

SELECTIVE LABELING WITH FALSE DISCOVERY RATE CONTROL

Huipeng Huang^{1*}, Wenbo Liao^{1,2*}, Huajun Xi¹, Hao Zeng¹, Mengchen Zhao³, Hongxin Wei^{1†}

¹Department of Statistics and Data Science, Southern University of Science and Technology

²Department of Mathematics, The Chinese University of Hong Kong

³School of Software Engineering, South China University of Technology

ABSTRACT

Obtaining high-quality labels for large datasets is expensive, requiring massive annotations from human experts. While AI models offer a cost-effective alternative by predicting labels, their label quality is compromised by the unavoidable labeling errors. Existing methods mitigate this issue through selective labeling, where AI labels a subset and human labels the remainder. However, these methods lack theoretical guarantees on the quality of AI-assigned labels, often resulting in unacceptably high labeling error within the AI-labeled subset. To address this, we introduce **Conformal Labeling**, a novel method to identify instances where AI predictions can be provably trusted. This is achieved by controlling the false discovery rate (FDR), the proportion of incorrect labels within the selected subset. In particular, we construct a conformal p -value for each test instance by comparing AI models' predicted confidence to those of calibration instances mislabeled by AI models. Then, we select test instances whose p -values are below a data-dependent threshold, certifying AI models' predictions as trustworthy. We provide theoretical guarantees that Conformal Labeling controls the FDR below the nominal level, ensuring that a predefined fraction of AI-assigned labels is correct on average. Extensive experiments demonstrate that our method achieves tight FDR control with high power across various tasks, including image and text labeling, and LLM QA.

1 INTRODUCTION

Large-scale, high-quality labeled data is crucial for the machine learning pipelines (Deng et al., 2009). While experts could provide high-quality labels for moderately sized datasets, the growing size of modern datasets has made this approach prohibitively expensive. AI models offer a cost-effective alternative by predicting labels, bypassing the need for human experts. However, AI models are prone to labeling error (Northcutt et al., 2021; Tan et al., 2024). For example, empirical evidence demonstrates that even state-of-the-art LLMs exhibit high labeling error when used for text annotation Baumann et al. (2025). The labeling error inherent to AI models significantly compromises their label quality, hindering the deployment of AI labeling for production. To balance the trade-off between labeling cost and error, selective labeling has been a promising solution (Geifman & El-Yaniv, 2017; Wang et al., 2023) by combining AI predictions with expert annotations.

Prior work on selective labeling primarily designed heuristic methods (Wang et al., 2021; Bernhardt et al., 2022; Wang et al., 2024a) that rely on AI models to label high-confidence instances while deferring the rest to human experts. However, these methods do not provide any theoretical guarantees on label quality. To address this, a recent work (Candès et al., 2025) proposes probably approximately correct (PAC) labeling, which guarantees that the overall labeling error is controlled with high probability. They achieve the guarantee by counterbalancing the error of AI labels with the zero error of expert labels. While the guarantee is theoretically appealing, the labeling error of AI models can be unacceptably high, even reaching 100%, while the overall labeling error is controlled. These limitations motivate us to investigate how to provably guarantee the quality of AI-assigned labels.

*Equal Contribution.

†Correspond to weihx@sustech.edu.cn.

In this work, we propose **Conformal Labeling**, which formulates the labeling problem as multiple hypotheses testing. This method allows us to select a subset of the unlabeled dataset for AI labeling and ensure the quality of AI-assigned labels by providing a rigorous guarantee on the false discovery rate (FDR). In particular, given a labeled calibration dataset, we compute conformal p -values for test instances through a rank-based comparison of their predicted confidence against those of calibration instances where AI predictions are incorrect. This ensures that p -values for mislabeled instances stochastically dominate the uniform distribution on $[0, 1]$, while p -values for correctly labeled instances are concentrated near zero, allowing selection procedures to distinguish between them easily. Then, we select test instances whose p -values are less than or equal to a data-driven threshold, which is delicately set with the calibration dataset to control FDR at the desired level. We further theoretically prove that Conformal Labeling achieves valid FDR control under mild assumptions, ensuring that the expected proportion of incorrect labels in the selected subset is below the desired level. While the label quality of existing methods depends heavily on model performance, Conformal Labeling guarantees label quality regardless of the underlying model’s performance by strictly controlling the FDR.

To validate our method, we conduct extensive experiments on various labeling tasks, including image labeling (ImageNet (Deng et al., 2009), ImageNet-V2 (Recht et al., 2019)), text labeling (stance on global warming (Luo et al., 2021), and misinformation (Gabriel et al., 2022)), and LLM QA (MedMCQA (Pal et al., 2022), MMLU (Hendrycks et al., 2021), MMLU-Pro (Wang et al., 2024a)) tasks. The results demonstrate that Conformal Labeling achieves high power with controlled FDR, indicating that AI models can label a large proportion of data with high quality. For example, Conformal Labeling can label 58.67% of the ImageNet dataset with ResNet-34 (He et al., 2016), while keeping the FDR below 10%. In comparison, a naive approach of using AI-assigned labels for the entire dataset results in a labeling error of over 25%. Moreover, through comprehensive ablation studies, we validate that Conformal Labeling is robust to the size of calibration datasets, and more powerful models enable better selection results.

We summarize our contributions as follows:

- We propose Conformal Labeling, a novel method for identifying instances where AI predictions could be provably trusted. Regardless of AI models’ performance, Conformal Labeling guarantees the quality of AI-assigned labels by strictly controlling the FDR.
- We theoretically prove that Conformal Labeling provides a strict quality guarantee for AI-assigned labels: it achieves valid FDR control, ensuring the expected proportion of incorrect labels is below a user-specified level.
- We empirically show that Conformal Labeling significantly reduces the labeling cost while tightly controlling the FDR, through extensive experiments conducted on image labeling, text labeling, and LLM QA tasks with various models.

2 PRELIMINARIES

Problem setup. In this work, we study the problem of identifying instances where AI predictions can be provably trusted. Here, we give a formulation of multi-class classification as an example. Let \mathcal{X} denote the feature space and $\mathcal{Y} = \{1, \dots, K\}$ denote the label space. The test dataset $\mathcal{D}_{\text{test}} = \{X_j\}_{j=1}^m$ consists of m data instances, sampled i.i.d. from a data distribution $\mathcal{P}_{\mathcal{X}}$. We denote the unseen ground-truth labels of instance X_j as Y_j . Besides, we consider a pre-trained AI model $f : \mathcal{X} \rightarrow \mathbb{R}^{|\mathcal{Y}|}$ used to generate labels for the test dataset. For a given test instance X , the AI model predicts the label with the largest estimated probability $\hat{Y} = \arg \max_{y \in \mathcal{Y}} f_y(X)$, where $f_y(X)$ denotes the estimated class probability for class $y \in \mathcal{Y}$.

Since AI models are typically prone to labeling errors, we aim to select a large subset from the test dataset $\mathcal{D}_{\text{label}}$ to control the portion of incorrect labels. Formally, our goal is to identify the largest subset of indices $\mathcal{R} \subseteq \{1, \dots, m\}$ that controls the false discovery rate (FDR), defined as below:

$$\text{FDR} = \mathbb{E} \left[\frac{|\mathcal{R} \cap \mathcal{H}_0|}{\max(|\mathcal{R}|, 1)} \right],$$

where $\mathcal{H}_0 = \{j \in \{1, \dots, m\} : Y_j \neq \hat{Y}_j\}$ is the set of indices with incorrect predictions. For notation shorthand, we denote $[m] = \{1, \dots, m\}$ in the following. The FDR metric measures the

expected proportion of mislabeled samples within the AI-labeled subset, illustrating the quality of AI-assigned labels by explicitly bounding the fraction of incorrect labels.

In addition to FDR control, we also expect AI models to label as many test instances as possible, which corresponds to maximizing power:

$$\text{Power} = \mathbb{E} \left[\frac{|\mathcal{R} \cap \mathcal{H}_1|}{\max(|\mathcal{H}_1|, 1)} \right], \quad (1)$$

where $\mathcal{H}_1 = \{j \in [m] : Y_j = \hat{Y}_j\}$ is the set of indices where the AI prediction is correct. It should be emphasized that our method prioritizes FDR control over power, in that we strictly enforce $\text{FDR} \leq \alpha$ while optimizing power under the constraints.

In this work, we assume access to a small labeled calibration dataset $\mathcal{D}_{\text{cal}} = \{(X_i, Y_i)\}_{i=1}^n$. For convenience, we denote the test dataset as $\mathcal{D}_{\text{test}} = \{(X_j, Y_j)\}_{j=n+1}^{n+m}$, where Y_j is not observed. Since the labeling cost of a small dataset by human annotators is typically affordable, this assumption is practical in the real world and is also adopted in prior work (Candès et al., 2025). Besides, we assume that examples of the test and calibration datasets are both drawn i.i.d. from the joint distribution $\mathcal{P}_{\mathcal{X}\mathcal{Y}}$, a common setting in selective labeling (Jung et al., 2024; Candès et al., 2025).

