WHEN WORDS FALL SHORT: ITERATIVE SYNERGY BETWEEN VERBALIZED REASONING AND HIDDEN FEATURES FOR LLM CONFIDENCE ESTIMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Confidence estimation is crucial for developing trustworthy large language models (LLMs), with most methods following estimator-based or verbalization-based paradigms. While recent research increasingly focuses on improving verbalized self-reports of confidence, we challenge the prevailing view that this approach surpasses independent confidence estimators. Our empirical study shows that a dedicated confidence estimator can substantially outperform verbalized confidence, indicating that LLMs' internal representations contain richer confidence signals. Building on this finding, we propose **I**terative **Po**licy-**E**stimator **T**raining (**IPoET**), a framework that synergizes the complementary strengths of verbalized reasoning traces and informative representations. IPoET alternates policy optimization with estimator updating, integrating estimator-derived confidence feedback into policy learning and refreshing the estimator on new policy rollouts. Experiments across diverse datasets and Qwen and Llama backbones demonstrate that, by iteratively exploiting richer hidden features and adapting to the evolving policy distribution, IPoET consistently outperforms both estimator- and verbalization-based baselines in-domain and achieves superior or comparable results across all out-of-domain metrics. For more details, refer to `https://anonymous.4open.science/r/ipoet`.

1 INTRODUCTION

Large language models (LLMs) have become increasingly capable across a wide range of challenging tasks, such as reasoning, code generation, and planning (Jaech et al., 2024; Guo et al., 2025; Jimenez et al., 2024; Xiao et al., 2024). However, higher accuracy does not necessarily mean that models can correctly assess whether their outputs should be trusted. LLMs frequently generate hallucinated or incorrect answers with unwarranted confidence (Xiong et al., 2024; Anh-Hoang et al., 2025). This lack of reliable confidence estimation severely limits the practical deployment of LLMs in high-stakes domains such as medicine, law, and finance, where decision-making errors carry significant consequences (Li et al., 2024; Siino et al., 2025; Li et al., 2023; Kumaran et al., 2026b).

Existing research on confidence estimation can generally be categorized into two paradigms: (i) *Estimator-based methods*. These methods train independent models or lightweight probing heads to exploit features embedded within the model's hidden states (Beigi et al., 2024; Mahaut et al., 2024). (ii) *Verbalization-based methods*. This paradigm leverages the linguistic reasoning of LLMs and derives verbalized confidence scores through prompting or training (Yang et al., 2024; Li et al., 2026b). Recent work on this task has increasingly focused on RL-based verbalization methods, with several studies reporting that verbalized confidence outperforms estimator-based methods (Damani et al., 2025; Bani-Harouni et al., 2025; Ma et al., 2026; Zhang et al., 2026a).

However, revisiting this comparison leads us to a different finding: with validation-based checkpoint selection, an independent confidence estimator can substantially outperform state-of-the-art verbalized confidence in both in-domain and out-of-domain evaluations, as shown in Figure 1a. We find that estimator overfitting can bias comparisons in favor of verbalized confidence, as continued optimization can reduce training error while degrading confidence estimation on unseen data.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

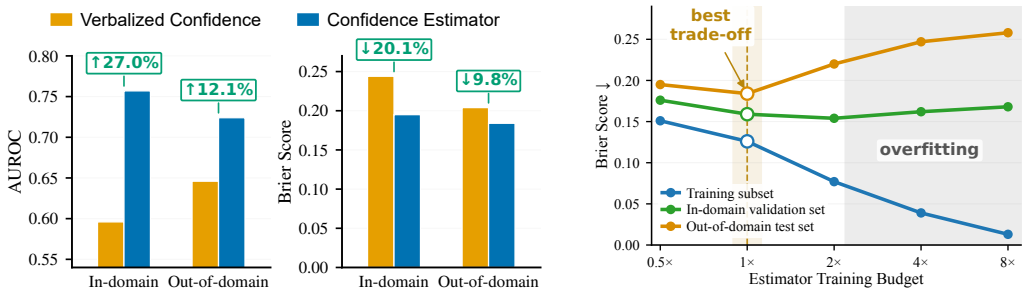

Figure 1: (a) In-domain and out-of-domain comparison of confidence estimator and verbalized confidence using AUROC and Brier score. Percentages denote relative improvements over verbalized confidence. (b) Estimator performance across training budgets: Brier scores on a training subset, an in-domain validation set, and an out-of-domain test set illustrate overfitting with extended training.

These results establish estimators trained with overfitting control as a strong confidence estimation paradigm that deserves further attention.

This finding also highlights the complementary strengths and limitations of estimator- and verbalization-based approaches. Hidden features within model representations are essential for accurate confidence estimation, yet they cannot be easily translated into natural language (Kumaran et al., 2026a). At the same time, decoupling the estimator from the main policy limits its ability to fully utilize the advanced reasoning capabilities of contemporary LLMs, which have proven effective for confidence estimation and various other tasks. This raises a central question: *Can we harness the advantages of both paradigms to achieve more reliable confidence estimation?*

To answer this, we propose **I**terative **Po**licy-**E**stimator **T**raining (**IPoET**) for confidence estimation. It alternates between two stages: (i) optimizing the policy with an objective that incorporates estimator feedback while preserving task accuracy, and (ii) updating the estimator to leverage hidden states encoding reasoning traces from the current policy. This process enables policy reasoning to provide richer confidence signals that the estimator can extract from its hidden representations, while keeping the estimator aligned with the evolving policy distribution. Consequently, this co-adaptation yields stronger confidence estimation than isolated optimization.

We evaluate IPoET across multiple mathematical reasoning and question-answering datasets using different backbone models. Compared to both estimator- and verbalization-based approaches, our method consistently achieves superior or competitive performance on confidence estimation metrics while preserving task performance. Furthermore, IPoET generalizes effectively to out-of-domain tasks. We also investigate algorithmic variants, such as joint training, to provide deeper insights into the underlying effectiveness of our method.

Our contributions are summarized as follows:

- We revisit the comparison between estimator- and verbalization-based paradigms, showing that estimators can substantially outperform verbalized confidence and revealing estimator overfitting as an important factor underlying prior discrepancies.
- We propose IPoET, an iterative training framework that harnesses the complementary advantages of LLM verbalized reasoning and rich hidden features for confidence estimation.
- Across diverse datasets and backbone models, IPoET outperforms baselines on in-domain confidence estimation while achieving superior or comparable out-of-domain performance.

2 Empirical Study

In this section, we revisit two confidence estimation paradigms: estimator-based and verbalization-based. We begin by highlighting their implementations in Section 2.1, before examining how estimator training dynamics affect their relative performance in Section 2.2.

