

CONCORD: Label-Free Calibration of Verbalized LLM Confidence via Rollout Consistency

Maximilian Schall¹ Gerard de Melo¹ Padhraic Smyth²

¹Hasso Plattner Institute, University of Potsdam, Germany

²University of California, Irvine, USA

{maximilian.schall, gerard.demelo}@hpi.de
smyth@uci.edu

Abstract

Calibrated confidence, where a model’s stated probability matches its actual expected accuracy, is often essential in contexts where confidence scores are relied upon by humans to make decisions or serve as inputs to downstream agents. Existing approaches for training models to verbalize calibrated confidence typically require ground-truth correctness labels, which are costly to collect and unavailable for many prompt distributions. Multi-pass consistency methods achieve good calibration but significantly increase inference costs. We introduce CONCORD, a label-free method for training language models to improve the calibration of verbalized confidence scores without adding a dedicated calibration stage. Adapting models with Group Relative Policy Optimization already yields multiple rollouts per prompt; we repurpose these rollouts as a calibration signal by deriving confidence targets from semantic agreement among sampled answers, using either Natural Language Inference-based answer clustering or semantic entropy. The resulting reward encourages the model to align its stated confidence with empirical semantic agreement, without requiring ground-truth labels. During inference, the model generates a reasoning trace, an answer, and a better calibrated confidence score in a single pass. Across six benchmarks and four models, CONCORD substantially reduces expected calibration error over unadapted verbalized confidence and over label-dependent adaptation baselines, while maintaining competitive AUROC.

1 Introduction

When a language model states that it is 80% confident in an answer, users and downstream systems naturally interpret that number as a probability. If answers that assign 80% confidence are correct only 40% of the time, the model is not merely uncertain; it is misleading. Calibration, the alignment between stated confidence and empirical accuracy, is therefore essential whenever model outputs inform trust-sensitive decisions, from weather forecasting (DeGroot & Fienberg, 1983; Gneiting et al., 2007) and clinical support systems (Kompa et al., 2021) all the way to autonomous agent pipelines that must decide whether to act, defer, or seek additional information (Han et al., 2024).

Despite this, large language models are often poorly calibrated out of the box (Jiang et al., 2021; Kadavath et al., 2022). Verbalized confidence is cheap and black-box accessible, but it tends to be systematically overconfident (Xiong et al., 2024; Yang et al., 2026; Tian et al., 2023). Logit-based scores can often distinguish correct from incorrect answers reasonably well, yet still remain miscalibrated in absolute terms, assigning high confidence to predictions that are far less reliable in practice (Guo et al., 2017; Xu et al., 2025; Huang et al., 2025b; Cohen et al., 2024). Multi-sample consistency methods can improve calibration substantially (see Figure 1) by estimating agreement across multiple generations (Kuhn et al., 2023; Fadeeva et al., 2023; Bouchard & Chauhan, 2025), but they require repeated decoding at inference time. Their cost scales with the number of rollouts and the length of each generated response,

which makes them particularly unattractive for chain-of-thought reasoning and for agentic systems that must make many confidence-aware decisions per query.

This leaves a clear gap. We would like a confidence signal that is both well calibrated and available from a single forward pass. A natural approach is to train the model to produce calibrated confidence directly. However, existing adaptation methods typically require ground-truth correctness labels during training, whether through supervised fine-tuning (Li et al., 2025; Kapoor et al., 2024), reward design, or preference construction (Bani-Harouni et al., 2026; Xu et al., 2024; Zhang et al., 2024). In practice, this turns calibration into a data problem: labels are expensive to collect, unavailable for many prompt distributions, and susceptible to obsolescence under distribution shift. A method that must be recalibrated with newly labeled data each time the deployment distribution changes is difficult to maintain in production.

We propose CONCORD (CONSistency-Calibrated Optimization via Rollout Distributions), a training-time method with two properties that distinguish it from prior calibration approaches: it is *label-free*, and, when incorporated into an existing post-training pipeline, it adds *no dedicated calibration stage*. CONCORD uses Group Relative Policy Optimization (GRPO Shao et al., 2024), which already samples multiple rollouts per prompt for its own optimization, and reuses exactly those rollouts to compute the calibration signal: a Natural Language Inference (NLI) model groups the sampled answers into semantically equivalent clusters, and each response’s cluster frequency becomes the target for its verbalized confidence. Because the signal comes entirely from the model’s own outputs, no reference answers are required, and the marginal cost over standard GRPO post-training is the reward computation alone. Where inference-heavy consistency methods pay for repeated sampling at test time and supervised calibration-tuning pays in correctness labels, CONCORD pays neither whenever a multi-rollout RL post-training is already in the pipeline.

Our contributions are:

- We introduce CONCORD, a *label-free* reinforcement learning method that trains models to produce calibrated verbalized confidence from semantic consistency alone, using the rollouts that GRPO-based training already generates and thus requiring no dedicated calibration stage.
- We develop two consistency-derived reward variants, based on semantic cluster frequency and semantic entropy, with complementary calibration–discrimination trade-offs.
- Across four models from the Qwen-2.5 and Llama-3 families and benchmarks spanning mathematical reasoning, multiple-choice question answering, and multi-hop question answering, CONCORD reduces expected calibration error relative to unadapted verbalized confidence and relative to prior adaptation methods that rely on labeled data, while maintaining competitive discrimination.
- We show that self-consistency, established as an inference-time uncertainty estimate (Kuhn et al., 2023; Fadeeva et al., 2023; Bouchard & Chauhan, 2025), also serves as a training signal that reduces the dependence on labeled calibration data and transfers to the unseen benchmarks.

2 Related Work

2.1 Single-Pass and Multi-Pass Confidence Estimation

Existing methods for uncertainty estimation in language models differ mainly in two respects: what form of access they require to the model, and how much inference-time computation they incur. One line of work uses *white-box* signals derived from internal model states, such as token-level log-probabilities, length-normalized sequence scores, or ensemble disagreement (Fadeeva et al., 2023; Shorinwa et al., 2025). Related white-box probes such as $P(\text{True})$ (Kadavath et al., 2022) obtain an additional uncertainty signal by asking the model to evaluate its own answer in a second pass and considering the associated

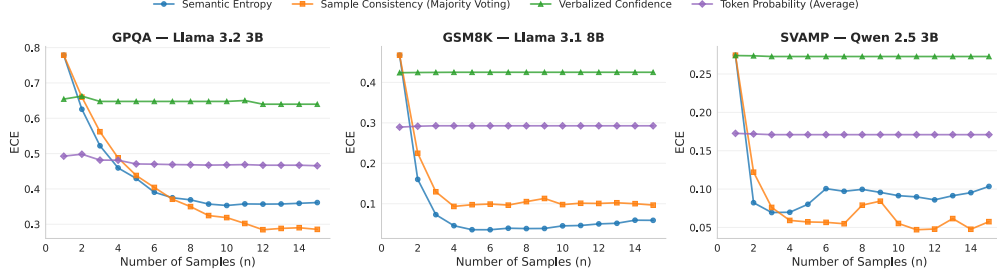


Figure 1: Expected calibration error (ECE, lower is better) as a function of the number of rollouts for three model–task pairs. Multi-sample consistency methods improve calibration as the number of rollouts increases, but they incur substantial inference-time cost; single-pass confidence methods are less calibrated.

softmax distribution, but this increases inference cost and remains susceptible to over-trust in incorrect model-generated answers (Stengel-Eskin & Van Durme, 2023; Li et al., 2024). These methods often provide useful discrimination between correct and incorrect predictions, but they are frequently miscalibrated in absolute terms (Guo et al., 2017). A model may therefore rank answers sensibly by confidence while still assigning probabilities that do not match empirical accuracy.

A second line of work focuses on black-box single-pass confidence estimation. The most direct form is verbalized confidence, in which the model states a numerical confidence score alongside its answer. This approach is attractive because it is cheap, model-agnostic, and easy to use, but empirical studies consistently find that verbalized confidence is systematically overconfident (Xiong et al., 2024; Yang et al., 2026; Tian et al., 2023). Prompting-based corrections, including steering and linear recalibration, can reduce this bias but do not eliminate it (Zhou et al., 2025; Ji et al., 2025). Recent mechanistic evidence further suggests that verbal confidence is not merely a post-hoc formatting artifact: LLMs appear to compute and cache confidence-related representations during answer generation, and these representations contain information beyond token log-probabilities (Kumaran et al., 2026).

A third line of work uses *multi-pass consistency* as the uncertainty signal. Here, confidence is estimated from agreement across multiple sampled generations, using exact match, semantic clustering, or semantic entropy (Kuhn et al., 2023; Fadeeva et al., 2023; Bouchard & Chauhan, 2025). These methods are often substantially better calibrated than single-pass alternatives, but they incur the highest inference-time cost because they require repeated decoding for each query. This overhead becomes especially burdensome for long chain-of-thought outputs and for agentic settings, in which confidence must repeatedly be estimated. Recent surveys provide a coherent overview of these approaches and emphasize the resulting trade-offs between access requirements, calibration quality, and computational cost (Xia et al., 2025; Shorinwa et al., 2025).

Our work targets the gap left by previous approaches: we use multi-sample consistency as a *training-time* supervision signal so that a model can output calibrated confidence in a single pass at inference time, without requiring white-box access.