Selective labeling methods and their limitations. Ensuring high-quality labels while reducing annotation costs has motivated extensive research on selective labeling. Prior work on selective labeling primarily focused on heuristic methods. For example, some studies design collaborative annotation frameworks that combine expert labels with LLM labels to streamline the annotation process (Li et al., 2023; Kim et al., 2024). Others propose domain-specific methods, such as meta-learning strategies for medical image labeling (Vrabc et al., 2022), annotation frameworks for text data (Duan & Lalor, 2023), and human–AI collaborative systems for object detection (Zhang et al., 2025). Although these heuristic approaches effectively reduce annotation costs, they lack formal guarantees on the label quality, which can result in unreliable labels when AI models perform poorly.

Probably approximately correct (PAC) labeling (Candès et al., 2025) addresses this limitation by providing a theoretical guarantee: the overall labeling error across the dataset is controlled with high probability. At its core, PAC labeling strategically collects zero-error expert labels for instances where the AI model exhibits the highest uncertainty, while relying on potentially imperfect AI predictions for more certain instances. This strategic allocation ensures that the dataset’s overall labeling error is small, as the zero-error expert annotations effectively counterbalance the error introduced by the AI-assigned labels. However, because the guarantee only applies at the aggregate dataset level, the subset labeled by AI may exhibit high labeling error. This limitation underscores a critical gap: existing selective labeling methods cannot ensure the quality of AI-assigned labels, hindering their reliable deployment in real-world applications. To address this, we explore methods to provably guarantee the quality of AI-assigned labels in selective labeling.

3 METHOD

3.1 CONFORMAL LABELING

Our previous section shows that existing methods cannot guarantee the quality of AI-assigned labels. To address this limitation, we propose **Conformal Labeling**, which identifies instances where AI predictions can be provably trusted by controlling the FDR. Our approach is composed of three primary steps: quantifying uncertainty, constructing conformal p -values, and thresholding.

Uncertainty quantification. Our approach builds on a key insight: we should select instances where the model exhibits high confidence in its predictions. To quantify the model confidence, we define an uncertainty score $\mathcal{S} : \mathcal{X} \rightarrow \mathbb{R}$, where a higher value indicates greater model uncertainty. We note that in the conformal inference framework, this score function is also known as the non-conformity score function. For example, we employ $\mathcal{S}(X) = 1 - \max_{y \in \mathcal{Y}} f_y(X)$ as our uncertainty score function (Hendrycks & Gimpel, 2016). This score is a valid measure of uncertainty, since prior works establish that misclassified samples generally receive lower probability scores (i.e., $\max_{y \in \mathcal{Y}} f_y(X)$) than correctly classified ones (Hendrycks & Gimpel, 2016; Tu et al., 2024).

Algorithm 1 Conformal Labeling

Require: Mislabeled Calibration set $\mathcal{D}_{\text{cal}}^0 = \{(X_i, Y_i)\}_{i=1}^{n_0}$, test instances $\{X_{n+j}\}_{j=1}^m$, pre-trained classifier f , calibration set size $n = |\mathcal{D}_{\text{cal}}|$, FDR target $\alpha \in (0, 1)$.

- 1: # 1. Compute uncertainty scores
- 2: **for** $i = 1, \dots, n+m$ **do**
- 3: Compute $S_i := 1 - \max_{y \in \mathcal{Y}} f_y(X_i)$.
- 4: **end for**
- 5: # 2. Construct conformal p-values
- 6: **for** $j = 1, \dots, m$ **do**
- 7: Construct \hat{p}_j according to equation 2.
- 8: **end for**
- 9: # 3. Thresholding
- 10: Compute $j^* = \max \left\{ j : \hat{p}_{(j)} \leq \frac{\alpha j(n+1)}{m(n_0+1)} \right\}$, where $\hat{p}_{(j)}$ is the j -th smallest p-value.
- 11: **Output:** $\mathcal{R} = \{j : \hat{p}_j \leq \hat{p}_{(j^*)}\}$.

Statistical guarantee via conformal p-value. To provide a statistical guarantee, we reformulate our problem as the following multiple hypothesis testing problem:

$$H_j^0 : Y_{n+j} \neq \hat{Y}_{n+j} \quad \text{v.s.} \quad H_j^1 : Y_{n+j} = \hat{Y}_{n+j}, \quad \forall j = 1, \dots, m$$

Rejecting the null hypothesis H_j^0 indicates that (X_{n+j}, \hat{Y}_{n+j}) should be included in the subset, as it is deemed to be classified correctly. To construct the selection subset, we employ *conformal p-value*, which builds upon conformal inference framework (Vovk et al., 1999; 2005). The underlying intuition is that: *a test instance X is likely a misclassification if its uncertainty score $S(X)$ is generally larger than the scores of instances that are known to be misclassified*. Leveraging this idea, conformal p-value is computed through a rank-based comparison of $S(X)$ against uncertainty scores of misclassified instances.

Formally, for the calibration dataset \mathcal{D}_{cal} , we identify the subset $\mathcal{D}_{\text{cal}}^0 \subseteq \mathcal{D}_{\text{cal}}$ where instances are misclassified by the AI model. For simplicity, we denote $\mathcal{D}_{\text{cal}}^0 = \{(X_i, Y_i)\}_{i=1}^{n_0}$, and thus $Y_i \neq \hat{Y}_i$ for $i = 1, \dots, n_0$. We compute the uncertainty scores for the entire dataset $\{(X_i, Y_i)\}_{i=1}^{n+m}$: $S_i = 1 - \max_{y \in \mathcal{Y}} f_y(X_i)$. Then, the conformal p-value for the instance X_{n+j} is computed by

$$\hat{p}_j = \frac{\sum_{i=1}^{n_0} \mathbf{1}\{S_i < S_{n+j}\} + (1 + \sum_{i=1}^{n_0} \mathbf{1}\{S_i = S_{n+j}\}) \cdot U_j}{n_0 + 1}, \quad (2)$$

where $U_j \sim \text{Uniform}[0, 1]$ are i.i.d. uniform random variable to randomize over ties when S_{n+j} equals some S_i , ensuring a continuous conformal p-value. Here, \hat{p}_j quantifies how extreme the uncertainty of X_{n+j} is compared to the scores of misclassified instances, with a small \hat{p}_j providing strong statistical evidence for correct prediction.

Standard results from conformal inference establish the validity of conformal p-value in practice: if the instance X_{n+j} is misclassified, then the conformal p-value stochastically dominates the uniform distribution on $[0, 1]$ (Bates et al., 2023; Jin & Candès, 2023): $\mathbb{P}\{\hat{p}_j \leq \alpha \mid H_j^0 \text{ is true}\} \leq \alpha$. This property indicates that the conformal p-values for misclassified instances are biased to be high, which allows us to set a threshold to flag potential correct instances, while controlling the FDR.

Thresholding. After obtaining conformal p-values for test instances, we apply a thresholding rule inspired by the Benjamini–Hochberg (BH) procedure (Benjamini & Hochberg, 1995) to select a maximal subset \mathcal{R} for AI labeling while controlling the FDR at level α . The key idea is that we gradually increase the acceptance threshold until including additional samples would risk exceeding the desired labeling error. In particular, let $p_{(1)} \leq \dots \leq p_{(m)}$ denote the ordered statistics of the p-values; the rejection set of our selection procedure applied to the conformal p-values is $\mathcal{R} = \{j : \hat{p}_j \leq \hat{p}_{(j^*)}\}$, where

$$j^* = \max \left\{ j : \hat{p}_{(j)} \leq \frac{\alpha j(n+1)}{m(n_0+1)} \right\},$$

with the convention that $\max \emptyset = 0$. We summarize the complete procedure of Conformal Labeling in Algorithm 1, which combines all three steps described above.

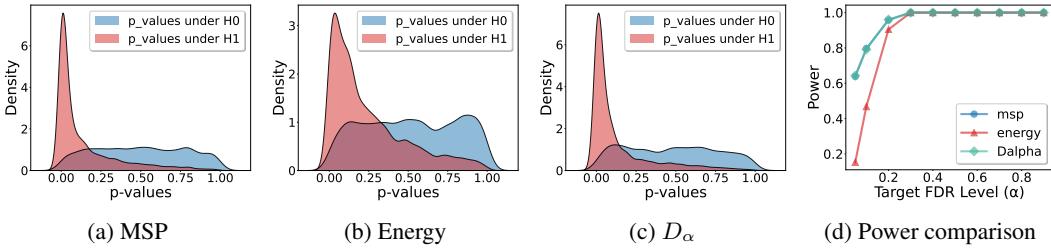


Figure 1: Empirical distributions and power (employed with our method) of conformal p -values under different uncertainty scores. The experiments include maximum softmax probability (MSP), energy, and DOCTOR- α score (D_α). The experiments are conducted on ImageNet with ResNet-34. The results show that both MSP and D_α score create a clear distinction between correct and incorrect predictions, thus achieving high statistical power. However, the energy score fails to provide this separation, leading to low power.