2.1 PRELIMINARIES

Task Definition: Given a prompt x from a dataset $\mathcal{D}$, an LLM π_θ generates a response y consisting of an optional reasoning trace τ and a final answer a. We define the correctness label as $z = \mathbb{I}[a \equiv a^*]$, where a^* is the ground-truth answer. Confidence estimation aims to assign a score $c \in [0, 1]$ representing the likelihood that a is correct.

Estimator-based approaches. These approaches utilize an independent confidence estimator s_ϕ. While prior work typically employs either a lightweight probing head or an entire model as the estimator, we follow Malladi et al. (2023) and adopt the latter for its superior performance (Ni et al., 2025). Specifically, s_ϕ processes the input x and response y to predict a confidence score c by applying a linear head to the hidden state of the last non-padding token, denoted as h: $c = s_\phi(x, y) = \mathrm{clip}(\mathbf{w}^\top h + b, 0, 1)$, where $\mathbf{w}$ and b denote the weights and bias. We train the estimator on collected rollouts $\mathcal{D}_{\text{est}} = \{(x, y, z)\}$ using a mean squared error (MSE) loss:

$$\mathcal{L}_{\text{est}} = \mathbb{E}_{(x,y,z)\sim\mathcal{D}_{\text{est}}} \left[(s_\phi(x, y) - z)^2 \right]. \tag{1}$$

Previous studies show that critical features for confidence estimation are distributed across intermediate layers (Azaria & Mitchell, 2023; Subramani et al., 2025). Through end-to-end gradient optimization, the final hidden state learns to distill and consolidate these confidence signals scattered throughout the network.

Verbalization-based approaches. These approaches leverage the strong capabilities of recent LLMs to generate an answer a alongside verbalized confidence c via a natural language reasoning trace τ. To jointly optimize both variables, this paradigm typically employs RL to search for reasoning traces τ that provide a more rigorous rationale for inferring a precise confidence c. Following RLCR (Damani et al., 2025), we augment the correctness reward in conventional RLVR (Shao et al., 2024) with a Brier-style term over the verbalized confidence:

$$\mathcal{J}_{\text{verb}}(\theta) = \mathbb{E}_{x\sim\mathcal{D},\,(y,c)\sim\pi_\theta(\cdot|x)} \left[z - (c - z)^2 \right]. \tag{2}$$

Recent studies have claimed that RL-trained verbalized confidence outperforms independently trained confidence estimators (Ma et al., 2026; Zhang et al., 2026a). For example, in comparisons with their respective estimator baselines, RLCR highlights calibration gains, particularly under distribution shift (Damani et al., 2025), while Rewarding Doubt emphasizes improved discrimination and generalization (Bani-Harouni et al., 2025).

2.2 ESTIMATOR-BASED VS. VERBALIZATION-BASED

We next revisit the comparison between estimator-based and verbalization-based confidence estimation. We hypothesize that estimator overfitting may explain the previously claimed advantage of verbalized confidence. An evaluation at a single training endpoint may obscure this effect, as it does not reveal how training and validation performance evolve during optimization. To investigate this possibility, we examine checkpoints throughout full-parameter tuning, testing whether continued optimization yields sustained improvements or reverses earlier gains in generalization. We use the HotpotQA training set and track Brier scores on a 5K training subset, a held-out in-domain validation set, and the out-of-domain benchmarks described in Section 4.1.

As shown in Figure 1b, the training Brier score decreases continuously, while the in-domain validation score first decreases and then rises. The out-of-domain score exhibits a more pronounced deterioration, increasing from an early minimum of 0.184 to 0.258 with extended training. This divergence provides evidence of overfitting. For subsequent estimator training, we adopt the $1\times$ configuration without additional checkpoint selection, as it achieves in-domain validation performance close to that of $2\times$ with half the budget. Consistent with this analysis, the full-scale experiments in Section 4.2 further demonstrate strong estimator performance with overfitting control.

Our comparison shows that estimator-based confidence estimation can substantially outperform RL-trained verbalized confidence. As shown in Figure 1a, the confidence estimator improves AUROC by 0.161 in-domain and 0.078 out-of-domain, while reducing Brier score by 0.049 and 0.020, respectively. These results challenge the view that verbalized confidence is inherently stronger than hidden-state estimation, highlighting the value of representation-level confidence evidence.

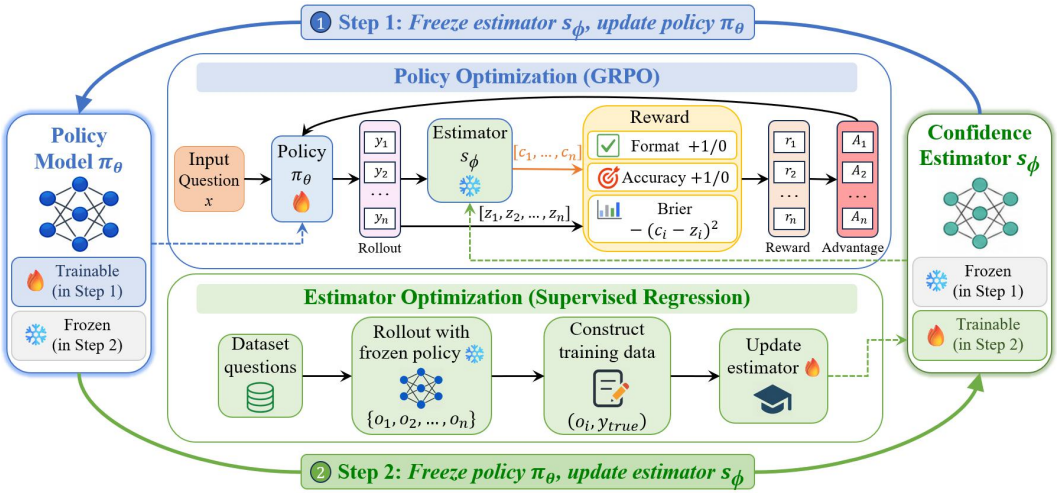

Figure 2: Overview of IPoET. The framework alternates between policy and estimator optimization. *In the policy-update stage*, the confidence estimator is frozen and provides confidence estimates to form an estimator-derived term, combined with format and accuracy rewards for GRPO. *In the estimator-update stage*, the policy is frozen to generate correctness-labeled rollouts for updating the estimator via supervised regression. Across iterations, the policy refines reasoning-oriented generation while the estimator adapts to the evolving policy distribution.

Beyond clarifying the relative performance of the two paradigms, this finding also leads us to examine their complementary capabilities. While estimators extract features from hidden states, RL-based verbalized methods connect confidence expression with the reasoning process (Tao et al., 2024). These distinct advantages suggest a natural opportunity to combine both approaches, which we explore through iterative updates of generation and confidence estimation in the next section.

