Calibrating LLMs for Selective Prediction: Balancing Coverage and Risk

Anonymous Author(s)
Affiliation
Address
email

Abstract

Despite the impressive capabilities of large language models (LLMs), their outputs often exhibit inconsistent correctness and unreliable factual accuracy. In high-stakes domains, overconfident yet incorrect predictions can lead to serious consequences, highlighting the need for robust uncertainty estimation. To address this, we introduce SelectLLM, an end-to-end method designed to enhance the ability of LLMs to recognize and express uncertainty effectively. By integrating selective prediction into finetuning, SelectLLM optimizes model performance over the covered domain, achieving a more balanced trade-off between predictive coverage and utility. Experimental results on TriviaQA, CommonsenseQA and MedConceptsQA show that SelectLLM significantly outperforms standard baselines, improving abstention behaviour while maintaining high accuracy.

1 Introduction

Large language models (LLMs) have rapidly become foundational components in natural language processing (NLP), driving progress across a wide range of tasks – from open-ended generation to complex reasoning. Despite their huge progress and impressive capabilities, LLMs still frequently produce outputs with varying levels of correctness and factual accuracy. A core challenge in deploying these models in real-world settings lies in balancing accuracy with calibrated confidence. While high accuracy remains a primary goal, it is equally critical for models to recognize and signal their own uncertainty, particularly in high-stakes scenarios such as healthcare [1, 2], finance [3, 4], and law [5, 6]. Overconfident incorrect responses can be significantly more harmful than abstentions or cautious, low-confidence responses. To address this, we leverage confidence modeling to enable selective prediction, allowing the system to abstain from answering when uncertainty is high [7], thereby trading off *coverage* for *reliability*. This trade-off is especially important in safety-critical applications or decision-support systems, where deferring uncertain cases to a human or fallback system is preferable to propagating potentially erroneous outputs. In this paper, we introduce a principled approach to enhancing safety of an LLM that allows a model to abstain from making a prediction when it is uncertain, thereby reducing the risk of harmful or misleading outputs. However, abstention introduces a secondary trade-off: while conservative behavior can reduce risk, excessive abstention diminishes the utility of the model by forgoing opportunities where correct responses are feasible. A model that abstains too frequently may be safe but ultimately useless. For example, in the "needle in-the-haystack" benchmark, LLMs become more uncertain when given the “nonexistent” option, even when capable of providing correct answers [8]. This highlights the challenge of balancing risk with utility (coverage): optimizing both the correctness of answers and the number of answered questions.

We formalize this challenge as a risk-coverage trade-off and categorize model outputs into four distinct cases following the previous literature [9, 10], as illustrated in Table 1: ❶ *Accepting a correct answer*

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

— the ideal case, contributing to both utility and reliability; ❷ *Rejecting an incorrect answer* — also desirable, as it avoids unreliable answers; ❸ *Rejecting a correct answer* — suboptimal, reducing the utility of the model; ❹ *Accepting an incorrect answer* — the most harmful case, compromising the accuracy of the model. Our objective is to maximize the occurrence of the first two cases while minimizing the occurrence of the latter two.

To illustrate the risk-coverage trade-off challenge, consider two medical AI assistants designed to help doctors interpret diagnostic test results. Assistant A, optimized solely for utility, studied all diagnostic topics uniformly but lacks the ability to accurately judge when to abstain. Consequently, it sometimes provides incorrect answers with high confidence or unnecessarily abstains even when it could have answered correctly. In contrast, Assistant B explicitly accounts for the risk-coverage trade-off by carefully distinguishing between cases it can confidently address and those it should avoid. When faced with ambiguous diagnostic cases, Assistant B appropriately abstains, whereas in clear-cut cases that Assistant A might wrongly skip, Assistant B reliably provides accurate answers. Consequently, Assistant B achieves the best average diagnostic performance, as illustrated in Figure 1.

To address this challenge, we propose a novel method, called SelectLLM, that explicitly produces confidence estimates and incorporates the task of confidence estimation into its training objectives. SelectLLM assigns confidence scores to questions rather than to generated answers, thereby quantifying the reliability of the LLM's response to specific queries independent from the multiple alternative answers generated. Questions can be classified into two categories based on a confidence threshold: those with confidence above a given threshold (covered by the model) and those below the threshold (not covered). Within the covered set of questions, we further distinguish between the questions the model is confident in answering correctly and those it confidently identifies as beyond its capability, corresponding to the first and second cases mentioned previously.

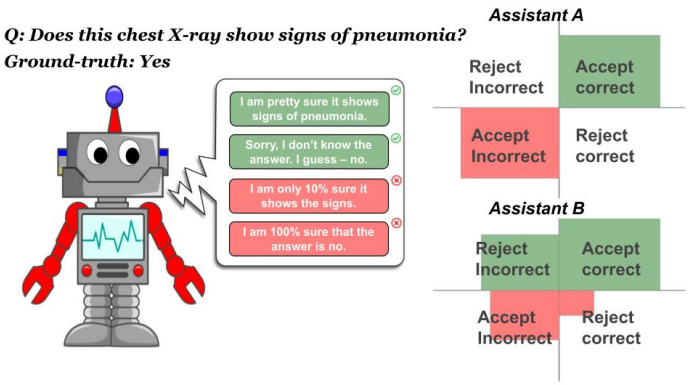

Figure 1: Illustration of risk-coverage trade-off. Given a question, Assistant A (base LLM), optimized solely for utility, often produces incorrect answers due to overconfidence. In contrast, Assistant B (with SelectLLM), which explicitly accounts for the risk–coverage trade-off, recognizes its limitations and abstains when uncertain. As a result, it avoids more errors and achieves better performance on diagnostic tasks.

SelectLLM is based on a well-trained LLM and jointly trains (fine-tunes the first and trains the second) two heads (shown in Figure 2): ❶ a *decoding head*, corresponding to the original LLM output layer for autoregressive token generation; ❷ a *selection head*, outputting a confidence score for the question. This two-head design is motivated by the known calibration deficiencies of trained LLMs. In a well-calibrated model, the decoding head's next-token probabilities could be used directly for confidence estimation. However, LLMs often exhibit overconfidence or underconfidence, making it necessary to learn a separate abstention signal. The selection head is explicitly optimized to improve the risk–coverage trade-off, allowing the model to balance utility with reliability. Our contributions are summarized as follows:

- We introduce SelectLLM, which incorporates risk–coverage trade-off control into the LLM training stage. It combines **Direct Preference Optimization (DPO)** [11] with confidence estimation to improve the risk-coverage trade-off;
- We **construct three high-quality benchmarks** for DPO fine-tuning based on open-sourced Question-and-Answer datasets, and conduct extensive experiments on **seven baselines with three different LLMs**, demonstrating that SelectLLM significantly outperforms state-of-the-art baselines in terms of risk and coverage metrics;
- We validate the confidence scores produced by SelectLLM by comparing their distribution to scores derived from the tone and phrasing of the generated responses, demonstrating that SelectLLM can natively output reliable confidence estimates for its predictions **without relying on any external models**.

