TEST-TIME CONFIDENCE ADAPTATION UNDER RL-INDUCED TRAJECTORY SHIFT

Anonymous authors
Paper under double-blind review

ABSTRACT

Trajectory-based confidence estimation has shown promising performance for large language models (LLMs) by leveraging rich signals from their reasoning processes. However, we find that such confidence estimators can degrade substantially after agentic reinforcement learning (Agentic RL) as it induces substantial shifts in the reasoning trajectories. More importantly, despite these shifts, we uncover a reliability structure that is largely preserved across Base and RL models: **more stable trajectories remain more likely to produce correct answers**. This preserved relationship provides a transferable, label-free reliability signal for post-RL confidence adaptation. Building on this finding, we propose **StaR-TTA**, a stability-guided framework for label-free test-time confidence adaptation. StaR-TTA jointly learns from correctness supervision and stability-derived reliability supervision on labeled Base trajectories, establishing the connection between trajectory stability and answer reliability. At test time, it reuses stability-induced reliability rankings from unlabeled RL trajectories to adapt confidence scores to the shifted trajectory distribution, without requiring target correctness labels. Across four QA benchmarks, StaR-TTA achieves an average AUROC of 0.721, with consistent improvements across unseen datasets, model families and scales, and different Agentic RL algorithms.

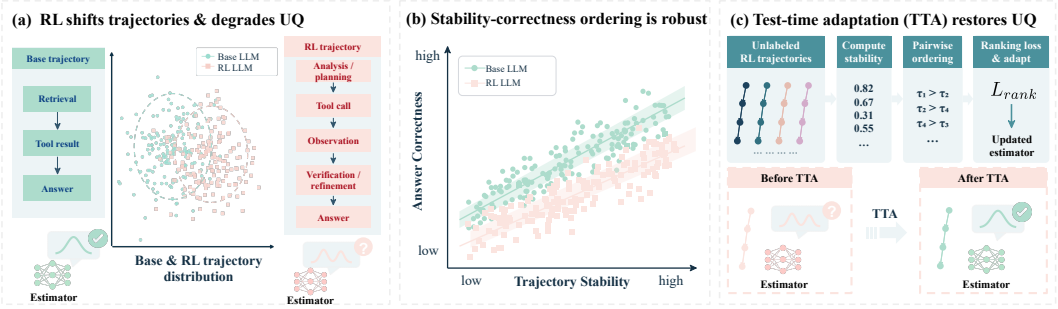

Figure 1: (a) Agentic RL shifts reasoning trajectories and degrades uncertainty quantification (UQ). (b) Stability-correctness ordering remains highly positive after RL. (c) Stability-guided test-time adaptation restores reliable confidence estimation. Using unlabeled RL trajectories, we exploit trajectory stability to construct pairwise ordering and adapt a confidence estimator at test time with a ranking loss, leading to more reliable confidence estimates.

1 INTRODUCTION

Reliable confidence estimation is essential for deploying large language models (LLMs) in real-world applications, particularly for reasoning tasks where plausible outputs can still be incorrect (Xiao et al., 2025; de Oliveira et al., 2025; Tian et al., 2023; Turpin et al., 2023; Guan et al., 2026; Xiao et al., 2026). Recent approaches increasingly estimate confidence from reasoning trajectories, which reveal how a model arrives at its answer (Khanmohammadi et al., 2025; Subramani et al., 2025). Meanwhile, LLMs are increasingly optimized through agentic reinforcement learning (Agentic RL), which can substantially reshape their reasoning and tool-use behavior (Cui et al.,

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.

2026; Wei et al., 2026; Jin et al., 2025c; Singh et al., 2025; Matsutani et al., 2026). This raises a critical question: **do confidence estimators learned from a base model remain reliable after Agentic RL?**

We find that they often do not. Confidence estimators that perform well on Base-model trajectories can degrade substantially when directly transferred to their RL-trained counterparts. Through controlled analysis, we identify post-RL trajectory shift as an important source of this degradation: even with tool results and final answers controlled, replacing RL trajectories with their Base counterparts recovers a substantial portion of the lost confidence-ranking performance. Retraining requires post-RL correctness labels, while direct transfer suffers from trajectory shift. We therefore ask: **can a confidence estimator adapt to post-RL trajectory shift without target correctness labels?**

Our key observation is that, although Agentic RL shifts trajectories, it does not destroy all reliability structure. Across both Base and RL models, trajectory stability remains positively associated with answer correctness: answers that remain more stable under repeated sampling are consistently more likely to be correct. Crucially, unlike correctness labels, trajectory stability can be obtained directly from model outputs without human annotation. This creates a useful supervision bridge across the RL shift: **correctness labels can ground the stability-reliability relationship on the source model, while stability remains available as a label-free reliability signal after RL**.

Building on this insight, we propose **StaR-TTA**, a label-free test-time confidence adaptation framework via **Sta**bility **R**anking. During source training, StaR-TTA jointly learns from correctness supervision and stability-derived reliability supervision, explicitly aligning confidence scores with the reliability ordering induced by trajectory stability. At test time, correctness labels are unavailable, but the preserved stability–correctness relationship provides a transferable supervisory signal. StaR-TTA extracts stability from unlabeled RL trajectories, constructs reliability rankings, and uses them to adapt the confidence scorer to the shifted trajectory distribution while preserving the source-learned confidence scale. The adapted scores are then mapped to correctness probabilities by a frozen confidence calibrator. In this way, the same stability-based reliability structure learned alongside correctness on the source model becomes the supervision for label-free adaptation after RL.

Extensive experiments on four QA benchmarks show that StaR-TTA consistently improves post-RL confidence estimation, achieving an average AUROC of 0.721. Its gains persist across unseen datasets, different model families and scales, and different Agentic RL algorithms, demonstrating the strong generalization of our method.

Our contributions are summarized as follows:

- We reveal that confidence estimation can degrade substantially after Agentic RL, and identify post-RL trajectory shift as an important contributor to the failure of trajectory-based confidence estimators.
- We uncover a transferable reliability structure under RL-induced trajectory shift: despite substantial changes in reasoning trajectories, the relationship between trajectory stability and answer correctness is largely preserved across Base and RL models.
- We propose **StaR-TTA**, a label-free test-time confidence adaptation framework that learns stability-based reliability from source correctness supervision and reuses the preserved reliability ordering to adapt confidence estimation to unlabeled post-RL trajectories.
- Extensive experiments show that StaR-TTA consistently improves post-RL confidence estimation and generalizes effectively across models, datasets, and Agentic RL algorithms.

2 Related Work

Confidence Estimation. Confidence estimation predicts the probability that an LLM answer is correct (Kadavath et al., 2022). $P(\mathrm{True})$ provides a self-evaluated signal (Kadavath et al., 2022). Beyond self-evaluation, CCPS (Khanmohammadi et al., 2025) predicts correctness from hidden-state perturbation features, while MICE (Subramani et al., 2025) uses logit-lens comparisons between intermediate and final predictions for tool-using agents. At the process level, HTC (Zhang et al., 2026b) estimates answer confidence for tool-using agents, whereas $P(\mathrm{Sufficient})$ (Zhang et al., 2026a) probes whether accumulated multi-turn information is sufficient for answering correctly. DACA (Luo et al., 2026) instead calibrates post-trained model confidence through agreement and

disagreement with its pre-trained counterpart, without correctness labels. These methods exploit hidden representations, intermediate predictions, trajectory behavior, and so on, but how reliably these signals transfer under model and trajectory distribution shifts remains an open question.