3 METHOD

We propose IPoET, an iterative framework that connects reasoning-oriented policy training with hidden-state confidence estimation, as illustrated in Figure 2. To better transfer confidence signals into policy learning, we adopt reinforcement learning as the optimization mechanism, enabling estimator-derived reward to directly shape the model's generated reasoning. We describe the training procedure in Section 3.1, and then define the estimator-derived reward in Section 3.2.

3.1 ITERATIVE POLICY-ESTIMATOR TRAINING

Let π_{θ_0} denote the base model. We first apply standard RLVR (Lambert et al., 2024) to obtain an initial policy π_{θ_1}. This warm-up stage provides a stable output format for answer extraction and correctness verification. Rollouts from π_{θ_1} are then used to train the confidence estimator s_{ϕ_1} with the supervised regression objective in Eq. 1.

After initialization, IPoET proceeds in alternating rounds. At iteration t, it performs two steps. **Step 1: *Freeze estimator* s_{ϕ_t}, *update policy* π_{θ_t}.** The estimator s_{ϕ_t} is held fixed while the policy π_{θ_t} is updated to $\pi_{\theta_{t+1}}$ with GRPO (Shao et al., 2024) using a reward that combines format, accuracy, and an estimator-derived confidence term, as detailed in Section 3.2. This step incorporates representation-level confidence feedback into policy learning, encouraging the policy to generate reasoning that makes answer correctness easier to infer from the estimator's hidden representations. **Step 2: *Freeze policy* $\pi_{\theta_{t+1}}$, *update estimator* s_{ϕ_t}.** Since the updated policy may induce a different response distribution, we sample fresh rollouts from $\pi_{\theta_{t+1}}$, assign correctness labels using task-specific answer verification, and update the estimator to $s_{\phi_{t+1}}$ with the same regression objective in Eq. 1. Across estimator updates, we use the fixed training budget established in Section 2.2 to reduce overfitting and maintain estimator reliability on the evolving policy distribution.

Rather than relying on a static estimator, IPoET alternates policy improvement with estimator refinement on current-policy rollouts, allowing the policy and estimator to co-adapt across rounds. The estimator serves both as a confidence predictor and as a source of feedback for reasoning generation.

3.2 Reward Function

Following the notation in Section 2.1, IPoET optimizes the policy using a reward with three components:

$$R_{\mathrm{IPoET}}(x, y, z; \phi) = R_{\mathrm{fmt}}(y) + R_{\mathrm{acc}}(z) + R_{\mathrm{est}}(x, y, z; \phi). \tag{3}$$

Here, R_{fmt} maintains the required output structure, and $R_{\mathrm{acc}}(z) = z$ rewards responses whose final answers match the ground-truth answer.

The key component is the estimator-derived reward R_{est}, which introduces hidden-state confidence estimation into policy training. For each rollout, the current estimator produces a confidence score from its internal representation. IPoET converts this representation-level estimate into a policy reward through a negative Brier-style objective:

$$R_{\mathrm{est}}(x, y, z; \phi) = -\left(s_\phi(x, y) - z\right)^2. \tag{4}$$

This term gives higher reward when estimated confidence matches correctness, thereby discouraging overconfidence and underconfidence. Thus, the estimator directly guides policy optimization using hidden-state confidence estimates, without requiring the policy to verbalize a confidence score.

4 Experiments

4.1 Experimental Setup

Datasets. We focus on tasks with well-defined correctness criteria to support reliable confidence evaluation across diverse reasoning and knowledge settings. Following the experimental setup of RLCR (Damani et al., 2025), we utilize the processed HotpotQA and Big-Math datasets as our training data, derived respectively from the modified HotpotQA distractor data (Yang et al., 2018) and the filtered Big-Math problems (Albalak et al., 2025). For evaluation, we consider benchmarks that reflect different sources of uncertainty in confidence estimation. They fall into three categories: (i) factual question answering, including HotpotQA, TriviaQA (Joshi et al., 2017), and SimpleQA (Wei et al., 2024); (ii) commonsense and expert-level knowledge reasoning, including CommonsenseQA (Talmor et al., 2019) and GPQA (Rein et al., 2023); and (iii) mathematical reasoning, including MATH-500 (Hendrycks et al., 2021), GSM8K (Cobbe et al., 2021), and Big-Math.

Evaluation Metrics. We evaluate each method along two dimensions (details in Appendix B). For task performance, we report Accuracy, computed using dataset-specific evaluation procedures, including exact match, *math_verify*, and LLM-as-a-judge (Zheng et al., 2023) evaluation. Given the short and objective answer formats of these benchmarks, we use automated evaluation without additional human annotation. For confidence estimation, we report Area Under the Receiver Operating Characteristic Curve (AUROC) (Bradley, 1997), Brier score (Glenn et al., 1950), and Expected Calibration Error (ECE) (Guo et al., 2017). AUROC evaluates how confidence scores order correct and incorrect predictions, reflecting the ranking quality of confidence estimates. Brier score averages the squared error between the predicted confidence and the binary correctness label, capturing how accurate each confidence estimate is at the instance level. ECE measures the mismatch between predicted confidence and empirical accuracy after grouping predictions into confidence bins.

Baselines. We compare with the initial model and the following representative baselines, focusing on single-pass confidence estimation without additional inference-time sampling. **RLVR** optimizes the policy with a binary answer reward (Shao et al., 2024) and verbalizes confidence only during evaluation. **Answer-Prob** averages the generation probabilities of all tokens in the final answer (Jiang et al., 2021), while **P(True)** prompts the model to judge whether its own answer is correct and uses the probability of the positive label as confidence (Kadavath et al., 2022). **RLVR + Estimator** trains a confidence estimator on RLVR-generated outputs and correctness labels, following the training protocol in Section 2. **RLCR** trains the model to generate both an answer and verbalized confidence by optimizing a reward that combines answer correctness with a negative

Table 1: Main results for models trained on HotpotQA and Big-Math. We report accuracy and confidence-estimation metrics, including AUROC, Brier score, and ECE, for Qwen and Llama backbones. For models trained on HotpotQA, we report results on HotpotQA and a six-dataset OOD average. For models trained on Big-Math, we report a Math average over MATH-500, GSM8K, and Big-Math, together with a five-dataset OOD average. Best and second-best results for all metrics under each backbone are marked in **bold** and underlined, respectively.