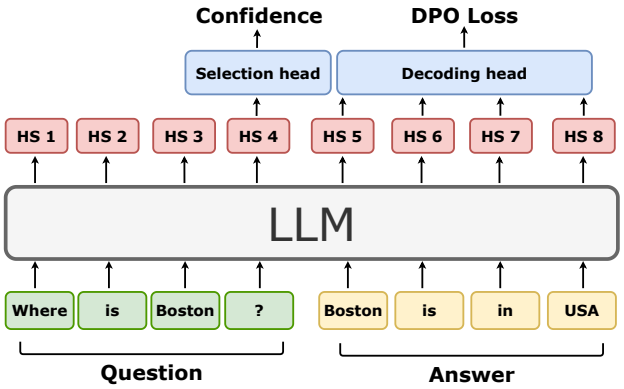

Figure 2: ***Overview of SelectLLM.*** *Given a question–answer input pair, the underlying LLM processes the full sequence and produces a hidden state (HS) for each token. The selection head operates on the hidden state corresponding to the last token of the question to estimate a confidence score for abstention; while the decoding head uses the answer-related hidden states to compute the DPO loss for LLM fine-tuning. This dual-head design enables SelectLLM to jointly optimize for utility and accuracy.*

Table 1: Four cases of the answer to a question: "In which branch of the arts does Allegra Kent work?".

	Accept (high confidence)	**Reject** (low confidence)
Correct	Allegra Kent is a ballet dancer. She worked as a principal dancer with the New York City Ballet.	I'm not entirely certain, but I think Allegra Kent might be involved in ballet.
Incorrect	Allegra Kent is a renowned opera singer who performed in major productions throughout Europe.	I'm not really sure, but maybe Allegra Kent is a painter?

2 Related work

Uncertainty Quantification in LLMs. Uncertainty estimation for large language models (LLMs) spans several complementary paradigms. and generally falls into two categories: (i) black-box approaches and (ii) white-box approaches. Black-box methods include verbalized uncertainty, where models are prompted to express confidence in natural language [12, 13, 14, 15], and sampling-based methods, which estimate predictive uncertainty from variability across multiple generations [16, 17, 18]. White-box approaches, in contrast, exploit model internals such as token-level probabilities, calibration of log-likelihoods, or hidden-state diagnostics to produce confidence scores. Related work includes TokenSAR [19], P(True) [20] and Semantic Entropy [21]. While many of these techniques primarily serve to identify uncertain predictions and guide abstention, there is also a growing line of work on uncertainty-aware training, where uncertainty estimates inform parameter updates [22, 23]. Our approach builds on these advances by directly incorporating selective prediction objectives into fine-tuning.

Alignment and Confidence in LLMs. Efforts to align LLMs with human preference, such as Proximal Policy Optimization (PPO) [24] and Direct Preference Optimization (DPO) [11], adjust model parameters to encourage desired behaviours. [25] proposed conservative reward modeling to encourage LLMs to be more cautious in their predictions, which relates to our objective of selective prediction. [26] introduced self-restraint fine-tuning, aiming to increase model confidence when appropriate while reducing overconfidence. Recent works such as [9] and [10] utilize DPO to align LLMs with human preference to guide the model to answer questions it knows and to avoid answering questions it does not know.

Selective Prediction in LLMs. Selective prediction has a rich history in machine learning [27, 28, 29, 30], and has recently been extended to LLMs [31, 32, 33]. However, none of these LLM-related works incorporates selective coverage into model training. SelectiveNet [34] provides a foundational framework for selective classification in deep networks. Our work extends this idea to the generative setting of LLMs, which poses unique challenges. SelectLLM differs from prior frameworks such as SelectiveNet in several critical ways. While SelectiveNet targets classification and regression, SelectLLM is designed for sequence generation. To enable this, we introduce a

new module that embeds the generated sequence before passing it to a confidence head, enabling reliable abstention decisions for natural language outputs. Moreover, SelectiveNet employs three heads—reward, selection, and auxiliary—to encourage shared representation learning. In contrast, SelectLLM adds only a single selection head $g(\cdot)$ to the original LLM and fine-tunes the entire framework to align with human preferences. This design enables SelectLLM to balance generation quality, prediction accuracy, and selective abstention, offering a principled framework for calibrated and trustworthy language generation.

In summary, by synthesizing advances from uncertainty quantification, fine-tuning, and selective prediction, SelectLLM introduces a principled framework that jointly optimizes predictive performance and uncertainty estimation, a contribution of particular significance for high-stakes applications.

3 Problem Formulation

We define *coverage* as the proportion of questions for which the model is confident enough to provide an answer:

$$\text{coverage} = \frac{1}{n}\sum_{i=1}^{n}(1 - a_i),$$

where n is the total number of questions, $a_i = 1$ if the model abstains on the ith question and $a_i = 0$ otherwise. While *risk* is defined as the error rate over the set of answered questions:

$$\text{risk} = \frac{\sum_{i=1}^{n} \mathbb{1}(\hat{y}_i \notin \mathcal{Y}_i \wedge a_i = 0)}{\sum_{i=1}^{n}(1 - a_i)},$$

where $\hat{y}_i$ is the model's output, $\mathcal{Y}_i$ is the set of correct answers for the ith question.

The goal is to ensure that LLMs can reliably estimate their predictive confidence and abstain when uncertainty is high, while also minimizing unnecessary abstentions to retain practical utility. Our approach is built on Direct Preference Optimization (DPO) [11], a human preference alignment method that fine-tunes language models using pairwise comparisons of answers without the need to explicitly model a reward function.

DPO [11] is a human preference alignment method that fine-tunes language models using comparisons of pairs of answers without the need to explicitly model a reward function. Specifically, in the **training** stage, we are given (1) a dataset $\mathcal{X} = \{x_1, x_2, \ldots, x_n\}$, where each x_i is a question posed to the model; (2) a corresponding set of human preference annotations $\mathcal{P} = \{(y_{i,+}, y_{i,-})\}$, where $y_{i,+}$ and $y_{i,-}$ denote the preferred and rejected answers to question x_i, respectively; and (3) a predefined coverage rate $0 < c < 1$, which represents the target proportion of questions for which the user expects the model to provide confident answers. Our goal is to maximize the likelihood of human-preferred answers relative to rejected ones given the coverage constraint c, yielding a fine-tuned model M_{select} and a selection head $g(\cdot)$ which outputs a confidence score c_i indicating the model's confidence in answering a specific question x_i.

In the **inference** stage, given (1) a dataset of input questions, $\mathcal{X} = \{x_1, x_2, \ldots, x_n\}$, where each x_i is a question; and (2) a trained model M_{select} and its selection head $g(\cdot)$, the model produces (1) a set of LLM-generated answers, $\hat{\mathcal{Y}} = \{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n\}$, where each $\hat{y}_i$ is the model's answer to x_i; and (2) a set of confidence scores, $\mathcal{C} = \{\text{conf}_1, \text{conf}_2, \ldots, \text{conf}_n\}$, where each conf_i represents the model's confidence that it can answer question x_i correctly.

Given the model's answer to a question, together with its confidence score to answer the question, the model abstains when its confidence score conf_i is below a given threshold τ. More formally, the abstention decision for question x_i is defined as $a_i = \begin{cases} 1 & \text{if } \text{conf}_i < \tau \\ 0 & \text{otherwise} \end{cases}$.

