and verbalized scores across different error tolerances  $\epsilon$ . Both scores achieve error control, but they show different levels of stability. The logits-based score shows smoother and more stable results, with lower standard deviations in both error and cost savings. Reliability diagrams in Figure 9 in Appendix G.7 show that logits-based scores have lower expected calibration error (Naeini et al., 2015), suggesting better reliability. In line with these findings, Table 1 shows that the logits-based score saves more costs on most datasets. These results show that while HyPAC works with various scores, the choice of score impacts performance stability.

**HyPAC provides strict error control while CoAnnotating fails.** We compare HyPAC with heuristic methods like CoAnnotating (Li et al., 2023) across multiple benchmarks. As shown in Figure 4, these heuristic methods fail to con-



**Figure 4. HyPAC controls the annotations error while CoAnnotating fails.** Comparison of annotation error rates between HyPAC and CoAnnotating (Li et al., 2023) across three benchmarks. HyPAC controls the error rate under the specified tolerance level 0.05 while CoAnnotating fails.

trol the annotation error. Even when these algorithms are calibrated with a target error during their process, they still cannot stay within the error target. In contrast, HyPAC controls the error below  $\epsilon$  because of the PAC guarantee.

**Additional results.** Due to space constraints, we place additional analyses in the appendix. Appendix G.1 presents the performance of HyPAC under the semantic loss function. Appendix G.2 reports results under API-based cost functions. Appendix G.3 conducts an ablation study on the impact of calibration set size, and shows the robustness of HyPAC on calibration set size. Appendix G.4 includes experiments with distribution-free, finite-sample UCBs, complementing our main results based on the CLT-based asymptotic UCB. Appendix G.5 shows HyPAC’s robustness against distribution shift.

## 6. Related work

**PAC-style error rate control** Modern PAC-style error rate control is mainly built on distribution-free risk control. Conformal prediction provides finite-sample coverage guarantees under exchangeability (Angelopoulos et al., 2025b). This idea has been extended to control the expected loss on prediction sets, enabling risk control for tasks such as multi-label classification and image segmentation (Bates et al., 2021). Risk control has been further reframed as multiple hypothesis testing, allowing simultaneous control of multiple risks, including false discovery rate (Angelopoulos et al., 2025a). These methods have been applied to various domains: conformal language modeling for text generation with guaranteed quality (Quach et al., 2024), and automatically adaptive conformal risk control that adjusts to the difficulty of test samples (Blot et al., 2025). Localized adaptive risk control has been introduced for online calibration with conditional guarantees (Zecchin & Simeone, 2024). Recently, distribution-free risk control has been applied to data labeling, producing labels that are probably approximately correct (Candès et al., 2025). However, this method only considers a single human-machine switch without optimiz-

ing the overall cost across multiple annotation sources. Our work extends this framework to a multi-source setting, optimizing the routing policy to minimize cost while controlling the labeling error.

**Hybrid annotation** Hybrid annotation combines automatic predictors with human-in-the-loop querying to reduce labeling costs. Recent advances use large language models as annotators, with studies showing that GPT-3 can serve as a data annotator across multiple NLP tasks (Ding et al., 2023) and that LLMs can match crowdsourced annotators with proper prompting (He et al., 2024). Several works explore human-LLM collaboration for annotation, including uncertainty-guided work allocation (Li et al., 2023), verifier-based quality assessment that routes low-confidence samples to humans (Wang et al., 2024), unsupervised collaboration between LLMs and small models (Xiao et al., 2023), and multi-agent systems that manage multiple annotation sources (Qin et al., 2025). However, these methods optimize label efficiency through empirical heuristics without providing distribution-free guarantees on the error rate of automatically assigned labels. Recent work has taken steps toward statistical validity by combining LLM confidence with human annotations (Gligorić et al., 2024), though this focuses on downstream inference rather than label quality control. Concurrent work proposes PAC labels that guarantee low overall labeling error with high probability (Candès et al., 2025), but uses a fixed confidence threshold without adaptive routing. Our work extends this direction by integrating distribution-free risk control into a hybrid annotation with multiple annotation sources, achieving label efficiency explicit finite-sample error bounds.

## 7. Conclusion

We presented HyPAC, a hybrid annotation method that labels inputs among multiple annotation sources with different cost-quality trade-offs. HyPAC calibrates two thresholds using upper confidence bounds, providing a PAC guarantee that the error rate stays below a user-specified tolerance with high probability. We proved that the selected thresholds achieve the minimum expected cost among all feasible solutions. Experiments on benchmarks show that HyPAC achieves large cost savings while maintaining the target error rate. By providing a practical and theoretically grounded approach to hybrid annotation, we hope this work will help researchers and practitioners build more cost-efficient annotation systems while maintaining quality guarantees.

## Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

## References

- Angelopoulos, A. N., Bates, S., Candès, E. J., Jordan, M. I., and Lei, L. Learn then test: Calibrating predictive algorithms to achieve risk control. *The Annals of Applied Statistics*, 19(2):1641–1662, 2025a. ISSN 1932-6157, 1941-7330.
- Angelopoulos, A. N., Bates, S., Fisch, A., Lei, L., and Schuster, T. Conformal risk control. *Preprint at arXiv:2208.02814*, 2025b.
- Bates, S., Angelopoulos, A., Lei, L., Malik, J., and Jordan, M. Distribution-free, risk-controlling prediction sets. *Journal of the ACM*, 68(6):1–34, 2021. ISSN 0004-5411, 1557-735X.
- Blot, V., Angelopoulos, A. N., Jordan, M., and Brunel, N. J.-B. Automatically adaptive conformal risk control. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, pp. 19–27. PMLR, 2025.
- Candès, E. J., Ilyas, A., and Zrnic, T. Probably approximately correct labels. *Preprint at arXiv:2506.10908*, 2025.
- Chen, Qin, Liu, Peng, Guan, Wang, and et al. Towards reasoning era: A survey of long chain-of-thought for reasoning large language models. *Preprint at arXiv:2503.09567*, 2025.
- DeepSeek-AI, Guo, Yang, Zhang, Song, Zhang, and et al. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *Preprint at arXiv:2501.12948*, 2025.
- Dekoninck, J., Baader, M., and Vechev, M. A unified approach to routing and cascading for LLMs. In *Forty-Second International Conference on Machine Learning*, 2025.
- Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.
- Ding, Qin, Liu, Chia, Li, Joty, and et al. Is GPT-3 a good data annotator? In Rogers, A., Boyd-Graber, J., and Okazaki, N. (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11173–11195, Toronto, Canada, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.626.
- Gema, Leang, Hong, Devoto, Mancino, Saxena, and et al. Are we done with MMLU? In Chiruzzo, L., Ritter, A., and Wang, L. (eds.), *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)*, pp. 5069–5096, Albuquerque, New Mexico, 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.nacl-long.262.
- Gligorić, K., Zrnic, T., Lee, C., Candès, E. J., and Jurafsky, D. Can unconfident LLM annotations be used for confident conclusions? *Preprint at arXiv:2408.15204*, 2024.
- Grattafiori, Dubey, Jauhri, Pandey, Kadian, Al-Dahle, and et al. The Llama 3 herd of models. *Preprint at arXiv:2407.21783*, 2024.
- Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 1321–1330. PMLR, 2017.
- He, Lin, Gong, Jin, Zhang, Lin, and et al. AnnoLLM: Making large language models to be better crowdsourced annotators. In Yang, Y., Davani, A., Sil, A., and Kumar, A. (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, pp. 165–190, Mexico City, Mexico, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.nacl-industry.15.
- He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.
- Hendrycks, Burns, Kadavath, Arora, Basart, Tang, and et al. Measuring mathematical problem solving with the MATH dataset. *Preprint at arXiv:2103.03874*, 2021.
- Hendrycks, D. and Dietterich, T. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.
- Hu, Liu, Feng, Zhao, Ng, Luu, and et al. Uncertainty of thoughts: Uncertainty-aware planning enhances information seeking in large language models. *Preprint at arXiv:2402.03271*, 2024.

- Huang, Song, Wang, Zhao, Chen, Juefei-Xu, and et al. Look before you leap: An exploratory study of uncertainty measurement for large language models. *IEEE Transactions on Software Engineering*, 51(2):413–429, 2025. ISSN 0098-5589, 1939-3520, 2326-3881.
- Jung, J., Brahman, F., and Choi, Y. Trust or escalate: LLM judges with provable guarantees for human agreement. *Preprint at arXiv:2407.18370*, 2024.
- Li, Shi, Ziems, Kan, Chen, Liu, and et al. CoAnnotating: Uncertainty-guided work allocation between human and large language models for data annotation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 1487–1505, 2023. doi: 10.18653/v1/2023.emnlp-main.92.
- Li, J. Human-LLM hybrid text answer aggregation for crowd annotations. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 15609–15622, Miami, Florida, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.874.
- Lightman, Kosaraju, Burda, Edwards, Baker, Lee, and et al. Let’s verify step by step. In *The Twelfth International Conference on Learning Representations*, 2023.
- Lin, Bras, Richardson, Sabharwal, Poovendran, Clark, and et al. ZebraLogic: On the scaling limits of LLMs for logical reasoning. In *Forty-Second International Conference on Machine Learning*, 2025.
- Liu, J., Xia, C. S., Wang, Y., and Zhang, L. Is your code generated by ChatGPT really correct? Rigorous evaluation of large language models for code generation. *Preprint at arXiv:2305.01210*, 2023.
- Naeini, M. P., Cooper, G., and Hauskrecht, M. Obtaining well calibrated probabilities using bayesian binning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1), 2015. ISSN 2374-3468.
- Owen, A. and Zhou Associate, Y. Safe and effective importance sampling. *Journal of the American Statistical Association*, 95(449):135–143, 2000. ISSN 0162-1459.
- Qin, Zhu, Xia, Chen, Zhu, Lin, and et al. CrowdAgent: Multi-agent managed multi-source annotation system. *Preprint at arXiv:2509.14030*, 2025.
- Quach, Fisch, Schuster, Yala, Sohn, Jaakkola, and et al. Conformal language modeling. In *The Twelfth International Conference on Learning Representations*, Vienna, Austria, 2024.
- Qwen, Yang, Yang, Zhang, Hui, Zheng, and et al. Qwen2.5 technical report. *Preprint at arXiv:2412.15115*, 2025.
- Snell, C., Lee, J., Xu, K., and Kumar, A. Scaling LLM test-time compute optimally can be more effective than scaling model parameters. *Preprint at arXiv:2408.03314*, 2024.
- Snow, R., O’Connor, B., Jurafsky, D., and Ng, A. Cheap and fast – but is it good? Evaluating non-expert annotations for natural language tasks. In Lapata, M. and Ng, H. T. (eds.), *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pp. 254–263, Honolulu, Hawaii, 2008. Association for Computational Linguistics.
- Tian, Mitchell, Zhou, Sharma, Rafailov, Yao, and et al. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5433–5442, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.330.
- Tokdar, S. T. and Kass, R. E. Importance sampling: A review. *WIREs Computational Statistics*, 2(1):54–60, 2010. ISSN 1939-0068.
- Wang, X., Kim, H., Rahman, S., Mitra, K., and Miao, Z. Human-LLM collaborative annotation through effective verification of LLM labels. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI ’24, pp. 1–21, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 979-8-4007-0330-0. doi: 10.1145/3613904.3641960.
- Waudby-Smith, I. and Ramdas, A. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society Series B (Statistical Methodology)*, 86(1):1–27, 2024. ISSN 1369-7412.
- Xiao, Dong, Zhao, Wu, Lin, Chen, and et al. FreeAL: Towards human-free active learning in the era of large language models. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 14520–14535, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.896.
- Xiong, Hu, Lu, Li, Fu, He, and et al. Can LLMs express their uncertainty? An empirical evaluation of confidence elicitation in LLMs. In *The Twelfth International Conference on Learning Representations*, 2023.
- Xu, S., Xie, W., Zhao, L., and He, P. Chain of draft: Thinking faster by writing less. *Preprint at arXiv:2502.18600*, 2025.