Theoretical results. We now provide a theoretical guarantee for Conformal Labeling. In the following theorem, we establish that Algorithm 1 controls the FDR below the desired level α . This theorem ensures that AI predictions can be provably trusted in the selected subset, as the expected proportion of incorrect labels is controlled below α . The proof is provided in Appendix B.2.

Theorem 3.1. Suppose calibration samples $(X_i, Y_i)_{i=1}^n$ and test samples $(X_{n+j}, Y_{n+j})_{j=1}^m$ are i.i.d. Let $\alpha \in (0, 1)$ be the target FDR level, and suppose the selection set \mathcal{R} is determined by Algorithm 1 applied at the target FDR level α . Define $p = \mathbb{E}[H_j^0]$, the probability that a test sample (X_{n+j}, Y_{n+j}) is incorrectly predicted. Then, the FDR of the selection set \mathcal{R} satisfies:

$$FDR \leq [1 - (1 - p)^{n+1}] \alpha \leq \alpha.$$

3.2 CHOICE OF UNCERTAINTY SCORE

In the above analysis, we establish that Conformal Labeling controls the FDR below the desired level. However, this guarantee alone is insufficient: a trivial procedure that simply labels nothing would achieve a perfect FDR of 0, yet offer no practical value. This highlights the need to also evaluate the method’s statistical power, which measures the method’s ability to identify as many correctly labeled instances as possible (see Eq. (1)).

As shown in prior work (Jin & Candès, 2023; Gui et al., 2024; Bai et al., 2025b), the statistical power of this method depends on the quality of the uncertainty score. In particular, a score that better separates correct from incorrect predictions directly increases statistical power (see Proposition 7 of Jin & Candès (2023)). To deliver practical recommendations, we empirically compare several uncertainty scores by visualizing their resulting p -value distributions and measuring the final statistical power employed with our method. We utilize a pre-trained ResNet-34 model on the ImageNet dataset, with three uncertainty scores: maximum softmax probability (MSP) (Hendrycks & Gimpel, 2016), energy score (Liu et al., 2020), and DOCTOR- α score (D_α) (Granese et al., 2021). We give an overview of these score functions in Appendix C. The results in Figure 1 show that both MSP and D_α score provide a clear distinction between correct and incorrect predictions, thus achieving high statistical power. However, the energy score fails to provide this separation, leading to low power. Given the comparable power performance of MSP and D_α score, we will use the more computationally efficient MSP in our main experiments.

4 EXPERIMENTS

In this section, we evaluate the effectiveness of Conformal Labeling on image labeling, text labeling, and LLM QA tasks with various models. We find that it achieves tight FDR control and high power, indicating that AI models can label a large proportion of data with high quality. We also conduct comprehensive ablation studies to provide practical guidance for applying our method.

Table 1: **Performance of Conformal Labeling on three labeling tasks.** \uparrow indicates that a larger value is better. We evaluate on Image Labeling (ImageNet, ImageNet-V2), text Labeling (Stance on global warming, Misinformation), and LLM QA (MedMCQA, MMLU) tasks. We report results for Conformal Labeling at $\alpha = 0.1$ and compare Conformal Labeling against two baselines: (i) a naive approach of labeling instances whose uncertainty score $S_{n+j} \leq 0.1$ with AI models, and (ii) labeling the entire dataset with AI models. The results show that Conformal Labeling consistently achieves tighter FDR control across all datasets and models compared with the baseline.

Task	Dataset	Model	Conformal Labeling ($\alpha=0.1$)			Naive ($S \leq 0.1$)			AI only
			FDR %	Power %(\uparrow)	Ratio %(\uparrow)	FDR %	Power %(\uparrow)	Ratio %(\uparrow)	
Image	ImageNet	ResNet-34	9.97	80.01	58.67	4.79	63.09	43.71	26.71
		DenseNet-161	9.99	85.03	65.56	5.58	72.08	52.98	22.89
		ResNeXt50	10.00	86.06	66.83	6.08	75.57	56.21	22.39
		CLIP-ViT-B/32	9.98	46.04	27.47	5.53	28.65	16.28	40.35
	ImageNet-V2	ResNet-34	10.00	61.99	37.87	7.25	56.03	33.15	39.03
		DenseNet-161	9.83	66.67	43.39	9.39	65.63	42.45	34.88
		ResNeXt50	9.95	67.86	44.75	10.17	68.71	45.38	34.08
		CLIP-ViT-B/32	9.96	34.56	18.10	7.86	25.83	13.18	47.78
Text	Stance	Llama-3.1-8B-Instruct	9.77	19.02	9.42	24.43	54.94	31.39	52.04
		Qwen3-32B	9.56	10.70	7.25	14.42	16.64	11.11	36.52
		Qwen2.5-72B-Instruct	9.71	26.97	17.84	26.27	74.49	59.01	35.09
	Misinformation	Llama-3.1-8B-Instruct	9.88	7.31	5.81	18.38	66.17	55.25	24.28
		Qwen3-32B	9.91	49.21	37.48	10.77	52.27	40.03	24.08
		Qwen2.5-72B-Instruct	9.58	44.36	34.81	17.81	87.13	74.72	21.69
QA	MedMCQA	Llama-3.1-8B-Instruct	9.70	31.44	18.90	15.01	46.74	29.53	40.35
		Qwen3-32B	9.75	49.80	33.27	13.79	65.27	45.36	33.44
		Llama-3.1-70B-Instruct	9.95	69.67	49.59	4.52	48.92	32.79	28.90
	MMLU	Llama-3.1-8B-Instruct	9.99	58.25	37.47	8.47	53.72	33.96	35.72
		Qwen3-32B	10.00	82.96	65.22	8.13	78.47	60.40	21.43
		Llama-3.1-70B-Instruct	9.96	88.20	72.10	4.18	67.94	52.17	18.24

4.1 EXPERIMENTAL SETUP

Tasks and datasets. We evaluate the effectiveness of Conformal Labeling across three labeling tasks, including image labeling, text labeling, and LLM question answering (QA). In appendix G, we also demonstrate how to extend Conformal Labeling to regression tasks. For the LLM QA task, the goal is to identify subsets of questions that LLMs can answer correctly. We employ common benchmark datasets for evaluations in each labeling task. For image classification, we use ImageNet (Deng et al., 2009) and its variant ImageNet-V2 datasets (Recht et al., 2019). For text labeling, we adopt two benchmark datasets. The first is Stance on Global Warming (Luo et al., 2021), which provides annotations ($Y_i \in \{\text{agree, neutral, disagree}\}$) to judge whether a headline agrees that global warming is a serious concern. The second is Misinformation (Gabriel et al., 2022), which contains binary annotations ($Y_i \in \{\text{misinfo, real}\}$) for identifying whether a given text contains misinformation. For the LLM QA task, we evaluate our method on MedMCQA (Pal et al., 2022), MMLU (Hendrycks et al., 2021), and MMLU-Pro (Wang et al., 2024b) datasets.

Models. We conduct extensive experiments on various open-sourced AI models. For image classification, we utilize three well-established deep image classifiers: ResNet-34 (He et al., 2016), DenseNet-161 (Huang et al., 2018), and ResNext50 (Xie et al., 2017). Additionally, we employ the Vision-Language Model CLIP (Radford et al., 2021), which is based on a Vision Transformer architecture (ViT-B/32) (Dosovitskiy et al., 2021). The above classifiers are provided by TorchVision (Paszke et al., 2019). For text labeling, we employ three LLMs: Llama-3.1-8B-Instruct (Grattafiori et al., 2024), Qwen3-32B, Qwen2.5-72B-Instruct (Qwen et al., 2025). For LLM QA tasks, we employ five LLMs: Qwen3-8B (Yang et al., 2025), Qwen3-14B, Qwen3-32B, Llama-3.1-8B-Instruct, and Llama-3.1-70B-Instruct. The above LLMs are provided by Hugging Face.