2.2 Calibration-Oriented Fine-Tuning

A complementary line of work seeks to improve calibration not by extracting a better uncertainty signal from a fixed model, but by adapting the model so that its outputs are better calibrated to begin with. Existing methods typically rely on ground-truth correctness labels during training, either through supervised fine-tuning, reinforcement learning, or preference-based objectives. Several recent approaches directly optimize confidence quality with label-based rewards. RLCR (Damani et al., 2026), REWARDING DOUBT (Bani-Harouni et al., 2026), and LOVEC (Zhang et al., 2026) all adapt the model during training to improve verbalized confidence behavior, while relying on externally specified supervision signals. CONFTUNER (Li et al., 2025) fine-tunes with a tokenized Brier score, a proper scoring rule that still uses answer correctness as its training target. NOVA (Liu et al., 2026) performs

rule-guided supervised fine-tuning for noise-aware confidence in retrieval-augmented settings. Related work also includes SAYSELF (Xu et al., 2024), which combines supervised fine-tuning with reinforcement learning to elicit confidence judgments and self-reflective rationales. Most closely related to our setting, SELF-CALIBRATION (Huang et al., 2025a) also produces single-pass confidence from self-consistency without ground-truth labels. It differs in how that signal is turned into a model: it constructs a synthetic calibration dataset offline and distills a fixed soft self-consistency target through a separate supervised fine-tuning stage. Although these methods generally differ in their optimization details, they share a common dependency on externally defined supervision signals, such as correctness labels or teacher confidence targets. CONCORD instead scores the rollouts that GRPO-based training already generates for policy optimization, requiring no dedicated calibration stage and no synthetic data.

Finally, recent work extends calibration beyond short-form answers to *long-form generation*. These approaches use supervised fine-tuning and reinforcement learning to elicit verbalized confidence expressions or sentence-level confidence tags in extended outputs (Band et al., 2024; Zhang et al., 2026). This broadens the scope of calibration, but does not remove the need for labeled supervision during training.

3 Approach

We adapt language models to produce better calibrated verbalized confidence scores using reinforcement learning, without access to ground-truth answers at any stage of training. The key observation is that multi-sample RL algorithms such as GRPO already generate multiple rollouts per prompt as part of their optimization procedure. We repurpose these rollouts to compute a semantic consistency signal that provides a calibration target, turning an existing computational cost into a supervisory signal at negligible additional expense.

This approach rests on the empirical correlation between self-consistency and accuracy, which is well-established across models and tasks (Kuhn et al., 2023; Fadeeva et al., 2023). On questions a model can reliably answer, independent rollouts tend to converge on the same response. On harder questions, rollouts disagree, producing low consistency scores that naturally translate to lower confidence. A potential failure mode is the consistently wrong model: if all N rollouts produce the same incorrect answer, consistency is 1 and the training signal encourages 100% confidence. Two factors mitigate this. First, the KL penalty anchors the policy near the pretrained model, which exhibits diversity on questions it cannot reliably answer; the policy must earn high consistency through genuine competence rather than mode collapse. Second, calibration is a statistical property across many prompts, and the consistent-error case is empirically rare enough that the aggregate signal remains informative.

3.1 Problem Formulation

Given a question q , we adapt a language model π_θ to generate a structured response containing: (1) chain-of-thought reasoning r , (2) a final answer a , and (3) a verbalized confidence score $c \in [0, 100]$ representing the model’s estimate of the probability (expressed as a percentage) that a is correct. A model is *calibrated* if, among all responses where it states confidence c , the fraction of correct answers equals $c/100$.

3.2 Training Framework

We use Group Relative Policy Optimization (GRPO; Shao et al., 2024) as the policy gradient algorithm. For each prompt q_i in a training batch, GRPO samples a group of N completions $\{(r_j, a_j, c_j)\}_{j=1}^N$ from the current policy π_θ , scores them with a reward function R , and updates the policy using group-normalized advantages.

Any RL algorithm that samples multiple completions per prompt (e.g., REINFORCE with leave-one-out baselines) could serve as the backbone; we choose GRPO for its simplicity and stable training dynamics.

3.3 Reward Design

3.3.1 NLI-Based Calibration Reward

The calibration reward encourages the verbalized confidence c to match the empirical agreement rate among the N sampled answers for the same prompt. Since semantically equivalent answers may differ on the surface (e.g., “0.5” vs. “1/2”), we estimate agreement in terms of their semantics using Natural Language Inference.

Given the N answers $\{a_1, \dots, a_N\}$, we construct a pairwise equivalence matrix $E \in \{0, 1\}^{N \times N}$, where $E_{jk} = 1$ if and only if the NLI model judges a_j and a_k as entailing in at least one direction. Applying transitive closure to E yields a partition into K semantic clusters $\{C_1, \dots, C_K\}$. For each response $j = 1, \dots, N$, where j belongs to cluster $C_{k(j)}$, the *semantic consistency score* is:

$$p_{\text{sem}}(j) = \frac{|C_{k(j)}|}{N} \quad (1)$$

This score approximates the probability that a random rollout would agree with response j . If the model produces the same answer in 8 out of 10 rollouts, a well-calibrated confidence for that answer should be near 80%.

The calibration reward scores how closely the verbalized confidence c_j matches this target using a Gaussian kernel with bandwidth σ_t :

$$R_{\text{nli}}(c_j, j) = \exp\left(-\frac{(c_j/100 - p_{\text{sem}}(j))^2}{2\sigma_t^2}\right) \quad (2)$$

3.3.2 Semantic Entropy Variant

We also implement an alternative calibration target based on semantic entropy (Kuhn et al., 2023). Given the same cluster partition $\{C_1, \dots, C_K\}$, we compute cluster probabilities $p_k = |C_k|/N$ and the normalized semantic entropy:

$$\tilde{S}_{\text{sem}} = \frac{-\sum_{k=1}^K p_k \log p_k}{\log N} \quad (3)$$

The entropy-based confidence target is then $p_{\text{se}} = 1 - \tilde{S}_{\text{sem}}$, mapping low response diversity to high confidence and vice versa. The reward R_{se} follows the same Gaussian kernel form as Equation 2, substituting p_{se} for $p_{\text{sem}}(j)$.

3.3.3 Bandwidth Annealing

Both R_{nli} and R_{se} use a Gaussian kernel whose bandwidth $\sigma_t, t = 0, \dots, T$ is linearly annealed from σ_0 to σ_T over training. Early in training, the model’s confidence outputs are poorly initialized and often far from the consistency target. A wide bandwidth provides a smooth gradient signal that rewards even coarse directional improvements. As training progresses, the narrowing bandwidth sharpens the reward peak around the exact target, enforcing precise calibration. We show the effects of annealing in Section 6.2.

4 Experimental Setup

We evaluate whether calibration learned from self-consistency transfers to held-out downstream benchmarks when adaptation is performed without ground-truth correctness labels. Our experiments focus on math reasoning, using GSM8K (Cobbe et al., 2021), SVAMP (Patel et al., 2021), and MATH-500 (Lightman et al., 2024) as evaluation benchmarks. Broader generalization across additional task families is reported separately in Section 5.2. Across all adapted methods, we keep the model, training data, optimizer, and update budget fixed, varying only the reward signal.

4.1 Models and Training Data

We experiment with four instruction-tuned models from two families at two parameter scales: Qwen-2.5-3B and Qwen-2.5-7B (Yang et al., 2025), and Llama-3.2-3B and Llama-3.1-8B-Instruct (Grattafiori et al., 2024). This setup lets us test whether calibration improvements are consistent across both architecture families and model sizes in the 3B–8B range.

For adaptation, we use prompts from OpenDataArena-ODA-Math-460k (Gao et al., 2026; Cai et al., 2025). We discard all reference answers and retain only the prompt text, so CONCORD is trained fully label-free. We subsample 1,600 prompts and train for a single epoch over this subset. None of the evaluation datasets are used during training or validation.

4.2 Training Configuration

All adapted models are trained with GRPO, with parameter-efficient fine-tuning through LoRA (Hu et al., 2022). We apply LoRA adapters to the query and value projection layers (q_proj, v_proj) with rank $r = 8$, scaling factor $\alpha = 32$, and dropout rate 0.1.

For each prompt, we sample $N = 10$ rollouts at temperature $T = 1.0$. We evaluate both CONCORD reward variants introduced in Section 3: the response-specific NLI-based consistency target (CONCORD-nli) and the prompt-level semantic entropy target (CONCORD-se). We use *DeBERTa-V3-Large* (He et al., 2023) as our NLI model. For both variants, the Gaussian bandwidth is annealed linearly from $\sigma_0 = 0.3$ to $\sigma_T = 0.08$ over training. We set the KL regularization coefficient to $\beta = 0.1$, use a constant learning rate of 3×10^{-6} , and train with per-device batch size 1 and without gradient accumulation. Prompts and completions are each truncated to 2,048 tokens. All runs use bfloat16 precision on a single GPU.

To ensure a fair comparison, all adapted baselines are trained with the identical setup: the same base model, same LoRA configuration, same prompt subset, and same number of training steps. The only difference across methods is the reward function.

4.3 Baselines

We compare against both unadapted single-pass confidence baselines and labeled adaptation baselines. Implementation details can be found in Appendix D.

