STRONG YET CONTROLLABLE LLM ABSTENTION WITHOUT LEARNING TO ABSTAIN

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) can reduce hallucination by abstaining from questions they would likely get wrong. Abstention can be implemented outside the model (selective answering) using a confidence threshold that controls the answer rate (coverage), or learned inside it (abstention training), which can perform better but fixes coverage once trained. To compare the two, we match their coverage and decompose risk (error rate) into answering ability and distinguishing ability, *i.e.*, how well the model answers questions and how accurately it abstains. The decomposition shows that abstention training performs better mainly by answering better: with strong confidence estimators, the two have nearly equal distinguishing ability. This motivates **T**rain for **A**ccuracy, then **P**ost-hoc **S**elect (**TAPS**), a modular approach that keeps coverage controllable, needs only standard post-training and an off-the-shelf confidence estimator, and benefits from advances in either component. Across three LLMs and three factual QA benchmarks, TAPS with a probing-based confidence estimator attains 1.5–9.4 pp lower mean risk than three representative abstention training methods at matched coverage. Within TAPS, accuracy training often weakens the model's own confidence signals, yet refitting the estimator largely preserves distinguishing ability. Together, these results establish TAPS as a strong yet controllable baseline and raise the bar for future abstention training: sacrificing control over coverage should yield gains beyond what can be achieved by optimizing answering and abstaining separately. The source code will be made publicly available upon acceptance.

1 INTRODUCTION

Hallucination remains a central challenge to the reliable deployment of large language models (LLMs) (Ji et al., 2023). A direct way to mitigate this problem is *abstention*: declining questions the model is likely to answer incorrectly (Wen et al., 2025). Abstention trades quantity for quality: the model answers fewer questions, but the answers it gives are more likely to be correct. Existing methods fall into two paradigms according to where the abstention decision is made. *Selective answering* decides outside the model, after generation: the model first generates an answer, and a separate confidence estimator withholds it if its confidence falls below a threshold (Xin et al., 2021; Yoshikawa & Okazaki, 2023). *Abstention training* instead internalizes the decision: during generation, the model itself learns both how to answer correctly and when to abstain (Zhang et al., 2024; Wei et al., 2026; Zhou et al., 2026).

The two paradigms exhibit complementary strengths and limitations. Selective answering is simple and flexible. It reduces abstention to confidence estimation without modifying the model, and adjusting the threshold controls how conservatively it answers, thereby allowing a single model to serve applications with different tolerances for error (Liu et al., 2022; Srinivasan et al., 2024). In contrast, abstention training has shown promising results in reducing hallucination (Wei et al., 2026), though each trained model is tied to a single, intrinsic answer rate. Moreover, recent studies find that even under different training incentives, the answer rates of such models tend to collapse into a narrow, conservative range (Che et al., 2026; Zeng et al., 2026), as we also observe (§6.2, Appendix B). Che et al. (2026) further attribute this collapse to the optimization dynamics of abstention training itself. This lack of control also persists at inference time, where prompting a trained model to vary its answer rate remains challenging (Wu et al., 2026). A natural question, then, is *whether abstention can be both strong and controllable*.

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.

Answering this question first requires comparing the two paradigms on common ground, since they have developed largely in parallel and are evaluated differently. Abstention training studies typically report a single score, such as accuracy, hallucination rate, their difference, or accuracy on answered questions alone (Zhang et al., 2024; Wei et al., 2026; Zeng et al., 2026). Consider one model that answers 5% of questions, 99% of them correctly, and another that answers 80%, 90% of them correctly. The first has the lower hallucination rate (0.05% vs. 8%) and the second the higher accuracy (72% vs. 5%). Yet neither score reveals which abstains better, since the two models operate at different answer rates: the second might also achieve 99% accuracy if limited to its 5% most confident answers. Thus, comparing abstention quality requires holding the answer rate fixed. We therefore adopt the evaluation protocol of selective prediction (Geifman & El-Yaniv, 2017), which separately tracks *coverage*, the fraction of questions answered, and *risk*, the error rate among them. In this risk–coverage plane, selective answering traces a curve as its threshold varies, whereas an abstention-trained model occupies a single point, so we compare their risks at that point's coverage, which we call *matched coverage*.

Our preliminary experiments confirm that TruthRL (Wei et al., 2026), a representative RL-based abstention training method, attains lower mean risk at matched coverage than selective answering applied to the same model before abstention training. We note that, at fixed coverage, lower risk can come from only two sources: generating fewer wrong answers or withholding answers that would be wrong rather than correct ones. This motivates an exact decomposition of risk into these two sources, which we term *answering ability* and *distinguishing ability*, respectively. The decomposition shows that TruthRL's advantage in distinguishing ability shrinks as selective answering uses stronger confidence estimators and even vanishes on average with the strongest. What remains comes entirely from answering ability. Despite its name, then, what sets abstention training apart from selective answering is how it answers, not how it abstains.

This raises a question: *what if selective answering were built on an equally strong answerer*? We therefore introduce ***T**rain for **A**ccuracy, then **P**ost-hoc **S**elect* (**TAPS**), a simple two-stage approach that first trains the model solely for answer accuracy and then leaves the abstention decision to a post-hoc confidence estimator. Because the decision stays outside the model, TAPS preserves controllability: coverage is adjusted through a threshold, just as in selective answering. It is also practical and modular: task-specific post-training is already standard practice, and abstention then requires only an off-the-shelf confidence estimator rather than a specialized training objective. Moreover, because the two stages draw on well-developed and largely independent fields, TAPS can directly incorporate advances in either.

We evaluate TAPS on three representative open-source LLMs and three factual QA benchmarks against three abstention training methods, R-Tuning (Zhang et al., 2024), TruthRL (Wei et al., 2026), and UCPO (Zeng et al., 2026), all trained from the same starting model on the same data as TAPS. At matched coverage, TAPS with a lightweight probing-based estimator (Orgad et al., 2025) attains 1.5 percentage points (pp) lower mean risk than TruthRL (winning 17 of 27 comparisons) and 8.1 and 9.4 pp lower than R-Tuning and UCPO, respectively. Crucially, it does so without sacrificing controllability, whereas TruthRL's coverage barely moves even when it is retrained with different abstention rewards. Further analysis highlights the complementary roles of TAPS's two stages. At the first stage, accuracy training emerges as the main contributor, lowering risk in 81 of 90 matched-coverage comparisons relative to selective answering without it, even when it weakens the model's own confidence signals. At the second stage, a supervised probing-based estimator refit on the trained model largely preserves distinguishing ability despite this weakening, and more generally, stronger estimators yield lower risk. These results do not argue against abstention training; rather, they establish TAPS as a strong baseline for isolating the gains specific to learning to abstain.

Our main contributions are summarized as follows:

- We compare abstention training and selective answering under a common risk–coverage evaluation protocol with controlled training conditions, and derive an exact decomposition of their risk into answering ability and distinguishing ability.
- This decomposition shows abstention training learns when to abstain and how to answer better, but only the latter sets it apart from selective answering with a strong confidence estimator.
- We establish TAPS as a strong and controllable baseline. Without any abstention-specific objective, it attains lower mean risk than three representative abstention training methods at matched coverage while keeping coverage adjustable.

2 Related Work

LLM abstention inherits a distinction between internalizing and externalizing the reject decision from classical selective prediction, which we review in Appendix A.

Training language models to abstain. Abstention training internalizes the abstention decision into the model policy. Early *supervised methods* train models on explicit, question-level abstain targets constructed from model-specific knowledge boundaries (Zhang et al., 2024; Yang et al., 2024; Cheng et al., 2024). More recent *reinforcement-learning methods* avoid this dataset construction cost by optimizing on-policy abstention behavior using outcome rewards (Wei et al., 2026; Mohamadi et al., 2025; Zhou et al., 2026), calibration rewards (Wu et al., 2026), or modified advantages (Zeng et al., 2026; Pan et al., 2026). Despite their success in mitigating hallucination, trained policies tend to collapse toward narrow ranges of coverage (Che et al., 2026; Zeng et al., 2026), a problem Che et al. (2026) trace to the optimization dynamics of abstention itself and thus find difficult to resolve.

Selective answering for language models. In contrast, selective answering externalizes abstention by thresholding answer confidence (Kamath et al., 2020; Xin et al., 2021; Yoshikawa & Okazaki, 2023). Works along this line focus on improving the confidence signal used to rank fixed answers. *Unsupervised methods* derive confidence from prompted self-evaluation, such as P(True) (Kadavath et al., 2022) and verbalized confidence (Tian et al., 2023), or from uncertainty across sampled responses (Kuhn et al., 2023; Farquhar et al., 2024; Cole et al., 2023). *Supervised methods* predict answer correctness using external calibrators (Kamath et al., 2020; Varshney et al., 2022) or hidden-state probes (Orgad et al., 2025). A separate line trains the LLM itself to express calibrated uncertainty (Lin et al., 2022; Xu et al., 2024; Damani et al., 2026) or directly optimizes selective performance (Luo et al., 2026). Closest to TAPS, ASPIRE (Chen et al., 2023) first adapts the answerer and then learns self-evaluation through separate soft prompts. Whereas ASPIRE pairs task adaptation with a learned self-evaluator to improve selection, TAPS treats accuracy training and confidence estimation as modular controls for isolating the value of learned abstention.

3 Abstention as Selective Prediction

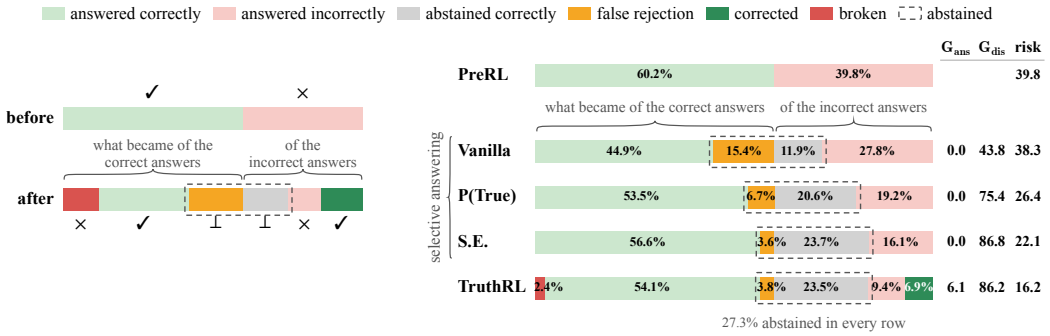
Abstention mitigates hallucination by withholding answers that are likely to be wrong. Formally, consider a dataset $\mathcal{D} = \{(x_i, y_i^*)\}_{i=1}^N$ of N questions, where x_i is a question and y_i^* is its reference answer. An abstention method takes a question x_i and either returns an answer $\hat{y}_i$ or abstains, partitioning the dataset into an answered set $\mathcal{A}$ and an abstained set $\mathcal{R}$. Ideally, all correct answers should remain in $\mathcal{A}$ and all others should fall into $\mathcal{R}$. Real methods fall short in both directions, causing a trade-off between the fraction of questions answered and the error rate among those answers.