Agentic Reinforcement Learning. Agentic reinforcement learning (Agentic RL) trains LLM policies through multi-step interactions with environments, tools, or other agents, using trajectory-level rewards (Cui et al., 2026; Wei et al., 2026). Common training algorithms include Proximal Policy Optimization (PPO) and Group Relative Policy Optimization (GRPO) (Ouyang et al., 2022; Havrilla et al., 2024; Guo et al., 2025; Shao et al., 2024). Unlike supervised fine-tuning on fixed reasoning traces, Agentic RL lets models learn when to reason, search, verify, and use tools through interaction. For example, Search-R1 (Jin et al., 2025b) trains a base LLM to interleave reasoning with search, issue multiple queries, and incorporate retrieved evidence before answering. It uses multi-turn rollouts, retrieved-token masking, and outcome-based rewards without step-level annotations. Other work applies Agentic RL to self-reflection and verification (Guo et al., 2025), tool use (Singh et al., 2025), and search without supervised reasoning traces (Chen et al., 2025). These methods can change action patterns, tool use behaviors, reasoning structure, and hidden representations, shifting the trajectory distributions on which confidence estimators rely.

3 Confidence Degradation under RL-Induced Trajectory Shift

3.1 Confidence Estimation Degrades after Agentic RL

We first examine whether confidence estimators learned on a Base LLM remain reliable after Agentic RL. Table 1 reports AUROC on NQ before and after RL. MICE-LR, MICE-RF, and CCPS explicitly exploit fine-grained trajectory information, while HTC summarizes structural properties of agent trajectories; DACA and $P(\text{Sufficient})$ rely less directly on trajectory representations. MICE-LR and MICE-RF drop by 23.5% and 27.7%, while CCPS drops by 34.1%. In contrast, DACA remains nearly unchanged and $P(\text{Sufficient})$ improves. These results reveal a clear vulnerability: confidence estimators that exploit trajectory information can suffer substantial degradation in confidence ranking after transfer to an RL-trained model. This motivates us to examine whether changes in the underlying trajectories are responsible for the degradation.

Method	Base AUROC↑	RL AUROC↑	Relative drop (%)
MICE-LR	0.741	0.567	23.5
MICE-RF	0.773	0.559	27.7
HTC	0.765	0.675	11.8
CCPS	0.745	0.491	34.1
DACA	0.659	0.656	0.5
$P(\text{Sufficient})$	0.627	0.689	−9.9

Table 1: Base-to-RL confidence transfer on NQ.

3.2 Isolating the Effect

To isolate the source of this degradation, we progressively control three factors that can change after Agentic RL: tool results, final answers, and reasoning trajectories. The results are summarized in Table 2.

Method	Tool control	Answer control	Trajectory replay	Recovery (%)
MICE-LR	.567 (−.174)	.514 (−.110)	.563 (+.049)	44.5
MICE-RF	.559 (−.215)	.577 (−.087)	.623 (+.045)	51.9
CCPS	.491 (−.255)	.498 (−.133)	.513 (+.015)	11.1

Table 2: Controlled AUROC analysis on NQ.

Under *tool control*, we replace RL tool results with their corresponding Base-model results, yet substantial AUROC gaps remain. Under *answer control*, we further restrict evaluation to examples for which the Base and RL models produce identical tool results and final answers. The degradation still persists. Finally, under *trajectory replay*, we keep the controlled tool results and final answers fixed but replace the RL reasoning trajectory with its Base-model counterpart. This recovers 44.5%, 51.9%, and 11.1% of the remaining AUROC loss for MICE-LR, MICE-RF, and CCPS, respectively. These results show that changes in tools or final answers alone cannot explain the transfer failure. Instead, the shift in reasoning trajectories itself contributes substantially to the degradation of trajectory-based confidence estimation.

This raises a key challenge: if the trajectory distribution changes after RL, what reliability information remains transferable from the Base model?

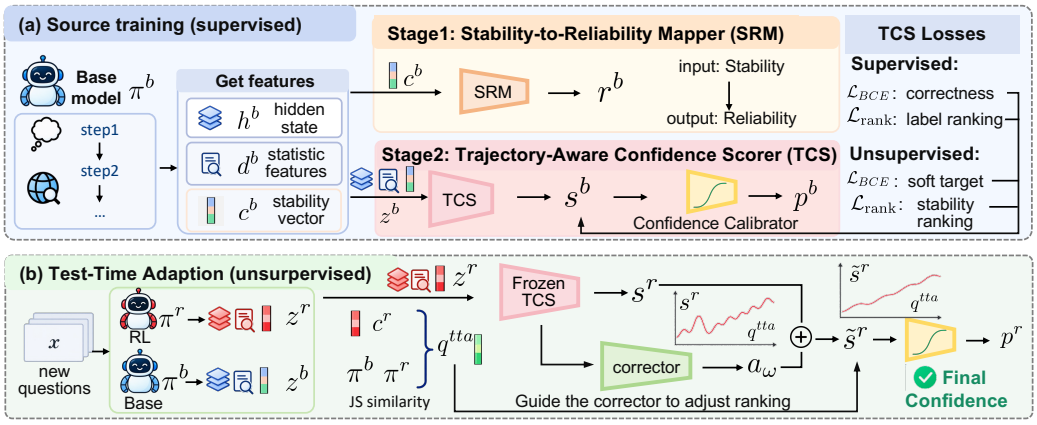

Figure 2: Overview of our method. (a) Firstly, an SRM model is trained to capture the relationship between trajectory stability and correctness. Then a scorer (TCS) is optimized, followed by a calibrator (CC) to convert scores into calibrated probabilities. (b) At test time, we adapt the results of TCS to Agentic-RL trajectories. We extract stability signals from both Agentic-RL and Base trajectories and use this information to adjust the score while keeping TCS and CC fixed .

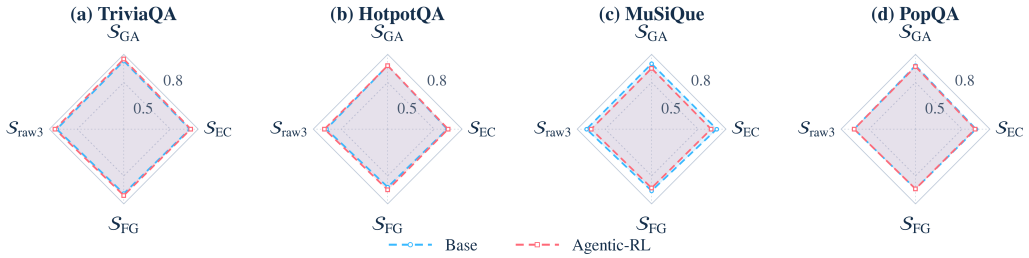

Figure 3: AUROC of sampling-stability signals for answer correctness.

3.3 Stability–Correctness Association Is Largely Preserved after Agentic RL

Despite the substantial trajectory shift, we find that not all reliability information is lost. In particular, the relationship between **trajectory stability** and answer correctness remains informative after Agentic RL. For each trajectory, we fix the prefix before final-answer generation and only sample multiple short answers without repeating interactions. We characterize their stability using three complementary signals: $\mathcal{S}_{\text{GA}}$, the fraction of sampled short answers matching the original answer; $\mathcal{S}_{\text{EC}}$, the concentration of the sampled short answer distribution; and $\mathcal{S}_{\text{FG}}$, the frequency gap between the two most common answers. We further consider their equal-weight average, $\mathcal{S}_{\text{raw3}}$.

Figure 3 evaluates each stability signal as a predictor of answer correctness on both Base and RL trajectories. Across all four datasets, the stability signals achieve AUROC well above 0.5 under both Base and Agentic-RL models, while $\mathcal{S}_{\text{raw3}}$ consistently reaches around 0.7–0.8. This predictive ability is largely retained after Agentic RL despite substantial trajectory shifts. Thus, despite substantial trajectory shifts after Agentic RL, more stable trajectories remain more likely to produce correct answers. This observation provides a transferable reliability signal across the RL-induced trajectory shift. Unlike correctness labels, trajectory stability can be obtained directly from unlabeled model outputs. This motivates us to leverage stability-derived reliability ordering as label-free supervision to adapt confidence estimation to post-RL trajectories.

4 Method

Our analysis in Sec. 3 shows that Agentic RL substantially shifts trajectory representations, while trajectory stability remains predictive of correctness. Based on this finding, we propose **StaR-TTA**, which uses stability as a transferable reliability signal for post-RL confidence adaptation.