Single-pass baselines. We evaluate four confidence signals that require no repeated sampling of the answer: (i) **Verbalized confidence**, where the unadapted base model outputs a numerical confidence score in text using the same answer-plus-confidence prompt applied to all methods; (ii) **Token Mean Probability**; (iii) **Length-Normalized sequence probability**; and (iv) **P(True)** (Kadavath et al., 2022), which reuses the single sampled answer but requires one additional scoring pass to read off the probability of the token True. The verbalized baseline is the only one of the four that is black-box and text-only at inference.

Multi-pass consistency baselines. As an inference-time reference we also report test-time consistency (TT), which applies the same NLI-based cluster-frequency (TT-NLI) and semantic-entropy (TT-SE) targets that CONCORD uses for training, but computes them at *inference* from $N=10$ freshly sampled rollouts per question rather than from a single pass. TT requires repeated decoding; it serves as the multi-pass reference point that CONCORD approximates with a single forward pass.

Adapted baselines. We further compare CONCORD against recent adaptation methods. Three use label-based rewards with reinforcement learning: RLCR (Damani et al., 2026), REWARDING DOUBT (Bani-Harouni et al., 2026), and LOVEC (Zhang et al., 2026); for these we implemented the published reward function and use our training pipeline. We additionally compare against two supervised-fine-tuning calibration methods, CONFUTER (Li et al., 2025), which trains against answer correctness, and SELF-CALIBRATION (Huang et al., 2025a), which like CONCORD uses no ground-truth labels but requires a separate calibration stage and synthetic data. Both are trained with their own recipes on the same prompt subset.

4.4 Evaluation Tasks

Our main results focus on three mathematical reasoning benchmarks: GSM8K (Cobbe et al., 2021), SVAMP (Patel et al., 2021), MATH-500 (Lightman et al., 2024). These datasets are distinct from the OpenDataArena prompts used for adaptation, but remain within the same broad domain. This setting tests whether calibration learned from self-consistency transfers across datasets rather than merely overfitting to the prompt distribution seen during training.

We report additional results on a broader set of tasks spanning multiple-choice question answering, and multi-hop question answering in Section 5.2. Further details are given in Appendix C.

4.5 Metrics and Inference Protocol

We evaluate all methods using the Expected Calibration Error (ECE), which measures the absolute mismatch between stated confidence and empirical accuracy, averaged over confidence bins; lower is better. We use $M=10$ equal-width bins. In Appendix A, we further show discrimination in terms of the Area Under the ROC Curve (AUROC), which measures how well confidence scores rank correct answers above incorrect ones; higher is better. We treat ECE as the primary metric because the main objective is numerical trustworthiness of the reported confidence, not just ranking quality. We also report task accuracy in Appendix E, which is largely stable across methods.

At inference time, all adapted models generate a single response using greedy decoding. The prompt and evaluation protocol are identical across every method. We prompt chain-of-thought reasoning, a final answer enclosed in <answer> tags, and a confidence score enclosed in <confidence> tags. For the baselines that do not verbalize confidence directly (Token Mean Probability and Length-Normalized probability), we extract the corresponding probabilities from the same response; P(True) reuses that response but requires one further scoring pass. The multi-pass consistency baselines (TT-NLI, TT-SE) draw $N=10$ samples at temperature $T=1.0$ per question. We report the prompt template in Appendix B.

5 Results

In this section, we discuss the main results of our extensive evaluation considering both in-domain (Section 5.1) and out-of-domain settings (Section 5.2). Additional results are presented in the Appendix.

5.1 In-Domain Transfer

		GSM8K ECE ↓				SVAMP ECE ↓				MATH-500 ECE ↓			
Method		Q3B	Q7B	L3B	L8B	Q3B	Q7B	L3B	L8B	Q3B	Q7B	L3B	L8B
Logits	Token Mean Prob.	0.133	0.073	0.150	0.130	0.119	0.080	0.119	0.115	0.396	0.274	0.486	0.355
	Length-Normalized	0.137	0.073	0.152	0.132	0.119	0.080	0.116	0.112	0.392	0.271	0.480	0.343
	P(True)	0.151	0.105	0.186	0.242	0.125	0.139	0.225	0.205	0.307	0.249	0.222	0.157
τ	Verbalized	0.141	0.062	0.155	0.096	0.130	0.073	0.107	0.094	0.395	0.277	0.533	0.431
TT	TT-NLI ($N=10$)	0.097*	0.028*	0.128*	0.072	0.060*	0.043	0.116	0.042*	0.067	0.068	0.058	0.119
	TT-SE ($N=10$)	0.127	0.035	0.163	0.091	0.085	0.037*	0.138	0.058	0.049*	0.040*	0.030*	0.104*
Ours	CONCORD-nli	0.118	0.091	0.142	0.129	0.112	0.094	0.382	0.111	0.204	0.295	0.450	0.392
	CONCORD-se	0.126	0.059	0.135	0.066*	0.113	0.062	0.103*	0.089	0.141	0.284	0.435	0.175
Labels	RLCR	0.668	0.071	0.743	0.256	0.659	0.096	0.825	0.213	0.447	0.140	0.466	0.274
	Rewarding Doubt	0.515	0.454	0.593	0.374	0.564	0.424	0.712	0.245	0.422	0.372	0.453	0.282
	LoVeC	0.547	0.062	0.526	0.326	0.610	0.099	0.559	0.235	0.379	0.153	0.461	0.234
SFT	ConfTuner	0.178	0.120	0.622	0.601	0.142	0.120	0.607	0.609	0.478	0.398	0.457	0.319
	Self-Calibration	0.146	0.085	0.210	0.125	0.123	0.095	0.123	0.095	0.406	0.282	0.545	0.395

Table 1: In-domain ECE (GSM8K, SVAMP, MATH-500), lower is better. Bold indicates the best single-pass black-box/text-only method in each column; * indicates best overall in each column. Q3B=Qwen-2.5-3B, Q7B=Qwen-2.5-7B, L3B=Llama-3.2-3B, L8B=Llama-3.1-8B.

[Table 1](#) reports ECE on GSM8K, SVAMP, and MATH-500. Among single-pass text-only methods, CONCORD attains the lowest ECE for 11 of the 12 model–dataset combinations, with CONCORD-se accounting for nine of them. The extent of the gain depends on how well calibrated the base model already is. On GSM8K and SVAMP, the unadapted verbalized baseline is comparatively strong (0.062 to 0.155) and CONCORD-se yields consistent but modest improvements, for example from 0.096 to 0.066 on Llama-3.1-8B for GSM8K. On the harder MATH-500, the baseline degrades considerably and the improvements are correspondingly larger, from 0.395 to 0.141 on Qwen-2.5-3B and from 0.431 to 0.175 on Llama-3.1-8B. MATH-500 with Qwen-2.5-7B is the only in-domain combination for which CONCORD does not improve on the verbalized baseline (0.277 to 0.284).

The white-box and multi-pass references behave as expected. Logit-based scores are competitive on the easier benchmarks but degrade on MATH-500, where length-normalized probability ranges from 0.271 to 0.480 and CONCORD-se stays below it on three of the four models; this is the calibration–discrimination gap discussed in [Appendix A](#). The TT baselines, which apply the same consistency signal CONCORD trains on but recompute it at inference from $N=10$ fresh samples, give the lowest ECE overall for 10 of the 12 combinations and act as the reference our single-pass method approximates. The remaining gap is often small, and on GSM8K with Llama-3.1-8B (0.066 against 0.072) and SVAMP with Llama-3.2-3B (0.103 against 0.116), CONCORD-se is better than TT from a single forward pass.

The adaptation baselines separate sharply by supervision type. The label-based RL methods degrade substantially on several models: RLCR reaches 0.743 on Llama-3.2-3B for GSM8K and 0.825 on the same model for SVAMP, against verbalized baselines of 0.155 and 0.107, with Rewarding Doubt and LoVeC following the same pattern. The supervised calibration baselines are more competitive: ConfTuner is reasonable on the Qwen models but collapses on both Llama models (0.622 and 0.601 on GSM8K), while Self-Calibration is the closest competitor overall, trailing CONCORD-se by small margins on GSM8K and SVAMP but falling well behind on MATH-500 for three of the four models. These results do not imply that label-based calibration is generally ineffective, but they do suggest that consistency-based supervision transfers more reliably across models.

The two variants differ mainly in stability. CONCORD-se is the stronger and more reliable of the two, giving the lowest text-only ECE in nine of the 12 cases and the higher AUROC in nine as well ([Appendix A](#)). CONCORD-nli is best on Qwen-2.5-3B for both GSM8K (0.118) and SVAMP (0.112), but is less stable on the Llama models, reaching 0.382 on SVAMP with Llama-3.2-3B and 0.392 on MATH-500 with Llama-3.1-8B. We therefore treat CONCORD-se as the in-domain default; the case for CONCORD-nli rests on its stronger ranking quality under domain shift ([Section 5.2](#), [Appendix A](#)). Qualitative examples are given in [Appendix F](#).

5.2 Out-of-Domain Transfer

[Table 2](#) extends the evaluation to MuSiQue ([Trivedi et al., 2022](#)), MMLU-Pro ([Wang et al., 2024](#)), and MMLU ([Hendrycks et al., 2021](#)), which differ substantially from the unlabeled mathematical prompts used during adaptation. None of these datasets appears in training or validation, so they test empirical transfer to previously unobserved distributions. We emphasize that these are domain shifts away from the mathematical post-training data, not adversarial or arbitrary shifts, and our claims regarding transfer are restricted to this setting.