Metrics commonly adopted by prior abstention training studies fail to fully characterize this trade-off. Abstention is neither correct nor incorrect, but any single score over all questions must assign it a value relative to the other two. Prior studies do so implicitly (Yang et al., 2024; Cheng et al., 2024): overall accuracy values an abstention as an incorrect answer, hallucination rate values it as a correct answer, and composites such as accuracy minus hallucination rate (Wei et al., 2026) value it midway between the two. Under any such valuation, a difference in score need not reflect better abstention decisions. It may simply reflect a different amount of abstention: withholding answers at random, for instance, lowers the hallucination rate without improving a single decision. Accuracy over the answered questions alone (Zhang et al., 2024; Zeng et al., 2026) sidesteps the valuation by ignoring abstentions, but then omits the coverage they cost. These metrics thus suffice to describe one policy at its native coverage, but cannot fairly compare policies with different coverages.

To fully characterize this trade-off and enable fair comparison, we keep its two sides separate, as *coverage*, *i.e.*, how many questions are answered, and *risk*, *i.e.*, how often those answers are wrong:

$$c = \frac{|\mathcal{A}|}{N}, \quad r = \frac{1}{|\mathcal{A}|} \sum_{i \in \mathcal{A}} e_i, \tag{1}$$

where e_i is the error indicator of answer $\hat{y}_i$. This matches the standard formulation of selective prediction (Geifman & El-Yaniv, 2017), in which a predictor is equipped with a reject option and evaluated by these two metrics jointly. LLM abstention can thus be viewed as selective prediction under a generative setup, with the answering model as the predictor and the abstention decision as



(a) Illustrative example **(b) Qwen2.5-7B-Instruct on TriviaQA**

Figure 1: **Behavioral decomposition of abstention performance at matched coverage.** (a) Illustration of outcome transitions under abstention. (b) On Qwen2.5-7B-Instruct/TriviaQA at matched coverage $c = 72.7\%$, stronger selective answering (S.E.) closes the distinguishing-ability gap with TruthRL (G_{dis}), but TruthRL's remaining edge in answering ability (G_{ans}) still yields lower risk.

the reject option. Within this space, different abstention behaviors induce different operating points (c, r). Let $r(c)$ denote the risk at coverage c. A better method attains lower $r(c)$ across coverage levels, which is summarized by lower **area under the risk–coverage curve (AURC)**:

$$\text{AURC} = \int_0^1 r(c)\, dc. \tag{2}$$

Since any abstention method, however implemented, reduces to points in this space, the formulation provides common ground for comparing the two paradigms. For selective answering methods, since sweeping the confidence threshold traces a full risk–coverage curve, they can be directly compared using AURC. An abstention-trained policy, in contrast, is fixed once trained and typically yields only a single native operating point $(c_{\text{train}}, r_{\text{train}})$. Therefore, instead of estimating AURC, we compare the two paradigms *at matched coverage*: we adjust the acceptance threshold of selective answering to match c_{train}, and compare the resulting risk with r_{train}, which we denote as **Risk@Cov**.

4 Understanding the Advantage of Abstention Training

With the two abstention paradigms placed on common ground, we ask whether abstention training outperforms selective answering, and if so, where this advantage comes from.

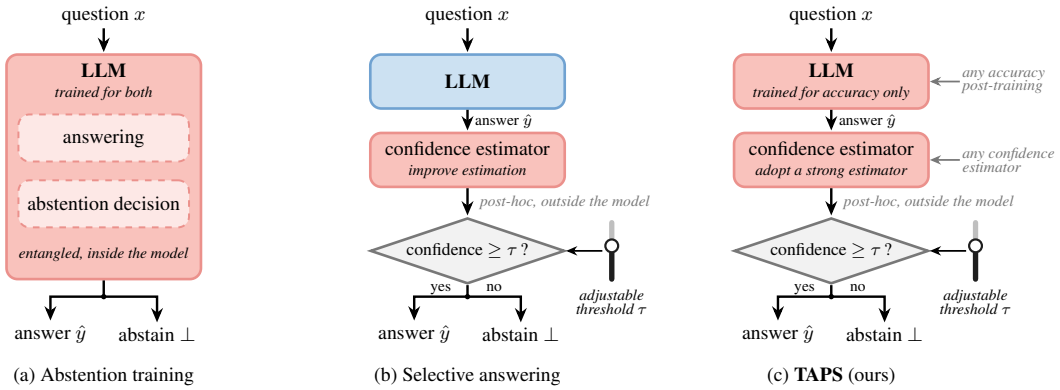
4.1 Behavioral Decomposition of Abstention Performance

Following the Risk@Cov comparison protocol established in Section 3, we analyze what determines risk at a fixed coverage through the behavioral transitions illustrated in Figure 1(a). Before applying an abstention method, each question is answered either correctly (✓) or incorrectly (×). After applying it, each answer flows into one of three outcomes: correct, incorrect, or abstained (⊥). Let $N_{a\to a'}$ denote the number of questions transitioning from $a \in \{\checkmark, \times\}$ to $a' \in \{\checkmark, \times, \perp\}$. These transitions reduce the number of remaining errors in two ways. First, answer generation may change: correcting an initially incorrect answer ($N_{\times\to\checkmark}$) removes an error, while breaking an initially correct answer ($N_{\checkmark\to\times}$) adds one, for a net effect of $N_{\times\to\checkmark} - N_{\checkmark\to\times}$. Second, abstention removes an error whenever an initially incorrect answer is withheld, contributing $N_{\times\to\perp}$. Starting from $N_{\times(\text{before})}$ incorrect answers, the final number of errors is thus

$$N_{\times(\text{after})} = N_{\times(\text{before})} - (N_{\times\to\checkmark} - N_{\checkmark\to\times}) - N_{\times\to\perp}. \tag{3}$$

At coverage c, exactly cN questions remain answered and $(1-c)N$ are abstained. Dividing Equation 3 by cN gives the risk under coverage c:

$$r(c) = \underbrace{\frac{N_{\times(\text{before})}}{cN}}_{r_0} - \underbrace{\frac{N_{\times\to\checkmark} - N_{\checkmark\to\times}}{cN}}_{G_{\text{ans}}} - \frac{1-c}{c} \cdot \underbrace{\frac{N_{\times\to\perp}}{(1-c)N}}_{G_{\text{dis}}}. \tag{4}$$



(a) Abstention training (b) Selective answering (c) **TAPS** (ours)

Figure 2: **Comparison of abstention approaches.** (a) Abstention training jointly optimizes answering and abstention decisions inside the model. (b) Selective answering externalizes abstention decision after generation using a confidence estimator and an adjustable threshold τ. (c) TAPS trains the answering model solely for accuracy while retaining the same post-hoc selection procedure.

Here, G_{ans} is the normalized net gain from answer correction (corrected minus broken in Figure 1), and G_{dis} is the fraction of abstained questions that were initially incorrect, *i.e.*, the precision of the abstention decisions (the proportion of correct rejections in the abstained block in Figure 1). We refer to them as *answering ability* and *distinguishing ability*, respectively. When methods are applied to the same model and are compared at the same coverage, r_0 and $(1-c)/c$ are fixed, so Equation 4 gives an exact behavioral decomposition of Risk@Cov into G_{ans} and G_{dis}.

4.2 Preliminary Findings

We next use this decomposition to examine where the advantage of abstention training comes from. We compare TruthRL (Wei et al., 2026), a representative RL-based abstention training method, against selective answering applied to the same starting model (*PreRL*). For selective answering, we consider Vanilla log-probability, P(True), and Semantic Entropy (S.E.) as confidence estimators, and adjust their thresholds to match TruthRL's native coverage. TruthRL uses its default abstention reward $r_{\text{abs}} = 0$. We perform this comparison across three models and three factual QA datasets. Figure 1(b) illustrates one setting, while Appendix E reports the complete results.

Stronger confidence estimation closes the distinguishing gap. Figure 1(b) shows that, against selective answering with Vanilla confidence, TruthRL achieves both a positive G_{ans} and a higher G_{dis}, so improvements in both abilities contribute to its lower risk. However, this apparent distinguishing advantage shrinks as the confidence estimator grows stronger. As reported in Appendix E, averaged across the nine model–dataset settings, G_{dis} rises from 59.47% with Vanilla to 84.64% with S.E., essentially matching TruthRL's 84.29% average. The distinguishing advantage of abstention training thus largely disappears once compared against a strong post-hoc selector.

The remaining gap lies in answering ability. As shown in Figure 1(b), selective answering leaves the generated answers unchanged and therefore has $G_{\text{ans}} = 0$ by construction, whereas TruthRL directly benefits from it and attains lower risk. As reported in Appendix E, TruthRL achieves a positive G_{ans} in every setting, averaging 5.07%. Together, these findings suggest that *the performance advantage of abstention training over strong selective answering methods primarily reflects improved answering ability rather than substantially better distinguishing ability.*

5 Train for Accuracy, then Post-hoc Select

Section 4 shows that stronger confidence estimation largely closes the gap in distinguishing ability, while leaving the gap in answering ability unaddressed. This naturally raises a question: what happens if selective answering is given an equally strengthened answerer? We construct this comparison by applying standard post-hoc confidence estimation to a model post-trained for answer accuracy. We refer to this two-stage recipe as **T**rain for **A**ccuracy, then **P**ost-hoc **S**elect (**TAPS**) (Figure 2(c)).

Stage 1: Accuracy training. Post-training for task accuracy has become a standard practice, with established methods directly applicable to this stage (Guo et al., 2025; Yu et al., 2025). Following common practice, we instantiate TAPS with outcome-reward RL:

$$r_{\text{ans}}(x, \hat{y}) = \begin{cases} +1, & \text{if } \hat{y} \text{ is a valid answer judged correct,} \\ -1, & \text{otherwise,} \end{cases} \tag{5}$$

We optimize this reward using the widely used GRPO (Shao et al., 2024) algorithm. Starting from *PreRL*, *i.e.*, the initial model used across abstention methods, we obtain the accuracy-trained model denoted *AccRL*.