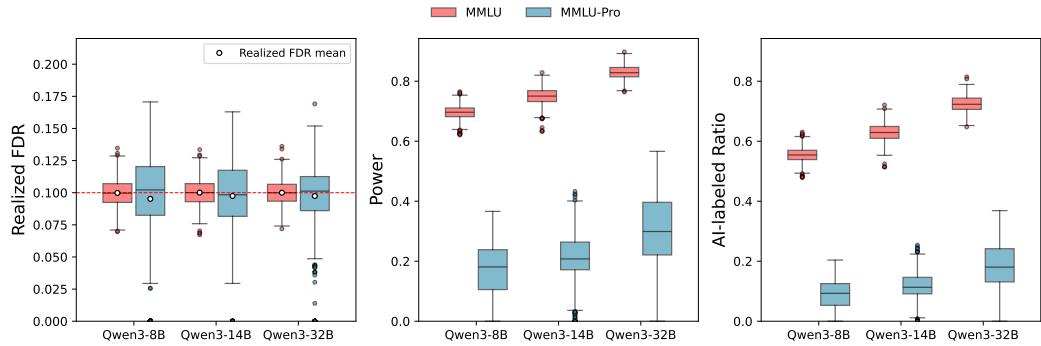


Figure 2: Performance comparison of Conformal Labeling across models of varying accuracy. We employ Qwen3-8B, Qwen3-14B, and Qwen3-32B (model accuracy increases with parameter count) on MMLU and MMLU-Pro with $\alpha = 0.1$. The results show that model with higher accuracy achieves greater power and AI-labeled ratio.

Baselines and evaluation metrics. We evaluate Conformal Labeling with two baseline methods: using AI models to label test instances with uncertainty scores $S_{n+j} \leq 0.1$, and applying AI predictions to the entire test dataset. We compare Conformal Labeling’s selection procedure against BH (Benjamini & Hochberg, 1995), Storey-BH (Storey, 2002), and Quantile-BH procedures (Benjamini et al., 2006). We evaluate the performance of our method and baseline using the following metrics: (1) FDR, the expected proportion of incorrect labels in the selected set \mathcal{R} ; (2) Power, the proportion of correctly labeled instances selected; (3) AI-labeled ratio, the number of data labeled by the AI models divided by the combined size of the calibration and test datasets.

Implementation details. To ensure the reliability of our results, we repeat each experiment 1000 times and report the average result. We randomly select 10% of the data as the calibration dataset. For all the experiments, we use the maximum softmax probability (MSP) as the uncertainty score function. In the LLM QA task, we adopt the standard multiple-choice evaluation pipeline: given a question and candidate answers (e.g., A, B, C, or D), the model estimates the probability of each option by extracting the logits corresponding to the option tokens and applying a softmax transformation, with the predicted label taken as the option with the highest probability. For text labeling, we reformulate each sample into a multiple-choice format (e.g., “positive,” “negative,” or “neutral”), enabling the same probability-extraction procedure as in the LLM QA task. More details of implementation are provided in Appendix F. We provide the code for reproducing our main experiments in this anonymous repository.

4.2 EXPERIMENTS RESULTS

Conformal Labeling achieves tight FDR control with high power. In table 1, we present the performance of Conformal Labeling against two baselines on three different labeling tasks: image labeling, text labeling, and LLM QA. A salient observation is that across all the labeling tasks and all the model architectures, Conformal Labeling successfully controls the FDR at or below the target FDR level. In comparison, both baseline methods lack FDR control, resulting in substantial labeling errors that compromise label quality when AI models are inaccurate. It is worth noting that the FDR is tightly controlled: for $\alpha = 10\%$, most experiments yield FDRs below 9.9%, with the largest deviation at 9.56%. This tight FDR control directly leads to high selection power. For example, on MMLU with Qwen3-32B at $\alpha = 10\%$, Conformal Labeling achieves a power of 82.96%, labeling 65.22% of the dataset with the AI model. Overall, empirical results show that Conformal Labeling consistently achieves tight FDR control with high power across different datasets and models.

Higher prediction accuracy enables better selection results. The performance of Conformal Labeling depends heavily on the underlying prediction accuracy, which is influenced by model capacity and dataset difficulty. We evaluate Conformal Labeling with Qwen3-8B, Qwen3-14B, and Qwen3-32B, whose increasing scales provide greater capacity. MMLU-Pro, more challenging than MMLU, results in lower prediction accuracy. Figure 2 presents the evaluation results. In all cases,

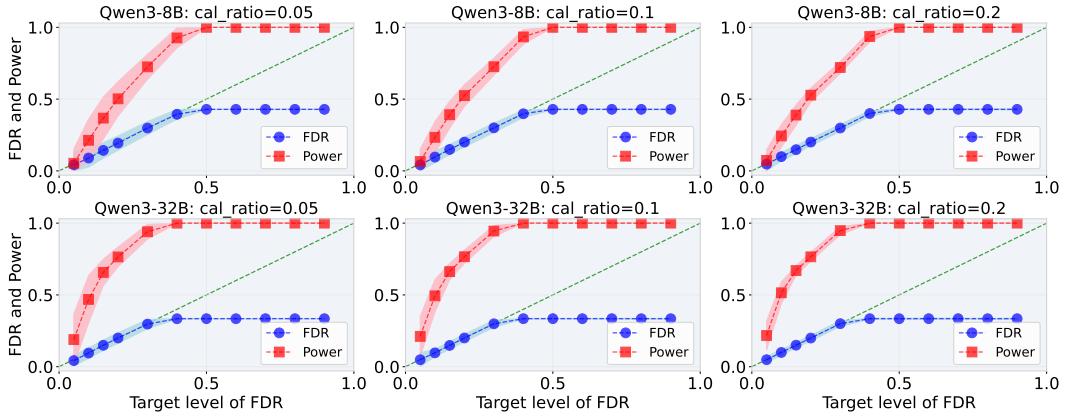


Figure 3: Performance of Conformal Labeling with varying calibration set sizes on the MedMCQA dataset. The top row corresponds to the results from Qwen3-8B and the bottom row to those from Qwen3-32B; each column corresponds to a value of calibration ratio. Shaded regions indicate one standard deviation around the average. The results show that a large calibration set slightly reduces the variance of FDR and power, leading to a more robust selection outcome. Overall, our method is robust to changes in the calibration set size.

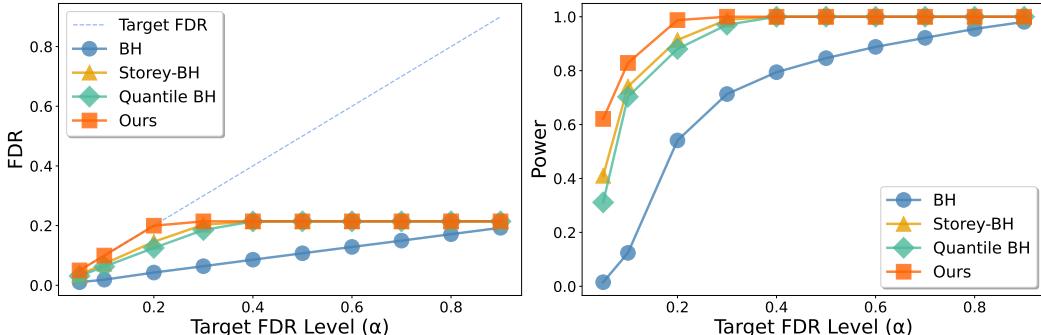


Figure 4: Ablation study on the selection procedure. The ablation study is conducted on the MMLU dataset with Qwen3-32B. We substitute Conformal Labeling’s selection procedure with alternative selection procedures, including BH, Storey-BH, and Quantile-BH procedures. The figure shows the Conformal Labeling’s selection procedure consistently achieves the tightest FDR control and the highest power among all the selection procedures.

the realized FDR remains below and close to $\alpha = 0.1$. Higher accuracy—achieved with stronger models or easier datasets—boosts power and the AI-labeled ratio. Specifically, for any given model, performance is superior on MMLU compared to MMLU-Pro. Similarly, for each dataset, larger models yield greater power and AI-labeled ratios. Overall, higher prediction accuracy leads to better selection results by achieving higher power and AI-labeled ratio with controlled FDR.

How many calibration samples are needed? The size of the calibration set plays a crucial role in constructing reliable conformal p -values. To study the effect of $|D_{\text{cal}}|$ on FDR and power, in Figure 3, we label $\{5\%, 10\%, 20\%\}$ of the unlabeled dataset as the calibration dataset. Our results demonstrate that Conformal Labeling is robust to calibration set size: even with a 5% calibration ratio, the FDR remains controlled with low standard deviation. Increasing the calibration ratio reduces the variance of both FDR and power, although the improvement from 10% to 20% is negligible. Based on this trade-off between variance reduction and labeling cost, we use a 10% calibration ratio for all the experiments. In summary, while Conformal Labeling is robust to calibration set sizes, larger calibration sets reduce the variance of FDR and power, thereby enhancing selection stability.

Ablation study on the selection procedure. To validate the effectiveness of Conformal Labeling’s selection procedure, we conduct an ablation study substituting it with alternative selection procedures, including BH, Storey-BH (Storey, 2002), and Quantile-BH (Benjamini et al., 2006) procedures (see Appendix D for details). The hyperparameters for the Storey-BH and Quantile-BH procedures are chosen using a bootstrap method detailed in the Appendix E. An ideal selection procedure aims to maximize power while keeping the FDR below the desired level.