4 SelectLLM

Our proposed method SelectLLM enhances pre-trained LLMs by introducing an additional head that explicitly estimates the model's confidence in answering a given question correctly. This selection head is trained or fine-tuned jointly with the base model. Specifically, given a pre-trained LLM π_θ, we augment it with a selection head $g(\cdot)$, which outputs a confidence score $\text{conf} \in (0, 1)$.

Unlike traditional confidence estimation methods that rely on token-level probabilities, our selection head operates on the last-layer hidden state of the final token in the input question. This design ensures that confidence estimation is based solely on the model and the input question.

4.1 Loss Function

The loss function of SelectLLM combines the **DPO loss**, which aligns the model's outputs with human preferences, and the **Select loss**, which manages the risk–coverage trade-off.

The **DPO loss** aligns the model's outputs with human preferences without requiring explicit reward modeling or reinforcement learning. Given a dataset of human preferences $\mathcal{P} = \{(x_i, y_{i,+}, y_{i,-})\}$, where $y_{i,+}$ is the preferred response and $y_{i,-}$ is the rejected response to question x_i, the DPO loss is defined as:

$$L_{\text{DPO}}(\pi_\theta, \pi_{\text{ref}}) = -\mathbb{E}_{(x,y_+,y_-)\sim\mathcal{P}}\left[\log\sigma\left(\beta\log\frac{\pi_\theta(y_+ \mid x)}{\pi_{\text{ref}}(y_+ \mid x)} - \beta\log\frac{\pi_\theta(y_- \mid x)}{\pi_{\text{ref}}(y_- \mid x)}\right)\right] \tag{1}$$

where:

- π_θ is the LLM we want to fine-tune.
- π_{ref} is a reference model, usually a frozen version of the original pre-trained language model.
- σ is the sigmoid function.
- β is a hyperparameter that controls the amount of divergence from the reference model π_{ref}.

Building on Section 3, we define the empirical selective risk for LLM fine-tuning as:

$$\hat{r} = \frac{1}{n}\sum_{i=1}^{n}(g(h_i)\cdot L_{\text{DPO}}) \tag{2}$$

where h_i denotes the hidden state of the last token in the question, $g(h_i) \in [0, 1]$ is the selection function that quantifies the model's confidence for the given question.

Notably, since the original DPO loss only boosts the margin between the chosen answer and the rejected answer, it may simultaneously decrease the probabilities of both chosen and rejected answers, compared to the reference model, which is not desirable. Therefore, we define a reward function measuring the difference in the probabilities between the answers of the fine-tuned model and the reference model, which is defined as follows:

$$w(y) = \beta\left(\log\pi_\theta(y) - \log\pi_{\text{ref}}(y)\right)$$

where β is a hyper-parameter, and π_θ, π_{ref} follow the same definitions as in the DPO loss.

Then we define the risk for generating chosen and rejected answers using this reward function:

$$\ell(\pi_\theta, \pi_{\text{ref}}, y) = \begin{cases}\log\sigma\left(\max(0, -w(y))\right) & \text{if } y \in y_+\\ \log\sigma\left(\max(0, w(y))\right) & \text{if } y \in y_-\end{cases}$$

The intuition behind this risk is as follows: a penalty is applied if the fine-tuned model assigns a lower probability to chosen answers than the reference model, or a higher probability to rejected answers. Therefore, SelectLLM incorporates this risk into its Select loss.

Building on the above, we define a modified empirical selective risk as follows:

$$\hat{r}_\ell(\pi_\theta, \pi_{\text{ref}}, g) = \frac{1}{n}\sum_{i=1}^{n}((1 - w_+ - w_-)\cdot L_{\text{DPO}} + w_+\cdot\ell(\pi_\theta, y_{i,+}) + w_-\cdot\ell(\pi_\theta, y_{i,-}))\cdot g(h_i) \tag{3}$$

where w_+ and w_- are hyper-parameters defined by the users. In the appendix, we include an ablation study to demonstrate the effectiveness of the two additional terms ($\ell(\pi_\theta, y_{i,+})$ and $\ell(\pi_\theta, y_{i,-})$).

The **Select loss** aims to minimize the selective risk while maintaining a predefined coverage level c. Formally, the Select objective is given by:

$$L_{\text{Select}} = \hat{r} + \lambda\cdot\Psi(c - \hat{\phi}(g)) \tag{4}$$

where $\hat{\phi}(g) = \frac{1}{n}\sum_{i=1}^{n} g(h_i)$ is the empirical coverage, $\lambda > 0$ is a regularization parameter, and $\Psi(a) = \max(0, a)^2$ penalizes deviations from the target coverage rate c defined by the user.

Finally, the **Combined loss** is defined as a weighted sum of the Select loss and the fine-tuning loss:

$$L_{\text{Combined}} = \alpha \cdot L_{\text{Select}} + (1 - \alpha) \cdot L_{\text{DPO}} \tag{5}$$

where $\alpha \in [0, 1]$ balances the weight of the two objectives. Following [35], we set $\alpha = 0.5$ without hyperparameter tuning in all experiments.

If we do not incorporate the Select loss, the model may produce outputs aligned with human preferences but lack effective confidence calibration, which could result in excessive abstention or incorrect responses overly confident. The use of the original DPO loss, L_{DPO}, is also essential to optimizing SelectLLM. Since the selection head is initialized randomly, without L_{DPO}, SelectLLM will focus on a fraction c of the training set, before accurate low level features are constructed. In such a case, SelectLLM will tend to overfit to the wrong subset of the training set. The L_{DPO} exposes the SelectLLM model to all training instances throughout the training process. Thus, integrating both losses ensures that the model achieves a balanced performance—producing high-quality, preference-aligned outputs while maintaining optimal coverage through calibrated confidence estimation.

5 Experiments

In this section, we first compare SelectLLM against seven baseline models on the TriviaQA [36] and CommonsenseQA [37] benchmarks, two widely used datasets for evaluating open-domain question-answering systems. We then demonstrate SelectLLM's ability to generalize across domains by fine-tuning on CommonsenseQA and testing on TriviaQA. Next, we validate the confidence scores produced by SelectLLM, followed by an ablation study to assess the impact of the reward loss terms and the coverage–risk trade-off.

5.1 Experimental Setup

We use Llama-3.1-8B-Instruct, Mistral-7B-Instruct-v0.2 and Qwen2.5-14B-Instruct in the experiments as the base models. We use QLoRA [38] with rank 16 to train all the models. For comparison, we use *base* (LLM without finetuning), *LACIE* [9] (DPO-based finetuning), *LARS* [39] (uses a well-trained score function), *MARS* [40] (uses a QA evaluator model), *TokenSAR* [19] (uses a sentence similarity model), *P(True)* [20] (a self-check method) and *Semantic Entropy (SE)* [21] (uses token probabilities) as our baselines. For all models, we report average performance across 5 seeds. We perform all the LLM fine-tuning on one A100-40GB GPU.