As illustrated in Figure 2, StaR-TTA jointly learns from correctness and stability-derived reliability supervision on labeled Base trajectories. At test time, stability-induced rankings provide label-free supervision for adapting confidence scores to RL trajectories, mitigating the impact of trajectory shift on confidence estimation.

4.1 Problem Setup

We study confidence transfer from a Base LLM M_b to its Agentic-RL counterpart M_r. Given a question q_i (hereafter, we omit the subscript i unless otherwise specified), model M_m, where $m \in \{b, r\}$, generates an interaction trajectory $\tau^m = M_m(q)$ containing its reasoning process, tool interactions, and final answer a^m. Let $y^m \in \{0, 1\}$ denote the correctness of a^m, and let $p^m = P(y^m = 1 \mid q, \tau^m)$ denote its confidence. During source training, correctness labels y^b are available for Base-model trajectories, whereas target correctness labels y^r are unavailable at test time. Our goal is to adapt a Base-trained confidence estimator to the post-RL trajectory distribution using only signals derived from unlabeled target trajectories.

4.2 Stability-Grounded Source Confidence Learning

Trajectory representation. We represent each trajectory using internal, interaction, and stability features. We extract a 64-D hidden representation h^m at the answer-generation position and a 24-D descriptor d^m of trajectory statistics, including tool interactions, invalid actions, and other related statistics. To measure stability without repeated agent rollouts, we fix the trajectory prefix before final-answer generation and sample five short answers. After answer normalization, we compute support $\mathcal{S}^m_{\mathrm{GA}} = f^m(a^m)$, entropy concentration $\mathcal{S}^m_{\mathrm{EC}} = 1 + \sum_{a:f^m(a)>0} f^m(a) \log f^m(a) / \log 5$, and frequency gap $\mathcal{S}^m_{\mathrm{FG}} = f^m_{(1)} - f^m_{(2)}$, where $f^m(a)$ is the empirical answer frequency, a^m is the original final answer, and $f^m_{(1)}, f^m_{(2)}$ are the two largest frequencies (zero-padded). Let $c^m = [\mathcal{S}^m_{\mathrm{GA}}, \mathcal{S}^m_{\mathrm{EC}}, \mathcal{S}^m_{\mathrm{FG}}] \in \mathbb{R}^3$. The complete representation is

$$z^m = [h^m; d^m; c^m] \in \mathbb{R}^{91}.$$

Stability-to-Reliability Mapper. Stability is available without correctness labels but provides only an indirect reliability signal. We therefore use labeled Base trajectories to learn a Stability-to-Reliability Mapper (SRM),

$$r^b = M_\phi(c^b) = \sigma(w^\top c^b + b), \qquad \mathcal{L}_{\mathrm{SRM}} = \mathrm{BCE}(r^b, y^b).$$

The SRM learns how the three stability signals relate to correctness.

Trajectory-Aware Confidence Scorer. To estimate the confidence from trajectory signals, we train a Trajectory-Aware Confidence Scorer (TCS).

$$s^b = g_\theta(z^b)$$

using both correctness and SRM-derived reliability supervision. Correctness provides direct supervision through BCE and pairwise ranking, while the SRM provides a transferable reliability ordering. We use index i and j to represent different source trajectories:

$$\mathcal{P}_y = \{(i, j) : y^b_i = 1,\ y^b_j = 0\}, \qquad \mathcal{P}_r = \{(i, j) : r^b_i - r^b_j > \delta\}.$$

For either pair set $\mathcal{P}$, we define

$$\mathcal{L}_{\mathrm{rank}}(\mathcal{P}) = \frac{1}{|\mathcal{P}|} \sum_{(i,j)\in\mathcal{P}} \log(1 + \exp[-(s^b_i - s^b_j)]).$$

The source objective is

$$\mathcal{L}_{\mathrm{score}} = \underbrace{\mathrm{BCE}(\sigma(s^b_i), y^b_i) + \lambda_h \mathcal{L}_{\mathrm{rank}}(\mathcal{P}_y)}_{\text{correctness supervision}} + \underbrace{\lambda_r \mathcal{L}_{\mathrm{rank}}(\mathcal{P}_r) + \lambda_s \mathrm{BCE}(\sigma(s^b_i), r^b_i)}_{\text{stability-derived supervision}}.$$

This joint training connects correctness-based confidence estimation with the stability-derived reliability signal available after RL.

Finally, we freeze the TCS and fit a monotonic Confidence Calibrator (CC) on Base-model scores and correctness labels:

$$p^b_i = C_\psi(s^b_i) = \sigma(\alpha s^b_i + \beta), \qquad \alpha > 0,$$

This converts the TCS score into a probability of correctness while preserving its ranking.

4.3 STABILITY-GUIDED TEST-TIME CONFIDENCE ADAPTATION

Directly applying the Base-trained TCS to RL trajectories yields $s^r = g_\theta(z^r)$, whose reliability can degrade under RL-induced trajectory shift. Although trajectory stability remains predictive of correctness, it is only an indirect reliability signal and cannot serve as a calibrated confidence estimate. We therefore use stability as *adaptation supervision*: it specifies how confidence scores should be reordered after RL, while the source-trained TCS provides the initial confidence estimate.

Target reliability ranking. RL stability alone characterizes the consistency of the post-RL model, but does not capture how strongly its behavior has changed from the Base model. We therefore complement c^r with Base–RL answer agreement. Given sampled-answer distributions π^b and π^r, we compute

$$A = 1 - \frac{\mathrm{JSD}(\pi^b, \pi^r)}{\log 2},$$

and construct the label-free target reliability signal

$$q^{\mathrm{tta}} = \frac{\mathcal{S}^r_{\mathrm{GA}} + \mathcal{S}^r_{\mathrm{EC}} + \mathcal{S}^r_{\mathrm{FG}} + A}{4}.$$

Rather than treating q^{tta} as confidence, we use it only to infer relative reliability. Specifically,

$$\mathcal{P}_{\mathrm{tta}} = \{(i,j) : q^{\mathrm{tta}}_i - q^{\mathrm{tta}}_j > \delta_{\mathrm{tta}}\}, \qquad t_{ij} = \sigma\left(\frac{q^{\mathrm{tta}}_i - q^{\mathrm{tta}}_j}{T}\right),$$

where i, j index different target trajectories and t_{ij} emphasizes larger reliability gaps.

Trajectory-conditioned residual adaptation. A direct correction based only on q^{tta} would apply similar adjustments to samples with similar stability, despite potentially different trajectory shifts. Instead, we use the stability-induced ordering to supervise an *instance-dependent* correction conditioned on the TCS trajectory representation. Specifically, we freeze the TCS and optimize a lightweight linear corrector a_ω:

$$\tilde{s}^r = s^r + a_\omega(e^r),$$

where e^r is the intermediate TCS representation. The residual formulation preserves the source-trained score while allowing trajectory-specific adaptation.

We use i, j to index trajectories and optimize

$$\mathcal{L}_{\mathrm{tta}} = \underbrace{\frac{1}{|\mathcal{P}_{\mathrm{tta}}|} \sum_{(i,j)\in\mathcal{P}_{\mathrm{tta}}} \mathrm{BCE}\big(\sigma(\tilde{s}^r_i - \tilde{s}^r_j), t_{ij}\big)}_{\text{stability-guided adaptation}} + \underbrace{\frac{\lambda_a}{N}\sum_i (\tilde{s}^r_i - s^r_i)^2}_{\text{source-score preservation}}.$$

The ranking term adapts confidence ordering to the target reliability structure, whereas the anchoring term prevents noisy stability supervision from overriding the correctness information learned on the source model. Only a_ω is updated during adaptation, while both LLMs, the SRM, TCS, and CC remain frozen. After adapting a_ω on a small set of unlabeled target samples, we freeze the corrector and apply the resulting confidence estimator to all subsequent RL trajectories without further optimization. For a new trajectory, the adapted score and final confidence are computed as

$$\tilde{s}^r = s^r + a_{\omega^*}(e^r), \qquad p^r = C_\psi(\tilde{s}^r),$$

where ω^* denotes the adapted corrector parameters.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

We use Qwen2.5-3B (Qwen et al., 2024) as our test model and Search-R1 (Jin et al., 2025a) as the Agentic RL framework. We evaluate effectiveness on TriviaQA (Joshi et al., 2017), HotpotQA (Yang et al., 2018), MuSiQue (Trivedi et al., 2021), and PopQA (Mallen et al., 2023). For cross-dataset generalization, we train on TriviaQA and test on NQ, 2WikiMultiHopQA (Ho et al., 2020),

Table 3: Main results of confidence estimation on the RL model's test trajectories.