The results of this ablation study are presented in Figure 4. While all the selection procedures successfully control the FDR below the desired level, Conformal Labeling’s selection procedure (red line) provides the tightest control, with realized FDR consistently closest to the desired level (dashed blue line). In comparison, alternative selection procedures are more conservative, especially at low target FDR levels, which reduces power. Overall, this ablation study highlights the superiority of Conformal Labeling’s selection procedure.

5 RELATED WORK

Selection with conformal p -values. Conformal p -values are widely studied for their distribution-free, model-agnostic properties in selection tasks. Existing methods fall into two main categories: conformal novelty detection (Bates et al., 2023; Bashari et al., 2023; Wu et al., 2025; Bashari et al., 2025; Lee et al., 2025; Huo et al., 2025) and conformal selection (Jin & Candès, 2023; Gui et al., 2024). Conformal novelty detection aims to identify out-of-distribution instances, while conformal selection aims to select data points that meet a specific quality criterion. For conformal novelty detection, recent advances primarily focus on improving the power of the selection procedure by using various forms of side information (Liang et al., 2024; Marandon et al., 2024; Zhao & Sun, 2025). Jin & Candès (2023) introduced the conformal selection framework for data selection in regression settings, which has since been extended to various settings: multivariate data selection (Bai et al., 2025b), online data selection (Xu & Ramdas, 2024; Liu et al., 2025), and human-in-the-loop adaptive data selection (Gui et al., 2025). Recent works have also applied CS to different tasks such as candidate screening (Lu et al., 2025), drug discovery (Jin & Candès, 2023; Bai et al., 2025a), and foundation model alignment (Gui et al., 2024). However, none of them has studied how to construct conformal p -values and design selection procedures in the selective labeling task.

Selective labeling. Our method builds on the idea of selective labeling, which prioritizes the collection of expert labels for instances where the AI model exhibits uncertainty, and relies on the model’s prediction where it is confident. (Gu et al., 2012; Vrabc et al., 2022). Prior works on selective labeling have explored various heuristic methods, including collaborative annotation frameworks that integrate expert and LLM efforts (Li et al., 2023), and domain-specific applications in text, vision, and medicine (Kim et al., 2024; Vrabc et al., 2022; Duan & Lalor, 2023; Zhang et al., 2025). More recent research aims to provide theoretical foundations, with methods designed to control the overall labeling error (Candès et al., 2025) or to enable valid statistical inference (Zrnic & Candes, 2024; Gligorić et al., 2024). Our method is also closely related to selective prediction, where models are allowed to abstain from making a prediction when uncertain about the output (Geifman & El-Yaniv, 2017; Mozannar & Sontag, 2020; Yang et al., 2023). For LLMs, recent works have also studied how to teach the model not to predict when the model is uncertain (Kamath et al., 2020; Yoshikawa & Okazaki, 2023). These selective prediction methods also lack theoretical guarantees for the AI predictions. Distinct from prior methods, our work is the first to provide a guarantee for the quality of AI-assigned labels, thereby enabling the reliable deployment of selective labeling.

6 CONCLUSION

In this work, we propose Conformal Labeling, a novel selective labeling method for identifying samples where AI predictions can be provably trusted. This is achieved by controlling the false discovery rate (FDR), which ensures that the expected fraction of incorrect labels in the selected subset is below a user-specified level. The key idea is to reformulate selective labeling as multiple hypothesis testing, which enables distinct theoretical guarantees and methodological advantages compared to prior approaches. In particular, we construct a conformal p -value for each test instance by comparing the AI model’s predicted confidence to those of mislabeled calibration instances. Then, we select all the test samples whose p -values are smaller than or equal to a data-dependent

threshold. We theoretically prove that Conformal Labeling successfully controls the FDR under mild assumptions. Extensive experiments demonstrate that Conformal Labeling achieves tight FDR control and high power across various tasks, including image and text labeling, and LLM QA. We hope the insight from this work will inspire future research to explore reliable selective labeling.

REFERENCES

- Tian Bai, Peng Tang, Yuting Xu, Vladimir Svetnik, Bingjia Yang, Abbas Khalili, Xiang Yu, and Archer Yang. Conformal selection for efficient and accurate compound screening in drug discovery. 2025a.
- Tian Bai, Yue Zhao, Xiang Yu, and Archer Y. Yang. Multivariate conformal selection. In *Forty-second International Conference on Machine Learning*, 2025b. URL <https://openreview.net/forum?id=g2tr7nA4pS>.
- Meshi Bashari, Amir Epstein, Yaniv Romano, and Matteo Sesia. Derandomized novelty detection with fdr control via conformal e-values. *Advances in Neural Information Processing Systems*, 36: 65585–65596, 2023.
- Meshi Bashari, Matteo Sesia, and Yaniv Romano. Robust conformal outlier detection under contaminated reference data. In *Forty-second International Conference on Machine Learning*, 2025. URL <https://openreview.net/forum?id=s55Af9Emyq>.
- Stephen Bates, Emmanuel Candès, Lihua Lei, Yaniv Romano, and Matteo Sesia. Testing for outliers with conformal p-values. *The Annals of Statistics*, 51(1):149–178, 2023.
- Joachim Baumann, Paul Röttger, Aleksandra Urman, Albert Wendsjö, Flor Miriam Plaza del Arco, Johannes B. Gruber, and Dirk Hovy. Large language model hacking: Quantifying the hidden risks of using llms for text annotation, 2025. URL <https://arxiv.org/abs/2509.08825>.
- Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.
- Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4):1165–1188, 2001.
- Yoav Benjamini, Abba M Krieger, and Daniel Yekutieli. Adaptive linear step-up procedures that control the false discovery rate. *Biometrika*, 93(3):491–507, 2006.
- Mélanie Bernhardt, Daniel C Castro, Ryutaro Tanno, Anton Schwaighofer, Kerem C Tezcan, Miguel Monteiro, Shruthi Bannur, Matthew P Lungren, Aditya Nori, Ben Glocker, et al. Active label cleaning for improved dataset quality under resource constraints. *Nature communications*, 13(1): 1161, 2022.
- Emmanuel J. Candès, Andrew Ilyas, and Tijana Zrnic. Probably approximately correct labels. *arXiv preprint arXiv:2506.10908*, 2025. URL <https://arxiv.org/abs/2506.10908>.
- Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.
- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL <https://openreview.net/forum?id=YicbFdNTTy>.
- Xiaojing Duan and John P Lalor. H-coal: Human correction of ai-generated labels for biomedical named entity recognition. *arXiv preprint arXiv:2311.11981*, 2023.
- Saadie Gabriel, Skyler Hallinan, Maarten Sap, Pemi Nguyen, Franziska Roesner, Eunsol Choi, and Yejin Choi. Misinfo reaction frames: Reasoning about readers’ reactions to news headlines, 2022. URL <https://arxiv.org/abs/2104.08790>.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. *Advances in neural information processing systems*, 30, 2017.

Kristina Gligorić, Tijana Zrnic, Cinoo Lee, Emmanuel J Candès, and Dan Jurafsky. Can unconfident ILM annotations be used for confident conclusions? *arXiv preprint arXiv:2408.15204*, 2024.

Federica Granese, Marco Romanelli, Daniele Gorla, Catuscia Palamidessi, and Pablo Piantanida. Doctor: A simple method for detecting misclassification errors. *Advances in Neural Information Processing Systems*, 34:5669–5681, 2021.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind That-tai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahaedokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jong-soo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasudevan Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhota, Lauren Rantala-Yearly, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidor, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparth, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mi-haylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia

Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghatham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wencheng Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhao-duo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL <https://arxiv.org/abs/2407.21783>.

Quanquan Gu, Tong Zhang, Jiawei Han, and Chris Ding. Selective labeling via error bound minimization. *Advances in neural information processing systems*, 25, 2012.

Yu Gui, Ying Jin, and Zhimei Ren. Conformal alignment: Knowing when to trust foundation models with guarantees. *Advances in Neural Information Processing Systems*, 37:73884–73919, 2024.