Metrics. Across all the experiments, we report the following evaluation metrics: the number of true positives **(TP)**, the number of true negatives **(TN)**, **Precision**, **Recall**, and **Coverage**. We also include the **TRUTH** metric introduced in [10], defined as the sum of TP and TN, which captures the number of correctly accepted and correctly abstained responses. Because the test dataset contains 1,000 samples, the upper bound of TRUTH is 1,000. As there are no ground-truth or reference confidence scores provided for each question, we cannot report AUROC or ECE scores.

For score-based methods (SelectLLM, LARS, MARS, TokenSAR, P(True), and SE), we tune a threshold on the validation set to maximize the TRUTH metric and then apply the same threshold to the test set for abstention. For non-score-based methods (base and LACIE), we use a rule-based evaluation strategy: a response is accepted as long as the model provides an answer and is rejected only if the model explicitly refuses or states that it does not know.

Datasets. We use the TriviaQA [36], CommonsenseQA [37], and MedConceptsQA [41] datasets. Following [9], for TriviaQA we randomly sample 10,000, 1,000, and 1,000 questions for the training, validation, and test sets, respectively. For CommonsenseQA, we randomly sample 8,000, 1,000, and 1,000 questions for the training, validation, and test sets, respectively. For MedConceptsQA, which is used solely for evaluation, we randomly sample 1,000 questions each for the validation and test sets.

To construct the chosen/rejected pairs used for LACIE and SelectLLM fine-tuning, we first augment each dataset with model-generated answers and their associated confidence scores. Specifically, we use the base models mentioned above to generate an answer for each question and then employ

DeepSeek-v3 to assign a confidence score based on the tone and phrasing of the generated response. We refer to this score as *tone-confidence*. The prompt provided to DeepSeek-v3 is: *"Rate how confident the response appears based solely on its tone and phrasing."*

We set a confidence threshold of 0.7: answers with scores above this threshold are accepted, while those below are rejected. If no correct answer exceeds the threshold, we default to a generic response—*"I don't know the answer."*—as the chosen answer. Such fallback responses occur in roughly 30% of the fine-tuning dataset. All remaining answers to the same question are treated as rejected. Finally, we construct the fine-tuning pairs for both LACIE and SelectLLM by sampling one chosen and one rejected answer for each question.

5.2 In-distribution Performance

Table 2: TriviaQA performance. ↑ indicates the higher the better, and ↓ indicates the lower the better. ***The TN value for both the base and LACIE is 0.0 (with a corresponding Recall of 1.0), since they do not abstain from any answers.***

Model	TP ↑	TN ↑	TRUTH ↑	Precision ↑	Recall ↑	Coverage (%)
Llama-3.1-8B-Instruct						
base	**601.7**$_{\pm 2.3}$	0.0$_{\pm 0.0}$	601.7$_{\pm 2.3}$	0.602$_{\pm 0.002}$	**1.000**$_{\pm 0.000}$	100.0$_{\pm 0.0}$
LARS	579.3$_{\pm 3.7}$	45.2$_{\pm 4.0}$	624.2$_{\pm 6.5}$	0.627$_{\pm 0.018}$	0.949$_{\pm 0.005}$	92.4$_{\pm 8.8}$
MARS	556.2$_{\pm 8.9}$	57.1$_{\pm 2.4}$	613.4$_{\pm 7.6}$	0.626$_{\pm 0.017}$	0.912$_{\pm 0.015}$	88.9$_{\pm 9.9}$
TokenSAR	559.2$_{\pm 9.3}$	62.3$_{\pm 6.2}$	621.1$_{\pm 7.9}$	0.630$_{\pm 0.006}$	0.916$_{\pm 0.022}$	88.7$_{\pm 14.6}$
P(True)	565.6$_{\pm 2.1}$	54.8$_{\pm 4.1}$	621.9$_{\pm 5.4}$	0.622$_{\pm 0.014}$	0.965$_{\pm 0.015}$	94.7$_{\pm 3.7}$
SE	589.5$_{\pm 7.4}$	32.1$_{\pm 5.8}$	619.3$_{\pm 7.5}$	0.627$_{\pm 0.010}$	0.926$_{\pm 0.011}$	90.1$_{\pm 12.8}$
LACIE (DPO)	579.3$_{\pm 23.6}$	0.0$_{\pm 0.0}$	579.3$_{\pm 23.6}$	0.579$_{\pm 0.024}$	**1.000**$_{\pm 0.000}$	100.0$_{\pm 0.0}$
SelectLLM	582.0$_{\pm 19.7}$	**170.0**$_{\pm 25.2}$	**752.0**$_{\pm 2.6}$	**0.773**$_{\pm 0.015}$	0.884$_{\pm 0.021}$	75.96$_{\pm 3.63}$
Mistral-7B-Instruct-v0.2						
base	**598.3**$_{\pm 4.0}$	0.0$_{\pm 0.0}$	598.3$_{\pm 9.0}$	0.598$_{\pm 0.009}$	**1.000**$_{\pm 0.000}$	100.0$_{\pm 0.0}$
LARS	587.4$_{\pm 7.5}$	48.2$_{\pm 3.4}$	635.3$_{\pm 8.2}$	0.626$_{\pm 0.010}$	0.977$_{\pm 0.008}$	93.8$_{\pm 12.9}$
MARS	558.5$_{\pm 8.1}$	40.2$_{\pm 4.2}$	598.1$_{\pm 2.9}$	0.608$_{\pm 0.013}$	0.928$_{\pm 0.010}$	91.7$_{\pm 4.7}$
TokenSAR	529.4$_{\pm 8.7}$	61.2$_{\pm 2.5}$	590.9$_{\pm 4.8}$	0.610$_{\pm 0.012}$	0.880$_{\pm 0.016}$	86.7$_{\pm 11.9}$
P(True)	532.8$_{\pm 10.9}$	81.2$_{\pm 5.1}$	613.1$_{\pm 6.4}$	0.626$_{\pm 0.009}$	0.885$_{\pm 0.015}$	85.0$_{\pm 8.3}$
SE	582.3$_{\pm 8.3}$	33.7$_{\pm 6.0}$	615.3$_{\pm 3.6}$	0.614$_{\pm 0.020}$	0.968$_{\pm 0.009}$	94.8$_{\pm 18.0}$
LACIE (DPO)	568.4$_{\pm 3.4}$	0.0$_{\pm 0.0}$	568.4$_{\pm 7.4}$	0.568$_{\pm 0.007}$	**1.000**$_{\pm 0.000}$	100.0$_{\pm 0.0}$
SelectLLM	522.0$_{\pm 19.9}$	**230.3**$_{\pm 24.7}$	**752.3**$_{\pm 12.3}$	**0.741**$_{\pm 0.019}$	0.891$_{\pm 0.039}$	70.87$_{\pm 4.21}$
Qwen2.5-14B-Instruct						
base	**636.2**$_{\pm 10.7}$	0.0$_{\pm 0.0}$	636.2$_{\pm 10.7}$	0.636$_{\pm 0.011}$	**1.000**$_{\pm 0.000}$	100.0$_{\pm 0.0}$
LARS	624.0$_{\pm 6.4}$	17.1$_{\pm 4.2}$	641.2$_{\pm 2.0}$	0.643$_{\pm 0.016}$	0.981$_{\pm 0.008}$	97.1$_{\pm 3.7}$
MARS	605.7$_{\pm 7.2}$	27.2$_{\pm 5.1}$	632.1$_{\pm 7.7}$	0.642$_{\pm 0.011}$	0.951$_{\pm 0.011}$	94.2$_{\pm 9.5}$
TokenSAR	580.4$_{\pm 2.3}$	72.2$_{\pm 11.8}$	652.6$_{\pm 3.6}$	0.665$_{\pm 0.015}$	0.912$_{\pm 0.012}$	87.2$_{\pm 7.4}$
P(True)	613.1$_{\pm 11.1}$	34.7$_{\pm 6.5}$	647.2$_{\pm 13.9}$	0.650$_{\pm 0.020}$	0.964$_{\pm 0.013}$	94.3$_{\pm 12.5}$
SE	624.2$_{\pm 9.5}$	30.3$_{\pm 2.4}$	654.7$_{\pm 5.8}$	0.651$_{\pm 0.011}$	0.981$_{\pm 0.008}$	95.8$_{\pm 14.6}$
LACIE (DPO)	646.7$_{\pm 3.3}$	0.0$_{\pm 0.0}$	646.7$_{\pm 3.3}$	0.647$_{\pm 0.003}$	**1.000**$_{\pm 0.000}$	100.0$_{\pm 0.0}$
SelectLLM	599.5$_{\pm 24.3}$	**141.8**$_{\pm 20.2}$	**741.3**$_{\pm 9.8}$	**0.745**$_{\pm 0.021}$	0.919$_{\pm 0.027}$	80.55$_{\pm 5.14}$