Stage 2: Post-hoc selection. At inference time, AccRL generates a candidate answer $\hat{y}$ for a question x, and an external confidence estimator assigns a score $s(x, \hat{y})$, with higher scores indicating greater confidence. Given an acceptance threshold τ, the selective answering function is

$$f_\tau(x) = \begin{cases} \hat{y}, & \text{if } s(x, \hat{y}) \geq \tau, \\ \perp, & \text{otherwise,} \end{cases} \tag{6}$$

where $\perp$ denotes abstention. This stage can directly reuse existing confidence estimation methods.

Beyond serving as a controlled comparison, this construction has two broader implications. ❶ TAPS is a straightforward enhancement of the selective answering paradigm, and therefore retains the controllability that abstention training gives up. If such a simple extension can achieve competitive performance, the added value of internalizing abstention decisions merits a more careful assessment. ❷ TAPS is also composed of two independently well-studied components: standard post-training and well-developed confidence estimation, so advances in either can be incorporated directly, making it an evolving baseline against which abstention training should be continually reassessed.

6 Experiments

We organize the empirical study around three questions:

- **RQ1:** Does accuracy training make TAPS competitive with abstention training?
- **RQ2:** How does accuracy training affect the two abilities, and thus abstention performance?
- **RQ3:** How does post-hoc selection affect distinguishing ability, and thus abstention performance?

6.1 Experimental Setup

Models and datasets. Following recent abstention studies (Wei et al., 2026; Zeng et al., 2026), we evaluate Qwen2.5-7B-Instruct (Qwen Team, 2024), Qwen3-8B (Yang et al., 2025), and Llama-3.1-8B-Instruct (Grattafiori et al., 2024) on three factual QA benchmarks: Natural Questions (NQ) (Kwiatkowski et al., 2019), PopQA (Mallen et al., 2023), and TriviaQA (Joshi et al., 2017), spanning a wide range of knowledge types and difficulty levels. All experiments use closed-book QA without retrieval or external tools. Dataset splits are provided in Appendix C.1.

Compared methods. We compare TAPS with the SFT-based R-Tuning (Zhang et al., 2024) and two recent RL-based methods, TruthRL (Wei et al., 2026) and UCPO (Zeng et al., 2026). For TruthRL, we evaluate $r_{\text{abs}} \in \{-0.5, 0, 0.5\}$, while UCPO follows its original advantage rules. TAPS uses four interchangeable confidence estimators: mean log-probability, P(True) (Kadavath et al., 2022), Semantic Entropy (S.E.) (Kuhn et al., 2023; Farquhar et al., 2024), and a supervised correctness probe (Orgad et al., 2025). Implementation details are provided in Appendices C.2–C.3.

Training and evaluation. We implement Stage 1 with `verl` (Sheng et al., 2025) and GRPO (Shao et al., 2024). Training and evaluation use chain-of-thought generation, and evaluation greedily decodes one candidate response per question. Qwen2.5-72B-Instruct (Qwen Team, 2024) judges correctness during training, while gpt-oss-20b (Agarwal et al., 2025) independently evaluates test outputs, avoiding exploitation of the training judge. Hyperparameters, generation protocols, prompts, and judge reliability are reported in Appendices C.4–C.5, I, and D, respectively.

Metrics. For each abstention-trained method m with native coverage C_m, we threshold TAPS to match C_m and report Risk@C_m. For curve-level analyses, we report AURC. We additionally use answer accuracy and AUROC as diagnostic proxies for answering and distinguishing abilities, respectively. Full metric definitions and computation details are provided in Appendix C.5.

Table 1: Risk (%, ↓) at matched coverage. Each group compares an abstention training method with selective answering and TAPS at the same coverage (%) shown in the shaded row. Avg. averages the nine model–dataset pairs within each group. For TruthRL, the value in parentheses is the abstention reward r_{abs}. The lowest risk in each group is in **bold**, and the second-lowest is underlined.

Risk@Cov. (%) ↓	Qwen2.5-7B-Instruct			Qwen3-8B			Llama-3.1-8B-Instruct			Avg.
	NQ	PopQA	TriviaQA	NQ	PopQA	TriviaQA	NQ	PopQA	TriviaQA	
Coverage of R-Tuning	32.51	15.16	48.63	47.23	25.18	60.16	57.22	24.67	73.86	
R-Tuning	41.11	25.60	12.23	46.29	35.77	18.94	42.85	25.56	12.46	28.98
Selective Answering (S.E.)	34.47	24.52	8.50	39.14	35.48	8.57	32.09	22.57	7.14	23.61
Selective Answering (Probing)	38.69	28.31	10.36	39.14	35.64	8.90	31.49	22.30	8.15	24.77
TAPS (S.E.)	40.57	**19.57**	**5.31**	39.41	32.12	**5.28**	34.80	**18.77**	9.66	22.83
TAPS (Probing)	**31.06**	25.99	7.60	**31.58**	**30.98**	5.79	**28.31**	20.48	**5.95**	**20.86**
Coverage of TruthRL (+0.5)	46.95	23.24	72.66	46.37	18.99	76.66	75.48	45.51	92.36	
TruthRL (+0.5)	37.66	29.31	**15.21**	36.63	34.30	15.05	34.72	38.78	**11.09**	28.08
Selective Answering (S.E.)	43.82	38.75	22.08	39.04	27.14	17.63	40.14	39.42	14.56	31.40
Selective Answering (Probing)	45.49	38.85	23.19	38.63	28.83	17.05	38.31	37.97	14.82	31.46
TAPS (S.E.)	42.20	**27.90**	17.62	39.52	**22.68**	14.74	36.36	32.72	13.37	27.46
TAPS (Probing)	**36.92**	31.84	17.61	**31.27**	23.99	**13.44**	**34.53**	**32.00**	11.96	**25.95**
Coverage of TruthRL (0)	50.60	19.69	72.70	50.45	20.07	78.13	77.48	43.37	90.10	
TruthRL (0)	40.54	29.06	**16.21**	37.27	31.45	15.44	**33.60**	**27.92**	10.84	26.93
Selective Answering (S.E.)	46.31	33.23	22.11	41.06	28.48	18.75	40.48	37.71	13.47	31.29
Selective Answering (Probing)	47.12	34.31	23.22	41.31	29.85	18.08	38.93	36.53	13.79	31.46
TAPS (S.E.)	43.31	**23.44**	17.66	40.81	**23.84**	15.66	36.97	32.42	12.37	27.39
TAPS (Probing)	**38.58**	28.95	17.64	**33.29**	24.97	**14.20**	35.15	30.81	**10.79**	**26.04**
Coverage of TruthRL (−0.5)	51.65	19.48	73.93	50.13	20.25	79.07	74.21	38.23	91.66	
TruthRL (−0.5)	40.34	27.15	**16.38**	37.76	34.77	16.83	34.38	27.81	**10.50**	27.33
Selective Answering (S.E.)	46.78	32.57	22.90	40.88	28.64	19.35	39.53	33.75	14.30	30.97
Selective Answering (Probing)	47.97	34.08	24.00	41.07	30.11	18.71	37.81	32.44	14.47	31.18
TAPS (S.E.)	43.95	**23.30**	18.28	40.56	**23.92**	16.14	35.82	**27.47**	13.25	26.97
TAPS (Probing)	**39.32**	28.90	18.38	**32.93**	25.10	**14.82**	**34.01**	27.85	11.57	**25.88**
Coverage of UCPO	14.71	36.72	55.98	17.43	41.22	46.03	60.73	25.87	74.21	
UCPO	27.38	61.42	13.74	34.73	71.55	11.16	32.50	27.74	7.38	31.96
Selective Answering (S.E.)	**19.84**	53.05	11.95	**17.55**	51.64	3.24	33.52	23.61	7.26	24.63
Selective Answering (Probing)	28.17	52.99	13.71	19.56	51.81	5.10	32.56	23.15	8.24	26.14
TAPS (S.E.)	36.90	44.24	**8.74**	27.42	46.89	**2.08**	34.48	**19.38**	9.61	25.53
TAPS (Probing)	23.81	**44.17**	9.77	19.20	**46.31**	3.14	**29.61**	20.96	**6.01**	**22.55**

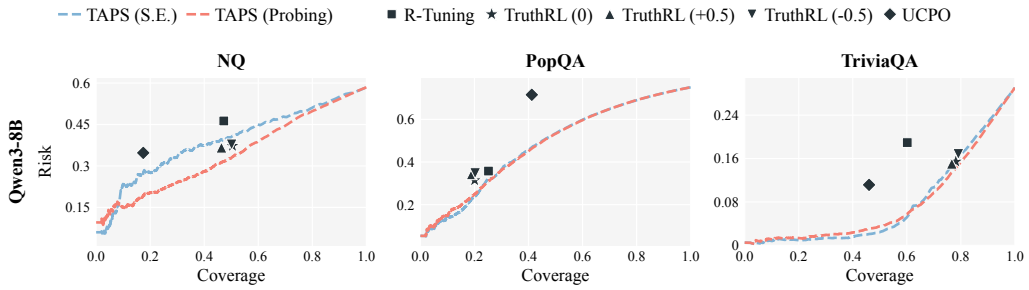


Figure 3: Risk–coverage comparison with abstention training. Each curve is obtained by varying the acceptance threshold τ for a fixed answering checkpoint and confidence estimator. Results for the other two models are reported in Appendix F.

6.2 Comparison with Abstention Training (RQ1)

Accuracy training strengthens selective answering. Table 1 isolates the contribution of accuracy training by comparing TAPS with its selective answering counterpart at matched coverage. Across 90 matched-coverage comparisons, accuracy training lowers risk in 81 cases, by 3.55 pp on average, showing that improving the underlying answerer directly strengthens abstention performance beyond post-hoc selection alone.