	HotpotQA				OOD Averaged			
	Acc.↑	AUROC↑	Brier↓	ECE↓	Acc.↑	AUROC↑	Brier↓	ECE↓
Qwen3-8B-Base	51.6%	0.555	0.363	0.343	56.5%	0.556	0.359	0.345
⌞ RLVR	62.2%	0.523	0.376	0.376	**62.3%**	0.518	0.357	0.358
⌞ Answer-Prob	62.2%	0.661	0.359	0.360	**62.3%**	0.548	0.340	0.327
⌞ P(True)	62.2%	0.543	0.308	0.260	**62.3%**	0.588	0.304	0.314
⌞ RLVR + Estimator	62.2%	0.757	0.195	0.087	**62.3%**	0.724	0.184	0.184
⌞ RLCR	61.0%	0.596	0.244	0.100	61.4%	0.646	0.204	0.171
⌞ RLCR (Step-aligned)	**65.3%**	0.606	0.263	0.205	61.6%	0.690	0.210	0.188
⌞ RLCR + Estimator	61.0%	0.753	0.201	**0.081**	61.4%	0.719	0.207	0.224
⌞ IPoET (ours)	63.3%	**0.800**	**0.184**	0.107	**62.3%**	**0.738**	**0.170**	**0.146**
Llama-3.1-8B-Instruct	45.9%	0.577	0.461	0.469	45.9%	0.606	0.413	0.424
⌞ RLVR + Estimator	61.6%	0.772	0.198	0.110	51.0%	**0.746**	**0.180**	0.161
⌞ RLCR	59.1%	0.603	0.248	0.121	50.2%	0.674	0.217	0.175
⌞ IPoET (ours)	**64.6%**	**0.781**	**0.188**	**0.101**	**51.5%**	0.737	0.182	**0.152**

	Math				OOD Averaged			
	Acc.↑	AUROC↑	Brier↓	ECE↓	Acc.↑	AUROC↑	Brier↓	ECE↓
Qwen3-8B-Base	62.8%	0.568	0.313	0.290	51.9%	0.562	0.389	0.382
⌞ RLVR	**74.4%**	0.517	0.255	0.255	**52.9%**	0.569	0.417	0.424
⌞ Answer-Prob	**74.4%**	0.709	0.235	0.239	**52.9%**	0.576	0.420	0.424
⌞ P(True)	**74.4%**	0.490	0.366	0.380	**52.9%**	0.645	0.275	0.237
⌞ RLVR + Estimator	**74.4%**	0.862	0.107	**0.037**	**52.9%**	0.684	0.217	0.197
⌞ RLCR	73.7%	0.625	0.184	0.115	52.8%	0.600	0.282	0.272
⌞ RLCR (Step-aligned)	73.8%	0.633	0.181	0.120	41.7%	0.536	0.361	0.347
⌞ RLCR + Estimator	73.7%	0.828	0.126	0.068	52.8%	0.662	0.249	0.249
⌞ IPoET (ours)	74.2%	**0.878**	**0.100**	0.040	50.9%	**0.698**	**0.206**	**0.183**
Llama-3.1-8B-Instruct	42.7%	0.591	0.501	0.512	43.7%	0.611	0.410	0.423
⌞ RLVR + Estimator	53.8%	0.789	0.167	0.085	**45.3%**	0.616	0.235	0.216
⌞ RLCR	50.6%	0.570	0.246	0.165	44.6%	0.616	**0.219**	**0.179**
⌞ IPoET (ours)	**56.5%**	**0.824**	**0.152**	**0.061**	45.2%	**0.622**	0.226	0.191

Brier score (Damani et al., 2025). **RLCR (Step-aligned)** uses the RLCR objective with the same number of training steps as IPoET, serving as a control for the effect of additional training. **RLCR + Estimator** applies a confidence estimator to RLCR-generated outputs. We also evaluate commercial frontier models and observe limitations in their verbalized confidence estimation; see Appendix C.

Implementation Details. Our method is built on the verl framework (Sheng et al., 2025), with GRPO as the RL algorithm (Shao et al., 2024). We use Qwen3-8B-Base (Yang et al., 2025) as the main backbone and evaluate Llama-3.1-8B-Instruct (Grattafiori et al., 2024) to assess generality. For a fair comparison, we reimplement the RL-based baselines in the same training pipeline while preserving their objectives and core design. All methods use the same training data within each setting. For each trainable baseline, we report results from the checkpoint with the best held-out in-domain validation performance. Following common practice in RL training for reasoning models (Guo et al., 2025; Hu et al., 2026), we initialize the policy from the backbone model and train it without KL regularization. Training details, including the prompt template, are provided in Appendix A.

4.2 Main Results

IPoET achieves strong overall in-domain performance. Table 1 shows that IPoET achieves the highest AUROC and the lowest Brier score across all four in-domain settings, covering both datasets

Table 2: Comparison between joint and iterative training on HotpotQA. Iterative training improves accuracy, AUROC, and Brier score in both in-domain and out-of-domain settings. Values in parentheses denote changes relative to joint training.

Setting	Strategy	Accuracy ↑	AUROC ↑	Brier ↓
In-domain	Joint	60.5%	0.773	0.209
	Iterative	**63.3%** (+2.8 pp)	**0.800** (+0.027)	**0.184** (-0.025)
Out-of-domain	Joint	58.4%	0.737	0.180
	Iterative	**62.3%** (+3.9 pp)	**0.738** (+0.001)	**0.170** (-0.010)

and backbone models. Compared with RLVR + Estimator, IPoET achieves AUROC gains and Brier score reductions of up to 0.043 and 0.015, respectively, highlighting its strengths in confidence discrimination and calibration. A paired bootstrap test on the Qwen3-8B-Base HotpotQA setting shows statistically significant improvements over RLVR + Estimator in both AUROC ($p = 0.0006$) and Brier score ($p = 0.0317$). IPoET also achieves the highest task accuracy and the lowest ECE in both Llama settings, with competitive results on both measures for Qwen3. Together, these results support the value of iterative policy–estimator interaction for improving confidence estimation. The baseline comparisons further corroborate our empirical study: with validation-based checkpoint selection, RLVR + Estimator outperforms RLCR in AUROC, Brier score, and ECE across all four in-domain settings, suggesting that hidden-state estimators can exploit correctness-related information not fully captured by explicit self-reports. Despite the additional verbalized confidence analysis in RLCR-generated outputs, RLCR + Estimator is worse than RLVR + Estimator on most metrics. One possible explanation is that explicit self-assessment may dilute correctness-related signals in the original reasoning trace and introduce extra noise into the estimator's hidden representations.

IPoET achieves competitive out-of-domain performance. On benchmarks outside the training domain, IPoET ranks first or second on all confidence-estimation metrics. Under both HotpotQA and Big-Math training, it achieves the highest out-of-domain AUROC and lowest Brier score and ECE on Qwen. On Llama, IPoET achieves the lowest ECE under HotpotQA training and the highest AUROC under Big-Math training, while remaining close to the best baseline results on the other metrics. These results suggest that the benefits of IPoET extend beyond the training domain, although improvements over existing methods vary across evaluation settings.