We conduct experiments on the TriviaQA and CommonsenQA datasets. As shown in Table 2&3, our method SelectLLM, consistently and substantially improves model truthfulness and precision across all three language models. It achieves the highest TRUTH score by a significant margin in every experiment—for instance, reaching 752.0 with Llama-3.1 compared to the base model's 601.7. This strong performance is primarily driven by its unique strength in correctly abstaining from providing an answer, as evidenced by its leading True Negative (TN) values (e.g., 230.3 for Mistral-7B on TriviaQA and 142.6 on CommonsenseQA). In contrast, all other score-based methods (LARS, MARS, TokenSAR, P(True), SE) fail to provide a reliable confidence score, since their low TN counts and only marginal precision gains over the base model demonstrate an inability to effectively identify and filter out incorrect answers. We further analyze the confidence scores generated by SelectLLM in Section 5.4.

Consequently, when SelectLLM does generate a response, its reliability is much higher, reflected in its top-ranking Precision scores (e.g., 0.745 for Qwen2.5 on TriviaQA vs. the base model's 0.636). This enhanced precision comes with a deliberate sacrifice of lower Coverage and Recall, as SelectLLM strategically answers fewer questions to avoid making errors. This demonstrates its effectiveness for applications where accuracy is more critical than providing an answer to every query.

Table 3: CommonsenseQA performance. ↑ indicates the higher the better, and ↓ indicates the lower the better. ***The TN value for both the base and LACIE is 0.0 (with a corresponding Recall of 1.0), since they do not abstain from any answers.***

Model	TP ↑	TN ↑	TRUTH ↑	Precision ↑	Recall ↑	Coverage (%)
Llama-3.1-8B-Instruct						
base	$627.3_{\pm 10.1}$	$0.0_{\pm 0.0}$	$627.3_{\pm 10.1}$	$0.627_{\pm 0.004}$	$\mathbf{1.000}_{\pm 0.000}$	$100.0_{\pm 0.0}$
LARS	$575.6_{\pm 4.4}$	$14.2_{\pm 6.2}$	$589.0_{\pm 9.1}$	$0.616_{\pm 0.019}$	$0.917_{\pm 0.027}$	$93.4_{\pm 12.8}$
MARS	$567.3_{\pm 7.2}$	$11.1_{\pm 6.1}$	$578.8_{\pm 8.9}$	$0.610_{\pm 0.011}$	$0.904_{\pm 0.010}$	$92.9_{\pm 11.0}$
TokenSAR	$554.3_{\pm 7.5}$	$21.1_{\pm 6.4}$	$575.7_{\pm 12.5}$	$0.612_{\pm 0.020}$	$0.884_{\pm 0.014}$	$90.6_{\pm 5.5}$
P(True)	$566.1_{\pm 6.9}$	$13.3_{\pm 5.7}$	$579.7_{\pm 4.7}$	$0.611_{\pm 0.013}$	$0.903_{\pm 0.018}$	$92.6_{\pm 9.2}$
SE	$559.4_{\pm 7.6}$	$20.0_{\pm 5.9}$	$579.1_{\pm 3.4}$	$0.613_{\pm 0.020}$	$0.891_{\pm 0.031}$	$91.2_{\pm 9.9}$
LACIE (DPO)	$\mathbf{733.7}_{\pm 12.2}$	$0.0_{\pm 0.0}$	$733.7_{\pm 12.2}$	$0.734_{\pm 0.012}$	$\mathbf{1.000}_{\pm 0.000}$	$100.0_{\pm 0.0}$
SelectLLM	$697.2_{\pm 23.1}$	$\mathbf{98.6}_{\pm 22.1}$	$\mathbf{795.8}_{\pm 11.2}$	$\mathbf{0.834}_{\pm 0.016}$	$0.915_{\pm 0.027}$	$83.28_{\pm 4.09}$
Mistral-7B-Instruct-v0.2						
base	$596.2_{\pm 12.9}$	$0.0_{\pm 0.0}$	$596.2_{\pm 10.9}$	$0.596_{\pm 0.009}$	$\mathbf{1.000}_{\pm 0.000}$	$100.0_{\pm 0.0}$
LARS	$595.9_{\pm 8.7}$	$19.5_{\pm 4.1}$	$614.5_{\pm 7.3}$	$0.607_{\pm 0.012}$	$0.998_{\pm 0.004}$	$98.0_{\pm 9.3}$
MARS	$582.3_{\pm 7.5}$	$26.8_{\pm 6.7}$	$608.1_{\pm 9.5}$	$0.606_{\pm 0.016}$	$0.976_{\pm 14.1}$	$96.0_{\pm 12.4}$
TokenSAR	$571.3_{\pm 6.8}$	$27.5_{\pm 2.7}$	$598.2_{\pm 6.2}$	$0.602_{\pm 0.019}$	$0.958_{\pm 0.016}$	$94.8_{\pm 21.9}$
P(True)	$563.6_{\pm 7.4}$	$51.7_{\pm 5.5}$	$614.1_{\pm 6.9}$	$0.614_{\pm 0.010}$	$0.945_{\pm 18.7}$	$91.6_{\pm 10.9}$
SE	$579.3_{\pm 11.2}$	$24.6_{\pm 10.3}$	$603.9_{\pm 7.1}$	$0.604_{\pm 0.012}$	$0.972_{\pm 15.4}$	$95.9_{\pm 13.3}$
LACIE (DPO)	$603.7_{\pm 9.0}$	$0.0_{\pm 0.0}$	$603.7_{\pm 9.0}$	$0.604_{\pm 9.9}$	$\mathbf{1.000}_{\pm 0.000}$	$100.0_{\pm 0.0}$
SelectLLM	$\mathbf{611.6}_{\pm 29.4}$	$\mathbf{142.6}_{\pm 27.9}$	$\mathbf{754.2}_{\pm 10.7}$	$\mathbf{0.775}_{\pm 0.028}$	$0.900_{\pm 0.026}$	$78.8_{\pm 6.43}$
Qwen2.5-14B-Instruct						
base	$800.0_{\pm 12.4}$	$0.0_{\pm 0.0}$	$800.0_{\pm 12.4}$	$0.800_{\pm 0.011}$	$\mathbf{1.000}_{\pm 0.000}$	$100.0_{\pm 0.0}$
LARS	$798.2_{\pm 13.8}$	$19.5_{\pm 7.1}$	$817.0_{\pm 8.4}$	$0.815_{\pm 0.010}$	$0.998_{\pm 0.005}$	$97.9_{\pm 8.2}$
MARS	$785.8_{\pm 5.9}$	$52.2_{\pm 15.3}$	$837.4_{\pm 6.8}$	$0.841_{\pm 0.011}$	$0.981_{\pm 0.008}$	$93.3_{\pm 7.7}$
TokenSAR	$713.6_{\pm 9.3}$	$62.2_{\pm 2.8}$	$775.8_{\pm 12.1}$	$0.838_{\pm 0.022}$	$0.891_{\pm 0.014}$	$85.1_{\pm 7.9}$
P(True)	$768.3_{\pm 7.7}$	$12.0_{\pm 5.2}$	$780.2_{\pm 7.6}$	$0.803_{\pm 0.018}$	$0.960_{\pm 0.010}$	$95.6_{\pm 10.9}$
SE	$777.3_{\pm 4.5}$	$41.7_{\pm 9.9}$	$818.5_{\pm 7.4}$	$0.830_{\pm 0.010}$	$0.971_{\pm 0.009}$	$93.6_{\pm 8.3}$
LACIE (DPO)	$\mathbf{823.7}_{\pm 4.0}$	$0.0_{\pm 0.0}$	$823.7_{\pm 4.0}$	$0.824_{\pm 0.004}$	$\mathbf{1.000}_{\pm 0.000}$	$100.0_{\pm 0.0}$
SelectLLM	$777.4_{\pm 9.0}$	$\mathbf{68.6}_{\pm 8.7}$	$\mathbf{846.0}_{\pm 3.0}$	$\mathbf{0.884}_{\pm 0.011}$	$0.938_{\pm 0.016}$	$88.01_{\pm 1.70}$