Yang, Li, Yang, Zhang, Hui, Zheng, and et al. Qwen3 technical report. *Preprint at arXiv:2505.09388*, 2025.

Yang, D., Tsai, Y.-H. H., and Yamada, M. On verbalized confidence scores for LLMs. *Preprint at arXiv:2412.14737*, 2024.

Yuan, M., Chen, J., Xing, Z., Mohammadi, G., and Quigley, A. A case study of scalable content annotation using multi-LLM consensus and human review. *Preprint at arXiv:2503.17620*, 2025.

Yue, M., Zhao, J., Zhang, M., Du, L., and Yao, Z. Large language model cascades with mixture of thought representations for cost-efficient reasoning. In *The Twelfth International Conference on Learning Representations*, 2023.

Zecchin, M. and Simeone, O. Localized adaptive risk control. In *Advances Neural Information Processing Systems*, volume 37, 2024.

Zeng, H., Huang, J., Jing, B., Wei, H., and An, B. PAC reasoning: Controlling the performance loss for efficient reasoning. *Preprint at arXiv:2510.09133*, 2025.

Zhou, Tan, Li, Yao, Guo, Li, and et al. A theoretical study on bridging internal probability and self-consistency for LLM reasoning. *Preprint at arXiv:2510.15444*, 2025.

## A. Proofs of theoretical results

This appendix contains the complete proofs of all theoretical results presented in Section 4.

### A.1. Proof of Lemma 4.1

In this section, we provide the complete proof of Lemma 4.1, which establishes the monotonicity of the risk function with respect to both threshold parameters. This result is fundamental to our theoretical framework.

*Proof of Lemma 4.1.* Recall that the risk function is:

$$R(u_1, u_2) = \mathbb{E}_{(x,y) \sim \mathcal{D}} [\ell(y, T_{u_1, u_2}(x))]$$

where  $T_{u_1, u_2}(x) = \tilde{f}_1(x)$  if  $U(x) \leq u_1$ ,  $T_{u_1, u_2}(x) = \tilde{f}_2(x)$  if  $u_1 < U(x) \leq u_2$ , and  $T_{u_1, u_2}(x) = y$  otherwise.

**Monotonicity in  $u_1$ .** Fix  $u_2$  and consider  $u_1 < u'_1 \leq u_2$ . The mappings  $T_{u_1, u_2}$  and  $T_{u'_1, u_2}$  differ only for examples with  $u_1 < U(x) \leq u'_1$ : these are routed to  $\tilde{f}_2$  under  $T_{u_1, u_2}$  but to  $\tilde{f}_1$  under  $T_{u'_1, u_2}$ . The risk difference is:

$$R(u'_1, u_2) - R(u_1, u_2) = \mathbb{P}(u_1 < U(x) \leq u'_1) \cdot \mathbb{E}_{u_1 < U(x) \leq u'_1} [\ell(y, \tilde{f}_1(x)) - \ell(y, \tilde{f}_2(x))]$$

Since  $\tilde{f}_2$  is more accurate than  $\tilde{f}_1$ , we have  $\mathbb{E}[\ell(y, \tilde{f}_1(x))] \geq \mathbb{E}[\ell(y, \tilde{f}_2(x))]$  for any conditional distribution. Thus  $R(u'_1, u_2) - R(u_1, u_2) \geq 0$ .

**Monotonicity in  $u_2$ .** Fix  $u_1$  and consider  $u_2 < u'_2$  with  $u_1 \leq u_2$ . The mappings differ only for examples with  $u_2 < U(x) \leq u'_2$ : these are routed to human annotation under  $T_{u_1, u_2}$  but to  $\tilde{f}_2$  under  $T_{u_1, u'_2}$ . Since  $\ell(y, y) = 0$ , the risk difference is:

$$R(u_1, u'_2) - R(u_1, u_2) = \mathbb{P}(u_2 < U(x) \leq u'_2) \cdot \mathbb{E}_{u_2 < U(x) \leq u'_2} [\ell(y, \tilde{f}_2(x))] \geq 0$$

where the inequality follows from the non-negativity of the loss function.  $\square$

### A.2. Proof of Lemma 4.2

*Proof.* We prove both the population version and the empirical version.

**Population version.** The expected cost can be written as:

$$\mathbb{E}_{x \sim \mathcal{D}_x} [C(u_1, u_2, x)] = \mathbb{E}[c_1(x) \cdot \mathbf{1}\{U(x) \leq u_1\}] + \mathbb{E}[c_2(x) \cdot \mathbf{1}\{u_1 < U(x) \leq u_2\}] + \mathbb{E}[c_h(x) \cdot \mathbf{1}\{U(x) > u_2\}]$$

Rearranging terms:

$$\mathbb{E}_{x \sim \mathcal{D}_x} [C(u_1, u_2, x)] = \mathbb{E}[c_h(x)] + \mathbb{E}[(c_1(x) - c_2(x)) \cdot \mathbf{1}\{U(x) \leq u_1\}] + \mathbb{E}[(c_2(x) - c_h(x)) \cdot \mathbf{1}\{U(x) \leq u_2\}]$$

Under the cost ordering assumption  $c_1(x) \leq c_2(x) \leq c_h(x)$ , we have:

$$c_1(x) - c_2(x) \leq 0 \quad \text{and} \quad c_2(x) - c_h(x) \leq 0 \quad \text{for almost all } x$$

For any  $u_1 < u'_1$  with  $u'_1 \leq u_2$ , the cost difference is:

$$\begin{aligned} \mathbb{E}[C(u'_1, u_2, x)] - \mathbb{E}[C(u_1, u_2, x)] &= \mathbb{E}[(c_1(x) - c_2(x)) \cdot (\mathbf{1}\{U(x) \leq u'_1\} - \mathbf{1}\{U(x) \leq u_1\})] \\ &= \mathbb{E}[(c_1(x) - c_2(x)) \cdot \mathbf{1}\{u_1 < U(x) \leq u'_1\}] \leq 0 \end{aligned}$$

where the inequality follows from  $c_1(x) - c_2(x) \leq 0$ .

Similarly, for any  $u_2 < u'_2$ , we have:

$$\mathbb{E}[C(u_1, u'_2, x)] - \mathbb{E}[C(u_1, u_2, x)] = \mathbb{E}[(c_2(x) - c_h(x)) \cdot \mathbf{1}\{u_2 < U(x) \leq u'_2\}] \leq 0$$

Therefore, the expected cost is non-increasing in both  $u_1$  and  $u_2$ .

**Empirical version.** For any finite sample  $\{x_i\}_{i=1}^m$ , the empirical cost is:

$$\widehat{C}(u_1, u_2) = \frac{1}{m} \sum_{i=1}^m C(u_1, u_2, x_i)$$

where  $C(u_1, u_2, x_i)$  is the cost for example  $x_i$  under the routing policy  $T_{u_1, u_2}$ .

By the definition of the routing policy, we have:

$$C(u_1, u_2, x_i) = \begin{cases} c_1(x_i) & \text{if } U(x_i) \leq u_1 \\ c_2(x_i) & \text{if } u_1 < U(x_i) \leq u_2 \\ c_h(x_i) & \text{if } U(x_i) > u_2 \end{cases}$$

For any  $u_1 < u'_1$  with  $u'_1 \leq u_2$ , the empirical cost difference is:

$$\begin{aligned} \widehat{C}(u'_1, u_2) - \widehat{C}(u_1, u_2) &= \frac{1}{m} \sum_{i=1}^m [C(u'_1, u_2, x_i) - C(u_1, u_2, x_i)] \\ &= \frac{1}{m} \sum_{i: u_1 < U(x_i) \leq u'_1} [c_1(x_i) - c_2(x_i)] \leq 0 \end{aligned}$$

where the inequality follows from  $c_1(x_i) \leq c_2(x_i)$  for all  $i$ .

Similarly, for any  $u_2 < u'_2$ :

$$\widehat{C}(u_1, u'_2) - \widehat{C}(u_1, u_2) = \frac{1}{m} \sum_{i: u_2 < U(x_i) \leq u'_2} [c_2(x_i) - c_h(x_i)] \leq 0$$

Therefore, the empirical cost is non-increasing in both  $u_1$  and  $u_2$ .