Method	TriviaQA				HotpotQA				MuSiQue				PopQA			
	AUROC↑	ECE↓	Brier↓	NLL↓	AUROC↑	ECE↓	Brier↓	NLL↓	AUROC↑	ECE↓	Brier↓	NLL↓	AUROC↑	ECE↓	Brier↓	NLL↓
DACA	.668	**.027**	.223	.638	.613	.155	.243	.679	.625	.310	.229	.650	.680	.110	.244	.681
P(Sufficient)	.661	.331	.337	1.131	.636	.526	.498	1.551	.648	.643	.551	1.584	.641	.459	.446	1.382
HTC	.653	.289	.316	.989	.507	.175	.268	.738	**.721**	.081	**.132**	**.423**	.678	.328	.334	.937
MICE-LR	.675	.125	.232	.670	.640	.112	.234	.662	.623	**.054**	.135	.443	.595	.282	.321	.882
MICE-RF	.629	.046	.229	.651	.615	.048	.219	.629	.610	.064	.136	.442	.590	.119	.254	.702
CCPS	.420	.405	.429	1.335	.438	.205	.284	.920	.494	.114	.149	.534	.425	.306	.357	1.407
Ours	**.800**	.054	**.181**	**.542**	**.715**	**.041**	**.200**	**.582**	.672	.120	.145	.532	**.696**	**.061**	**.223**	**.635**

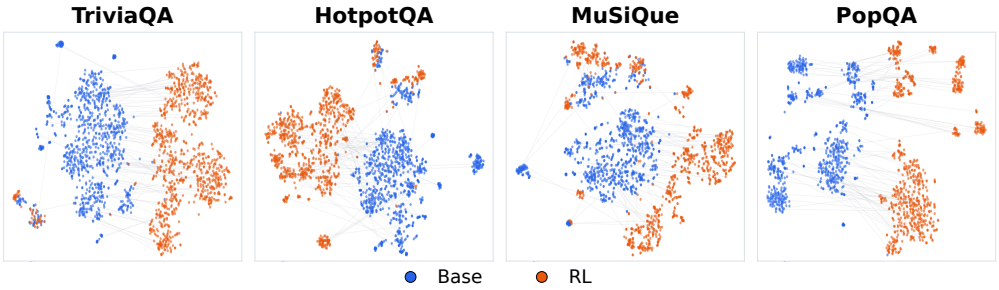

Figure 4: t-SNE visualization of the H64 hidden representations for Base and RL.

and Bamboogle (Press et al., 2023). We also test different model scales and families using Qwen2.5-7B-Instruct (Qwen et al., 2024) and Llama-3.2-3B-Instruct (Meta AI, 2024); and different RL algorithms: PPO (Havrilla et al., 2024), GRPO (Guo et al., 2025). We compare with five baselines: CCPS (Khanmohammadi et al., 2025), MICE (Subramani et al., 2025), HTC (Zhang et al., 2026b), P(Sufficient) (Zhang et al., 2026a), and DACA (Luo et al., 2026). MICE-LR and MICE-RF share internal features but use logistic regression and random forests, respectively. We report AUROC for ranking, ECE for calibration, and Brier score and NLL for probability quality (Fawcett, 2005; Guo et al., 2017; Brier, 1950; Gneiting & Raftery, 2007). We set $\lambda_a = 0.1$, $\lambda_h = 0.5$, $\lambda_r = 1.0$, and $\lambda_s = 0.5$ (details can be accessed in Appendix A.3.1 and Appendix A.3.2).

5.2 Main Results

We evaluate confidence transfer in Table 3. For each dataset, we train the confidence estimators on Base-LLM trajectories and evaluate them on Agentic RL trajectories. Across four datasets, our method achieves the best average performance: AUROC (0.721), ECE (0.069), Brier (0.187), and NLL (0.573). In terms of AUROC, only HTC outperforms our method on the MuSiQue dataset. Appendix A.4 compares the performance on both Base and RL trajectories, indicating that the poor performance of baselines is mainly due to the shift of Agentic RL. We further visualize the representation shift induced by Agentic RL. Fig. 4 shows that the Base and RL representations occupy visibly different regions of the feature space, indicating that RL training alters the internal representation adopted by the confidence estimator. We further report the baseline that directly uses stability information; the results are shown in Appendix A.2, which indicate that stability provides useful information but stability alone is insufficient. These results show that our method improves both confidence ranking and probability estimation under the influence of Agentic RL.

5.3 Generalization Analysis

5.3.1 Generalization across Datasets

We evaluate cross-dataset generalization in Table 4. We train the estimator on Base-LLM trajectories from TriviaQA and evaluate it on Agentic RL trajectories from 2WikiMultiHopQA, Bamboogle, and NQ, with TriviaQA serving as an in-domain reference. Our method achieves the highest AUROC and the lowest Brier score and NLL on all four datasets. It also achieves the lowest ECE on Bamboogle and NQ. These results demonstrate strong cross-dataset generalization.

Table 4: Cross-dataset generalization of confidence estimation.

	TriviaQA				2WikiMultiHopQA				Bamboogle				NQ			
Method	AUROC$^{\uparrow}$	ECE$^{\downarrow}$	Brier$^{\downarrow}$	NLL$^{\downarrow}$	AUROC$^{\uparrow}$	ECE$^{\downarrow}$	Brier$^{\downarrow}$	NLL$^{\downarrow}$	AUROC$^{\uparrow}$	ECE$^{\downarrow}$	Brier$^{\downarrow}$	NLL$^{\downarrow}$	AUROC$^{\uparrow}$	ECE$^{\downarrow}$	Brier$^{\downarrow}$	NLL$^{\downarrow}$
HTC	.653	.289	.316	.989	.482	.289	.337	1.333	.531	.212	.270	.909	.672	.131	.239	.737
MICE-LR	.675	.125	.232	.670	.502	.299	.346	.948	.585	.399	.379	1.027	.653	.271	.302	.837
MICE-RF	.629	.046	.229	.651	.503	.132	.268	.732	.520	.243	.289	.787	.607	.136	.250	.693
CCPS	.420	.405	.429	1.335	.453	.261	.324	1.368	.524	.220	.277	1.793	.491	.256	.320	1.791
DACA	.668	**.027**	.223	.638	.509	**.072**	.247	.686	.556	.218	.249	.692	.623	.165	.258	.710
P(Sufficient)	.661	.331	.337	1.131	.530	.373	.392	1.088	.653	.478	.432	1.204	.663	.485	.462	1.417
Ours	**.800**	.054	**.181**	**.542**	**.621**	.087	**.240**	**.673**	**.677**	**.204**	**.243**	**.686**	**.748**	**.099**	**.209**	**.605**

Table 5: Cross-model generalization: Llama-3.2-3B-Instruct with PPO .