5.3 Out-of-distribution Generalization

To further assess the generalizability of SelectLLM, we evaluate its performance on out-of-distribution (OOD) datasets. Specifically, the tested models are fine-tuned on CommonsenseQA, without any additional fine-tuning on the test datasets – TriviaQA and MedConceptsQA. The evaluation results are reported in Table 4&5. The results demonstrate that the learned abstention ability is transferable to OOD datasets. While the base and LACIE (DPO) models, which lack an abstention mechanism, are forced to answer every question, resulting in a True Negative (TN) of 0.0 and a low Precision, SelectLLM successfully transfers its learned skill of abstaining from uncertain queries to the unseen domains. This is clearly evidenced by its high TN counts: 74.0 on TriviaQA and a remarkable 172.0 on MedConceptsQA. By correctly identifying and abstaining from these challenging OOD questions, SelectLLM significantly boosts its Precision and surpasses the performance of both the base models and LACIE (DPO). The successful transfer of its capability results in a higher TRUTH score, showing that SelectLLM is not only more reliable in familiar settings but also exhibits robustness and generalizability when faced with novel data.

Table 4: TriviaQA (out-of-distribution) performance. ***The TN value for both the base and LACIE is 0.0 (with a corresponding Recall of 1.0), since they do not abstain from any answers.***

Model	TP ↑	TN ↑	TRUTH ↑	Precision ↑	Recall ↑	Coverage (%)
Llama-3.1-8B-Instruct						
base	$\mathbf{601.7}_{\pm 2.3}$	$0.0_{\pm 0.0}$	$601.7_{\pm 2.3}$	$0.602_{\pm 0.002}$	$\mathbf{1.000}_{\pm 0.000}$	$100.0_{\pm 0.0}$
LACIE (DPO)	$579.3_{\pm 23.6}$	$0.0_{\pm 0.0}$	$579.3_{\pm 23.6}$	$0.579_{\pm 0.024}$	$\mathbf{1.000}_{\pm 0.000}$	$100.0_{\pm 0.0}$
SelectLLM	$555.0_{\pm 12.7}$	$\mathbf{74.0}_{\pm 10.1}$	$\mathbf{629.0}_{\pm 13.6}$	$\mathbf{0.626}_{\pm 0.012}$	$0.933_{\pm 0.011}$	$86.72_{\pm 3.67}$

5.4 Validation of SelectLLM Confidence Scores

In this section, we validate the confidence scores generated by SelectLLM by comparing their distribution with the tone-confidence score (referred to Section 5.3) produced by DeepSeek-v3. To

Table 5: MedConceptsQA (out-of-distribution) performance. ***The TN value for both the base and DPO is 0.0 (with a corresponding Recall of 1.0), since they do not abstain from any answers.***

Model	TP ↑	TN ↑	TRUTH ↑	Precision ↑	Recall ↑	Coverage (%)
Llama-3.1-8B-Instruct						
base	$319.0_{\pm 5.13}$	$0.0_{\pm 0.00}$	$319.0_{\pm 5.13}$	$0.319_{\pm 0.05}$	$\mathbf{1.000}_{\pm 0.00}$	$100.0_{\pm 0.00}$
LACIE (DPO)	$\mathbf{465.0}_{\pm 37.48}$	$0.0_{\pm 0.00}$	$465.0_{\pm 37.48}$	$0.465_{\pm 0.04}$	$\mathbf{1.000}_{\pm 0.00}$	$100.0_{\pm 0.00}$
SelectLLM	$406.7_{\pm 22.23}$	$\mathbf{172.0}_{\pm 4.89}$	$\mathbf{578.7}_{\pm 17.62}$	$\mathbf{0.543}_{\pm 0.03}$	$0.839_{\pm 0.01}$	$75.0_{\pm 0.12}$

visualize these two distributions, we first divide the tone-confidence scores into five bins ([0.2, 0.36], [0.36, 0.52], [0.52, 0.68], [0.68, 0.84], [0.84, 1.00]). Each sample is assigned to a bin based on its tone-confidence score. We then compute the mean tone-confidence and the mean SelectLLM-generated confidence for the samples within each bin.

Figure 3 illustrates a small distribution difference between the confidence scores produced by SelectLLM and the tone-confidence scores generated by DeepSeek-v3 on two datasets. The close alignment of the mean SelectLLM confidence scores with the corresponding tone-confidence scores across all bins demonstrates that the selection head produces meaningful and well-calibrated confidence estimates. This evidence supports the conclusion that SelectLLM can internally and reliably estimate its own prediction confidence, without requiring external reference models.