**Strict monotonicity.** When the cost ordering is strict (i.e.,  $c_1(x) < c_2(x) < c_h(x)$  for almost all  $x$ ), the inequalities in both the population and empirical versions become strict whenever the corresponding indicator sets have positive measure (for population) or are non-empty (for empirical).  $\square$

### A.3. Proof of Theorem 4.3

In this section, we provide the complete proof of Theorem 4.3, which establishes that HyPAC provides a PAC guarantee for error control. We use the fixed-sequence testing framework to handle the fact that the threshold pair is selected based on the calibration data. The proof consists of three main steps. First, we show that the UCB provides valid coverage for any fixed threshold pair (Eq. (9)). Second, we extend this coverage to the data-dependent selected pair: the key observation is that  $\{R(\widehat{u}_1, \widehat{u}_2) > \epsilon\} \subseteq \{R(\widehat{u}_1, \widehat{u}_2) > \widehat{L}_{\widehat{u}_1, \widehat{u}_2}(\alpha)\}$ , since the algorithm only selects pairs with  $\widehat{L}_{\widehat{u}_1, \widehat{u}_2}(\alpha) \leq \epsilon$ . Third, we use the monotonicity of the risk function (Lemma 4.1) to apply fixed-sequence testing without Bonferroni correction. The monotonicity ensures that if  $R(u_1, u_2) \leq \epsilon$  for some pair, then  $R(u'_1, u'_2) \leq \epsilon$  for all  $u'_1 \leq u_1$  and  $u'_2 \leq u_2$ . This nested structure allows us to test threshold pairs in a fixed sequence (from largest to smallest) and stop at the first rejection, maintaining the significance level  $\alpha$  without multiple testing correction.

*Proof of Theorem 4.3.* By Assumption 1, for any fixed threshold pair  $(u_1, u_2)$  with  $0 \leq u_1 \leq u_2 \leq 1$ , the upper confidence bound  $\widehat{L}_{u_1, u_2}(\alpha)$  satisfies:

$$\mathbb{P}_{\mathcal{S}_{\text{cal}}} \left( R(u_1, u_2) \leq \widehat{L}_{u_1, u_2}(\alpha) \right) \geq 1 - \alpha \quad (9)$$

This means that with probability at least  $1 - \alpha$  over the random draw of the calibration set, the UCB is an upper bound on the true risk. The challenge is that the threshold pair  $(\widehat{u}_1, \widehat{u}_2)$  selected by Algorithm 1 is not fixed but depends on the calibration data  $\mathcal{S}_{\text{cal}}$ . Specifically, the algorithm selects the maximal feasible pair in the *lexicographic* order  $(u_2, u_1)$ :

$$(\widehat{u}_1, \widehat{u}_2) = \arg \max_{(u_1, u_2) \in \mathcal{F}} (u_2, u_1)$$

where  $\mathcal{F} = \{(u_1, u_2) : \widehat{L}_{u_1, u_2}(\alpha) \leq \epsilon, 0 \leq u_1 \leq u_2 \leq 1\}$  is the feasible set of threshold pairs whose UCB is below  $\epsilon$ . The lexicographic maximum first maximizes  $u_2$ , then maximizes  $u_1$  subject to  $u_1 \leq u_2$ .

The key insight is that the event  $\{R(\widehat{u}_1, \widehat{u}_2) > \epsilon\}$  implies the event  $\{R(\widehat{u}_1, \widehat{u}_2) > \widehat{L}_{\widehat{u}_1, \widehat{u}_2}(\alpha)\}$  because the algorithm only selects pairs with  $\widehat{L}_{\widehat{u}_1, \widehat{u}_2}(\alpha) \leq \epsilon$ . Formally, suppose  $R(\widehat{u}_1, \widehat{u}_2) > \epsilon$ . Since Algorithm 1 only selects pairs from the feasible set  $\mathcal{F}$ , we have  $\widehat{L}_{\widehat{u}_1, \widehat{u}_2}(\alpha) \leq \epsilon$ . Therefore:

$$R(\widehat{u}_1, \widehat{u}_2) > \epsilon \geq \widehat{L}_{\widehat{u}_1, \widehat{u}_2}(\alpha)$$

This means:

$$\{R(\widehat{u}_1, \widehat{u}_2) > \epsilon\} \subseteq \{R(\widehat{u}_1, \widehat{u}_2) > \widehat{L}_{\widehat{u}_1, \widehat{u}_2}(\alpha)\}$$

Then

$$\mathbb{P}_{\mathcal{S}_{\text{cal}}}(R(\widehat{u}_1, \widehat{u}_2) > \epsilon) \leq \mathbb{P}_{\mathcal{S}_{\text{cal}}}\left(R(\widehat{u}_1, \widehat{u}_2) > \widehat{L}_{\widehat{u}_1, \widehat{u}_2}(\alpha)\right).$$

Now, the event  $\{R(\widehat{u}_1, \widehat{u}_2) > \widehat{L}_{\widehat{u}_1, \widehat{u}_2}(\alpha)\}$  means that the UCB fails to upper bound the true risk for the selected pair. By the monotonicity of the risk function (Lemma 4.1), the null hypotheses  $H_{u_1, u_2} : R(u_1, u_2) > \epsilon$  have a nested structure: if  $H_{u_1, u_2}$  is true, then  $H_{u'_1, u'_2}$  is also true for all  $u'_1 \geq u_1$  and  $u'_2 \geq u_2$ . This enables fixed-sequence testing (Angelopoulos et al., 2025a).

The key property of fixed-sequence testing is that it controls the family-wise error rate at level  $\alpha$  without any correction. Formally, let  $(u_1^{(1)}, u_2^{(1)}), (u_1^{(2)}, u_2^{(2)}), \dots$  be the sequence of threshold pairs tested by the algorithm (ordered from largest to smallest). Let  $k^*$  be the index of the first pair whose null hypothesis is true (i.e.,  $R(u_1^{(k^*)}, u_2^{(k^*)}) > \epsilon$ ). Then:

$$\begin{aligned} & \mathbb{P}_{\mathcal{S}_{\text{cal}}}(\text{reject } H_{u_1^{(k^*)}, u_2^{(k^*)}}) \\ &= \mathbb{P}_{\mathcal{S}_{\text{cal}}}(\widehat{L}_{u_1^{(k^*)}, u_2^{(k^*)}}(\alpha) \leq \epsilon \mid R(u_1^{(k^*)}, u_2^{(k^*)}) > \epsilon) \\ &\leq \alpha \end{aligned}$$

Since the algorithm stops at the first rejection, it can only make an error if it rejects  $H_{u_1^{(k^*)}, u_2^{(k^*)}}$ . Therefore:

$$\mathbb{P}_{\mathcal{S}_{\text{cal}}}(R(\widehat{u}_1, \widehat{u}_2) > \epsilon) \leq \alpha$$

Finally, we verify that Algorithm 1 indeed only selects threshold pairs whose UCB is below  $\epsilon$ . By construction, the algorithm searches over the grid  $\mathcal{G}$  and selects:

$$\begin{aligned} \widehat{u}_2 &= \max\{u_2 : \exists u_1 \leq u_2 \text{ s.t. } \widehat{L}_{u_1, u_2}(\alpha) \leq \epsilon\} \\ \widehat{u}_1 &= \max\{u_1 : u_1 \leq \widehat{u}_2 \text{ and } \widehat{L}_{u_1, \widehat{u}_2}(\alpha) \leq \epsilon\} \end{aligned}$$

This ensures that  $\widehat{L}_{\widehat{u}_1, \widehat{u}_2}(\alpha) \leq \epsilon$  by construction. If no such pair exists (i.e.,  $\mathcal{F} = \emptyset$ ), the algorithm can return  $(0, 0)$ , which routes all examples to the thinking model and human annotation, guaranteeing zero error.

Combining the three steps, we have shown that:

$$\mathbb{P}_{\mathcal{S}_{\text{cal}}}(R(\widehat{u}_1, \widehat{u}_2) \leq \epsilon) \geq 1 - \alpha$$

This completes the proof of Theorem 4.3.  $\square$

*Remark A.1* (Role of monotonicity in avoiding multiple testing correction). A naive approach would apply a union bound over all threshold pairs in the grid  $\mathcal{G}$ , requiring a Bonferroni correction that sets the significance level for each UCB to  $\alpha/|\mathcal{G}|$ . This would make the UCBs wider and the selected thresholds more conservative. However, the monotonicity of the risk function (Lemma 4.1) allows us to avoid this correction entirely. The nested structure of the null hypotheses enables fixed-sequence testing: the algorithm tests threshold pairs from largest to smallest, and stops at the first rejection. By the monotonicity, all pairs with smaller thresholds also satisfy  $R \leq \epsilon$ , so we do not need to test them. This is why the PAC guarantee holds with significance level  $\alpha$ , not  $\alpha/|\mathcal{G}|$ .