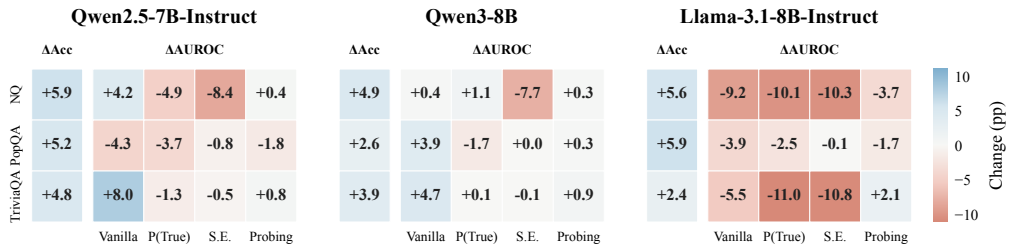


Figure 4: Effects of accuracy training on answering ability (Acc) and distinguishing ability (AUROC). Values show AccRL minus PreRL in percentage points (pp).

TAPS is competitive with abstention training. With this gain, TAPS becomes a strong competitor to methods that explicitly learn to abstain: both variants achieve lower mean risk than R-Tuning and UCPO, and against the more stringent TruthRL baseline, TAPS with probing achieves lower risk in 17 of 27 comparisons (mean advantage 1.49 pp) across its three abstention rewards, while TAPS with S.E. also remains comparable on average. Much of the performance gap to abstention training can therefore be closed and often reversed by post-hoc selecting on a strong answerer.

TAPS retains direct control over coverage. Figure 3 shows the other key advantage of leaving abstention decisions post-hoc on Qwen3-8B. Due to space limitation, results for the other two models are reported in Appendix F. For TAPS, varying the acceptance threshold traces a full risk–coverage curve, letting one trained model serve different coverage requirements. Abstention-trained policies, in contrast, offer far weaker control: each checkpoint yields a single operating point, and even across TruthRL's three separately trained reward variants, coverage spans less than 5 percentage points in 8/9 settings. Moreover, a lower abstention reward does not necessarily lead to a higher coverage, further complicating abstention control. Therefore, together with the performance results above, TAPS shows that competitive abstention performance need not come at the price of controllability. One might instead ask whether a post-hoc selector could be applied to abstention-trained models to obtain more competitive operating points. We discuss this question in detail in Appendix H.

6.3 Effects of Accuracy Training (RQ2)

Section 6.2 shows that accuracy training generally strengthens selective answering. We now use the answering–distinguishing decomposition in Section 4 to study how it changes these two abilities and how their changes translate into overall abstention performance, using Acc and AUROC as proxy metrics for these abilities and AURC for overall performance as specified in Section 6.1.

Accuracy training consistently improves answering ability, but its effect on distinguishing ability varies. Figure 4 shows that accuracy RL improves Acc across all settings, by 2.4–5.9 pp, while its effect on AUROC is far less uniform: most changes are modest for the two Qwen models, whereas unsupervised estimators often lose substantial AUROC for Llama-3.1-8B-Instruct, especially on NQ and TriviaQA. Probing, in contrast, remains comparatively stable, consistent with it being refit on each checkpoint's representations, which lets the selector adapt after accuracy training.

Answering gains can outweigh degraded distinguishing ability. Figure 5 relates the changes in Acc and AUROC to the resulting change in AURC for each model–dataset–estimator configuration. Accuracy training reduces AURC in 28 of the 36 configurations, consistent with the gains observed in Section 6.2. Notably, many configurations achieve lower AURC despite degraded AUROC, showing that stronger answering ability can improve the overall trade-off even as confidence ranking deteriorates. But this is not universal, as configurations with larger ranking degradation can still exhibit worse AURC. Neither improved accuracy nor improved AUROC alone therefore determines abstention performance. Whether answering gains translate into better performance depends on how much distinguishing ability is preserved after accuracy training.

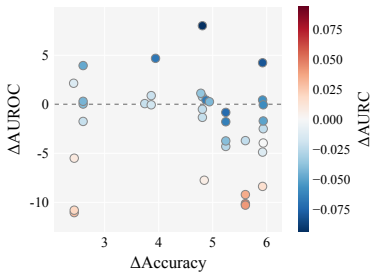


Figure 5: AURC Changes from PreRL to AccRL.

6.4 Effects of Confidence Estimation (RQ3)

Beyond accuracy training, TAPS also depends on the confidence estimator used at Stage 2. We compare how different estimators affect distinguishing ability and the resulting abstention performance.

Stronger confidence estimation generally improves abstention performance. Within each model–dataset setting, all estimators use the same AccRL checkpoint and greedy decoding configuration. Differences in AURC primarily reflect how effectively their confidence scores distinguish correct from incorrect answers. Figure 6 illustrates this relationship on Qwen3-8B, while Appendix G reports results for the other models. Across all nine model–dataset settings, P(True), S.E., and probing achieve higher AUROC than Vanilla, and higher AUROC generally coincides with lower AURC. The strongest estimator remains setting-dependent: probing is more consistently strong overall, whereas S.E. performs best on PopQA and TriviaQA for Qwen2.5-7B-Instruct. These results highlight the complementary roles of the two TAPS stages: Stage 1 improves the candidate answers, while Stage 2 determines how effectively the remaining errors are identified and rejected.

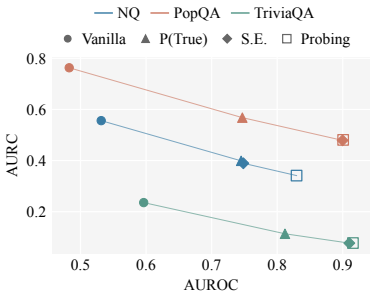


Figure 6: Relationship between AUROC and AURC.

6.5 Discussion

Implications for abstention training. Our results suggest a stricter way to assess the value of learning to abstain. Abstention training jointly changes two aspects of the model: how well it answers and how well it decides which answers to withhold. Consequently, improvements over the starting model conflate gains in answering ability with gains in distinguishing ability. Our decomposition separates these effects and shows that simply equipping selective answering with strong confidence estimators can match much of the distinguishing gain, while the remaining advantage primarily comes from improved answering. This makes an accuracy-trained answerer paired with a strong post-hoc selector a more informative reference for identifying what abstention training contributes beyond leaving decisions post-hoc.

Practical considerations. The deployment cost of TAPS depends largely on the confidence estimator chosen for Stage 2, creating a flexible performance–cost trade-off. Semantic entropy requires multiple sampled generations and semantic comparisons, whereas probing requires correctness-labeled responses and a one-time fitting step but adds only a lightweight classifier at inference time. Cheaper estimators such as Vanilla confidence can instead be used when computational cost is more important. Importantly, TAPS itself is not tied to any particular estimator: improved confidence estimators can be substituted directly without redesigning the answering model or introducing a new abstention-training objective. This modularity allows the selection mechanism to be chosen according to the deployment setting and improved independently over time.

7 Conclusion

In this paper, we revisited LLM abstention by decomposing matched-coverage risk into answering and distinguishing ability. The decomposition shows that abstention training's advantage over strong selective answering comes from better answering ability, which motivates TAPS: train for accuracy, then select post hoc. With a probing-based estimator, TAPS obtains lower mean risk than three representative abstention training methods at matched coverage while keeping coverage adjustable. More broadly, our results call for a more careful interpretation of progress in abstention training: better overall performance does not necessarily imply better abstention, and TAPS provides a simple way to disentangle the two.

For future work, several directions remain open. An important next step is to test whether TAPS scales to larger models and generalizes to other tasks, including retrieval-augmented QA and agentic tasks. Because accuracy training can weaken the model's own confidence signals, and preserving distinguishing ability requires refitting a supervised confidence estimator, calibration-preserving post-training could make TAPS more efficient and robust. Finally, developing methods that enable abstention-trained models to adjust their coverage after training is another promising direction.

AI USE STATEMENT

In this work, we used generative AI tools for implementing our methods, assisting with translation, creating or modifying scientific figures or images, summarizing or analyzing existing literature, and editing the paper to improve readability. We have not used generative AI tools for helping develop theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, or interpreting results. The following categories are not applicable to this work: generating synthetic datasets, cleaning and reformatting datasets, and supporting qualitative and thematic data analysis. We have reviewed all AI-assisted work. We manually verified every citation, AI-generated code, and experimental detail. No model-generated result is treated as experimental evidence, and all reported numbers were reproduced from the project artifacts and independently checked by the authors. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work aims to improve the reliability and controllability of LLM abstention using publicly available models and factual QA benchmarks. Abstention does not eliminate factual errors, and confidence estimates may reflect biases in the underlying models and data. Our results should therefore not be interpreted as guarantees of correctness or safety.

REPRODUCIBILITY STATEMENT

Section 6.1 summarizes the experimental setup. Appendix C details dataset splits, baseline implementations, confidence estimators, training hyperparameters, hardware, and evaluation protocols. Appendix D reports the reliability analysis of the evaluation judge, and Appendix I provides the generation and judging prompts. We will publicly release our code, including training scripts, evaluation pipelines, and prompts, upon acceptance.

REFERENCES

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, et al. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*, 2025.

Peter L. Bartlett and Marten H. Wegkamp. Classification with a reject option using a hinge loss. *Journal of Machine Learning Research*, 9(59):1823–1840, 2008.

Xujun Che, Yuchen Yuan, Weida Zhao, and Chenyang Yu. Abstention as an action can kill both the reward gradient and the kl anchor: Collapse law and repair for error-penalized reinforcement learning. *arXiv preprint arXiv:2608.00301*, 2026.

Jiefeng Chen, Jinsung Yoon, Sayna Ebrahimi, Sercan Arik, Tomas Pfister, and Somesh Jha. Adaptation with self-evaluation to improve selective prediction in LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 5190–5213, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.345.

Qinyuan Cheng, Tianxiang Sun, Xiangyang Liu, Wenwei Zhang, Zhangyue Yin, Shimin Li, Linyang Li, Zhengfu He, Kai Chen, and Xipeng Qiu. Can AI assistants know what they don't know? In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 8184–8202. PMLR, 21–27 Jul 2024.

C. K. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, 1970. doi: 10.1109/TIT.1970.1054406.

Jeremy Cole, Michael Zhang, Daniel Gillick, Julian Eisenschlos, Bhuwan Dhingra, and Jacob Eisenstein. Selectively answering ambiguous questions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 530–543, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.35.

Corinna Cortes, Giulia DeSalvo, and Mehryar Mohri. Learning with rejection. In *Algorithmic Learning Theory*, volume 9925 of *Lecture Notes in Computer Science*, pp. 67–82. Springer, 2016. doi: 10.1007/978-3-319-46379-7_5.

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 *International Conference on Learning Representations*, volume 2026, pp. 59456–59493, 2026.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(53):1605–1641, 2010.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024.