IPoET's gains stem from iterative policy-estimator coupling rather than additional training alone. RLCR (Step-aligned) extends RLCR to the same number of training updates as IPoET, providing a controlled comparison for assessing continued optimization. Additional RLCR training does not yield consistent gains and can degrade out-of-domain performance under Big-Math training. This contrast suggests that the benefit of additional optimization depends on how the policy and confidence estimator are jointly refined, rather than on the number of updates alone. IPoET therefore makes more effective use of the extra training budget for confidence estimation.

4.3 Ablation and Analysis

Effect of Estimator Feedback and Iterative Updates. We first analyze how the two stages of IPoET contribute to its performance. As shown in Figure 3, with the estimator s_{ϕ_1} fixed, updating the policy from π_{θ_1} to π_{θ_2} increases AUROC and reduces Brier score in both in-domain and out-of-domain evaluations. ECE also decreases in both settings, with the in-domain score falling from 0.087 to 0.055. Alongside these gains, in-domain accuracy increases by 1.1 percentage points, while out-of-domain accuracy is preserved. The first-stage results are consistent with the intended role of estimator feedback: rather than treating confidence estimation solely as a prediction task over given responses, IPoET uses the fixed estimator's prediction errors to shape response generation, encouraging outputs that better support reliable correctness assessment. Moreover, computing this reward from estimator predictions rather than verbalized confidence allows IPoET to guide policy optimization with the more accurate confidence estimates observed in our empirical study. With the policy held fixed, the subsequent estimator update further improves confidence quality, especially out of domain. These step-wise gains indicate complementary roles of the two stages: policy optimization

incorporates estimator feedback into generation behavior, while estimator refinement adapts confidence estimation to the updated policy distribution. Together, these results support the iterative interaction between policy and estimator underlying IPoET.

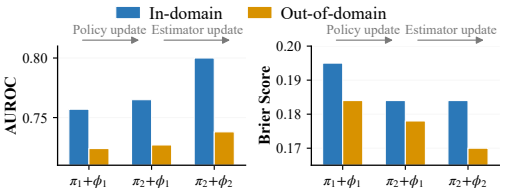

Figure 3: Step-wise effects of IPoET, showing that both policy and estimator updates progressively improve confidence estimation.

Table 3: Effect of alternating-update granularity under Big-Math training with the same update budget and data allocation. Best results are shown in **bold**.

Iteration Setting	In-domain		Out-of-domain	
	AUROC ↑	Brier ↓	AUROC ↑	Brier ↓
2×200	**0.878**	**0.100**	**0.698**	0.206
4×100	0.873	0.102	0.696	**0.198**

Joint vs. Iterative Training. We next examine how estimator feedback should be incorporated into policy optimization by comparing two training designs developed in our work. Table 2 shows that iterative training yields higher AUROC and lower Brier score than joint training across both evaluation settings, indicating better confidence estimation. More importantly, joint optimization degrades answer accuracy relative to the starting policy, whereas iterative training improves task performance, further supporting the design of IPoET. In joint training, the estimator must learn from a changing policy distribution while the policy is guided by an under-adapted estimator, potentially introducing interference between policy learning and estimator fitting. IPoET instead separates these roles across rounds: the estimator adapts to current policy outputs before being used to construct the confidence term in the next policy update, enabling more stable confidence-aware training.

Inference Efficiency and Estimator Overhead. IPoET introduces an additional confidence estimator with the same backbone architecture as the policy. To assess its practical computational cost, we measure total end-to-end inference time on 1,000 HotpotQA examples. IPoET requires 500.38 s, compared with 554.60 s for RLCR, representing a 9.8% reduction. Although the estimator increases the parameter count, it scores each response in a single forward pass, whereas RLCR performs additional autoregressive generation for confidence-related reasoning. Overall, this trade-off allows IPoET to remain inference-efficient in practice despite the additional model component.

Granularity of Alternating Updates. A natural question is whether IPoET benefits from more frequent alternation between policy and estimator updates. To isolate this factor, we compare two schedules with the same update budget and data allocation: 2×200 uses two longer update phases, while 4×100 splits each corresponding phase into two shorter ones. Table 3 shows that more frequent alternation does not consistently improve confidence estimation. The 2×200 schedule performs better in three of the four comparisons, including AUROC in both in-domain and out-of-domain evaluations and in-domain Brier score. These results suggest that sustained optimization within each phase better supports policy–estimator adaptation than switching more often. When each phase becomes too short, the policy may not fully absorb the estimator feedback before the estimator is updated again, yielding smaller gains in confidence estimation.

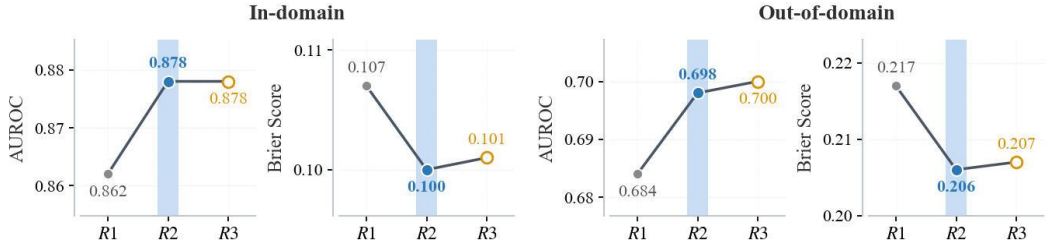

Figure 4: Round-wise comparison of in-domain and out-of-domain AUROC and Brier score under Big-Math training. Blue shading marks the configuration used in our main experiments (Round 2).

Impact of Iteration Rounds. We finally examine whether extending IPoET to additional rounds further improves confidence estimation. The trend in Figure 4 shows that the transition from Round 1 to Round 2 substantially improves both AUROC and Brier score on in-domain and out-of-domain evaluations, confirming the benefit of the first complete policy–estimator refinement cycle. Round 3 brings only a marginal improvement in out-of-domain AUROC, while in-domain AUROC remains unchanged and Brier score slightly increases in both settings. This pattern suggests diminishing returns from further alternation in this setting. We therefore adopt the Round 2 configuration for our main experiments, balancing confidence discrimination and calibration quality while avoiding an additional training round.

5 Related Work

Confidence Estimation from Implicit Information. These methods estimate confidence from implicit information produced during inference or internal computation. One line of work studies model-internal representations for confidence estimation. LLM hidden activations have been shown to encode factuality and truthfulness-related information (Chen et al., 2024), motivating lightweight probes or classifiers trained on internal states (Azaria & Mitchell, 2023; Servedio et al., 2025). Subsequent work extends this view to reasoning settings, including hidden-state verification of reasoning steps and perturbation-based confidence probing (Zhang et al., 2025; Khanmohammadi et al., 2025). Recent work exploits internal representations by aggregating hidden-state information across multiple layers during self-evaluation for uncertainty estimation (Xiao et al., 2026). Another line derives implicit confidence from output probabilities or self-evaluation signals: Answer-Prob uses token likelihoods or answer probabilities as confidence proxies (Jiang et al., 2021; Gupta et al., 2024), while P(True) asks the model to judge whether its own answer is correct and uses the probability of a positive judgment (Kadavath et al., 2022). However, these signals are usually used to score fixed outputs, whereas IPoET converts internal-representation estimates into feedback for policy training.