#### A.4. Proof of Theorem 4.4

*Proof.* Fix a calibration set  $\mathcal{S}_{\text{cal}}$ . By Lemma 4.2, both  $\widehat{C}(u_1, u_2)$  and  $\mathbb{E}_{x \sim \mathcal{D}_x}[C(u_1, u_2, x)]$  are non-increasing in both  $u_1$  and  $u_2$ . Algorithm 1 computes:

$$(\widehat{u}_1, \widehat{u}_2) = \arg \min_{(u_1, u_2) \in \mathcal{F}_{\text{PAC}}(\mathcal{S}_{\text{cal}})} \widehat{C}(u_1, u_2),$$

where  $\mathcal{F}_{\text{PAC}}(\mathcal{S}_{\text{cal}}) = \{(u_1, u_2) : \widehat{L}_{u_1, u_2}(\alpha) \leq \epsilon, 0 \leq u_1 \leq u_2 \leq 1\}$ . Since  $\widehat{C}(u_1, u_2)$  is non-increasing in both arguments, the minimum over  $\mathcal{F}_{\text{PAC}}(\mathcal{S}_{\text{cal}})$  is achieved at the maximum feasible thresholds. By the same monotonicity,  $\mathbb{E}_{x \sim \mathcal{D}_x}[C(u_1, u_2, x)]$  is also minimized at the maximum feasible thresholds. Therefore,

$$\mathbb{E}_{x \sim \mathcal{D}_x}[C(\widehat{u}_1, \widehat{u}_2, x)] = \min_{(u_1, u_2) \in \mathcal{F}_{\text{PAC}}(\mathcal{S}_{\text{cal}})} \mathbb{E}_{x \sim \mathcal{D}_x}[C(u_1, u_2, x)].$$

Note that  $\mathcal{F}_{\text{PAC}}(\mathcal{S}_{\text{cal}})$  depends on  $\mathcal{S}_{\text{cal}}$  via empirical bounds, but  $\mathbb{E}_{x \sim \mathcal{D}_x}[C(u_1, u_2, x)]$  is a population quantity independent of  $\mathcal{S}_{\text{cal}}$ .  $\square$

#### A.5. Finite-sample guarantee

While Theorem 4.3 provides a guarantee on the true risk  $R(\widehat{u}_1, \widehat{u}_2)$ , in practice we often want to evaluate the performance on a finite test set. We now provide a bound on the empirical risk that accounts for the finite test set size.

**Theorem A.2** (Empirical risk PAC guarantee). *Let  $(\widehat{u}_1, \widehat{u}_2)$  be the threshold pair selected by Algorithm 1 with error tolerance  $\epsilon$  and significance level  $\alpha$ . Let  $\mathcal{S}_{\text{test}} = \{x_j\}_{j=1}^N$  be a test set drawn independently from the same distribution as the calibration set. Define the empirical risk as:*

$$\widehat{R}(\widehat{u}_1, \widehat{u}_2) = \frac{1}{N} \sum_{j=1}^N \ell(y_j, T_{\widehat{u}_1, \widehat{u}_2}(x_j))$$

Assuming the loss function is bounded in  $[0, B]$ , for any  $t > 0$ :

$$\mathbb{P}_{\mathcal{S}_{\text{cal}}, \mathcal{S}_{\text{test}}} \left( \widehat{R}(\widehat{u}_1, \widehat{u}_2) \leq \epsilon + t \right) \geq 1 - \alpha - (1 - \alpha) \exp \left( -\frac{2Nt^2}{B^2} \right)$$

where the probability is taken over both the calibration set and the test set.

This theorem tells us how to interpret the empirical error rate on a finite test set: if we observe  $\widehat{R}(\widehat{u}_1, \widehat{u}_2)$ , the true error rate is close to this value, with deviation controlled by  $N$ .

*Proof of Theorem A.2.* The proof proceeds as follows. We first establish a concentration bound for the empirical risk around the true risk. We then combine this with the PAC guarantee from Theorem 4.3 using a union bound argument.

Fix a threshold pair  $(u_1, u_2)$  and consider the empirical risk:

$$\widehat{R}(u_1, u_2) = \frac{1}{N} \sum_{j=1}^N \ell(y_j, T_{u_1, u_2}(x_j))$$

Since the test examples are drawn independently from  $\mathcal{D}$ , the losses are i.i.d. random variables with:

$$\mathbb{E}_{x_j \sim \mathcal{D}}[\ell(y_j, T_{u_1, u_2}(x_j))] = R(u_1, u_2)$$

Since the loss is bounded in  $[0, B]$ , Hoeffding's inequality gives:

$$\mathbb{P} \left( \frac{1}{N} \sum_{i=1}^N X_i - \mathbb{E} \left[ \frac{1}{N} \sum_{i=1}^N X_i \right] \geq t \right) \leq \exp \left( -\frac{2N^2 t^2}{\sum_{i=1}^N (b_i - a_i)^2} \right)$$

With  $b_j - a_j = B$  for all  $j$ , we have  $\sum_{j=1}^N (b_j - a_j)^2 = NB^2$ , so:

$$\mathbb{P}_{\mathcal{S}_{\text{test}}} \left( \widehat{R}(u_1, u_2) - R(u_1, u_2) > t \right) \leq \exp \left( -\frac{2Nt^2}{B^2} \right)$$

Now we combine this with Theorem 4.3. We decompose the probability by conditioning on whether  $R(\widehat{u}_1, \widehat{u}_2) \leq \epsilon$ :

$$\begin{aligned} & \mathbb{P}_{\mathcal{S}_{\text{cal}}, \mathcal{S}_{\text{test}}} \left( \widehat{R}(\widehat{u}_1, \widehat{u}_2) \leq \epsilon + t \right) \\ &= \mathbb{P}_{\mathcal{S}_{\text{cal}}, \mathcal{S}_{\text{test}}} \left( \widehat{R} \leq \epsilon + t \mid R \leq \epsilon \right) \cdot \mathbb{P}_{\mathcal{S}_{\text{cal}}} (R \leq \epsilon) \\ &\quad + \mathbb{P}_{\mathcal{S}_{\text{cal}}, \mathcal{S}_{\text{test}}} \left( \widehat{R} \leq \epsilon + t \mid R > \epsilon \right) \cdot \mathbb{P}_{\mathcal{S}_{\text{cal}}} (R > \epsilon) \end{aligned}$$

where we write  $R = R(\widehat{u}_1, \widehat{u}_2)$  and  $\widehat{R} = \widehat{R}(\widehat{u}_1, \widehat{u}_2)$  for brevity.

Since the second term is non-negative, we have:

$$\mathbb{P}_{\mathcal{S}_{\text{cal}}, \mathcal{S}_{\text{test}}} \left( \widehat{R} \leq \epsilon + t \right) \geq \mathbb{P}_{\mathcal{S}_{\text{cal}}, \mathcal{S}_{\text{test}}} \left( \widehat{R} \leq \epsilon + t \mid R \leq \epsilon \right) \cdot \mathbb{P}_{\mathcal{S}_{\text{cal}}} (R \leq \epsilon)$$

When  $R \leq \epsilon$ , the event  $\{\widehat{R} \leq \epsilon + t\}$  is implied by  $\{\widehat{R} - R \leq t\}$ . By Hoeffding's inequality:

$$\mathbb{P}_{\mathcal{S}_{\text{test}}} \left( \widehat{R} \leq \epsilon + t \mid R \leq \epsilon \right) \geq \mathbb{P}_{\mathcal{S}_{\text{test}}} \left( \widehat{R} - R \leq t \right) \geq 1 - \exp \left( -\frac{2Nt^2}{B^2} \right)$$

By Theorem 4.3,  $\mathbb{P}_{\mathcal{S}_{\text{cal}}} (R \leq \epsilon) \geq 1 - \alpha$ .

Combining these:

$$\begin{aligned} \mathbb{P}_{\mathcal{S}_{\text{cal}}, \mathcal{S}_{\text{test}}} \left( \widehat{R} \leq \epsilon + t \right) &\geq \left( 1 - \exp \left( -\frac{2Nt^2}{B^2} \right) \right) \cdot (1 - \alpha) \\ &= 1 - \alpha - \exp \left( -\frac{2Nt^2}{B^2} \right) + \alpha \exp \left( -\frac{2Nt^2}{B^2} \right) \\ &= 1 - \alpha - (1 - \alpha) \exp \left( -\frac{2Nt^2}{B^2} \right) \end{aligned}$$

This completes the proof of Theorem A.2.  $\square$

## B. UCB validity examples

In this section, we provide concrete examples of how to construct valid upper confidence bounds (UCBs) that satisfy Assumption 1. The UCB must satisfy: for any fixed threshold pair  $(u_1, u_2)$ ,

$$\mathbb{P}_{\mathcal{S}_{\text{cal}}} \left( R(u_1, u_2) \leq \widehat{L}_{u_1, u_2}(\alpha) \right) \geq 1 - \alpha$$

**CLT-based UCB.** If we use the CLT-based UCB from Eq. (7):

$$\widehat{L}_{u_1, u_2}(\alpha) = \widehat{R}_{\text{IS}}(u_1, u_2) + z_{1-\alpha} \frac{\widehat{\sigma}_W(u_1, u_2)}{\sqrt{m}}$$

where  $\widehat{R}_{\text{IS}}(u_1, u_2)$  is the importance-weighted risk estimate and  $\widehat{\sigma}_W(u_1, u_2)$  is the sample standard deviation of the importance-weighted losses. By the central limit theorem, for large enough  $m$ :

$$\frac{\widehat{R}_{\text{IS}}(u_1, u_2) - R(u_1, u_2)}{\widehat{\sigma}_W(u_1, u_2)/\sqrt{m}} \xrightarrow{d} \mathcal{N}(0, 1)$$

This implies that:

$$\mathbb{P} \left( R(u_1, u_2) \leq \hat{R}_{\text{IS}}(u_1, u_2) + z_{1-\alpha} \frac{\hat{\sigma}_W(u_1, u_2)}{\sqrt{m}} \right) = 1 - \alpha + o(1).$$

**Hoeffding-based UCB.** For finite samples, we can use Hoeffding’s inequality to obtain a non-asymptotic bound. Since the loss is bounded, the importance-weighted losses  $W_i = \frac{Z_i}{p} \ell(y_i, T_{u_1, u_2}(x_i))$  are bounded in  $[0, B/p]$ . By Hoeffding’s inequality:

$$\mathbb{P} \left( \hat{R}_{\text{IS}}(u_1, u_2) - R(u_1, u_2) \geq t \right) \leq \exp \left( - \frac{2mt^2}{(B/p)^2} \right)$$

Setting the right-hand side equal to  $\alpha$  and solving for  $t$ :

$$t = \frac{B}{p\sqrt{2m}} \sqrt{\log(1/\alpha)}$$

This gives us a Hoeffding-based UCB:

$$\hat{L}_{u_1, u_2}^{\text{Hoeff}}(\alpha) = \hat{R}_{\text{IS}}(u_1, u_2) + \frac{B}{p\sqrt{2m}} \sqrt{\log(1/\alpha)}$$

which satisfies:

$$\mathbb{P}_{\mathcal{S}_{\text{cal}}} \left( R(u_1, u_2) \leq \hat{L}_{u_1, u_2}^{\text{Hoeff}}(\alpha) \right) \geq 1 - \alpha$$

**Empirical Bernstein bound.** A tighter bound can be obtained using the empirical Bernstein inequality, which adapts to the variance of the data. Let  $\hat{V} = \frac{1}{m-1} \sum_{i=1}^m (W_i - \hat{R}_{\text{IS}})^2$  be the sample variance. The empirical Bernstein UCB is:

$$\hat{L}_{u_1, u_2}^{\text{Bern}}(\alpha) = \hat{R}_{\text{IS}}(u_1, u_2) + \sqrt{\frac{2\hat{V} \log(2/\alpha)}{m}} + \frac{7B \log(2/\alpha)}{3(m-1)p}$$

This bound is tighter than Hoeffding when the variance is small.