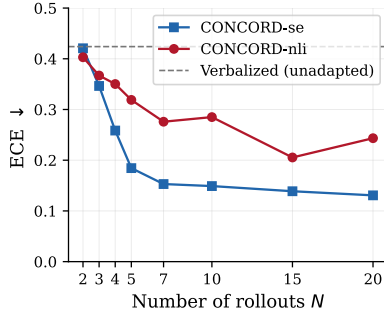
Absolute calibration degrades for every method under shift, including the multi-pass references: on MuSiQue the best result in three of the four columns is above 0.24, against in-domain values below 0.10. The ranking also becomes less favorable to CONCORD. Where CONCORD gave the best single-pass text-only ECE for 11 of 12 in-domain combinations, it does so for 5 of 12 here, with the remainder split between the label-based RL methods and ConfTuner. The label-based methods thus reverse their in-domain behavior: RLCR, which collapsed to 0.743 on GSM8K with Llama-3.2-3B, is now best on three columns, and LoVeC attains the overall best MMLU result on Qwen-2.5-7B.

		MuSiQue ECE ↓				MMLU-Pro ECE ↓				MMLU ECE ↓			
Method		Q3B	Q7B	L3B	L8B	Q3B	Q7B	L3B	L8B	Q3B	Q7B	L3B	L8B
Logits	Token Mean Prob.	0.740	0.721	0.616	0.552	0.619	0.558	0.605	0.522	0.385	0.249	0.383	0.263
	Length-Normalized	0.726	0.707	0.602	0.537	0.611	0.556	0.595	0.519	0.385	0.248	0.374	0.262
	P(True)	0.310	0.342*	0.149	0.242*	0.428	0.415	0.352	0.249	0.479	0.246	0.279	0.183
T	Verbalized	0.659	0.630	0.500	0.470	0.613	0.490	0.492	0.398	0.533	0.190	0.342	0.153
TT	TT-NLI (N=10)	0.367	0.381	0.355	0.478	0.199	0.272	0.253	0.189	0.131*	0.107	0.070	0.078
	TT-SE (N=10)	0.304*	0.359	0.317	0.445	0.182*	0.270*	0.229*	0.174*	0.152	0.113	0.063*	0.080
Ours	CONCORD-nli	0.416	0.437	0.346	0.454	0.346	0.458	0.342	0.391	0.246	0.205	0.359	0.184
	CONCORD-se	0.407	0.403	0.323	0.281	0.314	0.408	0.334	0.266	0.219	0.146	0.187	0.073*
Labels	RLCR	0.444	0.454	0.277	0.278	0.324	0.337	0.287	0.203	0.271	0.098	0.435	0.076
	Rewarding Doubt	0.451	0.486	0.330	0.334	0.485	0.456	0.322	0.317	0.206	0.281	0.396	0.120
	LoVeC	0.475	0.487	0.326	0.302	0.340	0.383	0.331	0.251	0.212	0.091*	0.360	0.086
SFT	ConfTuner	0.660	0.513	0.109*	0.492	0.658	0.570	0.230	0.412	0.361	0.269	0.517	0.164
	Self-Calibration	0.651	0.626	0.513	0.479	0.615	0.490	0.491	0.407	0.520	0.203	0.349	0.163

Table 2: Out-of-domain ECE; lower is better. Marks as in Table 1.

What holds is the comparison against the signal CONCORD replaces. CONCORD-se lowers ECE relative to unadapted verbalized confidence in all 12 settings, for example from 0.659 to 0.407 on MuSiQue with Qwen-2.5-3B and from 0.533 to 0.219 on MMLU with Qwen-2.5-3B; on MMLU with Llama-3.1-8B its 0.073 is the lowest of any method in that column, multi-pass baselines included. CONCORD-nli transfers less well, falling behind the verbalized baseline on three of the four MMLU columns.

6 Further Experiments

Figure 2: Effect of rollout count N on calibration quality, evaluated on GSM8K with Llama-3.1-8B.

We analyze the key design choices of CONCORD on Llama-3.1-8B evaluated on GSM8K, isolating the effect of each component while holding all other hyperparameters fixed.

6.1 Effect of Rollout Count

As shown in Figure 2, increasing the rollout count improves calibration because the self-consistency target becomes less noisy and therefore more useful for learning confidence. However, the cost of GRPO depends strongly on N , since each increase requires sampling and scoring more completions per prompt during training. The gains therefore diminish along a steep cost curve, making moderate N the most practical trade-off.

6.2 Effect of Bandwidth Annealing

The Gaussian calibration reward uses a bandwidth σ_t linearly annealed over the training duration (see Section 3). The intuition is that early-stage confidence outputs are poorly initialized and benefit from a forgiving reward landscape, while later stages benefit from a precise alignment with the consistency target. Table 3 shows that annealing improves calibration and also slightly strengthens discrimination relative to using a fixed bandwidth throughout training. Both phases of the schedule therefore matter: the wide phase prevents reward sparsity early in training, and the narrow phase tightens calibration near convergence.

	ECE ↓	AUROC ↑
No annealing	0.374	0.519
Annealing	0.066	0.628

Table 3: Bandwidth annealing ablation on CONCORD-se on GSM8K with Llama-3.1-8B

7 Discussion and Conclusion

Across Qwen-2.5 and Llama models in the 3B–8B range and six benchmarks spanning mathematical reasoning, multiple-choice QA, and multi-hop QA, CONCORD-se improves on unadapted verbalized confidence in all but one model–task pair while remaining single-pass and text-only at inference. The gains are largest where the base model is worst calibrated, on MATH-500 and on the shifted benchmarks, where verbalized confidence degrades sharply and CONCORD-se cuts the error by more than half in several settings. CONCORD-se is the more stable variant and gives the strongest single-pass text-only calibration in most in-domain settings; CONCORD-nli ranks examples better under domain shift, so the two offer a calibration–discrimination trade-off outside the training distribution.

The comparison to label-based adaptation depends on the setting. In domain, CONCORD is better calibrated than RLCR, Rewarding Doubt, LoVeC, and the supervised calibration baselines on all but one model–dataset combination despite never seeing a reference answer. Under domain shift that advantage disappears: the label-based methods recover, and each attains the best text-only ECE in at least one out-of-domain column. What survives is the comparison against the signal CONCORD replaces, since CONCORD-se lowers ECE relative to unadapted verbalized confidence on every shifted dataset we evaluate. Against test-time consistency, which recomputes the same signal from repeated sampling, CONCORD closes much of the gap in domain and is better in a few settings, but the gap stays wide on the multi-hop and harder multiple-choice benchmarks. Self-consistency is therefore a usable supervision signal for verbalized confidence.

CONCORD’s applicability is bounded in several ways. Hyperparameters were fixed across models and datasets rather than tuned per setting: bandwidth schedule, rollout count, and KL coefficient. Calibration also depends on the NLI model, whose clustering errors propagate into the training target. The consistently-wrong case from Section 3 is the sharpest limit: where a model repeats the same error across rollouts, consistency is 1 and the reward drives confidence toward 100%, which no label-free diagnostic can flag. Finally, we adapt 3B–8B models with LoRA over 1,600 prompts, leaving full fine-tuning and larger scales untested.

CONCORD trains language models to produce calibrated verbalized confidence by scoring the rollouts a GRPO run already generates against their own semantic consistency, using no ground-truth correctness labels at any stage. The NLI variant targets per-response cluster frequency; the semantic-entropy variant targets a smoother prompt-level quantity. Multi-rollout RL pipelines already pay for the samples CONCORD scores, so the method converts existing computation into calibrated confidence, with no extra supervision and no repeated sampling at deployment. We release our code at <https://github.com/Maxscha/concord>.

Acknowledgements

The project on which this report is based was funded in part by the Federal Ministry of Research, Technology and Space under the funding code “KI-Servicezentrum Berlin-Brandenburg” 16IS22092. Padhraic Smyth was supported in part by US National Science Foundation award NSF 2505006 and by funding support from Google and from SAP. Responsibility for the content of this publication remains with the authors.

Ethics Statement

This work improves LLM transparency by training models to verbalize calibrated confidence without human annotations. The NLI model used for semantic clustering may encode linguistic biases that affect confidence targets across dialects or non-standard phrasing. Although our method improves calibration considerably, LLM-reported confidence scores should never be trusted unquestioningly. We recommend domain-specific evaluation with held-out labeled data before deployment in high-stakes settings. While CONCORD removes multi-sample overhead at inference, the training phase generates multiple rollouts per prompt, incurring computational costs comparable to other multi-rollout RL pipelines.