	TriviaQA				HotpotQA				MuSiQue				PopQA			
Method	AUROC$^{\uparrow}$	ECE$^{\downarrow}$	Brier$^{\downarrow}$	NLL$^{\downarrow}$	AUROC$^{\uparrow}$	ECE$^{\downarrow}$	Brier$^{\downarrow}$	NLL$^{\downarrow}$	AUROC$^{\uparrow}$	ECE$^{\downarrow}$	Brier$^{\downarrow}$	NLL$^{\downarrow}$	AUROC$^{\uparrow}$	ECE$^{\downarrow}$	Brier$^{\downarrow}$	NLL$^{\downarrow}$
HTC	.683	.114	.233	.669	.631	**.071**	.202	.596	.688	.103	.090	.313	.611	.211	.274	.896
MICE-LR	.730	.047	.202	.593	.663	.108	.208	.597	.651	.071	.085	.300	.700	.098	.218	.622
MICE-RF	.715	**.029**	.202	.587	.681	.097	.197	.569	.661	.069	.080	.286	.713	.103	.212	.601
CCPS	.707	.105	.210	.607	.686	.126	.201	.580	.652	.103	.087	.308	.691	.083	.213	.619
DACA	.772	.204	.237	.943	.583	.152	.234	.661	.682	.394	.265	.753	.752	.297	.319	1.076
P(Sufficient)	.609	.135	.254	.703	.652	.285	.323	.982	.564	.354	.222	.634	.615	**.064**	.234	.659
Ours	**.833**	.047	**.167**	**.503**	**.746**	.074	**.179**	**.528**	**.744**	**.045**	**.074**	**.260**	**.830**	.093	**.174**	**.516**

5.3.2 Generalization across Models

We evaluate generalization across different model families using Llama-3.2-3B-Instruct (Table 5). We retrain the confidence estimator on the new model, while keeping all other configurations fixed. On Llama-3.2-3B-Instruct, we achieve the best AUROC, Brier score, and NLL on TriviaQA, HotpotQA, MuSiQue, and PopQA, with AUROC gains of 5.6–7.8 percentage points over the strongest baseline. We also evaluate generalization across model scales using Qwen2.5-7B-Instruct (Table 6) to test our metod. Under the 7B backbone, StaR-TTA achieves the highest AUROC on all four datasets, exceeding the strongest baseline by 1.1–10.2 percentage points, and obtains the lowest Brier score and NLL. These results demonstrate the effectiveness of our method across LLM scales and families.

5.3.3 Generalization across RL Algorithms

Finally, we verify that our findings are not specific to the PPO algorithm used for Agentic RL training. Table 7 reports results on GRPO algorithms, while keeping the same Base LLM and evaluation protocol. We achieve the best AUROC on all four datasets (0.756 on TriviaQA, 0.724 on HotpotQA, 0.676 on MuSiQue, and 0.750 on PopQA), improving over the strongest baseline by 3.4 to 9.2 points, and additionally attain the best Brier and NLL on HotpotQA, MuSiQue, and PopQA. These results confirm that trajectory shift is a general phenomenon of RL training and that trajectory-based confidence estimation consistently benefits across different RL algorithms.

5.4 Ablation Studies

5.4.1 Effect of Test-Time Adaption

We compare the full method with a variant that removes TTA and directly maps the original student scores through the same frozen calibrator. Table 8 reports the actual metrics for both settings. TTA improves AUROC on HotpotQA, MuSiQue, and PopQA, while slightly reducing it on TriviaQA. The average AUROC increases by approximately 1.73 percentage points. Average Brier score and NLL also improve, whereas average ECE increases

Dataset	Setting	AUROC↑	ECE↓	Brier↓	NLL↓
TriviaQA	w/o TTA	0.8018	0.0455	0.1794	0.5386
	w/ TTA	0.8000	0.0535	0.1808	0.5421
HotpotQA	w/o TTA	0.7112	0.0349	0.1997	0.5828
	w/ TTA	0.7154	0.0409	0.1996	0.5824
MuSiQue	w/o TTA	0.6296	0.1212	0.1478	0.5503
	w/ TTA	0.6721	0.1198	0.1450	0.5316
PopQA	w/o TTA	0.6714	0.0600	0.2273	0.6464
	w/ TTA	0.6956	0.0610	0.2226	0.6354
Average	w/o TTA	0.7035	0.0654	0.1885	0.5795
	w/ TTA	0.7208	0.0688	0.1870	0.5729

Table 8: Effect of removing TTA.

Table 6: Cross-model generalization: Qwen2.5-7B-Instruct with PPO.

	TriviaQA				HotpotQA				MuSiQue				PopQA			
Method	AUROC$^{\uparrow}$	ECE$^{\downarrow}$	Brier$^{\downarrow}$	NLL$^{\downarrow}$	AUROC$^{\uparrow}$	ECE$^{\downarrow}$	Brier$^{\downarrow}$	NLL$^{\downarrow}$	AUROC$^{\uparrow}$	ECE$^{\downarrow}$	Brier$^{\downarrow}$	NLL$^{\downarrow}$	AUROC$^{\uparrow}$	ECE$^{\downarrow}$	Brier$^{\downarrow}$	NLL$^{\downarrow}$
HTC	.717	.355	.353	2.684	.647	.152	.246	.710	.794	.149	.161	.477	.697	.075	.220	.630
MICE-LR	.641	.064	.220	.639	.614	.096	.236	.657	.673	.041	.140	.440	.581	.068	.242	.674
MICE-RF	.669	**.036**	.209	.607	.629	.087	.228	.638	.645	.061	.146	.456	.652	**.059**	.227	.639
CCPS	.655	.059	.202	.587	.659	**.078**	.224	.631	.672	.087	.147	.458	.625	.146	.253	.697
DACA	.725	.265	.268	2.005	.657	.372	.379	1.823	.697	.315	.314	1.155	.676	.389	.387	2.679
P(Sufficient)	.750	.258	.267	1.603	.704	.363	.371	2.249	.730	.369	.362	1.991	.691	.380	.381	2.487
Ours	**.822**	.112	**.180**	**.539**	**.769**	.081	**.197**	**.569**	**.805**	**.038**	**.121**	**.382**	**.799**	.103	**.193**	**.567**

Table 7: Generalization across RL algorithms: Qwen2.5-3B with GRPO.

	TriviaQA				HotpotQA				MuSiQue				PopQA			
Method	AUROC$^{\uparrow}$	ECE$^{\downarrow}$	Brier$^{\downarrow}$	NLL$^{\downarrow}$	AUROC$^{\uparrow}$	ECE$^{\downarrow}$	Brier$^{\downarrow}$	NLL$^{\downarrow}$	AUROC$^{\uparrow}$	ECE$^{\downarrow}$	Brier$^{\downarrow}$	NLL$^{\downarrow}$	AUROC$^{\uparrow}$	ECE$^{\downarrow}$	Brier$^{\downarrow}$	NLL$^{\downarrow}$
HTC	.663	.321	.331	1.171	.422	.245	.274	1.039	.551	.844	.767	2.479	.636	.579	.570	2.613
MICE-LR	.640	.150	.251	.716	.624	.149	.230	.657	.567	.051	.052	.217	.524	.267	.327	.890
MICE-RF	.564	.080	.244	.681	.605	**.068**	.204	.593	.587	.079	.057	.234	.521	.112	.254	.701
CCPS	.425	.536	.546	4.192	.437	.259	.275	1.661	.483	.210	.104	.434	.467	.318	.354	2.919
DACA	.692	.093	.232	.655	.611	.244	.261	.716	.642	.465	.263	.718	.697	.134	.252	.696
P(Sufficient)	.647	.368	.370	1.343	.632	.631	.604	2.091	.613	.840	.768	2.532	.671	.521	.504	1.713
Ours	**.756**	**.052**	**.199**	**.585**	**.724**	.114	**.192**	**.567**	**.676**	**.042**	**.048**	**.199**	**.750**	**.097**	**.204**	**.596**

slightly. These results show that TTA mainly benefits confidence ranking, with mixed effects on calibration. We also report ablation results on external descriptors d_i^m, stability vectors c_i^m, q^{tta}, and our training loss in Appendix A.1, where we find removing any of them will decrease the performance.