**Betting-based bound.** We also consider a betting-based upper confidence bound adapted from the finite-sample confidence sequence of [Waudby-Smith & Ramdas \(2024\)](#). Let

$$W_i = \frac{Z_i}{p} \ell(y_i, T_{u_1, u_2}(x_i)), \quad 0 \leq W_i \leq \frac{1}{p},$$

and define the normalized variables  $X_i = pW_i \in [0, 1]$ . Let  $\mu = \mathbb{E}[X_i] = pR(u_1, u_2)$ .

For  $t = 1, \dots, m$ , define

$$\hat{\mu}_t = \frac{0.5 + \sum_{j=1}^t X_j}{t+1}, \quad \hat{\sigma}_t^2 = \frac{0.25 + \sum_{j=1}^t (X_j - \hat{\mu}_t)^2}{t+1}, \quad \lambda_t = \sqrt{\frac{2 \log(2/\alpha)}{m \hat{\sigma}_{t-1}^2}}.$$

For any candidate mean  $m \in [0, 1]$ , define

$$M_t(m) = \frac{1}{2} \max \left\{ \prod_{j=1}^t \left( 1 + \min \left( \lambda_j, \frac{0.5}{m} \right) (X_j - m) \right), \prod_{j=1}^t \left( 1 - \min \left( \lambda_j, \frac{0.5}{1-m} \right) (X_j - m) \right) \right\}.$$

Define the confidence set

$$\mathcal{C} = \bigcap_{t=1}^m \left\{ m \in [0, 1] : M_t(m) < \frac{1}{\alpha} \right\}.$$

Let

$$\mathcal{C}_R := \{c/p : c \in \mathcal{C}\}.$$

Then,

$$\mathbb{P}(R(u_1, u_2) \in \mathcal{C}_R) \geq 1 - \alpha.$$

Intuitively, the supermartingale  $M_t(m)$  should be thought of as the amount of evidence against  $m$  being the true mean. That is,  $M_t(m)$  being big suggests that  $m$  is unlikely to be the true mean, so the final confidence set is the collection of all  $m$  for which the amount of such evidence is small.

## C. Extension to more than three annotation sources

Our framework naturally extends to systems with more than three annotation sources. Suppose we have  $K$  annotation sources  $\tilde{f}_1, \dots, \tilde{f}_{K-1}, y$  with cost functions  $c_1(x), c_2(x), \dots, c_{K-1}(x), c_K(x)$  satisfying  $c_1(x) \leq c_2(x) \leq \dots \leq c_K(x)$  for almost all  $x$ , and accuracies that increase with cost.

We use  $K - 1$  thresholds  $u_1 \leq u_2 \leq \dots \leq u_{K-1}$  to partition the input space into  $K$  regions. The deployment mapping becomes:

$$T_{u_1, \dots, u_{K-1}}(x) = \begin{cases} \tilde{f}_1(x) & \text{if } U(x) \leq u_1 \\ \tilde{f}_2(x) & \text{if } u_1 < U(x) \leq u_2 \\ \vdots & \\ \tilde{f}_{K-1}(x) & \text{if } u_{K-2} < U(x) \leq u_{K-1} \\ y & \text{if } U(x) > u_{K-1} \end{cases}$$

The risk function is:

$$R(u_1, \dots, u_{K-1}) = \mathbb{E}_{(x,y) \sim \mathcal{D}}[\ell(y, T_{u_1, \dots, u_{K-1}}(x))]$$

The monotonicity result extends naturally:  $R$  is monotone non-decreasing in each threshold  $u_k$ , since increasing any threshold routes more examples to cheaper (less accurate) sources. The calibration procedure also extends: we construct UCBs for all threshold tuples  $(u_1, \dots, u_{K-1})$  in a discretized grid, then select the optimal tuple that minimizes cost while satisfying the risk constraint. The PAC guarantee carries over: with probability at least  $1 - \alpha$ , the selected thresholds satisfy  $R(u_1, \dots, u_{K-1}) \leq \epsilon$ . The main challenge is computational: the number of threshold tuples grows exponentially with  $K$ , which increases calibration cost and may require larger calibration sets. In practice,  $K = 3$  or  $K = 4$  sources are typically sufficient.

## D. Score

A key component of our hybrid annotation system is the score  $U(x)$  that measures the uncertainty of the annotation system for input  $x$ . This score is used to decide which annotation source to use: examples with low scores can be handled by cheaper models, while examples with high scores should be sent to more accurate sources.

Intuitively, this uncertainty may relate to the difficulty of the input, but our theory does not require such an assumption. A more accurate score leads to a more efficient calibration process. In practice, model uncertainty serves as a natural proxy for this purpose (Huang et al., 2025; Xiong et al., 2023; Hu et al., 2024). When a model is uncertain about its prediction, this often indicates that the input is inherently ambiguous, lies near decision boundaries, or requires deeper reasoning that the model cannot perform reliably. Such examples are precisely those that benefit from more capable annotation sources.

We consider two main approaches to computing the score for large language models, both of which can be used within our framework.

**Logits-based score.** For autoregressive language models that generate sequences token by token, we can use the model’s output probabilities to compute the score (Guo et al., 2017). Let  $y = (y_1, y_2, \dots, y_l)$  be a generated sequence of length  $l$ . The model produces a probability distribution over the next token at each step:  $\mathbb{P}(y_j | y_1, \dots, y_{j-1}, x)$ . We define the logits-based score as

$$U_{\text{logits}}(x) = 1 - \frac{1}{l} \sum_{j=1}^l \mathbb{P}(y_j | y_1, \dots, y_{j-1}, x)$$

where  $y_j$  is the token actually generated by the model at position  $j$ . This score measures the average confidence of the model across all generated tokens. When the model is very confident about each token (high probabilities), the score is low.

When the model is less confident (low probabilities), the score is high, indicating an example that may benefit from a more capable annotation source.

This approach has several advantages: it is model-agnostic, requires no additional computation beyond standard inference, and provides a fine-grained measure at the token level. However, it requires access to the model’s logits, which may not always be available for API-based models.

**Verbalized score.** An alternative approach is to ask the model to explicitly express its confidence in natural language (Yang et al., 2024; Tian et al., 2023). We can prompt the model to output a confidence level along with its prediction. We then map these verbalized confidence levels to numerical scores in  $[0, 1]$ .

For example, we might use a prompt like: ‘Please provide your answer and indicate your confidence level (high/medium/low).’ We then parse the model’s response to extract the confidence level and map it to a score: high confidence  $\rightarrow$  low score, low confidence  $\rightarrow$  high score.

This approach has the advantage of working with black-box API models where we do not have access to logits. It also allows the model to use its reasoning capabilities to assess the input, potentially capturing semantic information that may not be reflected in token-level probabilities. However, it requires careful prompt engineering and may add overhead to the inference process.

**Alternative scores.** Our framework is flexible and can work with other methods for computing the score. One alternative is ensemble-based scoring, where we run the model multiple times with different random seeds or sampling parameters and measure the disagreement among the outputs, with higher disagreement indicating higher uncertainty. Another alternative is calibrated scoring, where we use a separate calibration model trained to predict the probability that the main model’s output is correct based on features of the input and output. We can also use hybrid approaches that combine multiple scoring methods to get a more robust estimate.

The choice of scoring method depends on the specific application, the available model APIs, and the computational budget. In our experiments, we focus on logits-based and verbalized scores as they are widely applicable and easy to implement.

## E. Compared baselines

In this work, we compare HyPAC against PAC labeling (Candès et al., 2025), Cascaded Selective Evaluation (CSE) (Jung et al., 2024), and a heuristic baseline. PAC Labeling leverages a confidence threshold to route each example either to an LLM for automatic labeling or to an expert annotator. CSE performs sequential selective prediction by escalating inputs to increasingly expensive models based on confidence thresholds, without providing an explicit finite-sample guarantee on the final annotation error. For the heuristic baseline, we select hyperparameters  $\hat{u}_1$  and  $\hat{u}_2$  to minimize annotation cost while ensuring that the empirical risk on the calibration data does not exceed  $\epsilon$ .

**Probably Approximately Correct (PAC) Labels (Candès et al., 2025).** PAC labeling is a selective annotation framework that combines inexpensive model-generated labels with expert annotations while providing a dataset-level accuracy guarantee. Given a target error level  $\epsilon$  and tolerance  $\alpha$ , the method selectively escalates high-uncertainty samples to human annotators, ensuring that the overall labeling error is bounded with high probability. The original PAC labeling formulation assumes a fixed dataset and does not rely on a separate calibration set. To make it compatible with our framework, we introduce a slight modification and reformulate it as a two-stage PAC labeling procedure.

**Cascaded Selective Evaluation (CSE).** The CSE guarantees a target level of loss control for LLM response while reducing evaluation cost by escalating only uncertain cases to larger models.<sup>1</sup> Under the i.i.d. data assumption, Cascaded Selective Evaluation (CSE) admits a rigorous theoretical guarantee: with probability  $1 - \alpha$ ,

$$\mathbb{P}\left(f_{\text{CSE}}(x) \neq y_{\text{human}} \mid x \text{ is labeled by } \tilde{f}_1 \text{ or } \tilde{f}_2\right) \leq \delta.$$

Note that in the original CSE implementation, expert annotations are not considered. All data instances that are not annotated by the LLM are directly escalated. To make the method compatible with our setup, we use expert annotations for escalated instances.