Leo Feng, Mohamed Osama Ahmed, Hossein Hajimirsadeghi, and Amir H. Abdi. Towards better selective classification. In *The Eleventh International Conference on Learning Representations*, 2023.

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

Yonatan Geifman and Ran El-Yaniv. SelectiveNet: A deep neural network with an integrated reject option. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2151–2159. PMLR, 09–15 Jun 2019.

Yonatan Geifman, Guy Uziel, and Ran El-Yaniv. Bias-reduced uncertainty estimation for deep neural classifiers. In *International Conference on Learning Representations*, 2019.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638, September 2025. doi: 10.1038/s41586-025-09422-z.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12), March 2023. ISSN 0360-0300. doi: 10.1145/3571730.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1147.

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.

Amita Kamath, Robin Jia, and Percy Liang. Selective question answering under domain shift. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5684–5696, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.503.

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*, 2023.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, et al. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019.

Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research*, 2022.

Jessie Liu, Blanca Gallego, and Sebastiano Barbieri. Incorporating uncertainty in learning to defer algorithms for safe computer-aided diagnosis. *Scientific Reports*, 12:1762, 2022. doi: 10.1038/s41598-022-05725-7.

Gaoxiang Luo, Yifan Wu, Sinian Zhang, Aryan Deshwal, and Ju Sun. Aligning language models with selective prediction. In *Forty-third International Conference on Machine Learning Second Workshop on Agents in the Wild: Safety, Security, and Beyond*, 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.

Mohamad Amin Mohamadi, Tianhao Wang, and Zhiyuan Li. Honesty over accuracy: Trustworthy language models through reinforced hesitation. *arXiv preprint arXiv:2511.11500*, 2025.

Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. Llms know more than they show: On the intrinsic representation of llm hallucinations. In *International Conference on Learning Representations*, volume 2025, pp. 66880–66913, 2025.

Muyu Pan, Shu Zhao, Nan Zhang, Philip Shin, Varun Parekh, Vijaykrishnan Narayanan, and Rui Zhang. Tiar: Trajectory-informed advantage reweighting for llm abstention learning. *arXiv preprint arXiv:2605.25850*, 2026.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

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

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. HybridFlow: A flexible and efficient RLHF framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, EuroSys '25, pp. 1279–1297, New York, NY, USA, 2025. Association for Computing Machinery. doi: 10.1145/3689031.3696075.

Tejas Srinivasan, Jack Hessel, Tanmay Gupta, Bill Yuchen Lin, Yejin Choi, Jesse Thomason, and Khyathi Chandu. Selective "selective prediction": Reducing unnecessary abstention in vision-language reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 12935–12948, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.767.

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 *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5433–5442, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.330.

Neeraj Varshney, Swaroop Mishra, and Chitta Baral. Towards improving selective prediction ability of NLP systems. In *Proceedings of the 7th Workshop on Representation Learning for NLP*, pp. 221–226, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.repl4nlp-1.23.

Zhepei Wei, Xiao Yang, Kai Sun, Jiaqi Wang, Rulin Shao, Jingxiang Chen, Mohammad Kachuee, Teja Gollapudi, Yiwei Liao, Nicolas Scheffer, Rakesh Wanga, Anuj Kumar, Yu Meng, Wen-tau Yih, and Xin Luna Dong. TruthRL: Incentivizing truthful LLMs via reinforcement learning. In *Forty-third International Conference on Machine Learning*, 2026.

Bingbing Wen, Jihan Yao, Shangbin Feng, Chenjun Xu, Yulia Tsvetkov, Bill Howe, and Lucy Lu Wang. Know your limits: A survey of abstention in large language models. *Transactions of the Association for Computational Linguistics*, 13:529–556, 2025. doi: 10.1162/tacl_a_00754.

Jiayun Wu, Jiashuo Liu, Zhiyuan Zeng, Tianyang Zhan, Tianle Cai, and Wenhao Huang. Mitigating LLM hallucination via behaviorally calibrated reinforcement learning. *arXiv preprint arXiv:2512.19920*, 2026.

Ji Xin, Raphael Tang, Yaoliang Yu, and Jimmy Lin. The art of abstention: Selective prediction and error regularization for natural language processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1040–1051, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.84.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, YIFEI LI, Jie Fu, Junxian He, and Bryan Hooi. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. In *International Conference on Learning Representations*, volume 2024, pp. 23650–23678, 2024.

Tianyang Xu, Shujin Wu, Shizhe Diao, Xiaoze Liu, Xingyao Wang, Yangyi Chen, and Jing Gao. SaySelf: Teaching LLMs to express confidence with self-reflective rationales. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 5985–5998, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.343.

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

Yuqing Yang, Ethan Chern, Xipeng Qiu, Graham Neubig, and Pengfei Liu. Alignment for honesty. In *Advances in Neural Information Processing Systems*, volume 37, pp. 63565–63598. Curran Associates, Inc., 2024. doi: 10.52202/079017-2030.

Hiyori Yoshikawa and Naoaki Okazaki. Selective-LAMA: Selective prediction for confidence-aware evaluation of language models. In *Findings of the Association for Computational Linguistics: EACL 2023*, pp. 2017–2028, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-eacl.150.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, juncai liu, LingJun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Ru Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Yonghui Wu, and Mingxuan Wang. Dapo: An open-source llm reinforcement learning system at scale. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 113222–113244. Curran Associates, Inc., 2025. doi: 10.52202/085713-3775.

Xianzhou Zeng, Jing Huang, Chunmei Xie, Gongrui Nan, Siye Chen, Mengyu Lu, Weiqi Xiong, Qixuan Zhou, Junhao Zhang, Qiang Zhu, Yadong Li, and Xingzhong Xu. UCPO: Uncertainty-aware policy optimization. In *Forty-third International Conference on Machine Learning*, 2026.

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 *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, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.394.

Xinyu Zhou, Chang Jin, Carsten Eickhoff, Zhijiang Guo, and Seyed Ali Bahrainian. When silence is golden: Can llms learn to abstain in temporal qa and beyond? In *International Conference on Learning Representations*, volume 2026, pp. 73183–73209, 2026.

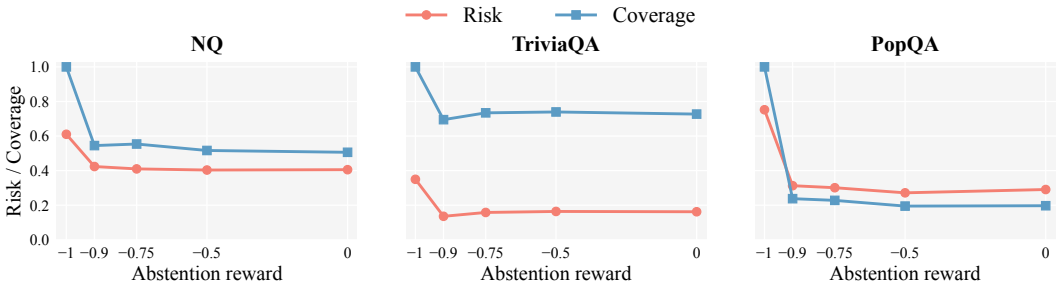


Figure 7: **Effect of sweeping the training-time abstention reward.** Each point represents a separately trained TruthRL policy based on Qwen2.5-7B-Instruct across NQ, PopQA, and TriviaQA. A common pattern emerges: moving r_{abs} from -1 to -0.9 causes an abrupt drop in coverage, while subsequent changes produce only limited or non-monotonic coverage variation.

A Selective prediction in classification.

Selective prediction improves reliability by allowing a predictor to withhold uncertain outputs (Chow, 1970; El-Yaniv & Wiener, 2010; Geifman & El-Yaniv, 2017). Prior work broadly follows two paradigms, distinguished by whether rejection is learned jointly with prediction. *Post-hoc selection* keeps the predictor fixed and accepts predictions whose confidence exceeds a threshold (Geifman & El-Yaniv, 2017; Geifman et al., 2019). *Learned rejection* instead jointly optimizes prediction and rejection within the classifier (Bartlett & Wegkamp, 2008; Cortes et al., 2016; Geifman & El-Yaniv, 2019). Modern LLM abstention inherits the same distinction: selective answering externalizes the abstention decision, whereas abstention training internalizes it. We find that the apparent advantage of abstention training largely reflects stronger answering ability rather than more accurate abstention decisions, paralleling the observation of Feng et al. (2023) that learned-rejection gains in classification can arise from a stronger classifier rather than a better rejection mechanism.

B Limited Coverage Control through Abstention-Reward Sweeps

The ternary reward in Equation 7 makes the abstention reward r_{abs} a natural training-time control signal. Increasing r_{abs} makes abstention more favorable relative to incorrect answering and might therefore be expected to reduce coverage. We conduct a preliminary study to examine whether varying this reward can produce policies spanning a controllable range of coverage levels.

We train separate TruthRL policies with $r_{\text{abs}} \in \{-1, -0.9, -0.75, -0.5, 0\}$, while keeping the rewards for correct and incorrect answers fixed at $+1$ and -1, respectively. Each value corresponds to a separate training run. Figure 7 shows a sharp behavioral change in coverage and risk as the abstention reward is swept. At $r_{\text{abs}} = -1$, abstention and incorrect answering receive the same reward, and the learned policies retain nearly full coverage. Increasing the reward slightly to $r_{\text{abs}} = -0.9$ triggers an abrupt transition to substantially lower coverage. After this transition, further increases in r_{abs} do not produce a sequence of progressively lower or well-separated coverage and risk levels. Instead, the resulting policies collapse within a relatively narrow range across all datasets. In the risk–coverage plane, this collapse appears as a cluster of operating points within a narrow range, preventing fine-grained coverage control.

C Experimental Details

C.1 Data Splits

We evaluate different approaches across three factual QA benchmarks: Natural Questions (NQ), PopQA and TriviaQA. For NQ and TriviaQA, whose original training splits each exceed 80,000 examples, we randomly sample 10,000 examples for training and reserve a small held-out portion for validation. Following common practice, we adopt their validation splits as test sets, since NQ lacks an official test set and the test annotations of TriviaQA are not publicly available. For PopQA, which provides only a single evaluation set, we randomly partition it into training, validation, and test subsets. The final statistics for all dataset splits are summarized in Table 2. Training is performed separately for every model–dataset pair, using only that dataset's training partition.

Table 2: Dataset partitions.