5.4.2 TTA Sample Efficiency

Our TTA procedure uses stability-based pairwise rankings to adjust confidence scores. More unlabeled target questions can provide more informative comparisons and make results more reliable. We therefore vary the number of questions used for adaptation, as shown in Table 9. With only a small number of adaptation samples (e.g., $N = 2$ or $N = 4$), the results become unstable and even decrease. Performance gradually recovers and stabilizes as we increase the number of unlabeled questions for adaptation. The average AUROC rises from 0.7035 without TTA to 0.7155 with 64 questions. We can find that only a small number of samples (e.g., 64) during TTA is sufficient. These results suggest that additional unlabeled examples improve the reliability of ranking-based adaptation, with diminishing gains as the sample budget increases. We further analyze the additional cost introduced by our method in Appendix A.5. The results show that the stability signal only introduces an extra token overhead of merely 0.8%.

N	Average	Trivia	Hotpot	MuSiQue	PopQA
0	0.7035	0.8018	0.7112	0.6296	0.6714
1	0.7035	0.8018	0.7112	0.6296	0.6714
2	0.6306	0.7537	0.5125	0.6383	0.6181
4	0.6680	0.7828	0.5744	0.6671	0.6477
8	0.7089	0.7911	0.6839	0.6683	0.6924
12	0.7072	0.7914	0.6893	0.6703	0.6779
16	0.7056	0.7943	0.6811	0.6734	0.6738
32	0.7120	0.8002	0.7112	0.6709	0.6658
64	0.7155	0.8014	0.7151	0.6734	0.6723
All	0.7208	0.8000	0.7154	0.6721	0.6956

Table 9: AUROC under different TTA sample sizes.

6 Conclusion

We studied why confidence estimators trained on Base LLM trajectories degrade after Agentic RL. We find that RL-induced trajectory shifts account for a substantial fraction of the degradation. Despite this shift, trajectory stability remains predictive of answer correctness in both Base and RL models. This preserved relationship motivates StaR-TTA, a label-free method that adjusts confidence scores on Agentic RL trajectories at test time. Across four datasets, StaR-TTA achieves an average AUROC of 0.721. It also generalizes across datasets, model families, and PPO- or GRPO-trained agents. These results suggest that StaR-TTA may provide a useful starting point for further research on confidence estimation in Agentic-RL trained LLM agents.

AI USE STATEMENT

In this work, we used generative AI tools for providing feedback on research methodology. Additionally, we used generative AI tools for edit a research paper to improve readability. We have reviewed all AI-assisted work. We checked LLM-generated feedback and we manually polished the content generated by the LLM. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

G. W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78, 1950.

Mingyang Chen, Linzhuang Sun, Tianpeng Li, Haoze Sun, Yijie Zhou, Chenzheng Zhu, Haofen Wang, Jeff Z Pan, Wen Zhang, Huajun Chen, et al. Learning to reason with search for llms via reinforcement learning. *arXiv preprint arXiv:2503.19470*, 2025.

Fangming Cui, Ruixiao Zhu, Cheng Fang, Sunan Li, and Jiahong Li. Rethinking agentic reinforcement learning in large language models. *arXiv preprint arXiv:2604.27859*, 2026.

Rodrigo de Oliveira, Matthew Garber, James M Gwinnutt, Emaan Rashidi, Jwu-Hsuan Hwang, William Gilmour, Jay Nanavati, Khaldoun Zine El Abidine, and Christina DeFilippo Mack. A study of calibration as a measurement of trustworthiness of large language models in biomedical natural language processing. *JAMIA open*, 8(4):ooaf058, 2025.

Tom Fawcett. An introduction to roc analysis. *Pattern Recognition Letters*, 27(8):861–874, 2005.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Publications of the American Statistical Association*, 102(March):359–378, 2007.

Jiannan Guan, Qiguang Chen, Libo Qin, Dengyun Peng, Jinhao Liu, Liangyu Huo, Jian Xie, and Wanxiang Che. Beware of reasoning overconfidence: Pitfalls in the reasoning process for multi-solution tasks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 30843–30851, 2026.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. 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.

Alex Havrilla, Yuqing Du, Sharath Chandra Raparthy, Christoforos Nalmpantis, Jane Dwivedi-Yu, Maksym Zhuravinskyi, Eric Hambro, Sainbayar Sukhbaatar, and Roberta Raileanu. Teaching large language models to reason with reinforcement learning. *arXiv preprint arXiv:2403.04642*, 2024.

Xanh Ho, Anh Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps. 2020.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. 2025a.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516*, 2025b.

Jiahe Jin, Abhijay Paladugu, and Chenyan Xiong. Beneficial reasoning behaviors in agentic search and effective post-training to obtain them. *arXiv preprint arXiv:2510.06534*, 2025c.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. 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.

Beier Luo, Shuoyuan Wang, Sharon Li, and Hongxin Wei. Your pre-trained llm is secretly an unsupervised confidence calibrator. *Advances in Neural Information Processing Systems*, 38: 163514–163543, 2026.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9802–9822, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.546. URL https://aclanthology.org/2023.acl-long.546/.

Kohsei Matsutani, Shota Takashiro, Gouki Minegishi, Takeshi Kojima, Yusuke Iwasawa, and Yutaka Matsuo. Rl squeezes, sft expands: A comparative study of reasoning llms. In *International Conference on Learning Representations*, volume 2026, pp. 141482–141526, 2026.

Meta AI. Llama 3.2: Revolutionizing edge ai and vision with open, customizable models. Meta AI Blog, September 2024. URL https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/. Accessed: 2026-09-14.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah Smith, and Mike Lewis. Measuring and narrowing the compositionality gap in language models. *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. doi: 10.48550/arXiv.2412.15115. URL https://arxiv.org/abs/2412.15115.

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, 2024. *URL https://arxiv. org/abs/2402.03300*, 2(3):5, 2024.

Joykirat Singh, Raghav Magazine, Yash Pandya, and Akshay Nambi. Agentic reasoning and tool integration for llms via reinforcement learning. *arXiv preprint arXiv:2505.01441*, 2025.

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.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5433–5442, 2023.

H. 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, 2021.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36:74952–74965, 2023.

Tianxin Wei, Ting-Wei Li, Zhining Liu, Xuying Ning, Ze Yang, Jiaru Zou, Zhichen Zeng, Ruizhong Qiu, Xiao Lin, Dongqi Fu, et al. Agentic reasoning for large language models. *arXiv preprint arXiv:2601.12538*, 2026.

Quan Xiao, Debarun Bhattacharjya, Balaji Ganesan, Radu Marinescu, Katsiaryna Mirylenka, Nhan H Pham, Michael Glass, and Junkyu Lee. The consistency hypothesis in uncertainty quantification for large language models. *arXiv preprint arXiv:2506.21849*, 2025.

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.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. 2018.

Caiqi Zhang, Ruihan Yang, Xiaochen Zhu, Chengzu Li, Tiancheng Hu, Yijiang River Dong, Deqing Yang, and Nigel Collier. Confidence estimation for llms in multi-turn interactions. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 25661–25676, 2026a.

Jiaxin Zhang, Caiming Xiong, and Chien-Sheng Wu. Agentic confidence calibration. *arXiv preprint arXiv:2601.15778*, 2026b.

A Appendix

Table 10: Independent ablations of Scorer input features. Each ablation removes only one feature group.