<sup>1</sup>Throughout this work, we use  $\epsilon$  to denote the target error level, and  $\alpha$  to denote the tolerance level controlling the probability with which this guarantee holds. This notation is different from the CSE paper.

**Algorithm 2** Two-Stage PAC Labeling with Uncertainty Thresholding**Input:** Calibration set  $\mathcal{I}_{\text{cal}}$ , test set  $\mathcal{I}_{\text{test}}$ , LLM  $\tilde{f}$ , expert  $f$ , loss  $\ell$ , error tolerance  $\epsilon$ , confidence level  $\alpha$ .**Output:** Composite predictor  $\hat{f}$ .1: For all  $i \in \mathcal{I}_{\text{cal}}$ , compute  $\tilde{y}_i = \tilde{f}(x_i)$  and uncertainty scores  $U_i$ .2: For each  $u \in \{U_i : i \in \mathcal{I}_{\text{cal}}\}$ , compute  $\hat{L}_u(\alpha)$  as in Eq.(7).

3: Set

$$\hat{u} = \max\{u : \hat{L}_u(\alpha) \leq \epsilon\}.$$

4: For all  $i \in \mathcal{I}_{\text{test}}$ , compute  $\tilde{y}_i = \tilde{f}(x_i)$  and  $U_i$ .

5: Define

$$\hat{f}(x_i) = \begin{cases} \tilde{f}(x_i), & U_i < \hat{u}, \\ f(x_i), & U_i \geq \hat{u}. \end{cases} \quad \forall i \in \mathcal{I}_{\text{test}}.$$

6: **Return**  $\hat{f}$ .

Table 2. Accuracy (%) of different models on annotation datasets.

Dataset / LLM	Qwen3-4B-Instruct-2507	Qwen3-4B-Thinking-2507	Llama-3.1-8B-Instruct	DeepSeek-R1-Distill-Llama-8B	Qwen2.5-32B-Instruct	DeepSeek-R1-Distill-Qwen-32B
MMLU-Redux	79.83	81.57	64.77	67.83	79.73	84.93
MATH-500	91.80	96.00	46.80	87.20	80.20	92.60
MATH-L5	91.12	94.73	20.53	80.72	56.17	93.62
ZebraLogic	79.10	89.00	13.2	34.80	24.50	68.60
HumanEval <sup>+</sup>	92.68	97.56	92.68	93.90	82.32	99.39

**CoAnnotating (Li et al., 2023).** We consider a cost-aware heuristic that selects thresholds by directly optimizing empirical performance on a calibration set, without formal risk control. The method partitions examples according to their uncertainty scores and searches for threshold pairs that minimize total annotation cost while keeping the empirical loss below a target level. The selected thresholds are then applied to the test set using the same escalation rule. The full procedure is summarized in Algorithm 3.

## F. Implementation details

**Hyperparameter settings of LLMs** In this study, we configure the decoding parameters as follows: for Qwen/Qwen3-4B-Instruct-2507, we set Temperature=0.7, TopP=0.8, TopK=20, and MinP=0; for Qwen/Qwen3-4B-Thinking-2507, we set Temperature=0.6, TopP=0.95, TopK=20, and MinP=0. For Llama/Llama-3.1-8B-Instruct, we set Temperature=0.7, TopP=0.8, and MinP=0; for DeepSeek/DeepSeek-R1-Distill-Llama-8B, we set Temperature=0.6, TopP=0.95, and MinP=0. For Qwen/Qwen2.5-32B-Instruct, we set Temperature=0.7, TopP=0.8, and MinP=0; for DeepSeek/DeepSeek-R1-Distill-Qwen-32B, we set Temperature=0.6, TopP=0.95, and MinP=0. For experiments on the MATH-500 dataset, we additionally apply a presence penalty of 1.5 to avoid repetitive generation. Experiments were run on NVIDIA GeForce RTX 4090, NVIDIA L40, and NVIDIA RTX Pro 6000.

**Dataset details** Table 3 summarizes the datasets employed in our experiments, together with their corresponding splitting strategies. For each dataset, we report its type, overall size, and the partitioning into calibration and test sets. In Table 2, we report the accuracy of LLMs on these annotation datasets.

**Prompts for verbalized confidence** In Table 4, we present the prompt used to elicit the verbalized confidence scores. After ten trials, we obtained the average confidence score and defined the verbalized uncertainty score as 1 minus this average confidence. For PAC labeling baseline on HumanEval<sup>+</sup>, we find that using this prompt, we fail to extract the verbalized confidence for Qwen3-4B-Thinking-2507. Therefore, on HumanEval<sup>+</sup>, we instead use Qwen3-4B-Instruct-2507 to obtain verbalized confidence for the outputs generated by Qwen3-4B-Thinking-2507.

**API pricing parameters** We evaluate API cost functions using two model pairs: (i) Qwen2.5-32B-Instruct and DeepSeek-R1-Distill-Qwen-32B, and (ii) Llama-3.1-70B-Instruct and DeepSeek-R1-Distill-Llama-70B. For the Qwen2.5-32B-Instruct and DeepSeek-R1-Distill-Qwen-32B models, both input and output token prices are identical under Aliyun’s pricing scheme,

**Algorithm 3** Heuristic Mean Baseline

---

**Input:** Calibration set  $\mathcal{D}_{\text{cal}} : \{(x_i, y_i)\}_{i \in [m]}$ , test set  $\mathcal{D}_{\text{test}} : \{(x_{m+j}, y_{m+j})\}_{j \in [n]}$ ;  
 Three annotators: non-thinking model  $\tilde{f}_1$ , thinking model  $\tilde{f}_2$ , and expert annotator;  
 Uncertainty scores  $U(x) \in [0, 1]$ ;  
 Per-annotator cost array  $\text{cost}_i(x)$ ;  
 Target error threshold  $\epsilon$

**Output:** Predicted labels on  $\mathcal{D}_{\text{test}}$

- 1: **Calibration phase:**
- 2: Construct a grid  $\mathcal{U}$  from the unique values of  $\{U(x) : x \in \mathcal{D}_{\text{cal}}\}$
- 3: Initialize an empty candidate set  $\mathcal{T}$
- 4: **for** each  $u_1 \in \mathcal{U}$  **do**
- 5:     Search for the largest  $u_2 \geq u_1$  such that
 
$$\text{Mean}(\{L(x_i, u_1, u_2)\}_{i \in [m]}) \leq \epsilon$$
- 6:     **if** such  $u_2$  exists **then**
- 7:         Add  $(u_1, u_2)$  to  $\mathcal{T}$
- 8:     **else**
- 9:         **break**
- 10:   **end if**
- 11: **end for**
- 12: Select  $(u_1^*, u_2^*) \in \mathcal{T}$  minimizing the total calibration cost
- 13: **Evaluation phase:**
- 14: **for** each  $(x, y) \in \mathcal{D}_{\text{test}}$  **do**
- 15:     **if**  $U(x) \leq u_1^*$  **then**
- 16:         Output prediction of  $\tilde{f}_1$
- 17:     **else if**  $U(x) \leq u_2^*$  **then**
- 18:         Output prediction of  $\tilde{f}_2$
- 19:     **else**
- 20:         Output expert label
- 21:     **end if**
- 22: **end for**

---

with input tokens priced at 0.002 Chinese Yuan per thousand tokens and output tokens priced at 0.006 Chinese Yuan per thousand tokens. For the 70B model pair, we adopt the API pricing provided by DeepInfra. The Llama-3.1-70B-Instruct model is priced at \$0.40 per thousand tokens for both input and output, while DeepSeek-R1-Distill-Llama-70B is priced at \$0.80 per thousand tokens for both input and output.

**Hyperparameter setting for empirical comparison between HyPAC and CSE.** We use dataset-specific hyperparameters for CSE to attain comparable errors to HyPAC. For logits-based uncertainty score, on MMLU-Redux, we set  $\alpha = 0.07$  and  $\epsilon = 0.15$ ; on MATH-500, we set  $\alpha = 0.08$  and  $\epsilon = 0.10$ ; on MATH-L5, we set  $\alpha = 0.1$  and  $\epsilon = 0.14$ ; on Zebra-Logic, we set  $\alpha = 0.10$  and  $\epsilon = 0.10$ ; on HumanEval<sup>+</sup>, we set  $\alpha = 0.10$  and  $\epsilon = 0.12$ . For verbalized uncertainty score, on MMLU-Redux, we set  $\alpha = 0.1$  and  $\epsilon = 0.15$ ; on MATH-500, we set  $\alpha = 0.1$  and  $\epsilon = 0.14$ ; on MATH-L5, we set  $\alpha = 0.1$  and  $\epsilon = 0.1$ ; on Zebra-Logic, we set  $\alpha = 0.10$  and  $\epsilon = 0.10$ ; on HumanEval<sup>+</sup>, we set  $\alpha = 0.05$  and  $\epsilon = 0.05$ .

## G. Extensive Study

### G.1. Performance of HyPAC under the semantic cosine-distance loss function

**HyPAC tightly controls the annotation error under the semantic cosine-distance loss function.** In our main experiment, we use the binary 0–1 loss function. In this subsection, we adopt the semantic cosine-distance loss function, which measures the semantic similarity between outputs in the embedding space. Formally, given the human-annotated labels  $y$  and the

Table 3. Details of datasets and splitting settings for annotation experiments.

Dataset	Dataset Type	Dataset Size	Split Setting	Size
MMLU-Redux	General Task	3000	Calibration Test	500 2500
MATH-500	Math Reasoning	500	Calibration Test	300 200
MATH-L5	Math Reasoning	721	Calibration Test	300 421
Zebra-Logic	Text Reasoning	750	Calibration Test	450 300
HumanEval+	Coding Task	164	Calibration Test	100 64

Table 4. Prompt for the verbalized confidence scores.