Yu Gui, Ying Jin, Yash Nair, and Zhimei Ren. Acs: An interactive framework for conformal selection. *arXiv preprint arXiv:2507.15825*, 2025.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

-
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations (ICLR)*, 2021. arXiv:2009.03300.
- Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks, 2018. URL <https://arxiv.org/abs/1608.06993>.
- Yuyang Huo, Xiaoyang Wu, Changliang Zou, and Haojie Ren. Unified conformalized multiple testing with full data efficiency. *arXiv preprint arXiv:2508.12085*, 2025.
- Ying Jin and Emmanuel J Candès. Selection by prediction with conformal p-values. *Journal of Machine Learning Research*, 24(244):1–41, 2023.
- Jaehun Jung, Faeze Brahman, and Yejin Choi. Trust or escalate: Llm judges with provable guarantees for human agreement. *arXiv preprint arXiv:2407.18370*, 2024.
- Amita Kamath, Robin Jia, and Percy Liang. Selective question answering under domain shift. *arXiv preprint arXiv:2006.09462*, 2020.
- Hannah Kim, Kushan Mitra, Rafael Li Chen, Sajjadur Rahman, and Dan Zhang. Meganno+: A human-llm collaborative annotation system, 2024. URL <https://arxiv.org/abs/2402.18050>.
- Junu Lee, Ilia Popov, and Zhimei Ren. Full-conformal novelty detection: A powerful and non-random approach. *arXiv preprint arXiv:2501.02703*, 2025.
- Minzhi Li, Taiwei Shi, Caleb Ziems, Min-Yen Kan, Nancy Chen, Zhengyuan Liu, and Diyi Yang. Coannotating: Uncertainty-guided work allocation between human and large language models for data annotation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.92. URL <http://dx.doi.org/10.18653/v1/2023.emnlp-main.92>.
- Ziyi Liang, Matteo Sesia, and Wenguang Sun. Integrative conformal p-values for out-of-distribution testing with labelled outliers. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(3):671–693, 01 2024. ISSN 1369-7412. doi: 10.1093/rssb/qkad138. URL <https://doi.org/10.1093/rssb/qkad138>.
- Kangdao Liu, Huajun Xi, Chi-Man Vong, and Hongxin Wei. Online conformal selection with accept-to-reject changes. *arXiv preprint arXiv:2508.13838*, 2025.
- Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in neural information processing systems*, 33:21464–21475, 2020.
- Lin Lu, Yuyang Huo, Haojie Ren, Zhaojun Wang, and Changliang Zou. Feedback-enhanced online multiple testing with applications to conformal selection. *arXiv preprint arXiv:2509.03297*, 2025.
- Beier Luo, Shuoyuan Wang, Yixuan Li, and Hongxin Wei. Your pre-trained llm is secretly an unsupervised confidence calibrator, 2025. URL <https://arxiv.org/abs/2505.16690>.
- Yiwei Luo, Dallas Card, and Dan Jurafsky. Detecting stance in media on global warming, 2021. URL <https://arxiv.org/abs/2010.15149>.
- Ariane Marandon, Lihua Lei, David Mary, and Etienne Roquain. Adaptive novelty detection with false discovery rate guarantee. *The Annals of Statistics*, 52(1):157–183, 2024.
- Hussein Mozannar and David Sontag. Consistent estimators for learning to defer to an expert. In *International conference on machine learning*, pp. 7076–7087. PMLR, 2020.
- Curtis G. Northcutt, Anish Athalye, and Jonas Mueller. Pervasive label errors in test sets destabilize machine learning benchmarks, 2021. URL <https://arxiv.org/abs/2103.14749>.
- Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa : A large-scale multi-subject multi-choice dataset for medical domain question answering, 2022. URL <https://arxiv.org/abs/2203.14371>.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library, 2019. URL <https://arxiv.org/abs/1912.01703>.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL <https://arxiv.org/abs/2412.15115>.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL <https://arxiv.org/abs/2103.00020>.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet?, 2019. URL <https://arxiv.org/abs/1902.10811>.

John D Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 64(3):479–498, 2002.

Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. Large language models for data annotation and synthesis: A survey. *arXiv preprint arXiv:2402.13446*, 2024.

Weijie Tu, Weijian Deng, Liang Zheng, and Tom Gedeon. What does softmax probability tell us about classifiers ranking across diverse test conditions? *arXiv preprint arXiv:2406.09908*, 2024.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer, 2005.

Volodya Vovk, Alexander Gammerman, and Craig Saunders. Machine-learning applications of algorithmic randomness. 1999.

Damir Vrabac, Akshay Smit, Yujie He, Andrew Y Ng, Andrew L Beam, and Pranav Rajpurkar. Medselect: Selective labeling for medical image classification using meta-learning. In *International Conference on Medical Imaging with Deep Learning*, pp. 1301–1310. PMLR, 2022.

Haoran Wang, Qiuye Jin, Shiman Li, Siyu Liu, Manning Wang, and Zhijian Song. A comprehensive survey on deep active learning in medical image analysis. *Medical Image Analysis*, 95:103201, 2024a.

Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. Want to reduce labeling cost? gpt-3 can help. *arXiv preprint arXiv:2108.13487*, 2021.

Xudong Wang, Long Lian, and Stella X. Yu. Unsupervised selective labeling for more effective semi-supervised learning, 2023. URL <https://arxiv.org/abs/2110.03006>.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *arXiv preprint arXiv:2406.01574*, 2024b.

Xiaoyang Wu, Lin Lu, Zhaojun Wang, and Changliang Zou. Conditional testing based on localized conformal Sp\$-values. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=Ip6UwB35uT>.

Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks, 2017. URL <https://arxiv.org/abs/1611.05431>.

Ziyu Xu and Aaditya Ramdas. Online multiple testing with e-values. In *International Conference on Artificial Intelligence and Statistics*, pp. 3997–4005. PMLR, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengan Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL <https://arxiv.org/abs/2505.09388>.

Qi Yang, Shreya Ravikumar, Fynn Schmitt-Ulms, Satvik Lolla, Ege Demir, Iaroslav Elistratov, Alex Lavaee, Sadhana Lolla, Elaheh Ahmadi, Daniela Rus, et al. Uncertainty-aware language modeling for selective question answering. *arXiv preprint arXiv:2311.15451*, 2023.

Hiyori Yoshikawa and Naoaki Okazaki. Selective-lama: Selective prediction for confidence-aware evaluation of language models. In *Findings of the Association for Computational Linguistics: EACL 2023*, pp. 2017–2028, 2023.

He Zhang, Xinyi Fu, and John M Carroll. Augmenting image annotation: A human-lmm collaborative framework for efficient object selection and label generation. *arXiv preprint arXiv:2503.11096*, 2025.

Zinan Zhao and Wenguang Sun. A conformalized empirical bayes method for multiple testing with side information, 2025. URL <https://arxiv.org/abs/2502.19667>.

Tijana Zrnic and Emmanuel Candes. Active statistical inference. In *Forty-first International Conference on Machine Learning*, 2024. URL <https://openreview.net/forum?id=GKMCCtWC7H>.

A USE OF LARGE LANGUAGE MODEL

This paper uses large language models solely to polish specific sentences or paragraphs, without further use of LLMs for other purposes.

B TECHNICAL PROOFS

B.1 A LEMMA FOR PROVING THEOREM 3.1

Lemma B.1. *Under the assumptions of Theorem 3.1, where the calibration samples $(X_i, Y_i)_{i=1}^n$ and the test samples $(X_{n+j}, Y_{n+j})_{j=1}^m$ are independently and identically distributed (i.i.d.), and conditional on n_0 (the number of true null hypotheses in the calibration set) and m_0 (the number of true null hypotheses in the test set), the expected false discovery proportion (FDP) satisfies*

$$\mathbb{E}[FDP \mid n_0, m_0] \leq \frac{1+n}{1+n_0} \frac{m_0}{m} \alpha,$$

where $\alpha \in (0, 1)$ is the target false discovery rate (FDR) level, and the selection set \mathcal{R} is determined by Algorithm 1.

Proof. Suppose the first n_0 are true null samples in the calibration set and the first m_0 samples are true null samples in the test set. By the standard result in conformal inference (Vovk et al., 2005), we have

$$\mathbb{P}\{\hat{p}_j \leq \alpha \mid n_0, m_0\} \leq \alpha, j = 1, \dots, m_0$$

Let j^* denote the number of rejections from Algorithm 1, and recall the false discovery proportion (FDP) as

$$FDP = \frac{\sum_{j=1}^{m_0} \mathbf{1}\left\{\hat{p}_{n+j} \leq \frac{1+n}{1+n_0} \cdot \frac{\alpha j^*}{m}\right\}}{\max(j^*, 1)}.$$

Since $FDP = 0$ whenever $j^* = 0$, we have

$$\begin{aligned} \mathbb{E}[FDP \mid n_0, m_0] &= \mathbb{E}\left[\frac{1}{j^*} \sum_{j=1}^{m_0} \mathbf{1}\left\{\hat{p}_{n+j} \leq \frac{1+n}{1+n_0} \cdot \frac{\alpha j^*}{m}\right\} \middle| n_0, m_0, j^* > 0\right] \mathbb{P}\{j^* > 0\} \\ &\quad + \mathbb{E}\left[\sum_{j=1}^{m_0} \mathbf{1}\left\{\hat{p}_{n+j} \leq \frac{1+n}{1+n_0} \cdot \frac{\alpha j^*}{m}\right\} \middle| n_0, m_0, j^* = 0\right] \mathbb{P}\{j^* = 0\} \quad (3) \\ &\leq \mathbb{E}\left[\frac{1}{j^*} \sum_{j=1}^{m_0} \mathbf{1}\left\{\hat{p}_{n+j} \leq \frac{1+n}{1+n_0} \cdot \frac{\alpha j^*}{m}\right\} \middle| n_0, m_0, j^* > 0\right] \\ &= \mathbb{E}[FDP \mid n_0, m_0, j^* > 0]. \end{aligned}$$