References

- Neil Band, Xuechen Li, Tengyu Ma, and Tatsunori Hashimoto. Linguistic calibration of long-form generations. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 2732–2778. PMLR, 21–27 Jul 2024. URL <https://proceedings.mlr.press/v235/band24a.html>.
- 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. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 117927–117943, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/bf8065446507b0d3842838564ac4f1f3-Paper-Conference.pdf.
- Dylan Bouchard and Mohit Singh Chauhan. Uncertainty quantification for language models: a suite of black-box, white-box, LLM judge, and ensemble scorers. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856.
- Mengzhang Cai, Xin Gao, Yu Li, Honglin Lin, Zheng Liu, Zhuoshi Pan, Qizhi Pei, Xiaoran Shang, Mengyuan Sun, Zinan Tang, Xiaoyang Wang, Zhanping Zhong, Yun Zhu, Dahua Lin, Conghui He, and Lijun Wu. Opendataarena: A fair and open arena for benchmarking post-training dataset value, 2025. URL <https://arxiv.org/abs/2512.14051>.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *ArXiv preprint*, abs/2110.14168, 2021. URL <https://arxiv.org/abs/2110.14168>.
- Roi Cohen, Konstantin Dobler, Eden Biran, and Gerard de Melo. I Don’t Know: Explicit modeling of uncertainty with an [IDK] token. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 10935–10958. Curran Associates, Inc., 2024. doi: 10.52202/079017-0349. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/14c018d2e72c521605b0567029ef0efb-Paper-Conference.pdf.
- 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. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 59456–59493, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/615675cc6e94ddb1a783904fb178b5f6-Paper-Conference.pdf.
- Morris H. DeGroot and Stephen E. Fienberg. The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 32(1):12–22, 1983. ISSN 00390526, 14679884. doi: 10.2307/2987588. URL <https://doi.org/10.2307/2987588>.
- Ekaterina Fadeeva, Roman Vashurin, Akim Tsvigun, Artem Vazhentsev, Sergey Petrakov, Kirill Fedyanin, Daniil Vasilev, Elizaveta Goncharova, Alexander Panchenko, Maxim Panov, Timothy Baldwin, and Artem Shelmanov. LM-polygraph: Uncertainty estimation for language models. In Yansong Feng and Els Lefever (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 446–461, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-demo.41. URL <https://doi.org/10.18653/v1/2023.emnlp-demo.41>.
- Xin Gao, Xiaoyang Wang, Yun Zhu, Mengzhang Cai, Conghui He, and Lijun Wu. Closing the data loop: Using opendataarena to engineer superior training datasets. *CoRR*, abs/2601.09733, January 2026. URL <https://doi.org/10.48550/arXiv.2601.09733>.

- Tilman Gneiting, Fadoua Balabdaoui, and Adrian E Raftery. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 69 (2):243–268, 2007. doi: 10.21236/ada454827. URL <https://doi.org/10.21236/ada454827>.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, and Alex plus co-authors Vaughan. The llama 3 herd of models. *ArXiv preprint*, 2024. URL <https://arxiv.org/abs/2407.21783>.
- Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 2017. URL <http://proceedings.mlr.press/v70/guo17a.html>.
- Jiuzhou Han, Wray Buntine, and Ehsan Shareghi. Towards uncertainty-aware language agent. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the association for computational linguistics: ACL 2024*, pp. 6662–6685. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.398. URL <https://doi.org/10.18653/v1/2024.findings-acl.398>.
- Pengcheng He, Jianfeng Gao, and Weizhu Chen. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing, 2023. URL <https://arxiv.org/abs/2111.09543>.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022.
- Chengsong Huang, Langlin Huang, Jixuan Leng, Jiacheng Liu, and Jiaxin Huang. Efficient test-time scaling via self-calibration. In *NeurIPS 2025 Workshop on Efficient Reasoning*, 2025a. URL <https://openreview.net/forum?id=RvMjxGpV0a>.
- Jerry Huang, Peng Lu, and Qiuhaio Zeng. Calibrated language models and how to find them with label smoothing. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 25593–25612. PMLR, 13–19 Jul 2025b. URL <https://proceedings.mlr.press/v267/huang25w.html>.
- Ziwei Ji, Lei Yu, Yeskendir Koishekenov, Yejin Bang, Anthony Hartshorn, Alan Schelten, Cheng Zhang, Pascale Fung, and Nicola Cancedda. Calibrating verbal uncertainty as a linear feature to reduce hallucinations. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 3769–3793, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.187. URL <https://aclanthology.org/2025.emnlp-main.187/>.
- 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. doi: 10.1162/tacl_a.00407. URL <https://aclanthology.org/2021.tacl-1.57/>.
- Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai,

- Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language models (mostly) know what they know. *ArXiv preprint*, abs/2207.05221, 2022. URL <https://arxiv.org/abs/2207.05221>.
- Sanyam Kapoor, Nate Gruver, Manley Roberts, Katherine Collins, Arka Pal, Umang Bhatt, Adrian Weller, Samuel Dooley, Micah Goldblum, and Andrew Wilson. Large language models must be taught to know what they don't know. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 85932–85972. Curran Associates, Inc., 2024. doi: 10.52202/079017-2729. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/9c20f16b05f5e5e70fa07e2a4364b80e-Paper-Conference.pdf.
- Benjamin Kompa, Jasper Snoek, and Andrew L. Beam. Second opinion needed: communicating uncertainty in medical machine learning. *npj Digital Medicine*, 4(1):4, 2021. ISSN 2398-6352. doi: 10.1038/s41746-020-00367-3. URL <https://doi.org/10.1038/s41746-020-00367-3>.
- Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.
- Dharshan Kumaran, Arthur Conmy, Federico Barbero, Simon Osindero, Viorica Patraucean, and Petar Velickovic. How do LLMs compute verbal confidence? In *Forty-third International Conference on Machine Learning*, 2026. URL <https://openreview.net/forum?id=t7oQT3uX66>.
- Moxin Li, Wenjie Wang, Fuli Feng, Fengbin Zhu, Qifan Wang, and Tat-Seng Chua. Think twice before trusting: Self-detection for large language models through comprehensive answer reflection. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 11858–11875, Miami, Florida, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.693. URL <https://aclanthology.org/2024.findings-emnlp.693/>.
- Yibo Li, Miao Xiong, Jiaying Wu, and Bryan Hooi. Conftuner: Training large language models to express their confidence verbally. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, pp. 53484–53513. Curran Associates, Inc., 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/4d4f7cf206bb00f9a38a5b6ae92cf79a-Paper-Conference.pdf.
- Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 39578–39601, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/aca97732e30bcf1303bc22ac3924fd16-Paper-Conference.pdf.
- Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL <https://openreview.net/forum?id=8s8K2UZGTZ>.
- Jiayu Liu, Rui Wang, Qing Zong, Yumeng Wang, Cheng Qian, Qingcheng Zeng, Tianshi Zheng, Haochen Shi, Dadi Guo, Baixuan Xu, Chunyang Li, and Yangqiu Song. NOVA: Noise-aware verbal confidence calibration for robust large language models in rag systems. *ArXiv preprint*, 2026. URL <https://arxiv.org/abs/2601.11004>.
- Arkil Patel, Satwik Bhattamishra, and Navin Goyal. Are NLP models really able to solve simple math word problems? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp.

- 2080–2094, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.168. URL <https://aclanthology.org/2021.naacl-main.168>.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *CoRR*, abs/2402.03300, 2024. URL <https://doi.org/10.48550/arXiv.2402.03300>.
- Ola Shorinwa, Zhiting Mei, Justin Lidard, Allen Z. Ren, and Anirudha Majumdar. A survey on uncertainty quantification of large language models: Taxonomy, open research challenges, and future directions. *Acm Computing Surveys*, 58(3):1–38, 2025. ISSN 0360-0300. doi: 10.1145/3744238. URL <https://doi.org/10.1145/3744238>.
- Elias Stengel-Eskin and Benjamin Van Durme. Calibrated interpretation: Confidence estimation in semantic parsing. *Transactions of the Association for Computational Linguistics*, 11:1213–1231, 2023. doi: 10.1162/tacl.a.00598. URL <https://aclanthology.org/2023.tacl-1.69/>.
- Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5433–5442, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.330. URL <https://doi.org/10.18653/v1/2023.emnlp-main.330>.
- Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. MuSiQue: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022. doi: 10.1162/tacl.a.00475. URL <https://doi.org/10.1162/tacl.a.00475>.
- Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyang Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhui Chen. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, pp. 95266–95290. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2024. doi: 10.52202/079017-3018. URL <https://doi.org/10.52202/079017-3018>.
- Zhiqiu Xia, Jinxuan Xu, Yuqian Zhang, and Hang Liu. A survey of uncertainty estimation methods on large language models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the association for computational linguistics: ACL 2025*, pp. 21381–21396. Association for Computational Linguistics, 2025. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.1101. URL <https://doi.org/10.18653/v1/2025.findings-acl.1101>.
- 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 *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024.
- Chenjun Xu, Bingbing Wen, Bin Han, Robert Wolfe, Lucy Lu Wang, and Bill Howe. Do language models mirror human confidence? exploring psychological insights to address overconfidence in LLMs. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 25655–25672. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.1316. URL <https://doi.org/10.18653/v1/2025.findings-acl.1316>.
- Tianyang Xu, Shujin Wu, Shizhe Diao, Xiaozhe Liu, Xingyao Wang, Yangyi Chen, and Jing Gao. SaySelf: Teaching LLMs to express confidence with self-reflective rationales. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024*

- conference on empirical methods in natural language processing*, pp. 5985–5998. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.343. URL <https://doi.org/10.18653/v1/2024.emnlp-main.343>.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL <https://arxiv.org/abs/2412.15115>.
- Daniel Yang, Yao-Hung Hubert Tsai, and Makoto Yamada. On verbalized confidence scores for llms. *ArXiv preprint*, 2026. URL <https://arxiv.org/abs/2412.14737>.
- Caiqi Zhang, Xiaochen Zhu, Chengzu Li, Nigel Collier, and Andreas Vlachos. LoVeC: Reinforcement learning for better verbalized confidence in long-form generation. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 33336–33363, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.1539. URL <https://aclanthology.org/2026.acl-long.1539/>.
- Hanning Zhang, Shizhe Diao, Yong Lin, Yi Fung, Qing Lian, Xingyao Wang, Yangyi Chen, Heng Ji, and Tong Zhang. R-tuning: Instructing large language models to say ‘I don’t know’. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 7113–7139, Mexico City, Mexico, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.394. URL <https://doi.org/10.18653/v1/2024.naacl-long.394>.
- Ziang Zhou, Tianyuan Jin, Jieming Shi, and Qing Li. SteerConf: Steering LLMs for confidence elicitation. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, pp. 19460–19486. Curran Associates, Inc., 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/1bfd87d2d92f0556819467dc08034f76-Paper-Conference.pdf.