Dataset	Training	Validation	Test
NQ	10,000	512	3,610
PopQA	5,000	256	9,011
TriviaQA	10,000	512	11,313

C.2 Abstention Training Methods

R-Tuning. R-Tuning (Zhang et al., 2024) learns answer–certainty responses through SFT. We first generate answers from the starting model on the training questions and label their correctness using the judge in Appendix C.5. Each supervised target contains the reference answer, followed by a certainty question and either `I am sure.` or `I am unsure.`, depending on whether the starting model answered correctly. Questions answered incorrectly remain in the training set, but their incorrect generated answers are replaced with the reference answers in the supervised targets. At evaluation, following their standard interpretation, we treat unsure declarations as abstentions.

TruthRL. TruthRL (Wei et al., 2026) uses RL to distinguish correct answering, incorrect answering, and abstention through a ternary reward. The GRPO reward implementation uses

$$r_{\mathrm{TruthRL}}(x, \hat{y}) = \begin{cases} +1, & \text{if } \hat{y} \text{ is a valid correct answer,} \\ r_{\mathrm{abs}}, & \text{if } \hat{y} \text{ is a valid abstention,} \\ -1, & \text{otherwise.} \end{cases} \tag{7}$$

We evaluate $r_{\mathrm{abs}} \in \{-0.5, 0, 0.5\}$ in the main experiments and use $r_{\mathrm{abs}} = 0$ in the preliminary analysis. Prompts are provided in Appendix I.

UCPO. UCPO (Zeng et al., 2026) uses the same three outcome categories but replaces standard GRPO advantages with uncertainty-aware estimates. We retain the original TAD/DURA rules with $\alpha = 2$, $\lambda = 0.5$, $w = 1$, and $\epsilon = 10^{-6}$. The implementation uses the sample standard deviation and computes its batch-level statistic over the full rollout batch. Groups without both correct and incorrect responses receive zero advantages. Abstention advantages are computed dynamically rather than by normalizing a fixed zero reward. UCPO uses the same rollout budget, abstention-enabled prompts, correctness judge, and response-length limit as the other RL conditions.

C.3 Confidence Estimation

For each answering checkpoint, we adopt the same zero-temperature greedy decoding setup for answer generation, and apply different confidence-estimation methods to its generated answers. Self-evaluation labels and additional sampled answers are used only to construct confidence scores. They do not replace the candidate answer or introduce majority-vote answering.

Vanilla mean log-probability. Vanilla uses the average conditional log-probability of the final-answer tokens. We extract the answer inside the final complete `\boxed{}` expression and identify the tokens whose decoded spans overlap this text, denoted by $\mathcal{I}_{\mathrm{ans}}$. The confidence score is

$$s_{\mathrm{Vanilla}}(x, \hat{y}) = \frac{1}{|\mathcal{I}_{\mathrm{ans}}|} \sum_{t \in \mathcal{I}_{\mathrm{ans}}} \log \pi_\theta(y_t \mid x, y_{<t}). \tag{8}$$

Reasoning tokens are excluded from the average but remain part of the conditioning context. Higher scores indicate greater confidence.

P(True). Prompting-based confidence estimation asks the model to assess the reliability of its own answers (Kadavath et al., 2022; Tian et al., 2023; Xiong et al., 2024). We use the P(True) formulation of Kadavath et al. (2022). The answering checkpoint receives the question and its proposed answer in a separate prompt that asks whether the answer is true. No reference answer is supplied. We score the label sequences `True` and `False` and normalize their probabilities:

$$s_{\mathrm{P(True)}}(x, \hat{y}) = \frac{\exp \ell_{\mathrm{True}}}{\exp \ell_{\mathrm{True}} + \exp \ell_{\mathrm{False}}}, \tag{9}$$

where ℓ_v is the sum of conditional token log-probabilities for label sequence v.

Semantic entropy. Semantic entropy measures uncertainty over answer meanings by grouping semantically equivalent responses (Kuhn et al., 2023; Farquhar et al., 2024). In our implementation, we draw 10 additional responses per question from the answering checkpoint with temperature 1.0, top-$p = 0.9$, and top-$k = 50$. After extracting the final answers and combining normalized duplicates, we group answers using bidirectional semantic entailment. Two answers are grouped together only when each entails the other as an answer to the question. With M valid samples and n_c samples in cluster c, we compute

$$p_c = \frac{n_c}{M}, \qquad H_{\mathrm{sem}}(x) = -\sum_c p_c \log p_c, \qquad s_{\mathrm{SE}}(x) = -H_{\mathrm{sem}}(x). \tag{10}$$

We use natural logarithms and exclude malformed samples from the cluster-frequency calculation.

Probing. Probing predicts answer correctness from the answering model's hidden representations (Orgad et al., 2025). We fit an ℓ_2-regularized logistic regression on the final normalized hidden state at the last response-content token, excluding trailing special tokens and padding. The answering model remains frozen. Correct answers receive positive labels, while incorrect answers receive negative labels. We remove training questions whose normalized text appears in validation or test. The remaining questions are partitioned into 80% fitting and 20% development subsets using a class-stratified split. Duplicate questions are kept within the same subset, and feature standardization is fitted only on the fitting subset. We search inverse regularization strengths $C \in \{0.001, 0.01, 0.1, 1, 10\}$ and use L-BFGS with at most 1,000 iterations. The classifier with the lowest development log loss is applied to validation and test without refitting. Its predicted probability of correctness serves as the confidence score.

C.4 Training Configuration

RL Optimization. We use the verl framework (Sheng et al., 2025) for RL training. TAPS accuracy training and TruthRL use GRPO (Shao et al., 2024), while UCPO uses the advantage estimator described in Appendix C.2. For Qwen2.5-7B-Instruct and Llama-3.1-8B-Instruct, we use a total response budget of 256 tokens, which is sufficient for the short reasoning traces in single-hop factual QA with negligible truncation. For Qwen3-8B, we enable native thinking and run with a larger response budget of 2048 tokens. All RL training experiments were conducted on a single compute node equipped with 8 NVIDIA A100 (80 GB) GPUs. Four GPUs are allocated to model training, and the other four serve the judge model. Table 3 lists the shared hyperparameters.

SFT optimization. R-Tuning updates all model parameters for four epochs with a learning rate of 10^{-5}, a global batch size of 64, and a maximum sequence length of 1,024 tokens. We use a micro-batch of one example per training GPU and 16 gradient-accumulation steps. The loss covers response tokens, including EOS, and excludes prompt tokens.

C.5 Judging and Evaluation Protocol

Correctness judging. To support large-scale judging and facilitate experimental reproducibility, we use an LLM-as-a-judge protocol for both training rewards and evaluation. We deploy different judge models in these two stages to avoid hacking, with Qwen2.5-72B-Instruct for training and gpt-oss-20b for evaluation. Both judges receive the question, reference answer, and extracted final answer, without the reasoning trace. Each judge returns `A`, `B`, or `C` for correct, incorrect, or abstention, respectively. The full judging prompt, including all examples, is provided in Appendix I. We also assess the judge's reliability, with results reported in Appendix D.

Table 3: RL training hyperparameters.

Setting	Qwen2.5-7B-Instruct / Llama-3.1-8B-Instruct	Qwen3-8B
Training budget	8 epochs	8 epochs
Questions per rollout batch	128	128
Responses per question	5	5
Rollout temperature / top-p	1.0 / 1.0	1.0 / 1.0
Rollout top-k truncation	Disabled	Disabled
Maximum prompt length	512 tokens	512 tokens
Native thinking	Not applicable	Enabled
Maximum response length	256 tokens	2048 tokens
Optimizer	AdamW	AdamW
Learning rate	10^{-6}	10^{-6}
Schedule / warmup	Constant / none	Constant / none
Weight decay	0.01	0.01
AdamW (β_1, β_2)	$(0.9, 0.999)$	$(0.9, 0.999)$
PPO mini-batch size	64 prompts	64 prompts
Micro-batch size	16 trajectories per GPU	2 trajectories per GPU
PPO epochs per rollout batch	1	1
Clipping ratio	0.2	0.2
Gradient-norm clipping	1.0	1.0
Loss aggregation	Token mean	Token mean
KL coefficient	0.001	0.001
Entropy-bonus coefficient	0	0
Checkpoint / validation interval	Every 100 steps	Every 50 steps
Used Checkpoint	The last checkpoint	The last checkpoint

Matched-coverage comparison. Let N denote the number of total test questions of a dataset. For an abstention-trained policy m, let c_m, w_m, and a_m denote its correct-answer, incorrect-attempt, and abstention counts on these questions. Its accepted count, coverage, and risk are

$$k_m = c_m + w_m, \qquad C_m = \frac{k_m}{N}, \qquad R_m = \frac{w_m}{k_m}. \tag{11}$$

We select the k_m highest-confidence TAPS answers, randomly sampling from answers tied at the cutoff to fill the remaining slots.

Curve-level comparison. Since selective answering methods can elicit an entire risk–coverage curve, we use AURC to compare their overall performance. We rank the N answers by decreasing confidence, using the same random tie-breaking rule, and compute

$$R_k = \frac{1}{k}\sum_{i=1}^{k} e_{(i)}, \qquad \mathrm{AURC} = \frac{1}{N}\sum_{k=1}^{N} R_k, \tag{12}$$

where $e_{(i)}$ is the error indicator of the i-th ranked answer and R_k is the risk at coverage k/N.

Measuring answering ability. We use overall answer accuracy as a complementary diagnostics of the answering ability. For a dataset of N test questions, let e_i denote the error indicator of the candidate answer generated for question i. Answer accuracy is

$$\mathrm{Acc} = 1 - \frac{1}{N}\sum_{i=1}^{N} e_i. \tag{13}$$

This measures the fraction of questions answered correctly under the specified generation protocol. Because it is computed before any answers are withheld, it characterizes the quality of the generated answers without being affected by the confidence estimator or acceptance threshold.