By the super-uniform property of conformal p -values,

$$\mathbb{E}\left[\sum_{j=1}^{m_0} \mathbf{1}\left\{\hat{p}_{n+j} \leq \frac{1+n}{1+n_0} \cdot \frac{\alpha j^*}{m}\right\} \middle| n_0, m_0, j^*\right] \leq m_0 \cdot \frac{1+n}{1+n_0} \cdot \frac{\alpha j^*}{m}.$$

Thus,

$$\mathbb{E}[FDP \mid n_0, m_0, j^* > 0] \leq \frac{1}{j^*} \cdot m_0 \cdot \frac{1+n}{1+n_0} \cdot \frac{\alpha j^*}{m} = \frac{m_0}{m} \cdot \frac{1+n}{1+n_0} \alpha.$$

Combining with (3) yields

$$\mathbb{E}[FDP \mid n_0, m_0] \leq \mathbb{E}[FDP \mid n_0, m_0, j^* > 0] \leq \frac{m_0}{m} \cdot \frac{1+n}{1+n_0} \alpha,$$

which completes the proof. \square

B.2 PROOF OF THEOREM 3.1

Proof of theorem 3.1. Under the i.i.d. assumption of Theorem 3.1, the calibration samples $(X_i, Y_i)_{i=1}^n$ and the test samples $(X_{n+j}, Y_{n+j})_{j=1}^m$ are independently and identically distributed. This implies that each hypothesis, whether from the calibration or test set, has an equal probability p of being a true null, where p represents the expected probability of incorrect prediction under the null hypothesis.

Consequently, the number of true null hypotheses in the calibration set, denoted n_0 , follows a binomial distribution $n_0 \sim \text{Binomial}(n, p)$, and the number of true null hypotheses in the test set, denoted m_0 , follows $m_0 \sim \text{Binomial}(m, p)$. The independence across the calibration and test sets arises from the i.i.d. structure, ensuring that n_0 and m_0 are independent random variables.

Using the law of total expectation, we express FDR as

$$\begin{aligned}
\text{FDR} &= \mathbb{E}[\text{FDP}] \\
&= \mathbb{E}[\mathbb{E}[\text{FDP} \mid n_0, m_0]] \\
&\leq \mathbb{E}\left[\frac{m_0}{m} \frac{1+n}{1+n_0} \alpha\right] \quad \text{by Lemma B.1} \\
&= \mathbb{E}\left[\frac{m_0}{m}\right] \cdot \mathbb{E}\left[\frac{1+n}{1+n_0}\right] \cdot \alpha \quad \text{since } n_0 \text{ and } m_0 \text{ are independent} \\
&= p \cdot \mathbb{E}\left[\frac{1+n}{1+n_0}\right] \cdot \alpha \quad \text{since } m_0 \sim \text{Binomial}(m, p)
\end{aligned} \tag{4}$$

Now it suffices to show that $p \cdot \mathbb{E}\left[\frac{1+n}{1+n_0}\right] = [1 - (1-p)^{n+1}]$.

Since n_0 has probability mass function

$$\mathbb{P}(n_0 = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, \dots, n,$$

we compute

$$\begin{aligned}
p \cdot \mathbb{E}\left[\frac{1+n}{1+n_0}\right] &= p \cdot \sum_{k=0}^n \frac{1+n}{1+k} \binom{n}{k} p^k (1-p)^{n-k} \\
&= \sum_{k=0}^n \frac{(n+1)!}{(k+1)!(n-k)!} p^{k+1} (1-p)^{n-k} \\
&= \sum_{k=0}^n \binom{n+1}{k+1} p^{k+1} (1-p)^{n-k} \\
&= \sum_{l=1}^{n+1} \binom{n+1}{l} p^l (1-p)^{n+1-l} \quad \text{by letting } l = k + 1 \\
&= \sum_{l=0}^{n+1} \binom{n+1}{l} p^l (1-p)^{n+1-l} - \binom{n+1}{0} p^0 (1-p)^{n+1-0} \\
&= \sum_{l=0}^{n+1} \mathbb{P}(X = l) - (1-p)^{n+1} \quad \text{where } X \sim \text{Binomial}(n+1, p) \\
&= 1 - (1-p)^{n+1}
\end{aligned} \tag{5}$$

This completes the proof, establishing the desired bound on the FDR. □

C OVERVIEW OF DIFFERENT UNCERTAINTY SCORE FUNCTIONS

In this section, we provide an overview of three representative uncertainty score functions that are widely used in misclassification detection: Maximum Softmax Probability (MSP) (Hendrycks & Gimpel, 2016), the energy-based score (Liu et al., 2020), and the DOCTOR- α score (Granese et al., 2021). Each of these functions captures predictive uncertainty from a different perspective.

Maximum Softmax Probability (MSP). The Maximum Softmax Probability (MSP) baseline (Hendrycks & Gimpel, 2016) proposes to use confidence of the AI model as an uncertainty score.

$$S_{\text{MSP}}(x) = 1 - \max_{y \in \mathcal{Y}} p_y(x),$$

where $p_y(x)$ is the softmax probability assigned to class y for input x . The key idea is straightforward: if the model assigns a high probability to its most likely class, the prediction is considered confident. Although MSP is simple and effective, it has notable limitations: it only reflects the confidence in the top-1 prediction and ignores the structure of the remaining probability distribution, which may contain useful information about uncertainty.

Energy-based score. The energy-based score (Liu et al., 2020) is defined as

$$S_{\text{Energy}}(x) = \log \sum_{y \in \mathcal{Y}} \exp(f_y(x)),$$

where $f_y(x)$ denotes the logit value for class y and $T > 0$ is a temperature parameter. This score is derived from the concept of energy in statistical physics and leverages the log-sum-exp operator over all logits. Unlike MSP, which only considers the maximum probability, the energy score integrates information from the full logit vector, thereby providing a smoother and more informative confidence measure.

DOCTOR- α score. The DOCTOR- α score (Granese et al., 2021) is defined as

$$S_\alpha(x) = \sum_{y \in \mathcal{Y}} p_y(x)^2,$$

where $p_y(x)$ denotes the softmax probability for class y . This score is inspired by information-theoretic measures of uncertainty, as it is closely related to the quadratic Rényi entropy. The intuition is that if the predictive distribution is sharp (i.e., one class has probability close to one), then $\sum_y p_y(x)^2$ will be large, indicating high confidence. Conversely, if the distribution is flat (i.e., the model is uncertain and spreads probability mass across many classes), then $S_\alpha(x)$ will be small. Compared to MSP, the DOCTOR- α score leverages information from the entire probability distribution rather than only the top prediction, making it a richer measure of uncertainty for misclassification detection.

D BH PROCEDURE AND ITS ADAPTIVE VARIANTS

Consider testing m null hypotheses H_0^1, \dots, H_0^m based on their corresponding p -values $\{p_1, p_2, \dots, p_m\}$. For a true null hypothesis H_0^j , the corresponding p -value p_j is a random variable that is super-uniform on $[0, 1]$ under the null hypothesis. Formally, for any $u \in [0, 1]$,

$$\mathbb{P}(p_j \leq u \mid H_0^j \text{ is true}) \leq u.$$

Define $p_{(j)}$ as the j -th smallest p -value among a set of p -values $\{p_1, p_2, \dots, p_m\}$. Given a set of p -values $\{p_1, p_2, \dots, p_m\}$, the BH algorithm returns $S = \{j \in \{0, \dots, m\} : p_j \leq \frac{\alpha j^*}{m}\}$, where α is the target FDR level and

$$j^* = \max\{j \in \{1, \dots, m\} : p_{(j)} \leq \frac{\alpha j}{m}\}$$

When the null p -values $\{p_j : j \in \mathcal{H}_0\}$ are independent, the BH procedure is proved to control the FDR at level $\pi_0 \alpha$ in finite samples (Benjamini & Hochberg, 1995), where $\pi_0 = \frac{|\mathcal{H}_0|}{m}$ is the

proportion of true nulls. The independence assumption can be further relaxed to the PRDS condition (Benjamini & Yekutieli, 2001).