Appendix

A Comparing Calibration and Discrimination

Our method is designed to improve a model’s calibration, i.e., the accuracy of its verbalized confidence. In other words, if a method claims to be $p\%$ confident, ideally it would be correct $p\%$ of the time. In this section, we study to what extent such improved calibration also accords with the discriminative ability of a model’s output scores. To this end, we compute the AUROC for the same in-domain and out-of-domain evaluations as reported in the main paper (Table 4 and Table 5). Whereas the main paper emphasizes the ECE metric because the primary goal is numerically calibrated confidence, AUROC characterizes how well each method ranks correct answers above incorrect ones.

Calibration and discrimination separate clearly in our results. The multi-pass TT baselines give the best AUROC in every in-domain column and in eight of the twelve out-of-domain columns, consistent with their strong ECE. Among the remaining methods, P(True) is the strongest baseline: it leads ten of the twelve in-domain columns outside TT and attains the best overall AUROC in three out-of-domain columns, despite being uncompetitive on ECE in most of those same settings. This is consistent with prior work showing that white-box confidence signals can be effective for discrimination even when they are not well calibrated in absolute terms. A model may produce accurate rankings despite being highly overconfident or underconfident. The token-probability baselines illustrate the opposite failure: on GSM8K and SVAMP they fall below 0.45 AUROC for both Llama models, near or below chance, while their in-domain ECE is comparable to the verbalized baseline.

The two CONCORD variants trade off differently in the two regimes. In domain, CONCORD-se is the stronger variant on both metrics, giving the higher AUROC of the two in nine of the twelve settings and the best text-only AUROC in six; CONCORD-nli leads only on Qwen-2.5-3B, and there on SVAMP and MATH-500 rather than GSM8K. Out of domain the ordering reverses, with CONCORD-nli higher in eight of the twelve settings and best among text-only methods in five, while CONCORD-se retains its ECE advantage. The calibration–discrimination trade-off between the variants that we describe in the main text is therefore an out-of-domain effect rather than a uniform one.

Relative to the signal CONCORD replaces, adaptation does not come at the cost of ranking quality: CONCORD-se improves AUROC over the unadapted verbalized baseline in eleven of twelve in-domain settings, and at least one CONCORD variant does so in eleven of twelve out-of-domain settings. To further check that these changes reflect confidence quality rather than answer quality, we report detailed task accuracy in Appendix E, where accuracy is largely stable across methods.

B Prompt Template

All models are prompted with a fixed output structure so that both the final answer and the verbalized confidence can be parsed reliably. Unlike an answer-only prompt, the template prompts explicit step-by-step reasoning before the tagged answer, which we found necessary for the models to reach their full task accuracy. The template used in our experiments is shown below.

Question: {question}

Solve the problem step by step, then report your final answer and your confidence.

Requirements:

- Show your reasoning step by step. Take as many steps as you need to get it right.
- Then give the final answer as a single value (a number, no units or extra words) inside <answer></answer>.

	Method	GSM8K AUROC $\uparrow$				SVAMP AUROC $\uparrow$				MATH-500 AUROC $\uparrow$			
		Q3B	Q7B	L3B	L8B	Q3B	Q7B	L3B	L8B	Q3B	Q7B	L3B	L8B
Logits	Token Mean Prob.	0.567	0.632	0.438	0.349	0.404	0.611	0.446	0.363	0.654	0.680	0.530	0.496
	Length-Normalized	0.567	0.632	0.438	0.372	0.404	0.611	0.446	0.381	0.653	0.681	0.530	0.500
	P(True)	0.702	0.667	0.753	0.654	0.782	0.702	0.794	0.824	0.701	0.661	0.712	0.749
T	Verbalized	0.599	0.576	0.580	0.542	0.596	0.532	0.653	0.580	0.596	0.593	0.513	0.649
TT	TT-NLI ($N=10$)	0.912*	0.858*	0.900*	0.891*	0.913*	0.800*	0.859*	0.897*	0.823*	0.867*	0.781	0.752*
	TT-SE ($N=10$)	0.901	0.855	0.895	0.880	0.900	0.799	0.849	0.888	0.822	0.866	0.782*	0.749
Ours	CONCORD-nli	0.571	0.564	0.619	0.537	0.684	0.614	0.635	0.624	0.649	0.585	0.575	0.584
	CONCORD-se	0.569	0.605	0.633	0.628	0.672	0.642	0.703	0.740	0.646	0.729	0.598	0.680
Labels	RLCR	0.477	0.545	0.509	0.600	0.547	0.555	0.508	0.606	0.505	0.630	0.506	0.697
	Rewarding Doubt	0.512	0.629	0.565	0.621	0.616	0.550	0.546	0.578	0.529	0.557	0.537	0.693
	LoVeC	0.544	0.519	0.568	0.631	0.496	0.555	0.621	0.597	0.545	0.573	0.534	0.749
SFT	ConfTuner	0.515	0.506	0.501	0.578	0.514	0.500	0.634	0.567	0.518	0.505	0.502	0.587
	Self-Calibration	0.592	0.575	0.602	0.578	0.610	0.572	0.653	0.641	0.616	0.614	0.527	0.601

Table 4: In-domain AUROC (GSM8K, SVAMP, MATH-500); higher is better. Bold = best single-pass black-box/text-only method (excl. white-box logit and multi-pass TT baselines);

* = best overall. Model abbreviations as in Table 1.

	Method	MuSiQue AUROC $\uparrow$				MMLU-Pro AUROC $\uparrow$				MMLU AUROC $\uparrow$			
		Q3B	Q7B	L3B	L8B	Q3B	Q7B	L3B	L8B	Q3B	Q7B	L3B	L8B
Logits	Token Mean Prob.	0.673	0.691*	0.669	0.704	0.615	0.635	0.535	0.468	0.569	0.597	0.488	0.467
	Length-Normalized	0.672	0.689	0.665	0.700	0.615	0.635	0.538	0.468	0.569	0.597	0.488	0.467
	P(True)	0.693	0.690	0.724*	0.743*	0.620	0.657	0.671*	0.639	0.550	0.658	0.703	0.714
T	Verbalized	0.596	0.579	0.648	0.656	0.573	0.643	0.581	0.643	0.512	0.637	0.571	0.604
TT	TT-NLI ($N=10$)	0.669	0.646	0.583	0.605	0.699*	0.693*	0.652	0.699*	0.677*	0.776	0.766	0.805
	TT-SE ($N=10$)	0.703*	0.646	0.620	0.623	0.688	0.692	0.662	0.694	0.675	0.778*	0.768*	0.806*
Ours	CONCORD-nli	0.619	0.592	0.588	0.660	0.601	0.649	0.609	0.658	0.616	0.643	0.589	0.616
	CONCORD-se	0.620	0.539	0.643	0.687	0.591	0.621	0.617	0.613	0.598	0.609	0.581	0.593
Labels	RLCR	0.615	0.610	0.543	0.663	0.498	0.558	0.563	0.621	0.562	0.587	0.559	0.651
	Rewarding Doubt	0.608	0.624	0.550	0.658	0.473	0.523	0.580	0.618	0.557	0.578	0.581	0.611
	LoVeC	0.620	0.599	0.560	0.668	0.490	0.565	0.593	0.622	0.571	0.568	0.607	0.645
SFT	ConfTuner	0.581	0.631	0.593	0.623	0.509	0.520	0.551	0.654	0.533	0.538	0.498	0.657
	Self-Calibration	0.603	0.585	0.648	0.654	0.566	0.647	0.595	0.642	0.518	0.638	0.541	0.622

Table 5: Out-of-domain AUROC (MuSiQue, MMLU-Pro, MMLU); higher is better. Marks as in Table 4.

- Then ALWAYS give `<confidence></confidence>` as an integer from 0 to 100 - include it even if you are unsure.
- Write nothing after `</confidence>`.

End your response in exactly this format:

```
<answer>
[final answer value only]
</answer>
<confidence>
[integer 0-100]
</confidence>
```

For multiple-choice benchmarks (MMLU, MMLU-Pro), we use the same template except that the answer instruction requests the single option letter ("Put ONLY the single letter (A, B, C, or D) inside `<answer></answer>`"), so that the parsed answer matches the lettered gold label.