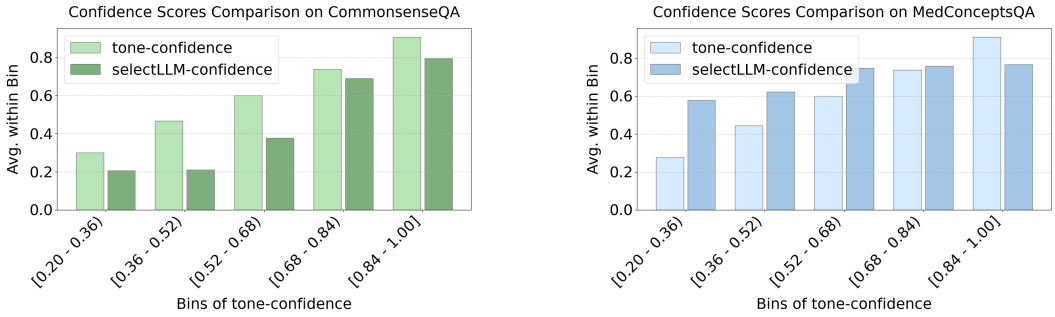

Figure 3: Distribution Difference between tone-confidence and SelectLLM-confidence for in-distribution (left, CommonsenseQA) and out-of-distribution (right, MedConceptsQA)

6 Conclusion

In this paper, we have introduced an alignment-based method, called SelectLLM, that explicitly produces confidence estimates and incorporates the task of confidence estimation into its training objectives. Our extensive empirical evaluations on three QA benchmark datasets, using three different LLMs, demonstrate that SelectLLM consistently achieves better risk-coverage tradeoffs than seven baselines. SelectLLM's notable strengths include superior uncertainty calibration, robust cross-domain generalization, and flexible, tunable performance. These experimental outcomes confirm SelectLLM as an effective and principled solution for enhancing model reliability and practical utility in uncertainty-sensitive, real-world scenarios.

References

[1] Felix Busch, Lena Hoffmann, Christopher Rueger, Elon HC van Dijk, Rawen Kader, Esteban Ortiz-Prado, Marcus R Makowski, Luca Saba, Martin Hadamitzky, Jakob Nikolas Kather, et al. Current applications and challenges in large language models for patient care: a systematic review. *Communications Medicine*, 5(1):26, 2025.

[2] Kerstin Denecke, Richard May, LLMHealthGroup, and Octavio Rivera Romero. Potential of large language models in health care: Delphi study. *Journal of Medical Internet Research*, 26:e52399, 2024.

[3] Minji Yoo. How much should we trust llm-based measures for accounting and finance research? *Available at SSRN*, 2024.

[4] Yuqi Nie, Yaxuan Kong, Xiaowen Dong, John M Mulvey, H Vincent Poor, Qingsong Wen, and Stefan Zohren. A survey of large language models for financial applications: Progress, prospects and challenges. *arXiv preprint arXiv:2406.11903*, 2024.

[5] Rajaa El Hamdani, Thomas Bonald, Fragkiskos D Malliaros, Nils Holzenberger, and Fabian Suchanek. The Factuality of Large Language Models in the Legal Domain. In *ACM International Conference on Information and Knowledge Management*, 2024.

[6] Pierre Colombo, Telmo Pessoa Pires, Malik Boudiaf, Dominic Culver, Rui Melo, Caio Corro, Andre FT Martins, Fabrizio Esposito, Vera Lúcia Raposo, Sofia Morgado, et al. SaulLM-7B: A pioneering Large Language Model for Law. *arXiv:2403.03883*, 2024.

[7] Bingbing Wen, Jihan Yao, Shangbin Feng, Chenjun Xu, Yulia Tsvetkov, Bill Howe, and Lucy Lu Wang. Know your limits: A survey of abstention in large language models. *arXiv preprint arXiv:2407.18418*, 2024.

[8] Yekyung Kim, Jenna Russell, Marzena Karpinska, and Mohit Iyyer. One ruler to measure them all: Benchmarking multilingual long-context language models. *arXiv preprint arXiv:2503.01996*, 2025.

[9] Elias Stengel-Eskin, Peter Hase, and Mohit Bansal. Lacie: Listener-aware finetuning for confidence calibration in large language models. *arXiv preprint arXiv:2405.21028*, 2024.

[10] Qinyuan Cheng, Tianxiang Sun, Xiangyang Liu, Wenwei Zhang, Zhangyue Yin, Shimin Li, Linyang Li, Zhengfu He, Kai Chen, and Xipeng Qiu. Can ai assistants know what they don't know? *arXiv preprint arXiv:2401.13275*, 2024.

[11] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741, 2023.

[12] Sree Harsha Tanneru, Chirag Agarwal, and Himabindu Lakkaraju. Quantifying uncertainty in natural language explanations of large language models. *arXiv preprint arXiv:2311.03533*, 2023.

[13] Gal Yona, Roee Aharoni, and Mor Geva. Can large language models faithfully express their intrinsic uncertainty in words? *arXiv preprint arXiv:2405.16908*, 2024.

[14] Zeming Wei, Yifei Wang, Ang Li, Yichuan Mo, and Yisen Wang. Jailbreak and guard aligned language models with only few in-context demonstrations. *arXiv preprint arXiv:2310.06387*, 2023.

[15] Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. Generating with confidence: Uncertainty quantification for black-box large language models. *arXiv preprint arXiv:2305.19187*, 2023.

[16] Jeremy R Cole, Michael JQ Zhang, Daniel Gillick, Julian Martin Eisenschlos, Bhuwan Dhingra, and Jacob Eisenstein. Selectively answering ambiguous questions. *arXiv preprint arXiv:2305.14613*, 2023.

[17] Jiabao Ji, Bairu Hou, Alexander Robey, George J Pappas, Hamed Hassani, Yang Zhang, Eric Wong, and Shiyu Chang. Defending large language models against jailbreak attacks via semantic smoothing. *arXiv preprint arXiv:2402.16192*, 2024.

[18] Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. *arXiv preprint arXiv:2306.13063*, 2023.

[19] Jinhao Duan, Hao Cheng, Shiqi Wang, Alex Zavalny, Chenan Wang, Renjing Xu, Bhavya Kailkhura, and Kaidi Xu. Shifting attention to relevance: Towards the predictive uncertainty quantification of free-form large language models. *arXiv preprint arXiv:2307.01379*, 2023.

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

[21] Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*, 2023.

[22] Ranganath Krishnan, Piyush Khanna, and Omesh Tickoo. Enhancing trust in large language models with uncertainty-aware fine-tuning. *arXiv preprint arXiv:2412.02904*, 2024.

[23] Ruijia Niu, Dongxia Wu, Rose Yu, and Yi-An Ma. Functional-level uncertainty quantification for calibrated fine-tuning on llms. *arXiv preprint arXiv:2410.06431*, 2024.

[24] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[25] Katie Kang, Eric Wallace, Claire Tomlin, Aviral Kumar, and Sergey Levine. Unfamiliar finetuning examples control how language models hallucinate. *arXiv preprint arXiv:2403.05612*, 2024.