Dataset	Setting	AUROC↑	ECE↓	Brier↓	NLL↓
TriviaQA	w/o Stability input c^m	0.7413	0.1529	0.2250	0.6408
	w/o Statistics d^m	0.7871	0.0473	0.1850	0.5526
	Ours	0.8000	0.0535	0.1808	0.5421
HotpotQA	w/o Stability input c^m	0.6170	0.1222	0.2337	0.6679
	w/o Statistics d^m	0.6992	0.0724	0.2079	0.6030
	Ours	0.7154	0.0409	0.1996	0.5824
MuSiQue	w/o Stability input c^m	0.6578	0.1241	0.1478	0.5422
	w/o Statistics d^m	0.6659	0.1061	0.1411	0.5024
	Ours	0.6721	0.1198	0.1450	0.5316
PopQA	w/o Stability input c^m	0.6645	0.2693	0.3009	0.8716
	w/o Statistics d^m	0.6675	0.0311	0.2259	0.6438
	Ours	0.6956	0.0610	0.2226	0.6354

A.1 Ablation Studies

To analyze the contribution of different components in our framework, we conduct ablation studies on both the trajectory representation and the test-time adaptation strategy. For trajectory feature construction, we incorporate three types of signals: internal representations h^m, external descriptors d^m, and stability vectors c^m. Since using hidden-layer representations h^m as trajectory features is a natural choice, we further investigate whether the additional stability and statistical signals provide complementary benefits.

During TTA, we construct q^{tta} to guide the ranking-based adaptation, which leverages both the stability vector c^m and the agreement between Agentic-RL and Base trajectories. This differs from the training stage, where only c^m is used. Therefore, we examine whether introducing the Base–RL agreement brings additional improvements.

During training, we use correctness and ranking losses based on ground-truth labels, along with two corresponding losses constructed from label-free stability signals. We further conduct ablation experiments to assess the individual contributions of these four loss terms.

A.1.1 Scorer Input Features

We independently remove stability features, and the 24-dimensional interaction statistics from the TCS input. Each ablation retains the other input features and follows the same source-training and calibration protocol. Stability signals remain available to the SRM, training losses, and TTA. Table 10 shows that removing stability inputs worsens all four metrics on every dataset. Removing interaction statistics also lowers AUROC on all four datasets, although ECE improves on three datasets and MuSiQue obtains lower Brier score and NLL. These results support retaining both feature groups, while showing that their benefits vary across metrics.

A.1.2 Base–RL Agreement during Adaptation

Table 11: Effect of removing Base–RL JS agreement from TTA.

Dataset	Setting	AUROC↑	ECE↓	Brier↓	NLL↓
TriviaQA	w/o JS agreement	0.8000	0.0521	0.1816	0.5441
	Ours	0.8000	0.0535	0.1808	0.5421
HotpotQA	w/o JS agreement	0.7067	0.0402	0.2022	0.5885
	Ours	0.7154	0.0409	0.1996	0.5824
MuSiQue	w/o JS agreement	0.6743	0.1198	0.1449	0.5310
	Ours	0.6721	0.1198	0.1450	0.5316
PopQA	w/o JS agreement	0.6927	0.0633	0.2238	0.6383
	Ours	0.6956	0.0610	0.2226	0.6354

We remove Base–RL JS agreement from the target quality signal and use only the mean RL stability, $q^{\mathrm{tta}} = S_{\mathrm{raw3}}$, where $\mathcal{S}_{\mathrm{raw3}}$ is the equal-weight average of the stability vector. The SRM, confidence calibrator, and other TTA settings remain unchanged. As shown in Table 11, removing agreement lowers AUROC on HotpotQA and PopQA, while leaving TriviaQA nearly unchanged and slightly improving MuSiQue. Brier score and NLL increase on TriviaQA, HotpotQA, and PopQA. These results indicate that Base–RL agreement provides useful information beyond RL stability, although its benefit varies across datasets.

A.1.3 Training Loss Ablation

To examine the contribution of each training loss, we remove one loss term at a time and evaluate the resulting variants on PopQA. Table 12 reports the average relative degradation across the four metrics, using the full method as the reference. Removing Label BCE produces the largest average degradation of 25.03%, followed by removing Stability Ranking at 21.22%. These results suggest that combining direct correctness supervision with stability-based ranking supervision is important for the aggregate performance of the confidence estimator. Removing Label Ranking and SRM BCE

Table 12: Ablation of individual training losses on PopQA. Average degradation denotes the mean relative degradation across the four metrics compared with the full method.

Setting	AUROC↑	ECE↓	Brier↓	NLL↓	Average degradation (%)↓
w/o Label BCE	0.6952	0.1151	0.2372	0.6659	25.03
w/o Label Ranking	0.6886	0.0647	0.2248	0.6408	2.23
w/o Stability Ranking	0.7180	0.1131	0.2250	0.6457	21.22
w/o SRM BCE	0.6972	0.0634	0.2223	0.6345	0.86
Full method	0.6956	0.0610	0.2226	0.6354	0.00

leads to smaller average degradations of 2.23% and 0.86%, respectively, indicating more modest contributions under this aggregate measure. Although the effects differ in magnitude, all ablated variants exhibit positive average degradation, supporting the combined training objective in this setting.

A.2 Direct Confidence Estimation from Stability

We investigate whether stability signals alone suffice for confidence estimation, potentially providing a simpler alternative to our full framework. We compare five stability-only predictors, each using the three-dimensional stability vector with no hidden layers: logistic regression, probit regression, complementary log-log regression, a sigmoid output trained with Brier loss, and ridge regression with predictions clipped to the probability range. Each predictor is trained on labeled Base LLM trajectories, with regularization selected by Base-validation NLL, and evaluated on the RL test trajectories. Table 13 reports equal-weight averages over TriviaQA, HotpotQA, MuSiQue, and PopQA. Our method achieves the best average performance on all four metrics, supporting the integration of stability signals with trajectory information.

Table 13: Confidence estimation averaged equally across four datasets.

Method	AUROC↑	ECE↓	Brier↓	NLL↓
Logistic regression	0.6803	0.0783	0.1950	0.5811
Probit regression	0.6807	0.0781	0.1958	0.5855
Complementary log-log	0.6802	0.0791	0.1953	0.5825
Sigmoid (Brier loss)	0.6821	0.0736	0.1942	0.5775
Clipped ridge regression	0.6803	0.0699	0.1944	0.5782
Ours	**0.7208**	**0.0688**	**0.1870**	**0.5729**

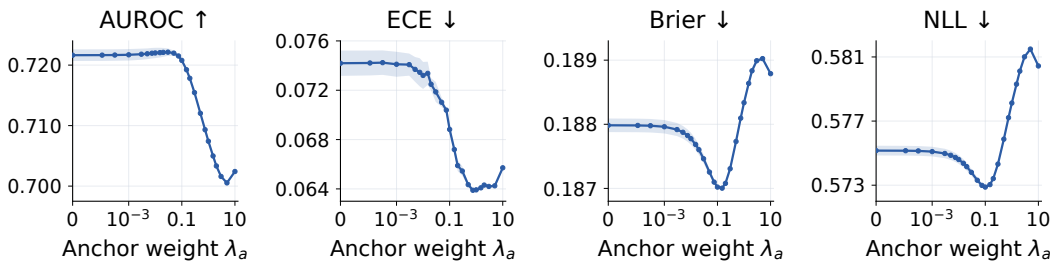

Figure 5: Sensitivity to the anchoring weight λ_a. Curves show the equal-weight average over four datasets.

A.3 Hyperparameter Analysis

A.3.1 Robustness to the Anchoring Weight

During TTA, the corrector is trained with a ranking loss and an anchoring loss. The ranking loss encourages scores to follow the stability-based ordering, while the anchoring loss limits their deviation from the initial student scores. Their balance is controlled by λ_a in $\mathcal{L}_{\mathrm{tta}} = \mathcal{L}_{\mathrm{rank}} + \lambda_a \mathcal{L}_{\mathrm{anchor}}$. A smaller weight allows stronger corrections, whereas a larger weight favors the original predictions. Figure 5 varies λ_a while keeping the other settings fixed on training set. AUROC peaks near 0.03, while Brier score and NLL reach their lowest values around 0.1–0.15. Larger weights generally reduce AUROC and eventually worsen Brier score and NLL, although ECE remains lower. We use $\lambda_a = 0.1$ as a compromise between ranking and probability quality.