Verbalized Confidence Estimation. Verbalized methods require the model to express confidence explicitly through natural language. Prompting-based approaches elicit numerical or linguistic confidence alongside generated answers or reasoning (Xiong et al., 2024; Yang et al., 2024). Training-based approaches further teach models to produce uncertainty rationales or calibrated confidence scores, such as self-reflective rationales in SaySelf (Xu et al., 2024), listener-aware confidence markers in LACIE (Stengel-Eskin et al., 2024), and calibration-aware on-policy distillation (Zhang et al., 2026b). Studies of reasoning models show that chain-of-thought generation can improve confidence expression (Yoon et al., 2026), motivating recent RL-based methods that optimize verbalized confidence with calibration-oriented rewards, such as Brier-style or logarithmic scoring objectives (Damani et al., 2025; Bani-Harouni et al., 2025). However, expressed confidence remains sensitive to prompt formulation, answer-dependent self-assessment, and placement of confidence statements relative to the answer (Xia et al., 2025; Seo et al., 2025; Guo et al., 2026; Li et al., 2026a). IPoET avoids relying on such explicit confidence expression by deriving confidence from an internal-representation estimator, while still benefiting from reasoning-oriented generation.

6 Conclusion

This work revisits two major paradigms for LLM confidence estimation: estimator-based and verbalization-based methods. We demonstrate that estimators can substantially outperform verbalized confidence when overfitting is controlled, partly explaining the differing conclusions in prior comparisons. Nevertheless, traditional estimators fail to utilize the strong reasoning capabilities of contemporary LLMs. To combine the strengths of both paradigms, we propose IPoET (Iterative Policy-Estimator Training). This framework enables the estimator to extract richer features from its hidden states using policy-generated reasoning traces, while the policy learns to produce more informative content for the estimator. Compared with the baselines, IPoET shows improved in-domain confidence estimation, comparable or better out-of-domain performance, and competitive task accuracy across multiple reasoning and question-answering datasets. These results highlight the value of internal representations for reliable confidence estimation and suggest that iterative policy-estimator training is a promising direction for building more trustworthy reasoning models.

AI USE STATEMENT

We used generative AI tools to improve the wording, grammar, and fluency of the manuscript and refine the presentation of scientific figures. These tools were not used to implement the proposed method, generate datasets, or prove mathematical claims. All reported numerical results and the data underlying the figures were obtained from experiments conducted by the authors. We reviewed all AI-assisted work. Specifically, we checked textual revisions for technical accuracy and consistency with our intended meaning. The revised figures were also verified against the underlying experimental data. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

We are committed to adhering to the ICLR Code of Ethics. Our study relies solely on publicly available benchmark datasets for mathematical reasoning and question answering, without recruiting human participants or collecting private personal data. We believe this work raises no direct ethical risks beyond standard concerns associated with LLM research and use.

REPRODUCIBILITY STATEMENT

To support reproducibility, we provide our source code at `https://anonymous.4open.science/r/ipoet`. The training procedure and reward function are described in Sections 3.1 and 3.2, respectively. Our experiments use publicly available backbone models and benchmark datasets, with the experimental setup described in Section 4.1. Additional training details and the prompt template are provided in Appendix A, while evaluation procedures and metric definitions are detailed in Appendix B.

REFERENCES

Alon Albalak, Duy Phung, Nathan Lile, Rafael Rafailov, Kanishk Gandhi, Louis Castricato, Anikait Singh, Chase Blagden, Violet Xiang, Dakota Mahan, et al. Big-math: A large-scale, high-quality math dataset for reinforcement learning in language models. *arXiv preprint arXiv:2502.17387*, 2025.

Dang Anh-Hoang, Vu Tran, and Le-Minh Nguyen. Survey and analysis of hallucinations in large language models: attribution to prompting strategies or model behavior. *Frontiers in Artificial Intelligence*, 8:1622292, 2025.

Anthropic. The claude 4 model family: Opus, sonnet, and haiku. `https://www.anthropic.com/research/claude-haiku-4-5`, 2025.

Amos Azaria and Tom Mitchell. The internal state of an llm knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 967–976, 2023.

David Bani-Harouni, Chantal Pellegrini, Paul Stangel, Ege Özsoy, Kamilia Zaripova, Nassir Navab, and Matthias Keicher. Rewarding doubt: A reinforcement learning approach to calibrated confidence expression of large language models. *arXiv preprint arXiv:2503.02623*, 2025.

Mohammad Beigi, Ying Shen, Runing Yang, Zihao Lin, Qifan Wang, Ankith Mohan, Jianfeng He, Ming Jin, Chang-Tien Lu, and Lifu Huang. Internalinspector i2: Robust confidence estimation in llms through internal states. In *Findings of the association for computational linguistics: EMNLP 2024*, pp. 12847–12865, 2024.

Michael Bereket and Jure Leskovec. Uncalibrated reasoning: Grpo induces overconfidence for stochastic outcomes. *arXiv preprint arXiv:2508.11800*, 2025.

Andrew P Bradley. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7):1145–1159, 1997.

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. Inside: Llms' internal states retain the power of hallucination detection. *arXiv preprint arXiv:2402.03744*, 2024.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Mehul Damani, Isha Puri, Stewart Slocum, Idan Shenfeld, Leshem Choshen, Yoon Kim, and Jacob Andreas. Beyond binary rewards: Training lms to reason about their uncertainty. *arXiv preprint arXiv:2507.16806*, 2025.

DeepSeek. Deepseek v4 preview release. `https://api-docs.deepseek.com/news/news260424`, 2026.

W Brier Glenn et al. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3, 1950.

Google DeepMind. Gemini 3 flash. `https://deepmind.google/models/gemini/flash/`, 2025. Model description page.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Junyu Guo, Shangding Gu, Ming Jin, Costas Spanos, and Javad Lavaei. Llms should express uncertainty explicitly. *arXiv preprint arXiv:2604.05306*, 2026.

Neha Gupta, Harikrishna Narasimhan, Wittawat Jitkrittum, Ankit Singh Rawat, Aditya Krishna Menon, and Sanjiv Kumar. Language model cascades: Token-level uncertainty and beyond. *arXiv preprint arXiv:2404.10136*, 2024.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021.