[26] Alexandre Piché, Aristides Milios, Dzmitry Bahdanau, and Chris Pal. Llms can learn self-restraint through iterative self-reflection. *arXiv preprint arXiv:2405.13022*, 2024.

[27] Giorgio Fumera and Fabio Roli. Support vector machines with embedded reject option. In *International workshop on support vector machines*, pages 68–82. Springer, 2002.

[28] Yair Wiener and Ran El-Yaniv. Agnostic pointwise-competitive selective classification. *Journal of Artificial Intelligence Research*, 52:171–201, 2015.

[29] Corinna Cortes, Giulia DeSalvo, and Mehryar Mohri. Boosting with abstention. *Advances in neural information processing systems*, 29, 2016.

[30] Corinna Cortes, Giulia DeSalvo, and Mehryar Mohri. Learning with rejection. In *International conference on algorithmic learning theory*, pages 67–82. Springer, 2016.

[31] 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*. Association for Computational Linguistics, 2023.

[32] Minjae Lee, Kyungmin Kim, Taesoo Kim, and Sangdon Park. Selective generation for controllable language models. *Advances in Neural Information Processing Systems*, 37:50494–50527, 2024.

[33] Tejas Srinivasan, Jack Hessel, Tanmay Gupta, Bill Yuchen Lin, Yejin Choi, Jesse Thomason, and Khyathi Raghavi Chandu. Selective" selective prediction": Reducing unnecessary abstention in vision-language reasoning. *arXiv preprint arXiv:2402.15610*, 2024.

[34] Yonatan Geifman and Ran El-Yaniv. Selectivenet: A deep neural network with an integrated reject option. In *International conference on machine learning*, pages 2151–2159. PMLR, 2019.

[35] Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. *Advances in Neural Information Processing Systems*, 30, 2017.

[36] Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*, 2017.

[37] Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *arXiv preprint arXiv:1811.00937*, 2018.

[38] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115, 2023.

[39] Duygu Nur Yaldiz, Yavuz Faruk Bakman, Baturalp Buyukates, Chenyang Tao, Anil Ramakrishna, Dimitrios Dimitriadis, Jieyu Zhao, and Salman Avestimehr. Do not design, learn: A trainable scoring function for uncertainty estimation in generative llms. *arXiv preprint arXiv:2406.11278*, 2024.

[40] Jannis Bulian, Christian Buck, Wojciech Gajewski, Benjamin Börschinger, and Tal Schuster. Tomayto, tomahto. beyond token-level answer equivalence for question answering evaluation. *arXiv preprint arXiv:2202.07654*, 2022.

[41] Ofir Ben Shoham and Nadav Rappoport. Medconceptsqa: Open source medical concepts qa benchmark. *Computers in Biology and Medicine*, 182:109089, 2024.

A Qualitative Analysis

In the following table, we provide the qualitative examples of the base and SelectLLM fine-tuned Llama3.1-8B models, together with the SelectLLM confidence score to the questions. Correct answers are highlighted in green. These examples demonstrate the effectiveness of SelectLLM in two key aspects: it enables the model to express high confidence when it truly knows the answer, and to avoid overconfidence when uncertain. Moreover, SelectLLM produces accurate confidence scores without relying on any external model. This demonstrates SelectLLM 's ability to achieve a better risk–coverage trade-off.

Question	**base (Llama3.1-8B)**	**SelectLLM**
While people just throw coins down them now, what originally had a pail to be lowered for it's intended use? Choose the answer from the following choices. (a) garage. (b) utility room. (c) slide. ***(d) wishing well.*** *(e) garden.*	I'm about 50% confident in my answer. I think the answer is d. I choose this because in a wishing well, a pail or bucket is often lowered down into the well so people can collect water. [Correct & Unconfident]	(d) wishing well. A wishing well is a type of stone-lined well that was originally built with a pail lowered down a rope for people to make a wish and retrieve water. The pail made it possible for people to make a wish and then retrieve water. This is the origin of the wishing well. The answer choice is (d) wishing well. I am 95% sure. [Correct & Confident, SelectLLM-confidence=0.968]
Miss Grady took a stick from Bob because he was playing with it during class. She wanted to make sure that he couldn't get to it so she put it where? Choose the answer from the following choices. ***(a) desk drawer.*** *(b) kitchen drawer. (c) classroom. (d) pocket. (e) office.*	Miss Grady put the stick where so that Bob could not get to it. In order to do that, she would most likely have put it in a location where she would have control over it. The classroom is where Miss Grady would have the most control over the stick, so that's where she would have put it. Answer: C. [Incorrect & Confident]	(a) or (d) possibly (d) perhaps a pencil or supply box, but (d) a desk drawer or (d) supply cabinet would be most likely. Explanation: Since Miss Grady took the stick from Bob to prevent him from using it, she would likely put it in a secure location where he would not be able to get to it. [Abstention & Unconfident, SelectLLM-confidence=0.051]
What is the description of the medical code S04.019S in ICD10CM? A. Unspecified fracture of unspecified thoracic vertebra, subsequent encounter for fracture with nonunion. B. Terrorism involving fires, conflagration and hot substances, civilian injured, subsequent encounter. ***C. Injury of optic nerve, unspecified eye, sequela.*** *D. Laceration of abdominal wall with foreign body, epigastric region without penetration into peritoneal cavity, sequela.*	I am 95% confident that my answer is B. Terrorism involving fires, conflagration and hot substances, civilian injured, subsequent encounter. This is because the code S04.019S starts with S̈04ẅhich is the code for terrorism, and the Ö19Sr̈epresents the subsequent encounter for a civilian injured. This makes option B the most likely correct answer. [Incorrect & Confident]	None of the options describe a code that is the same as T32.98 in ICD10CM. The options provided don't seem to be correct. The only way to actually answer this question would be to look at the ICD10CM for the code T32.98. [Abstention & Unconfident, SelectLLM-confidence=0.233]

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We illustrate the setting of LLM selective prediction in Section 1 and Section 3, provide a framework targeting this setting in Section 4, and demonstrate the framework's utility by experiments in Section 5.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: We discussed the limitation of our framework in Section 6, in which we inherit the issue that our proposed method is not examined in larger LLMs, which is due to the limited computational resources we have.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: We don't have theoretical result in the paper.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We specify the key steps and elements to reproduce the experimental results in Section 5.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We will include the code and dataset in the supplementary material.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.

- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We specify the fine-tuning and evaluation procedures and metrics calculations in Section 5.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report the means and standard deviations for all the experiment results across 5 different seeds. We show the statistical results in Section 5.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We state that our experiments can be done on two A100-40G GPUS in Section 5.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed the Code of Ethics and think our work follows the ethical requirements.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: This work proposes a framework to fine-tune the LLM for selective prediction, with potential positive societal impacts, which has been discussed in Section 1. We don't think there is any crucial potential societal consequence of our work.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: This work does not release any models or data that have a high risk for misuse.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We properly cite the three datasets used in our paper in the references. They are available publicly with a license.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: This work does not release new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.

- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This work does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This work does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: This work does not involve LLMs as any important, original, or non-standard component.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.