A.3.2 Robustness to Training Loss Weights

In our experiments, we set $\lambda_{\mathrm{h}} = 0.5$, $\lambda_r = 1.0$, and $\lambda_s = 0.5$. To assess robustness to these hyperparameters, we substantially vary each weight individually: doubling λ_{h} to 1.0, reducing λ_r to 0.2, and doubling λ_s to 1.0, while keeping all other settings fixed. As shown in Table 14, the mean absolute change across the three perturbations and four evaluation metrics is below 0.003, suggesting that our method is relatively insensitive to the tested variations in loss weights.

Table 14: Robustness to training loss weightss.

λ_{h}	λ_r	λ_s	AUROC↑	ECE↓	Brier↓	NLL↓	Change
0.5	1.0	0.5	0.7208	0.0688	0.1870	0.5729	-
1.0↑	1.0	0.5	0.7227	0.0656	0.1860	0.5662	-0.0023
0.5	0.2↓	0.5	0.7229	0.0711	0.1870	0.5785	$+0.0025$
0.5	1.0	1.0↑	0.7193	0.0710	0.1879	0.5786	$+0.0018$

A.4 Confidence Estimation on Base and RL Trajectories

Table 15 compares confidence estimation on held-out Base and RL trajectories. We evaluate our source-trained TCS and frozen calibrator on Base trajectories, and report the full TTA method on RL trajectories. Several baselines achieve strong performance on Base trajectories but deteriorate substantially when transferred to RL trajectories. For example, CCPS achieves AUROC values of 0.7868 on TriviaQA and 0.7329 on HotpotQA before transfer, which drop to 0.4197 and 0.4381 after transfer. Similarly, MICE-LR and MICE-RF exhibit lower AUROC on RL trajectories across all four datasets, despite their strong performance on Base trajectories. These comparisons show existing estimators can successfully estimate confidence on held-out trajectories from the Base model, yet their learned confidence signals become less reliable after Agentic RL changes the trajectory distribution.

A.5 Cost Analysis

We distinguish standard source training and confidence scoring from the additional costs of stability estimation and test-time adaptation. At test time, we freeze TCS and CC then update only a linear corrector using unlabeled target examples. Paired Base and RL trajectories are required only for training the corrector. Once corrector is trained, confidence scoring uses RL trajectories alone, without running the Base model on subsequent questions. Table 9 shows that 64 adaptation examples improve average AUROC from 0.7035 to 0.7155, compared with 0.7208 using all available examples. Thus, a small adaptation subset can provide useful gains while limiting the cost of Base-model execution. Other confidence estimators may need to perform retraining on a large collection of RL trajectories (e.g., tens of thousands). Without this retraining, they undergo substantial performance degradation.

For stability estimation, we fix each trajectory prefix and sample five short final answers, avoiding repeated reasoning and tool calls. We estimate the additional answer tokens cost using 100 randomly selected Base trajectories per dataset. As reported in Table 16, the five answers contain 46.1 tokens

Table 15: Confidence estimation on Base and RL test trajectories.

	Base trajectories				RL trajectories			
Method	AUROC↑	ECE↓	Brier↓	NLL↓	AUROC↑	ECE↓	Brier↓	NLL↓
TriviaQA								
DACA†	0.6979	0.2257	0.2625	0.7223	0.6675	0.0265	0.2232	0.6376
P(Sufficient)†	0.6815	0.4786	0.4497	1.3340	0.6605	0.3306	0.3372	1.1313
HTC	0.7976	0.0162	0.1765	0.5391	0.6529	0.2887	0.3161	0.9893
MICE-LR	0.7723	0.0223	0.1871	0.5540	0.6754	0.1249	0.2324	0.6703
MICE-RF	0.7885	0.0301	0.1812	0.5312	0.6290	0.0461	0.2294	0.6506
CCPS	0.7868	0.0224	0.1813	0.5370	0.4197	0.4045	0.4286	1.3352
Ours	0.7840	0.0210	0.1806	0.5371	0.8000	0.0535	0.1808	0.5421
HotpotQA								
DACA†	0.6517	0.3868	0.2737	0.7441	0.6128	0.1547	0.2432	0.6793
P(Sufficient)†	0.6496	0.6698	0.5740	1.6223	0.6363	0.5262	0.4983	1.5509
HTC	0.7397	0.0304	0.1104	0.3619	0.5067	0.1750	0.2684	0.7382
MICE-LR	0.7268	0.0321	0.1146	0.3726	0.6403	0.1122	0.2341	0.6623
MICE-RF	0.7530	0.0504	0.1101	0.3571	0.6149	0.0475	0.2190	0.6292
CCPS	0.7329	0.0347	0.1098	0.3634	0.4381	0.2048	0.2842	0.9195
Ours	0.7482	0.0293	0.1084	0.3551	0.7154	0.0409	0.1996	0.5824
MuSiQue								
DACA†	0.6813	0.4619	0.2616	0.7165	0.6253	0.3103	0.2285	0.6497
P(Sufficient)†	0.6402	0.7456	0.6080	1.6726	0.6475	0.6425	0.5505	1.5842
HTC	0.7055	0.0186	0.0314	0.1398	0.7209	0.0808	0.1317	0.4229
MICE-LR	0.7026	0.0174	0.0331	0.1428	0.6231	0.0540	0.1350	0.4434
MICE-RF	0.6618	0.0409	0.0368	0.1599	0.6101	0.0637	0.1360	0.4420
CCPS	0.6640	0.0337	0.0345	0.1540	0.4936	0.1140	0.1485	0.5337
Ours	0.7610	0.0088	0.0303	0.1293	0.6721	0.1198	0.1450	0.5316
PopQA								
DACA†	0.6420	0.2220	0.2519	0.6987	0.6804	0.1098	0.2438	0.6806
P(Sufficient)†	0.6615	0.5181	0.4706	1.3192	0.6405	0.4589	0.4459	1.3821
HTC	0.7439	0.0658	0.1766	0.5338	0.6779	0.3277	0.3337	0.9366
MICE-LR	0.7361	0.0477	0.1830	0.5431	0.5953	0.2816	0.3211	0.8815
MICE-RF	0.7937	0.0384	0.1636	0.4918	0.5903	0.1186	0.2539	0.7015
CCPS	0.7432	0.0392	0.1803	0.5307	0.4249	0.3060	0.3573	1.4068
Ours	0.7311	0.0643	0.1791	0.5347	0.6956	0.0610	0.2226	0.6354

on average, or approximately 9.2 tokens per answer, compared with 5,824 tokens in the original trajectory. The resulting 0.8% ratio measures the token increment from answer resampling.

Table 16: Token usage of original trajectories and five sampled answers.

Dataset	Total original tokens	Five answers tokens	Five answers / Total original (%)
TriviaQA	5,034.5	45.0	0.9
HotpotQA	5,289.3	45.8	0.9
MuSiQue	6,330.2	47.0	0.7
PopQA	6,642.5	46.8	0.7
Overall	5,824.1	46.1	0.8

A.6 Experimental Details

The source-training/source-validation/RL-test counts are 1,502/475/11,122 for TriviaQA, 1,486/478/6,687 for HotpotQA, 1,458/457/1,755 for MuSiQue, and 1,475/463/11,712 for PopQA.

As for 64-dimensional hidden representation, we use layers 28 and 36 for Qwen2.5-3B, and layers 24 and 28 for Qwen2.5-7B-Instruct and Llama-3.2-3B-Instruct. The SRM uses three-input logistic regression. The TCS is a two-branch MLP with two hidden layers per branch, whose scalar outputs are summed. It is trained using AdamW with learning rate 10^{-3}, weight decay 10^{-4}, batch size 128, and source-ranking threshold 0.3. The CC uses Platt scaling, $p = \sigma(as + b)$, fitted on source OOF scores. We sample five terminal answers with temperature 1.0, top-$p = 0.9$, top-$k = 50$. The corrector is a zero-initialized $288 \to 1$ linear layer which is trained for 20 Adam steps with learning rate 0.03. We judge whether the answer is correct by using exact match against any reference after lowercasing, removing punctuation and articles, and normalizing whitespace.