Jingcheng Hu, Yinmin Zhang, Qi Han, Daxin Jiang, Xiangyu Zhang, and Heung-Yeung Shum. Open-reasoner-zero: An open source approach to scaling up reinforcement learning on the base model. *Advances in Neural Information Processing Systems*, 38:162239–162262, 2026.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977, 2021.

Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. Swe-bench: Can language models resolve real-world github issues? In *International Conference on Learning Representations*, volume 2024, pp. 54107–54157, 2024.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611, 2017.

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.

Reza Khanmohammadi, Erfan Miahi, Mehrsa Mardikoraem, Simerjot Kaur, Ivan Brugere, Charese Smiley, Kundan S Thind, and Mohammad M Ghassemi. Calibrating llm confidence by probing perturbed representation stability. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 10459–10525, 2025.

Dharshan Kumaran, Arthur Conmy, Federico Barbero, Simon Osindero, Viorica Patraucean, and Petar Velickovic. How do llms compute verbal confidence. *arXiv preprint arXiv:2603.17839*, 2026a.

Dharshan Kumaran, Stephen M Fleming, Larisa Markeeva, Joe Heyward, Andrea Banino, Mrinal Mathur, Razvan Pascanu, Simon Osindero, Benedetto De Martino, Petar Veličković, et al. Competing biases underlie overconfidence and underconfidence in llms. *Nature Machine Intelligence*, pp. 1–14, 2026b.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, et al. Tulu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*, 2024.

Chen Li, Xiaoling Hu, Songzhu Zheng, Jiawei Zhou, and Chao Chen. Orce: Order-aware alignment of verbalized confidence in large language models. *arXiv preprint arXiv:2605.12446*, 2026a.

Junkai Li, Yunghwei Lai, Weitao Li, Jingyi Ren, Meng Zhang, Xinhui Kang, Siyu Wang, Peng Li, Ya-Qin Zhang, Weizhi Ma, et al. Agent hospital: A simulacrum of hospital with evolvable medical agents. *arXiv preprint arXiv:2405.02957*, 2024.

Yibo Li, Miao Xiong, Jiaying Wu, and Bryan Hooi. Conftuner: Training large language models to express their confidence verbally. *Advances in Neural Information Processing Systems*, 38: 53484–53513, 2026b.

Yinheng Li, Shaofei Wang, Han Ding, and Hang Chen. Large language models in finance: A survey. In *Proceedings of the fourth ACM international conference on AI in finance*, pp. 374–382, 2023.

Zhengzhao Ma, Xueru Wen, Boxi Cao, Yaojie Lu, Hongyu Lin, Jinglin Yang, Min He, Xianpei Han, and Le Sun. Decoupling reasoning and confidence: Resurrecting calibration in reinforcement learning from verifiable rewards. *arXiv preprint arXiv:2603.09117*, 2026.

Matéo Mahaut, Laura Aina, Paula Czarnowska, Momchil Hardalov, Thomas Müller, and Lluís Màrquez. Factual confidence of llms: on reliability and robustness of current estimators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4554–4570, 2024.

Sadhika Malladi, Tianyu Gao, Eshaan Nichani, Alex Damian, Jason D Lee, Danqi Chen, and Sanjeev Arora. Fine-tuning language models with just forward passes. *Advances in Neural Information Processing Systems*, 36:53038–53075, 2023.

Shiyu Ni, Keping Bi, Jiafeng Guo, Minghao Tang, Jingtong Wu, Zengxin Han, and Xueqi Cheng. Annotation-efficient universal honesty alignment. *arXiv preprint arXiv:2510.17509*, 2025.

OpenAI. Gpt-5 system card. `https://openai.com/index/gpt-5-system-card`, 2025.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. Gpqa: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv:2311.12022*, 2023.

Ki Jung Seo, Sehun Lim, and Taeuk Kim. Advice: Answer-dependent verbalized confidence estimation. *arXiv preprint arXiv:2510.10913*, 2025.

Giovanni Servedio, Alessandro De Bellis, Dario Di Palma, Vito Walter Anelli, and Tommaso Di Noia. Are the hidden states hiding something? testing the limits of factuality-encoding capabilities in llms. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6089–6104, 2025.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, pp. 1279–1297, 2025.

Marco Siino, Mariana Falco, Daniele Croce, and Paolo Rosso. Exploring llms applications in law: A literature review on current legal nlp approaches. *IEEE Access*, 13:18253–18276, 2025.

Elias Stengel-Eskin, Peter Hase, and Mohit Bansal. Lacie: Listener-aware finetuning for calibration in large language models. *Advances in Neural Information Processing Systems*, 37:43080–43106, 2024.

Nishant Subramani, Jason Eisner, Justin Svegliato, Benjamin Van Durme, Yu Su, and Sam Thomson. Mice for cats: Model-internal confidence estimation for calibrating agents with tools. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 12362–12375, 2025.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4149–4158, 2019.

Shuchang Tao, Liuyi Yao, Hanxing Ding, Yuexiang Xie, Qi Cao, Fei Sun, Jinyang Gao, Huawei Shen, and Bolin Ding. When to trust llms: Aligning confidence with response quality. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 5984–5996, 2024.

Benjamin Turtel, Danny Franklin, Kris Skotheim, Luke Hewitt, and Philipp Schoenegger. Outcome-based reinforcement learning to predict the future. *arXiv preprint arXiv:2505.17989*, 2025.

Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. Measuring short-form factuality in large language models. *arXiv preprint arXiv:2411.04368*, 2024.

Yuxi Xia, Pedro Henrique Luz De Araujo, Klim Zaporojets, and Benjamin Roth. Influences on llm calibration: A study of response agreement, loss functions, and prompt styles. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3740–3761, 2025.

Ruixuan Xiao, Wentao Ma, Ke Wang, Yuchuan Wu, Junbo Zhao, Haobo Wang, Fei Huang, and Yongbin Li. Flowbench: Revisiting and benchmarking workflow-guided planning for llm-based agents. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 10883–10900, 2024.

Zeguan Xiao, Diyang Dou, Boya Xiong, Yun Chen, and Guanhua Chen. Enhancing uncertainty estimation in llms with expectation of aggregated internal belief. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 34043–34051, 2026.

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. In *International Conference on Learning Representations*, volume 2024, pp. 23650–23678, 2024.

Tianyang Xu, Shujin Wu, Shizhe Diao, Xiaoze Liu, Xingyao Wang, Yangyi Chen, and Jing Gao. Sayself: Teaching llms to express confidence with self-reflective rationales. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 5985–5998, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Daniel Yang, Yao-Hung Hubert Tsai, and Makoto Yamada. On verbalized confidence scores for llms. *arXiv preprint arXiv:2412.14737*, 2024.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 2369–2380, 2018.