If π_0 is small, the FDR control will be overly conservative. When π_0 is known, we can apply BH procedure at level $\frac{\alpha}{\pi_0}$ to close the gap. In practice, π_0 is typically unknown. Several adaptive BH procedures attempt to address this issue by estimating π_0 and adjusting the target FDR level α accordingly. These procedures are often called the π_0 -adaptive versions of the BH algorithm. Two most famous estimators are Storey-BH (Storey, 2002) and Quantile-BH (Benjamini et al., 2006):

$$\hat{\pi}_0^{Storey}(\lambda) = \frac{1 + \sum_{i=1}^m \mathbf{1}\{p_i \geq \lambda\}}{m(1 - \lambda)}, \lambda \in (0, 1)$$

$$\hat{\pi}_0^{Quant}(k_0) = \frac{m - k_0 + 1}{m(1 - p_{(k_0)})}, k_0 \in \{1, \dots, m\}$$

λ and k_0 are hyperparameters determined by users.

E HYPERPARAMETER SELECTION FOR BH ADAPTIVE VARIANTS

Both Storey-BH and Quant-BH require careful hyperparameter selection— λ for Storey-BH and k_0 for Quant-BH—as this choice significantly impacts their performance. Following Storey (2002), we employ a bootstrap-based method to select the optimal λ (and analogously, k_0 for the Quantile BH procedure). Further details can be found in Section 9 of Storey (2002). The algorithm proceeds as follows:

1. Define a grid R for the hyperparameter, i.e. $R = \{0.1, 0.2, \dots, 0.9\}$ for Storey-BH.
2. For each $\lambda \in R$, compute:

$$\widehat{\text{pFDR}}_\lambda(\gamma) = \frac{\widehat{\pi}_0(\lambda)\gamma}{\widehat{\Pr}(p \leq \gamma)\{1 - (1 - \gamma)^m\}} \quad (6)$$

where $\widehat{\pi}_0(\lambda) = \widehat{\pi}_0^{Storey}(\lambda)$ or $\widehat{\pi}_0^{Quant}(\lambda)$ and $\widehat{\Pr}(p \leq \gamma)$ is the empirical estimate of $\Pr(p \leq \gamma)$

3. Generate B bootstrap replicates $\{p_1^{*,b}, \dots, p_m^{*,b}\}_{b=1}^B$ and compute $\widehat{\text{pFDR}}_\lambda^{*,b}(\gamma)$ for each b .
4. Estimate the MSE for each λ :

$$\widehat{\text{MSE}}(\lambda) = \frac{1}{B} \sum_{b=1}^B \left(\widehat{\text{pFDR}}_\lambda^{*,b}(\gamma) - \min_{\lambda' \in R} \widehat{\text{pFDR}}_{\lambda'}(\gamma) \right)^2 \quad (7)$$

5. Select $\hat{\lambda} = \arg \min_{\lambda \in R} \widehat{\text{MSE}}(\lambda)$

F IMPLEMENTATION DETAILS

Experiment details. We run our experiments on NVIDIA GeForce RTX 4090 and NVIDIA L40 GPU, and implement all methods by PyTorch and vLLM.

Dataset details. For the LLM QA datasets (MedMCQA, MMLU, and MMLU-Pro), we adopt the same prompts as in Luo et al. (2025). For the other datasets, we design our own prompts following the style of LLM QA prompts, as summarized in Table 2. In text labeling and LLM QA tasks, we merge the calibration and test sets to form the unlabeled dataset, except for MedMCQA. For MedMCQA, because test labels are unavailable, we use the calibration set as the unlabeled dataset. For image labeling, we use the validation sets of ImageNet and ImageNet-V2 as the unlabeled datasets.

Table 2: Prompts for different datasets

Dataset	Prompts
LLM QA	The following are multiple-choice questions. Give ONLY the correct option, no other words or explanation: [Question] A: [Option 1] B: [Option 2] C: [Option 3] D: [Option 4] Answer: [Mask]
Stance on global warming	You are given a statement about climate change. Determine the stance that a human would take towards this statement. Respond with ONLY the letter (A, B, or C) of the correct stance. Do not include any explanation. [Input Headline] A: agrees B: neutral C: disagrees Answer: [Mask]
Misinformation	You are a fact-checking assistant. Classify the following news headline as either real (A) or misinfo (B). Respond with ONLY the letter A or B. Do not include any explanation. Headline: [Input text] A: real B: misinfo Answer: [Mask]

Algorithm 2 Conformal Labeling Regression Tasks

Require: Calibration set $\mathcal{D}_{\text{cal}} = \{(X_i, Y_i)\}_{i=1}^n$; test instances $\{X_{n+j}\}_{j=1}^m$; pre-trained predictor f ; loss function L ; loss threshold ϵ ; target FDR level $\alpha \in (0, 1)$; nonconformity scores $\{S_i\}_{i=1}^{n+m}$

- 1: Identify calibration samples that exceed the loss threshold: $\mathcal{D}_{\text{cal}}^0 = \{(X_i, Y_i) : L(Y_i, \hat{Y}_i) > \epsilon\}_{i=1}^n$, and set $n_0 = |\mathcal{D}_{\text{cal}}^0|$.
- 2: **for** $j = 1, \dots, m$ **do**
- 3: Compute the conformal p -value \hat{p}_j according to equation 2.
- 4: **end for**
- 5: Apply the step-up procedure: $j^* = \max \left\{ j : \hat{p}_{(j)} \leq \frac{\alpha j(n+1)}{m(n_0+1)} \right\}$, where $\hat{p}_{(j)}$ is the j -th smallest p -value.
- 6: **Output:** The selected set $\mathcal{S} = \{j : \hat{p}_j \leq \hat{p}_{(j^*)}\}$.

G EXTENSION TO REGRESSION TASKS

While our primary focus is on classification, Conformal Labeling can be naturally extended to regression settings. Consider a loss function $L(Y, \hat{Y})$ that quantifies prediction error—for instance, the squared error $L(Y, \hat{Y}) = (Y - \hat{Y})^2$ —alongside a user-specified tolerance level ϵ . For each test sample, we define the null and alternative hypotheses as:

$$H_0^j : L(Y_{n+j}, \hat{Y}_{n+j}) > \epsilon \quad \text{versus} \quad H_1^j : L(Y_{n+j}, \hat{Y}_{n+j}) \leq \epsilon.$$

This framework generalizes the classification setting, which corresponds to the special case where $L(Y, \hat{Y}) = \mathbf{1}\{Y \neq \hat{Y}\}$ and $\epsilon = 0$. To apply Conformal Labeling in regression, we need an uncertainty score that reflects the model’s predictive uncertainty. While uncertainty score functions are straightforward in classification tasks (e.g., $1 - \max_{y \in \mathcal{Y}} f_y(X)$), regression requires alternative approaches to get an uncertainty score function. For example, when using LLMs, we can leverage their verbalized confidence or prompt them to output prediction intervals, using the interval width as the uncertainty score. The complete procedure for Conformal Labeling in the regression task is outlined in Algorithm 2.

We evaluate Conformal Labeling on two regression problems: sentiment analysis with GPT-4o and protein structure prediction with AlphaFold. We use the data provided by Candès et al. (2025). For sentiment analysis, we prompt GPT-4o to output a prediction interval $[a_i, b_i]$ for each target Y_i , set the predicted value as $\hat{Y}_i = (a_i + b_i)/2$, and use the interval length $U_i = b_i - a_i$ as the uncertainty score. For protein structure prediction, we take experimentally derived structures as ground truth and AlphaFold predictions as \hat{Y}_i . Uncertainty scores are obtained from AlphaFold’s internal confidence measure, the average predicted local distance difference test (pLDDT). For both experiments, we employ the L2 loss function.

In Table 3, we report the performance of Conformal Labeling on regression tasks with $\alpha = 0.1$. In all cases, Conformal Labeling consistently controls the realized FDR below the target level. A

Table 3: Performance of Conformal Labeling on regression tasks at $\alpha = 0.1$. Results are shown for sentiment analysis (top) and protein folding (bottom) under different values of the tolerance parameter ϵ . In all cases, Conformal Labeling controls the realized FDR below the target level.

Dataset	Metric	Method		
		CL ($\epsilon = 0.05$)	CL ($\epsilon = 0.06$)	CL ($\epsilon = 0.07$)
Sentiment analysis	FDR (%)	8.48%	8.60%	7.89%
	Power (%)	5.04%	53.14%	94.92%
	AI-labeled ratio	4.55%	47.84%	85.43%
		CL ($\epsilon = 1$)	CL ($\epsilon = 4$)	CL ($\epsilon = 9$)
Protein folding	FDR (%)	9.67%	9.90%	9.04%
	Power (%)	27.24%	49.73%	97.90%
	AI-labeled ratio	10.06%	36.80%	88.00%

key observation is that Conformal Labeling’s selection result in regression task is highly sensitive to the choice of the tolerance parameter ϵ . For instance, in sentiment analysis, setting $\epsilon = 0.05$ yields an AI-labeled ratio of only 4.55%, whereas a slightly larger value $\epsilon = 0.06$ increases the ratio to 47.84%. In practice, selecting an appropriate ϵ may require domain expertise or prior knowledge about the task.