**System prompt:** You are a reasoning assistant. For each question and proposed answer, you must estimate how likely the proposed answer is correct.

**User prompt:**

Question: {QUESTION}

Answer: {ANSWER}

Provide a probability (between 0.0 and 1.0) that your answer is correct. Only output the probability.

LLM-annotated label  $\hat{y}$ , we obtain their embeddings  $v_y$  and  $v_{\hat{y}}$  using a semantic embedding model, and define the loss as:

$$\ell(y_i, \hat{y}_i) = 1 - \frac{v_y \cdot v_{\hat{y}}}{\|v_y\| \|v_{\hat{y}}\|}. \quad (10)$$

For the semantic embedding model, we adopt “Qwen/Qwen3-Embedding-8B” (Yang et al., 2025). We conduct experiments on MATH-500, which provides ground-truth chains of thought for each question. We compute the cosine distance between the LLM-generated output and the ground-truth chain of thought to measure their semantic similarity.

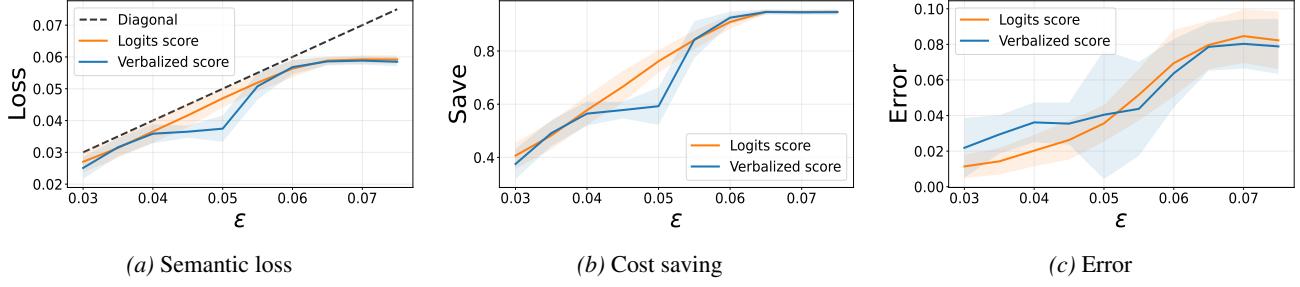
In Figure 5, we present the annotation results of HyPAC on MATH-500 under the semantic cosine-distance loss function. The results demonstrate that HyPAC tightly controls the annotation loss while substantially reducing the annotation cost under the semantic loss. Notably, the annotation error is also low, indicating that the semantic loss provides a reliable measure of annotation quality. Overall, these empirical results show that HyPAC achieves effective cost reduction without sacrificing annotation accuracy under the semantic loss function.

## G.2. Performance of HyPAC under the API cost function

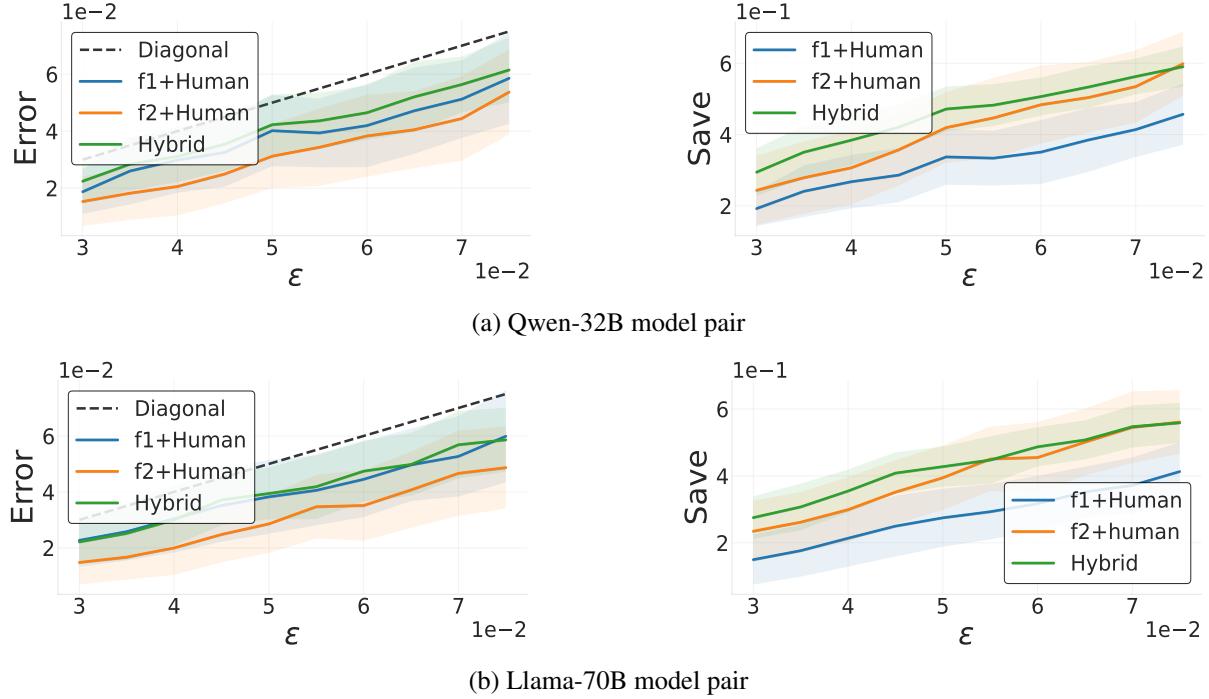
**HyPAC reduces the annotation cost under the API cost function.** We evaluate whether HyPAC effectively reduces annotation cost under API-based cost functions on MMLU-Redux. Figure 6 presents the annotation results of HyPAC and PAC Labeling under the API cost function. The results show that HyPAC consistently achieves substantially higher cost savings while maintaining controlled annotation error compared to PAC Labeling, demonstrating the superiority of HyPAC. Overall, these results demonstrate that HyPAC effectively reduces annotation cost under the API cost function while preserving tight error control.

## G.3. Ablation Study on Calibration Set Size

**Robust to calibration set size** We conduct experiments to investigate the stability of error and cost savings under different calibration set sizes. Specifically, we repeat the experiments with varying calibration ratios to examine how the size of the calibration set influences performance. The results are presented in Figure 7. The results show that HyPAC tightly controls



**Figure 5. HyPAC achieves higher cost savings while maintaining controlled semantic loss and error on MATH-500.** Loss, cost savings, and error of HyPAC with semantic loss on MATH-500 at confidence level  $\alpha = 0.05$ . We use Qwen3-4B-Instruct-2507 as the non-thinking model, Qwen3-4B-Thinking-2507 as the thinking model, and Qwen3-8B-Embedding as the embedding model.

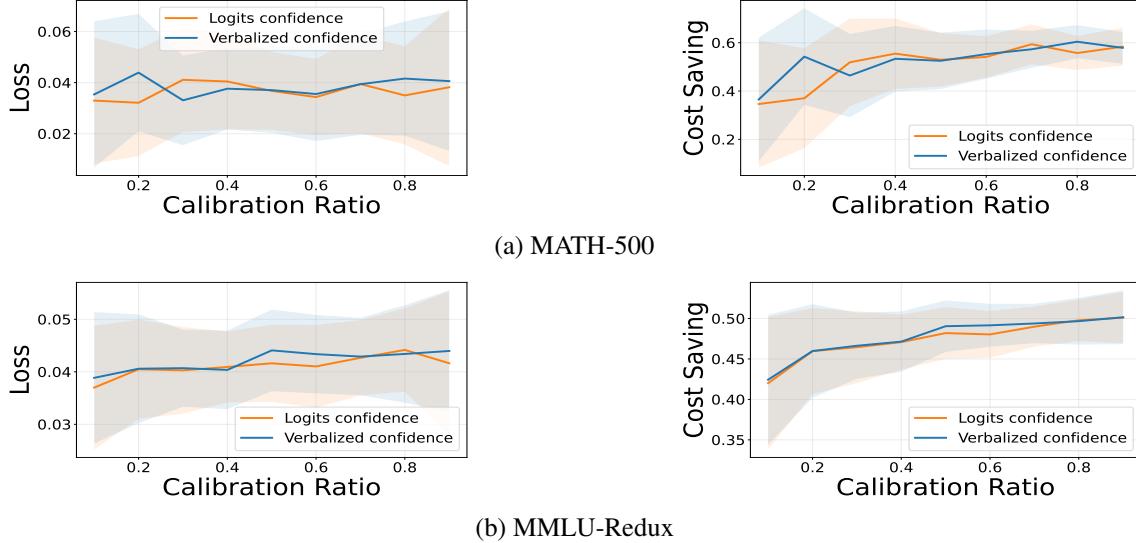


**Figure 6. HyPAC achieves higher cost savings than PAC Labeling while maintaining controlled error under the API cost function.** Error and cost savings of HyPAC compared against PAC Labeling baselines under the API cost function. Experiments are conducted on MMLU-Redux with logits-based uncertainty score under a confidence level of  $\alpha = 0.05$ . The shaded areas represent standard deviations.

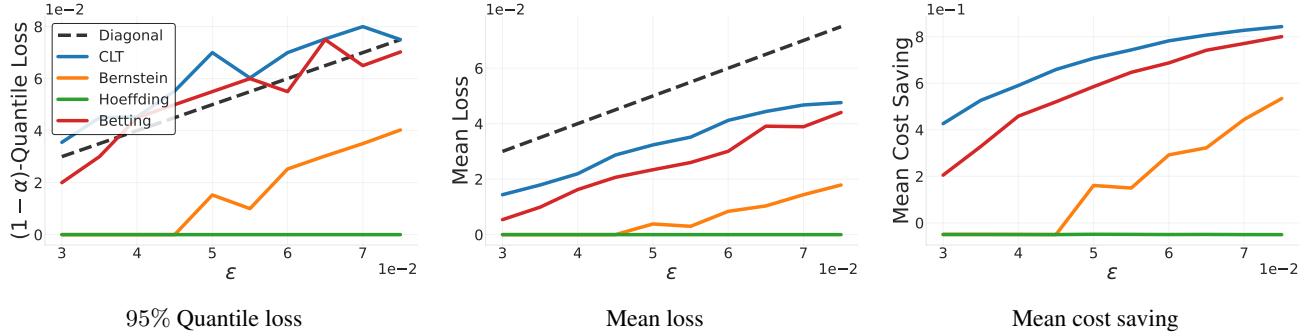
the annotation error below the target level, while a larger calibration size leads to slightly smaller shaded areas, indicating more stable annotation results. In addition, we observe that using a larger calibration set size can slightly increase the cost savings.