Dongkeun Yoon, Seungone Kim, Sohee Yang, Sunkyoung Kim, Soyeon Kim, Yongil Kim, Eunbi Choi, Yireun Kim, and Minjoon Seo. Reasoning models better express their confidence. *Advances in Neural Information Processing Systems*, 38:103869–103896, 2026.

Anqi Zhang, Yulin Chen, Jane Pan, Chen Zhao, Aurojit Panda, Jinyang Li, and He He. Reasoning models know when they're right: Probing hidden states for self-verification. *arXiv preprint arXiv:2504.05419*, 2025.

Chuang Zhang, Zizhen Zhu, Yihao Wei, Bing Tian, Junyi Liu, Henan Wang, Wang Xavier, and Yaxiao Liu. Confidence-calibrated small-large language model collaboration for cost-efficient reasoning. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4480–4501, 2026a.

Jiaxin Zhang, Xiangyu Peng, Qinglin Chen, Qinyuan Ye, Caiming Xiong, and Chien-Sheng Wu. The illusion of certainty: Decoupling capability and calibration in on-policy distillation. *arXiv preprint arXiv:2604.16830*, 2026b.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

A Training Details

All experiments are conducted on a server equipped with 8 NVIDIA A100 GPUs, each with 80GB of memory. For both HotpotQA and Big-Math, the overall training time is approximately 15 hours.

For policy training, we use the same hyperparameters in both rounds. Specifically, we use GRPO with 8 sampled responses per prompt and a sampling temperature of 0.7. The policy learning rate is set to 1e-6 for both HotpotQA and Big-Math. In line with Damani et al. (2025), we use a training batch size of 64 for HotpotQA and 72 for Big-Math. The maximum prompt and response lengths are set to 3072 and 1536 for HotpotQA, and 1024 and 4096 for Big-Math, respectively. We do not normalize GRPO advantages by their group standard deviation, following prior practice (Bereket & Leskovec, 2025; Turtel et al., 2025), thereby preserving reward-scale information for examples with large calibration errors.

For estimator training, we use a per-device batch size of 1 with gradient accumulation over 8 steps. We reduce the second-round learning rate to support stable refinement and avoid overfitting or excessive drift from the first-round estimator. Specifically, the learning rate decreases from 5×10^{-6} to 1×10^{-6} across the two HotpotQA rounds and from 5×10^{-6} to 5×10^{-7} across the two Big-Math rounds.

The prompt template used for policy training is shown in Figure 5.

> A conversation between User and Assistant. The user asks a question, and the Assistant solves it. The assistant first thinks about the reasoning process in the mind and then provides the user with the answer. The reasoning process and answer are enclosed within `<reason> </reason>` and `<answer> </answer>` tags, respectively, i.e., `<reason>` reasoning process here `</reason><answer>` answer here `</answer>`.

Figure 5: Policy training prompt template.

B Details of Evaluation Metrics

We follow the notation defined in the main text.

Accuracy. We compute accuracy using dataset-specific correctness criteria. For HotpotQA and HotpotQA-Modified, we use exact match. For GSM8K, MATH-500, and Big-Math, we use the *math_verify* library. For CommonsenseQA, GPQA, SimpleQA, and TriviaQA, we use an LLM-as-a-judge (Zheng et al., 2023) evaluation. To keep the judge model family separate from the evaluated backbone, we use Llama-3.1-8B-Instruct for Qwen-based methods and Qwen3-8B for Llama-based methods. Following Damani et al. (2025), we run the judge with temperature set to 0 and provide it with the question, the ground-truth answer, and the extracted answer. The judge is instructed to output only "YES" or "NO" according to whether the answer is correct. Since these datasets contain short and objective answers, we do not condition the judge on thinking traces, avoiding potential bias from generated rationales.

AUROC. We compute AUROC by sweeping the confidence threshold and integrating the resulting ROC curve:

$$\mathrm{AUROC} = \int_0^1 \mathrm{TPR}\left(\mathrm{FPR}^{-1}(x)\right)\, dx,$$

where TPR and FPR denote the true positive rate and false positive rate, respectively.

Brier score. Given confidence scores c_i and correctness labels z_i, we compute

$$\text{Brier score} = \frac{1}{N}\sum_{i=1}^{N}(c_i - z_i)^2.$$

ECE. We partition predictions into $M = 10$ confidence bins $\{B_m\}_{m=1}^{M}$. For each bin, we compute the average correctness and average confidence:

$$\bar{z}_m = \frac{1}{|B_m|} \sum_{i \in B_m} z_i.$$

$$\bar{c}_m = \frac{1}{|B_m|} \sum_{i \in B_m} c_i.$$

The ECE is then computed as

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{N} \left| \bar{z}_m - \bar{c}_m \right|.$$

Table 4: Comparison between IPoET and frontier commercial models. IPoET is trained on HotpotQA for the HotpotQA results and on Big-Math for the math average over MATH-500, GSM8K, and Big-Math. Best and second-best results are marked in **bold** and underlined, respectively.

(a) HotpotQA

Model	AUROC↑	Brier↓	ECE↓
GPT-5 mini	0.603	0.306	0.298
Claude Haiku 4.5	<u>0.671</u>	<u>0.279</u>	<u>0.266</u>
DeepSeek-V4-Flash	0.537	0.332	0.328
Gemini 3 Flash	0.509	0.357	0.357
IPoET (ours)	**0.800**	**0.184**	**0.107**

(b) Math average

Model	AUROC↑	Brier↓	ECE↓
GPT-5 mini	0.638	0.189	0.181
Claude Haiku 4.5	0.584	0.309	0.300
DeepSeek-V4-Flash	<u>0.686</u>	<u>0.137</u>	<u>0.136</u>
Gemini 3 Flash	0.557	0.150	0.150
IPoET (ours)	**0.878**	**0.100**	**0.040**

C Commercial Model Comparison

We further compare IPoET with commercial models in Table 4. We evaluate GPT-5 mini (OpenAI, 2025), Claude Haiku 4.5 (Anthropic, 2025), DeepSeek-V4-Flash (DeepSeek, 2026), and Gemini 3 Flash (Google DeepMind, 2025).[1] For the HotpotQA results, IPoET is trained on HotpotQA; for the math results, IPoET is trained on Big-Math and evaluated using the Math average over MATH-500, GSM8K, and Big-Math. Across both settings, IPoET obtains the best results on AUROC, Brier score, and ECE among the evaluated models. Commercial models demonstrate reasonable ability to estimate confidence, but their verbalized confidence remains less reliable in these settings. These results indicate that strong general-purpose models can still struggle to align expressed confidence with correctness on reasoning tasks, highlighting the need for task-oriented training objectives that explicitly target confidence reliability.

[1]Model versions used: `gpt-5-mini-2025-08-07`, `claude-haiku-4-5-20251001`, `deepseek-v4-flash`, and `gemini-3-flash-preview-nothinking`.