Measuring distinguishing ability. To characterize distinguishing ability, we evaluate whether confidence scores prioritize correct answers over incorrect ones. Let s_i denote the confidence

Table 4: Detailed reliability analysis of LLM judges against human annotations ($n = 100$ per stage). TruthRL uses zero abstention reward. Agreement: the percentage of exact label matches; False acceptance: human $\times/\perp$ labeled ✓ by the judge; False rejection: human ✓ labeled $\times/\perp$. $\times/\perp$ confusion: swaps between incorrect and abstention.

		gpt-oss-20b	DeepSeek-V4.1-Flash	GPT 5.5	Gemini 3.6 Flash	Claude Sonnet 5
PreRL	Agreement (%)	91.0	92.0	95.0	92.0	97.0
	False acceptance	**0**	3	3	7	2
	False rejection	8	3	2	0	1
	$\times/\perp$ confusion	1	2	0	1	0
AccRL	Agreement (%)	90.0	94.0	91.0	91.0	96.0
	False acceptance	**1**	5	2	8	2
	False rejection	9	1	7	1	2
	$\times/\perp$ confusion	0	0	0	0	0
TruthRL	Agreement (%)	92.0	97.0	92.0	94.0	93.0
	False acceptance	**0**	3	3	6	4
	False rejection	8	0	5	0	3
	$\times/\perp$ confusion	0	0	0	0	0

score of the i-th candidate answer, with higher scores indicating greater confidence, and define $\mathcal{I}_+ = \{i : e_i = 0\}$ and $\mathcal{I}_- = \{i : e_i = 1\}$. The area under the receiver operating characteristic curve (AUROC) can be computed as

$$\text{AUROC} = \frac{1}{|\mathcal{I}_+|\,|\mathcal{I}_-|} \sum_{i \in \mathcal{I}_+} \sum_{j \in \mathcal{I}_-} \left[\mathbf{1}\{s_i > s_j\} + \frac{1}{2}\mathbf{1}\{s_i = s_j\} \right], \tag{14}$$

Equivalently, AUROC is the probability that a randomly chosen correct answer receives a higher confidence score than a randomly chosen incorrect answer, with half credit for ties. An AUROC of 1 indicates perfect separation, while a random ranking has expected AUROC 0.5. Since selective answering retains higher-scoring answers, this ranking measure provides a threshold-independent proxy for distinguishing ability without requiring confidence scores to be calibrated probabilities.

D Reliability Analysis of LLM-as-a-Judge

To validate the reliability of gpt-oss-20b as our unified evaluation judge and mitigate potential reward-hacking concerns, we conduct a human agreement analysis alongside comparisons with frontier models, including DeepSeek-V4.1-Flash, GPT 5.5, Gemini 3.6 Flash, and Claude Sonnet 5. We randomly sample 300 model outputs covering PreRL, AccRL, and TruthRL(0) stages, two datasets (NQ and TriviaQA), and two model families (Qwen3 and Llama), with 25 samples per cell.

Table 4 shows that gpt-oss-20b achieves 91.0% overall agreement with human annotations, competitive with frontier judges such as DeepSeek-V4.1-Flash (94.3%) and Gemini 3.6 Flash (92.3%). Agreement remains consistent across PreRL, AccRL, and TruthRL, at 91.0%, 90.0%, and 92.0%, respectively. More importantly, gpt-oss-20b makes the fewest false acceptances at every stage, labeling an incorrect answer or abstention as correct only once across all 300 examples, compared with 8–21 cases for the other judges. This single case occurs on AccRL, with none on PreRL or TruthRL. Its remaining disagreements consist of 25 false rejections, where correct answers are labeled incorrect or as abstentions, and one incorrect–abstention confusion. Thus, its errors overwhelmingly withhold credit from correct answers rather than grant credit to non-correct outputs. This conservative error profile provides evidence against judge leniency as an explanation for the reported low risk and supports the reliability of our evaluation.

E Full Behavioral Decomposition Results

Table 5 reports the complete decomposition across all nine model–dataset settings. In theory, selective answering leaves the generated answers unchanged and thus has $G_{\text{ans}} = 0$ (small nonzero values in table arise from run-to-run generation variation across estimators and are negligible), so different estimators primarily affect G_{dis}. On average, G_{dis} increases from 59.47% with Vanilla to

Table 5: Answering and distinguishing ability at matched coverage for every model–dataset setting. All entries are percentages. Selective answering matches TruthRL's coverage within each setting. The highest value for each metric within each row is in **bold**.

Model	Dataset	Cov.	TruthRL		Selective Answering					
					Vanilla		P(True)		S.E.	
			G_{ans}	G_{dis}	G_{ans}	G_{dis}	G_{ans}	G_{dis}	G_{ans}	G_{dis}
Qwen2.5-7B-Instruct	NQ	50.60	**6.46**	87.36	0.00	67.93	0.00	82.63	0.00	**88.07**
	PopQA	19.69	**5.30**	91.75	0.00	78.91	0.00	88.43	0.00	**92.03**
	TriviaQA	72.70	**6.13**	86.23	0.00	43.75	0.00	75.42	0.00	**86.83**
Qwen3-8B	NQ	50.45	**4.74**	84.89	0.00	65.34	0.00	78.96	0.25	**85.59**
	PopQA	20.07	**1.19**	88.96	0.00	76.01	0.00	86.45	0.00	**90.00**
	TriviaQA	78.13	**3.29**	**83.66**	0.00	39.69	0.32	65.96	0.00	83.57
Llama-3.1-8B-Instruct	NQ	77.48	**6.61**	**78.17**	0.00	59.39	0.00	68.21	0.00	77.25
	PopQA	43.37	**9.12**	**86.89**	0.00	69.69	0.00	80.91	0.00	86.37
	TriviaQA	90.10	**2.78**	70.67	0.00	34.49	0.00	46.61	0.00	**72.02**
Avg.		55.84	**5.07**	84.29	0.00	59.47	0.04	74.84	0.03	**84.64**

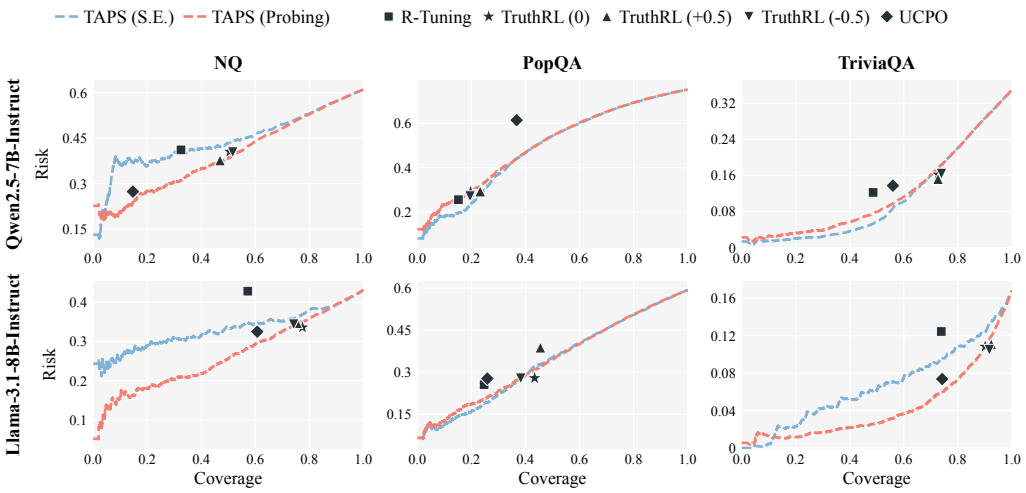


Figure 8: **Extended risk–coverage comparisons.** Results on Qwen2.5-7B-Instruct and Llama-3.1-8B-Instruct. TAPS curves vary the acceptance threshold. Markers denote the native operating points of abstention-trained checkpoints.

74.84% with P(True) and 84.64% with S.E., essentially matching TruthRL's 84.29%. TruthRL nevertheless achieves positive G_{ans} in every setting, averaging 5.07%. These results confirm that strong post-hoc selection closes the distinguishing gap, while abstention training's remaining advantage primarily comes from improved answering ability.

F Extended Results for Risk-Coverage Curves (RQ1)

Figure 8 extends the risk–coverage comparison to Qwen2.5-7B-Instruct and Llama-3.1-8B-Instruct. The same pattern holds across models: varying the TAPS threshold traces a full risk–coverage curve, whereas each abstention-trained checkpoint provides only one native operating point. Across all nine model–dataset settings, TruthRL's three reward variants span less than five percentage points of coverage in eight settings, confirming that training-time reward sweeps do not provide fine-grained coverage control.

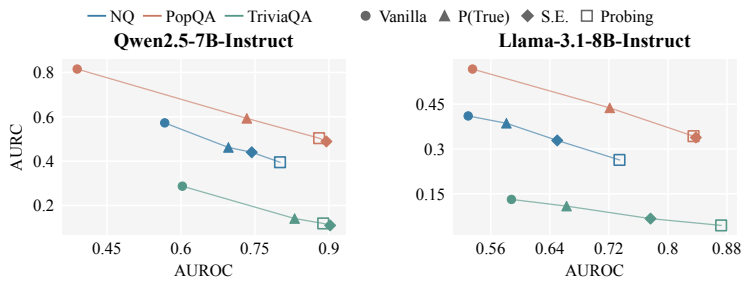


Figure 9: **Extended confidence-estimation results.** AUROC–AURC relationships on Qwen2.5-7B-Instruct and Llama-3.1-8B-Instruct. Each line connects confidence estimators for one dataset in increasing AUROC order.

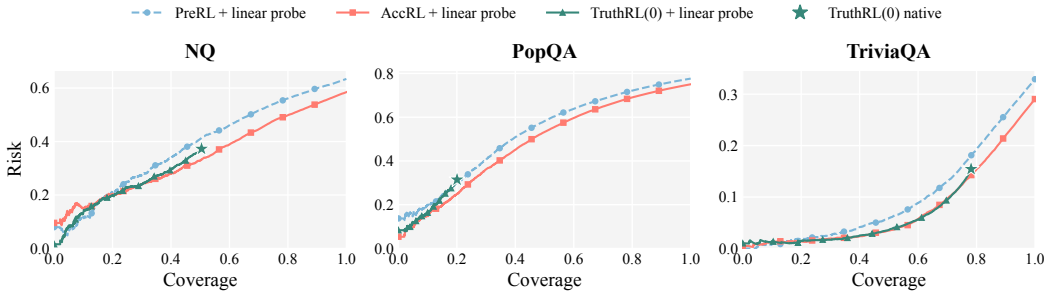


Figure 10: **Post-hoc selection on TruthRL.** The results show Qwen3-8B across three datasets. Because the probe can only filter TruthRL's native answers, its curve ends at the native coverage (star). Its relative advantage over TAPS varies across datasets and coverage levels.