We parse the model's predicted answer from the `<answer>` field and the verbalized confidence from the `<confidence>` field. Answer correctness is scored with symbolic equivalence checking (via `math.verify`) on the mathematical benchmarks and exact letter matching on multiple-choice tasks, rather than surface string matching. The instruction to end generation immediately after `</confidence>` keeps outputs machine-readable and prevents trailing text from interfering with the evaluation.

C Dataset Descriptions

This section provides brief descriptions of the benchmark datasets used in our evaluation.

C.1 Mathematical Reasoning

GSM8K (Cobbe et al., 2021) Grade School Math 8K is a dataset of 8,500 linguistically diverse grade school mathematical word problems requiring multi-step arithmetic reasoning. Problems involve 2–8 steps of computation and span topics such as basic arithmetic, fractions, percentages, and simple algebra. The dataset splits into 7,473 training and 1,319 test examples, with solutions provided as step-by-step natural language explanations followed by a final numerical answer.

SVAMP (Patel et al., 2021) Simple Variations on Arithmetic Math word Problems contains 1,000 one-unknown arithmetic word problems designed to test whether models genuinely understand mathematical reasoning or rely on superficial pattern matching. Problems are created by applying simple variations to existing math problems (e.g., changing question locations or operations), with difficulty comparable to elementary school mathematics. Each problem requires 1–2 arithmetic operations to solve.

MATH-500 (Lightman et al., 2024) A 500-problem subset of the MATH competition benchmark, spanning seven subjects (algebra, geometry, number theory, counting and probability, precalculus, intermediate algebra, and prealgebra) at five difficulty levels. Problems require multi-step symbolic derivations and a boxed final answer, and are substantially harder than grade-school word problems. We report MATH-500 as the harder in-domain math benchmark on which we assess whether self-consistency calibration holds up under greater task difficulty.

C.2 Multiple-Choice Question Answering

MMLU (Hendrycks et al., 2021) Massive Multitask Language Understanding is a comprehensive benchmark covering 57 diverse subjects spanning STEM, humanities, social sciences, and professional domains. The dataset contains 15,908 multiple-choice questions with four options each, ranging from elementary to advanced professional level. Questions test both

world knowledge and problem-solving ability across domains such as mathematics, history, law, medicine, and computer science.

MMLU-Pro (Wang et al., 2024) An enhanced version of MMLU designed to be more challenging and discriminative. Compared to MMLU, it increases the number of answer choices from 4 to 10, incorporates more reasoning-focused questions, and filters out questions solvable by simple pattern matching. The dataset contains questions across the same subject areas as MMLU but with increased difficulty and reduced potential for models to exploit spurious correlations.

C.3 Multi-Hop Question Answering

MuSiQue (Trivedi et al., 2022) Multihop Questions via Single-hop Question Composition is a multi-hop QA dataset with 25,000 questions requiring 2–4 hops of reasoning. Questions are systematically composed from single-hop questions, ensuring answering requires multi-hop reasoning rather than shortcuts. The dataset includes both answerable and unanswerable questions, with unanswerable variants created by replacing key entities, testing whether models truly perform multi-hop reasoning or exploit dataset artifacts.

C.4 Evaluation Metrics

ECE. The Expected Calibration Error measures the mismatch between predicted confidence and empirical accuracy:

$$\text{ECE} = \sum_{m=1}^M \frac{|B_m|}{n} |\text{acc}(B_m) - \text{conf}(B_m)|, \quad (4)$$

where

$$\text{acc}(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \mathbb{1}(\hat{y}_i = y_i), \quad \text{conf}(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} p_i. \quad (5)$$

We use $M = 10$ equally spaced confidence bins throughout. Lower ECE indicates better calibration.

AUROC. The AUROC measures how well confidence scores discriminate between correct and incorrect predictions:

$$\text{AUROC} = \int_0^1 \text{TPR}(t) d\text{FPR}(t), \quad (6)$$

where

$$\text{TPR}(t) = \frac{\text{TP}(t)}{\text{TP}(t) + \text{FN}(t)}, \quad \text{FPR}(t) = \frac{\text{FP}(t)}{\text{FP}(t) + \text{TN}(t)}. \quad (7)$$

An AUROC of 1.0 indicates perfect discrimination, while 0.5 corresponds to chance-level ranking.

D Baseline Definitions

This section summarizes all baseline methods considered in our experiments. Table 6 provides an overview. In the following, we describe them in more detail, distinguishing single-pass confidence estimation methods from label-dependent adaptation methods.

D.1 Single-Pass Baselines

Verbalized Confidence (VC; Lin et al., 2022; Tian et al., 2023; Xiong et al., 2024) The model is prompted to produce both an answer a and a numerical confidence score $c \in [0, 100]$ as part of its output. The confidence estimate is taken directly from the generated text:

$$\hat{p}(a) = \frac{c}{100}. \quad (8)$$

Method	Type	Inference Access	Extra Passes	Train Labels	Confidence Form	Single-Pass at Test
Verbalized Confidence	Single-pass	Black-box	No	No	Textual scalar	Yes
P(True)	Single-pass	White-box	Yes	No	Token probability	No
Token-Mean-Prob	Single-pass	White-box	No	No	Token / sequence probability	Yes
TT-NLI	Multi-pass	Black-box	Yes ($N=10$)	No	Cluster frequency	No
TT-SE	Multi-pass	White-box	Yes ($N=10$)	No	Semantic entropy	No
RLCR	Adapted (RL)	Black-box at inference	No	Yes	Textual scalar	Yes
Rewarding Doubt	Adapted (RL)	Black-box at inference	No	Yes	Textual scalar	Yes
LoVeC	Adapted (RL)	Black-box at inference	No	Yes	Textual scalar	Yes
ConfTuner	Adapted (SFT)	Black-box at inference	No	Yes	Textual scalar	Yes
Self-Calibration	Adapted (SFT)	Black-box at inference	No	No	Textual scalar	Yes
CONCORD-nli	Adapted (RL)	Black-box at inference	No	No	Textual scalar	Yes
CONCORD-se	Adapted (RL)	Black-box at inference	No	No	Textual scalar	Yes

Table 6: Summary of all baseline and comparison methods used in this work. Multi-pass reference methods (TT-NLI, TT-SE) draw $N=10$ samples per question at test time; all other methods produce single-pass confidence. “Train Labels” indicates whether the training objective uses ground-truth answer correctness.

P(True) (Kadavath et al., 2022) This method first generates a candidate answer a , then asks the model whether that answer is correct. The confidence is defined as the probability assigned to the token “True”:

$$\hat{p}(a) = P(\text{“True”} \mid q, a). \quad (9)$$

Token Probabilities (Token-Prob) (Kadavath et al., 2022; Tian et al., 2023) This baseline uses token-level probabilities as a proxy for confidence. For a generated answer sequence $a = (a_1, \dots, a_T)$, we use the length-normalized sequence probability

$$\hat{p}(a) = \exp \left(\frac{1}{T} \sum_{t=1}^T \log P(a_t \mid q, a_{<t}) \right). \quad (10)$$

For multiple-choice tasks, we use the probability assigned to the answer label:

$$\hat{p}(a) = P(a_{\text{label}} \mid q). \quad (11)$$

In the results tables, we report two variants of this signal: *Token Mean Prob.*, the arithmetic mean of per-token probabilities of the answer span, and *Length-Normalized*, the geometric-mean form given by the equation above. Both are white-box signals requiring access to output logits.

D.2 Adaptation Baselines

LoVeC (R_{lovec} ; Zhang et al., 2026) Let $c \in [0, 1]$ denote normalized confidence and $f = \mathbb{1}[a = a^*]$ the correctness indicator. The reward is

$$R_{\text{lovec}}(a, a^*, c) = \lambda \left(1 + \frac{f \log(c) + (1-f) \log(1-c)}{R_{\text{max}}} \right), \quad (12)$$

where $R_{\text{max}} = |\log(\epsilon)|$, $\epsilon = 10^{-7}$, and $\lambda = 1.0$. Invalid or missing confidence values receive a penalty of -2λ .

Rewarding Doubt (Bani-Harouni et al., 2026) Let $p = \text{clamp}(c/100, 0.001, 0.999)$. The raw log score is

$$S(p, a, a^*) = \begin{cases} \log(p) & \text{if } a = a^*, \\ \log(1-p) & \text{if } a \neq a^*. \end{cases} \quad (13)$$

The normalized reward is

$$R_{\text{doubt}}(a, a^*, c) = \lambda \left(\frac{S(p, a, a^*) - S_{\min}}{S_{\max} - S_{\min}} + \frac{1}{4} \mathbb{1}[a = a^*] \right), \quad (14)$$

where $S_{\max} = \log(0.999)$, $S_{\min} = \frac{1}{2} \log(0.001)$, and $\lambda = 10$. Invalid confidence values receive a penalty of -3λ .

RLCR Calibration (R_{rlcr}) (Damani et al., 2026) Following the RLCR framework, we use the Brier-score-style reward

$$R_{\text{rlcr}}(a, a^*, c) = \mathbb{1}[a = a^*] - \left(\frac{c}{100} - \mathbb{1}[a = a^*] \right)^2. \quad (15)$$

This directly encourages high confidence on correct answers and low confidence on incorrect answers.

ConfTuner (Li et al., 2025) A supervised-fine-tuning baseline that trains the model to emit calibrated verbalized confidence using a tokenized Brier score. We train it on the same prompt subset as CONCORD with the authors’ recipe.

Self-Calibration (Huang et al., 2025a) A label-free baseline that produces single-pass confidence from self-consistency. It constructs a synthetic calibration dataset offline and distills a fixed Soft Self-Consistency target through a separate supervised-fine-tuning stage. Unlike CONCORD, which scores the rollouts a GRPO run already generates, it requires a dedicated calibration stage and synthetic data.

D.3 Multi-Pass Reference Baselines

Test-Time Consistency (TT-NLI, TT-SE) (Kuhn et al., 2023) As an inference-time reference we report test-time consistency, which applies the same NLI-based cluster-frequency (TT-NLI) and semantic-entropy (TT-SE) signals that CONCORD distills, but computes them at inference from $N=10$ freshly sampled rollouts per question (temperature $T=1.0$) rather than from a single forward pass. These baselines are *not* deployment-equivalent because they require repeated decoding; they serve as the multi-pass upper bound that CONCORD approximates with a single pass.

E Task Accuracy

We report task accuracy in addition to calibration metrics to verify that changes in ECE and AUROC are not driven by large shifts in answer quality. The results are given in [Table 7](#) and [Table 8](#). Across methods, accuracy remains comparatively stable, which suggests that most differences are due to changes in confidence quality rather than underlying task performance.

		GSM8K Acc. (%) $\uparrow$				SVAMP Acc. (%) $\uparrow$				MATH-500 Acc. (%) $\uparrow$			
Method		Q3B	Q7B	L3B	L8B	Q3B	Q7B	L3B	L8B	Q3B	Q7B	L3B	L8B
Logits	Token Mean Prob.	85.9	92.7	80.6	89.6	87.6	92.0	83.2	90.3	58.5	71.4	45.6	51.9
	Length-Normalized	85.9	92.7	80.6	89.6	87.6	92.0	83.2	90.3	58.5	71.4	45.6	51.9
	P(True)	85.9	92.7	80.6	89.6	87.6	92.0	83.2	90.3	58.5	71.4	45.6	51.9
T	Verbalized	85.9	92.7	80.6	89.6	87.6	92.0	83.2	90.3	58.5	71.4	45.6	51.9
TT	TT-NLI ($N=10$)	90.5	93.8	83.3	91.4	91.3	93.0	87.7	89.3	59.1	73.2	47.4	50.2
	TT-SE ($N=10$)	90.5	93.8	83.3	91.4	91.3	93.0	87.7	89.3	59.1	73.2	47.4	50.2
Ours	CONCORD-nli	83.8	90.5	75.3	86.5	87.6	89.9	82.9	88.0	58.4	69.9	45.2	53.2
	CONCORD-se	83.7	90.8	75.2	86.2	86.3	90.9	84.0	87.2	60.2	68.3	43.7	48.8
Labels	RLCR	83.7	90.8	76.8	57.0	86.3	91.3	83.9	75.1	59.3	69.9	46.3	53.8
	Rewarding Doubt	84.0	90.8	75.3	53.7	89.0	91.0	85.1	75.6	58.3	72.0	46.4	54.4
	LoVeC	84.0	90.8	75.0	54.5	87.7	91.3	83.7	76.2	56.0	69.9	45.3	53.8
SFT	ConfTuner	81.7	87.8	68.6	85.2	86.0	88.0	71.3	86.3	51.7	59.8	43.9	54.6
	Self-Calibration	82.9	90.4	76.2	86.7	86.6	89.3	85.3	89.6	55.9	70.7	44.3	57.2

Table 7: In-domain task accuracy (%) on GSM8K, SVAMP, MATH-500. Accuracy is largely stable across methods, confirming that calibration gains do not come at the cost of task performance. Logit rows share the base model’s greedy answer (hence match Verbalized); TT uses the majority-vote answer over $N=10$ samples. Model abbreviations as in [Table 1](#).

		MuSiQue Acc. (%) $\uparrow$				MMLU-Pro Acc. (%) $\uparrow$				MMLU Acc. (%) $\uparrow$			
Method		Q3B	Q7B	L3B	L8B	Q3B	Q7B	L3B	L8B	Q3B	Q7B	L3B	L8B
Logits	Token Mean Prob.	21.0	22.0	22.6	33.3	34.5	42.2	29.8	43.4	59.9	74.6	55.4	71.2
	Length-Normalized	21.0	22.0	22.6	33.3	34.5	42.2	29.8	43.4	59.9	74.6	55.4	71.2
	P(True)	21.0	22.0	22.6	33.3	34.5	42.2	29.8	43.4	59.9	74.6	55.4	71.2
T	Verbalized	21.0	22.0	22.6	33.3	34.5	42.2	29.8	43.4	59.9	74.6	55.4	71.2
TT	TT-NLI ($N=10$)	7.4	26.6	8.6	8.2	35.1	44.6	22.1	43.7	51.1	76.9	57.6	70.9
	TT-SE ($N=10$)	7.4	26.6	8.6	8.2	35.1	44.6	22.1	43.7	51.1	76.9	57.6	70.9
Ours	CONCORD-nli	22.1	22.9	21.4	35.3	35.5	46.8	32.3	47.3	61.6	74.4	55.4	70.0
	CONCORD-se	22.3	21.9	20.2	32.4	34.8	46.1	29.3	45.8	59.7	73.8	54.5	70.2
Labels	RLCR	20.7	22.6	20.7	31.6	32.6	43.2	29.1	43.9	58.5	74.2	54.9	70.8
	Rewarding Doubt	20.9	23.2	21.9	33.2	30.6	43.1	28.0	42.9	54.0	74.1	53.8	70.3
	LoVeC	20.1	23.8	21.3	32.2	32.7	42.5	27.8	43.1	63.7	73.7	53.2	70.1
SFT	ConfTuner	21.2	29.0	14.2	32.1	31.9	41.4	26.8	43.2	61.7	72.4	55.7	70.4
	Self-Calibration	20.8	22.6	20.5	31.7	27.7	42.2	33.4	42.8	41.6	73.5	56.5	70.2

Table 8: Out-of-domain task accuracy (%) on MuSiQue, MMLU-Pro, MMLU. Notes as in [Table 7](#).

F Qualitative Examples — Pretrained vs. CONCORD

The following two examples are drawn from **Qwen-2.5-3B-Instruct** evaluated on **GSM8K**. For each case, the outputs are shown side-by-side for the pretrained baseline and the **CONCORD-se** checkpoint (seed 42); both reasoning traces are extracted from the actual model generations. Across all examples the key observation is the same: the pretrained model defaults to $p = 1.0$ (maximum confidence) regardless of whether its answer is correct, while the CONCORD model produces lower confidence on problems it gets wrong and retains high confidence on problems it solves correctly.

Case 1 · Same Wrong Answer, Very Different Confidence

Question.

Charlie wants to sell beeswax candles. For every pound of beeswax, he can make 10 tapered candles. One pound of beeswax *and the wicks* cost \$10.00 in supplies. If he sells each candle for \$2.00, what is his net profit if he makes and sells 20 candles? (Gold answer: \$20)

	Pretrained	CONCORD-se
Reasoning	20 candles need $20/10 = 2$ lb of beeswax, costing $2 \times \$10 = \$20$. Then adds \$10 for the wicks: total cost $\$20 + \$10 = \$30$. Revenue $20 \times \$2 = \$40$. Net profit $\$40 - \$30 = \$10 \dots$ [Error: double-counts the wicks — the \$10 supply price already includes them, so the true cost is \$20 and the profit is \$20.]	Identical derivation: 2 lb of beeswax at \$20, an extra \$10 for wicks, \$30 total cost, \$40 revenue, profit $\$40 - \$30 = \$10 \dots$ [Same double-counted wick cost, same wrong answer.]
Answer	$\$10 \times$	$\$10 \times$
Confidence	$p = 1.00$	$p = 0.20$

Interpretation. Both models make the same subtle mistake, double-counting the wick cost, and arrive at the identical wrong answer through near-identical reasoning. The pretrained model exhibits no awareness of its error and expresses maximum confidence. The CONCORD-se model, whose confidence reflects the self-consistency of its own rollouts, correctly reduces its confidence on this error-prone problem. A downstream system relying on verbalized confidence would *correctly* flag the CONCORD-se answer for verification while *incorrectly* trusting the pretrained model’s wrong answer.

Case 2 · Correct Answer — Confidence Preserved After RL Training

Question.

A football team played 22 games. They won 8 more than they lost. How many did they win?
(Gold answer: 15)

	Pretrained	CONCORD-se
Reasoning	Let L be the losses; wins $W = L + 8$. Since $W + L = 22$, $(L + 8) + L = 22 \Rightarrow 2L + 8 = 22 \Rightarrow L = 7$, so $W = 7 + 8 = 15$.	Identical derivation: $W = L + 8$ and $W + L = 22$ give $2L + 8 = 22$, $L = 7$, hence $W = 15$.
Answer	15 ✓	15 ✓
Confidence	$p = 1.00$	$p = 1.00$

Interpretation. On a problem they can solve reliably, both models produce identical reasoning and both correctly express maximum confidence. CONCORD-se training does *not* deflate confidence indiscriminately; the finetuned model retains $p = 1.0$ where its rollouts agree, showing that the calibration improvement is discriminative rather than global.

G Usage of LLMs

LLMs were used for grammar correction and controlled paraphrasing of the manuscript text. All such edits were manually reviewed by the authors.