#### G.4. Performance of HyPAC under finite-sample UCBs

**Betting-based UCB achieves the best finite-sample trade-off.** Figure 8 compares the annotation performance of HyPAC under four different upper confidence bounds (UCBs): the CLT-based UCB, Hoeffding-based UCB, empirical Bernstein UCB, and betting-based UCB. A detailed description of these UCB constructions is provided in Appendix B. Among all the finite-sample UCBs, the betting-based UCB achieves the best cost-quality trade-off, as it yields the tightest control of the  $(1 - \alpha)$ -quantile loss and the highest cost savings. In addition, we observe that the  $(1 - \alpha)$ -quantile loss under the CLT-based UCB occasionally exceeds the target threshold  $\epsilon$ . This is because the CLT-based UCB is only asymptotically valid and does not have a guarantee in finite samples. Overall, these results show the advantage of betting-based UCB in finite-sample regimes.



**Figure 7. HyPAC maintains stable performance across different calibration set sizes.** Error and cost savings of HyPAC under different calibration set sizes with confidence level  $\alpha = 0.05$ . The shaded areas indicate one standard deviation.



**Figure 8. The CLT-based UCB provides tighter bounds and better cost-performance tradeoff than the Hoeffding-based UCB.** Annotation performance of HyPAC under different upper confidence bounds. The experiments are conducted on the MATH-500 using Qwen3-4B-Instruct as the non-thinking model and Qwen3-4B-Thinking as the thinking model with logits-based uncertainty score under a confidence level of  $\alpha = 0.05$ .

### G.5. Performance of HyPAC under distribution shift

We emphasize that the calibration set is sampled i.i.d. from a large unlabeled dataset, leaving the remaining unlabeled data for testing. Given a large unlabeled dataset, we first annotate a small subset as the calibration set, then apply HyPAC to the remaining data. This process ensures that the calibration and test datasets are naturally i.i.d., satisfying the standard assumptions of HyPAC. Thus, HyPAC generally does not encounter distribution shifts in annotation tasks.

Although our method does not provide theoretical guarantees under distribution shift, we evaluate the empirical robustness of HyPAC using ImageNet (Deng et al., 2009) and ImageNet-C (Brightness) (Hendrycks & Dietterich, 2019). Specifically, we use ImageNet as the calibration set and ImageNet-C Brightness with varying severity levels as the test set. We conduct experiments on image annotation tasks instead of LLM open-ended generation tasks because there are currently no benchmark datasets with different levels of covariate shift for open-ended LLM generation, as far as we know. We use ResNet-34 (He et al., 2016) as the small model  $f_1$ , and ResNet-152 as the large model  $f_2$ . We define the cost of using ResNet-34, ResNet-152, and expert annotation to be  $c_1 = 1$ ,  $c_2 = 2$ , and  $c_h = 8$ , respectively. We use one minus the maximum softmax probability as the uncertainty score function.

We present the performance of HyPAC across various testing sets in Table 5. The results show that the cost saving of HyPAC is getting worse with a higher severity of distribution shift (may be due to the degraded accuracy), while the annotation error is relatively insensitive. Across all the severity levels, HyPAC successfully controls the annotation error below the target level, demonstrating that HyPAC is empirically robust to distribution shift.

Table 5. Performance of HyPAC under varying severity levels of ImageNet-C Brightness corruption. HyPAC maintains stable error control under different levels of shift, while the cost saving decreases with increasing corruption severity.

Test Dataset	$\epsilon = 0.05$		$\epsilon = 0.1$	
	Error (%)	Save (%)	Error (%)	Save (%)
Severity 1	4.51	40.75	9.75	55.19
Severity 2	4.54	38.27	9.88	52.88
Severity 3	4.52	34.25	9.87	48.99
Severity 4	4.25	27.98	9.94	42.81
Severity 5	4.11	20.69	9.78	34.96

Table 6. HyPAC achieves high cost savings across different model pairs while maintaining controlled annotation error. Annotation performance of HyPAC across four benchmark datasets using two model pairs: Llama-3.1-8B-Instruct (non-thinking model  $\tilde{f}_1$ ) with DeepSeek-R1-Distill-Llama-8B (thinking model  $\tilde{f}_2$ ), and Qwen2.5-32B-Instruct (non-thinking model  $\tilde{f}_1$ ) with DeepSeek-R1-Distill-Qwen-32B (thinking model  $\tilde{f}_2$ ). We report results for the annotation error and cost saving at  $\epsilon = 0.05$  and  $\alpha = 0.05$  using the logits-based uncertainty score function and token-based cost function. **Bold** numbers indicate the best results.

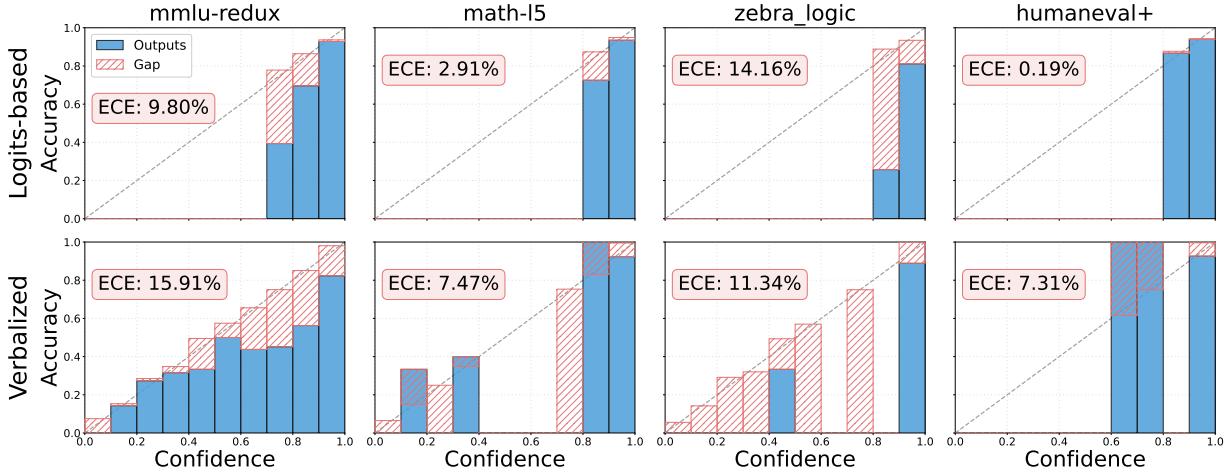
Datasets	Metric	Llama-8B model pair			Qwen-32B model pair		
		$\tilde{f}_1 + \tilde{f}_2 + \text{Human}$	$\tilde{f}_1 + \text{Human}$	$\tilde{f}_2 + \text{Human}$	$\tilde{f}_1 + \tilde{f}_2 + \text{Human}$	$\tilde{f}_1 + \text{Human}$	$\tilde{f}_2 + \text{Human}$
MMLU-Redux	Error (%)	3.95	3.87	4.08	4.00	3.83	4.09
	Save (%)	<b>16.83</b>	7.19	13.40	45.29	32.93	<b>50.57</b>
MATH-500	Error (%)	3.51	2.82	3.14	3.96	3.46	3.50
	Save (%)	<b>10.25</b>	-5.37	10.18	<b>56.89</b>	20.49	36.35
MATH-L5	Error (%)	3.21	3.17	3.55	3.70	3.24	3.32
	Save (%)	-5.39	-12.03	<b>11.38</b>	<b>72.67</b>	9.75	58.80
ZebraLogic	Error (%)	3.55	3.68	3.55	4.30	3.71	3.57
	Save (%)	-20.11	<b>-19.94</b>	-25.14	<b>6.35</b>	4.10	-9.14
HumanEval <sup>+</sup>	Error (%)	3.78	3.80	2.78	3.21	3.38	0.78
	Save (%)	<b>47.64</b>	31.95	5.97	<b>90.41</b>	16.05	85.34

## G.6. More annotation results across diverse models and datasets

Table 6 reports the annotation performance of HyPAC under two model pair configurations: (i) Llama-3.1-8B-Instruct as the non-thinking model paired with DeepSeek-R1-Distill-Llama-8B as the thinking model, and (ii) Qwen2.5-32B-Instruct as the non-thinking model paired with DeepSeek-R1-Distill-Qwen-32B as the thinking model. HyPAC achieves the highest cost savings across most datasets. However, on the Zebra-Logic dataset with the Llama-8B model pair, HyPAC incurs a higher annotation cost than using expert annotation alone. This behavior arises because the Llama-8B model pair attains low accuracy on Zebra-Logic (13.2% and 34.8%), and HyPAC relies heavily on expert annotations to satisfy the accuracy constraint. This behavior shows that the cost savings of HyPAC are closely tied to the capability of the employed LLMs. Higher-capacity LLMs require fewer expert annotations to satisfy accuracy constraints, thus yielding greater cost savings.

## G.7. More reliability diagrams of logits-based uncertainty score and verbalized uncertainty score

We present the reliability diagrams of Qwen3-4B-Instruct-2507 on MMLU-redux, MATH-L5, Zebra-Logic and HumanEval<sup>+</sup> in Figure 9. The results show that logits-based confidence achieves lower ECE than verbalized confidence in all datasets except for Zebra-Logic.



*Figure 9.* Reliability diagrams of Qwen3-4B-Instruct-2507 across four benchmark datasets (MMLU-Redux, MATH-L5, Zebra-Logic, and HumanEval<sup>+</sup>). Each column corresponds to a dataset. The top row uses logits-based confidence, while the bottom row uses verbalized confidence.