G Extended Results for Stage 2 Analysis (RQ3)

Figure 9 extends the Stage 2 analysis to Qwen2.5-7B-Instruct and Llama-3.1-8B-Instruct. Within each model–dataset setting, all estimators use the same AccRL checkpoint and greedy decoding configuration, with AURC differences primarily reflecting confidence ranking. P(True), S.E., and probing improve AUROC over Vanilla in all six settings shown, and higher AUROC generally corresponds to lower AURC. The relationship observed on Qwen3-8B therefore generalizes across models.

H Diagnostic Post-hoc Selection on TruthRL

Abstention training and post-hoc selection embody opposing design choices: the former internalizes the answer-or-abstain decision within the model policy, whereas the latter externalizes this decision to a confidence estimator. Pairing TruthRL with an external selector therefore departs from the intended form of abstention training and is not a natural primary comparison. Nevertheless, for completeness and to control for the effect of the external selector, we conduct a diagnostic analysis on Qwen3-8B across all three datasets.

A TruthRL policy produces no candidate answer for questions on which it natively abstains. We therefore leave these abstentions unchanged and fit a separate linear probe only to the subset of questions that TruthRL(0) chooses to answer. Sweeping the probe threshold further removes low-confidence native answers, yielding a partial risk–coverage curve over $c \in [0, c_{\mathrm{TruthRL}}]$. In contrast, PreRL and AccRL generate a candidate answer for every question, so post-hoc selection can trace their full risk–coverage curves.

Figure 10 shows no uniform ordering between TruthRL with post-hoc selection and TAPS. On TriviaQA, their curves largely overlap throughout the shared coverage range. On NQ and PopQA, TruthRL with probing achieves lower risk at more stringent coverage levels, whereas TAPS per-

forms better toward the upper end of the shared range. Thus, supplementing TruthRL with an external selector can expose additional benefits in some low-coverage regimes, but these benefits are neither uniform across datasets nor consistent across coverage levels.

Crucially, this construction can only reduce coverage from TruthRL's native operating point. It cannot recover candidate answers for questions already rejected by the internal policy and therefore cannot access the higher-coverage portion of the curve. This one-sided control is particularly limiting because abstention-trained policies can collapse toward narrow and conservative native coverage ranges (Che et al., 2026; Zeng et al., 2026). Consequently, the hybrid cannot serve as a general coverage-control mechanism comparable to post-hoc selection over an answer-generating model. We therefore use it only as a diagnostic control for isolating the effect of the selector. In the main experiments, we compare the two paradigms in their intended forms: native answer-or-abstain decisions for abstention training, and adjustable post-hoc selection for TAPS.

I PROMPTS

This subsection collects the important prompts used in our experiments. Braced fields such as `{question}` are replaced with the corresponding example text. Line wrapping and indentation are adjusted for presentation; the wording, examples, and output labels are preserved.

Prompts for Qwen2.5-7B-Instruct and Llama-3.1-8B-Instruct. PreRL and AccRL use the following answer-only prompt.

Answer-only QA Prompt for Qwen2.5-7B-Instruct and Llama-3.1-8B-Instruct

```
You are a question answering model, your task is to answer the question based on factual
    information in your own knowledge.

Please answer with a single entity, as concise as possible.

Please reason step by step and then provide the final answer. The reasoning process must be
    enclosed within <think> </think> tags. The final answer MUST be put in \boxed{}. For
    example, \boxed{3 times}, \boxed{New York}, etc.

Question: {question}
Answer:
```

TruthRL and UCPO use an abstention-enabled version to elicit model's internal abstention behavior:

Abstention-enabled QA Prompt for Qwen2.5-7B-Instruct and Llama-3.1-8B-Instruct

```
You are a question answering model, your task is to answer the question based on factual
    information in your own knowledge.

Please adhere to the following guidelines when formulating the answer:
1. Answer with a single entity, as concise as possible.
2. If you are uncertain or don't know the answer, answer "I don't know".

Please reason step by step and then provide the final answer. The reasoning process must be
    enclosed within <think> </think> tags. The final answer MUST be put in \boxed{}. For
    example, \boxed{I don't know}, \boxed{3 times}, \boxed{New York}, etc.

Question: {question}
Answer:
```

Prompts for Qwen3-8B. Since thinking tags are provided in the chat template of Qwen3-8B and prompting it to generate these tags can violate its generation format, we omit explicit requirements to generate thinking tags.

Answer-only QA Prompt for Qwen3-8B

```
You are a question answering model, your task is to answer the question based on factual
    information in your own knowledge.

Please answer with a single entity, as concise as possible.

Think carefully before answering. Output only the final answer in \boxed{}. For example, \
    boxed{3 times}, \boxed{New York}, etc.

Question: {question}
Answer:
```

Its abstention-enabled counterpart is:

Abstention-enabled QA Prompt for Qwen3-8B

```
You are a question answering model, your task is to answer the question based on factual
    information in your own knowledge.

Please adhere to the following guidelines when formulating the answer:
1. Answer with a single entity, as concise as possible.
2. If you are uncertain or don't know the answer, answer "I don't know".

Think carefully before answering. Output only the final answer in \boxed{}. For example, \
    boxed{I don't know}, \boxed{3 times}, \boxed{New York}, etc.

Question: {question}
Answer:
```

Correctness and abstention judge. The following is the complete user message sent to the judge model. It is shared by reward computation and evaluation.

Prompt for LLM-as-a-Judge

```
Your job is to look at a question, a gold target, and a predicted answer, and then assign a
    grade of either ["CORRECT", "INCORRECT", "ABSTENTION"].
The predicted answer you receive is the final extracted answer only, not the model's chain-of-
    thought.
Non-negotiable rule: if the predicted answer is a placeholder, template slot, anonymized
    variable, or missing-value marker instead of the actual requested answer, grade it as
    INCORRECT. It does not matter whether the placeholder has the right answer type.
First, I will give examples of each grade, and then you will grade a new example.


The following are examples of CORRECT predicted answers.
```
Question: What are the names of Barack Obama's children?
Gold target: Malia Obama and Sasha Obama
Predicted answer 1: sasha and malia obama
Predicted answer 2: Barack Obama has two daughters. Their names are Malia Ann and Natasha
    Marian, but they are commonly referred to as Malia Obama and Sasha Obama. Malia was born
    on July 4, 1998, and Sasha was born on June 10, 2001.
```
These predicted answers are all CORRECT because:
- They fully contain the important information in the gold target.
- They do not contain any information that contradicts the gold target.
- Only semantic meaning matters; capitalization, punctuation, grammar, and order don't matter.
- Hedging and guessing are not allowed.
```

The following are examples of INCORRECT predicted answers.
```
Question: What are the names of Barack Obama's children?
Gold target: Malia and Sasha
Predicted answer 1: Malia.
Predicted answer 2: Malia, Sasha, and Susan.
Predicted answer 3: Barack Obama does not have any children.
Predicted answer 4: I think it's either Malia and Sasha. Or it could be Malia and Jackie. Or
    it could be Joey and Malia.
Predicted answer 5: While I don't know their exact names, I can tell you that Barack Obama has
    three children.
Predicted answer 6: <NAME>
Predicted answer 7: [person]
Predicted answer 8: <the answer>
```
These predicted answers are all INCORRECT because:
- A factual statement in the answer contradicts the gold target.
- The response is not an answer, or is an answer to another question.
- Incorrect statements that have some hedging (e.g., "it is possible that", "although i'm not sure, i think") are also considered incorrect.
- A placeholder, template slot, anonymized variable, or missing-value marker is not an answer, even if it has the right type.

The following are examples of ABSTENTION predicted answers.
```
Question: What are the names of Barack Obama's children?
Gold target: Malia and Sasha
Predicted answer 1: I don't know.
Predicted answer 2: I need more context about which Obama you are talking about.
Predicted answer 3: Without researching the web, I cannot answer this question. However, I can
    tell you that Barack Obama has two children.
Predicted answer 4: Barack Obama has two children. I know that one of them is Malia, but I'm
    not sure about the other one.
```
These predicted answers are all ABSTENTION because:
- They explicitly state inability to answer, and no statements in the answer contradict the gold target.
- They explicitly state that the question is unanswerable or fictional.

Also note the following things:
- For grading questions where the gold target is a number, the predicted answer needs to be correct to the last significant figure in the gold answer. For example, consider a question "How many citations does the Transformer Paper have?" with gold target "120k".
- Predicted answers "120k", "124k", and "115k" are all CORRECT.
- Predicted answers "100k" and "113k" are INCORRECT.
- The gold target may contain more information than the question. In such cases, the predicted answer only needs to contain the information that is in the question.
- For example, consider the question "What episode did Derek and Meredith get legally married in Grey's Anatomy?" with gold target "Season 7, Episode 20: White Wedding". Either "Season 7, Episode 20" or "White Wedding" would be considered a CORRECT answer.
- Do not punish predicted answers if they omit information that would be clearly inferred from the question.
- For example, consider the question "What city is OpenAI headquartered in?" and the gold target "San Francisco, California". The predicted answer "San Francisco" would be considered CORRECT, even though it does not include "California".
- Consider the question "What award did A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality, & toxicity win at NAACL '24?", the gold target is "Outstanding Paper Award". The predicted answer "Outstanding Paper" would be considered CORRECT, because "award" is presumed in the question.

- For the question "What is the height of Jason Wei in meters?", the gold target is "1.73 m". The predicted answer "1.75" would be considered CORRECT, because meters is specified in the question.
- For the question "What is the name of Barack Obama's wife?", the gold target is "Michelle Obama". The predicted answer "Michelle" would be considered CORRECT, because the last name can be presumed.
- Do not punish for typos in people's name if it's clearly the same name.
- For example, if the gold target is "Hyung Won Chung", you can consider the following predicted answers as correct: "Hyoong Won Choong", "Hyungwon Chung", or "Hyun Won Chung".
- Always grade placeholders as INCORRECT, not CORRECT or ABSTENTION. A placeholder is text that stands in for a missing answer instead of providing the actual entity, value, title, place, genre, date, or number.

Grade the predicted answer of this new question as one of:
A: CORRECT
B: INCORRECT
C: ABSTENTION
Respond with exactly one of the letters "A", "B", or "C", with no punctuation, asterisks or text before or after it.

Here is a new example. Simply reply with either "A", "B", "C". Don't apologize or correct yourself if there was a mistake; we are just trying to grade the answer.
Before giving the verdict, apply this check: if the predicted answer is a placeholder or template slot rather than a concrete answer, the verdict must be B.
```
Question: {question}
Gold target: {target}
Predicted answer: {predicted_answer}
```
Verdict: