TIAR: Trajectory-Informed Advantage Reweighting for LLM Abstention Learning

Anonymous ACL submission

Abstract

This paper investigates large language model (LLM) abstention learning, specifically using ternary reward, which incentivize truthfulness in large language models. This paper extends that idea by moving from a ternary reward to a Trajectory-Informed advantage reweighting, dynamically re-weights the abstention reward during Group Relative Policy Optimization (GRPO) training. The objective of this work focuses on abstention learning instead of improving truthfulness, serving as an exploration into hallucination reduction. The novelty of this paper lies in methodological innovation, advantage re-weighting, and benchmark selection. Leveraging GRPO's multiple trajectories as a natural abstention signal, this method uses a reward signal to explore knowledge boundaries and encourage consistency. By demonstrating that trajectories can be used as a confidence indicator of the policy relative to the query, they are then used to dynamically calculate the abstention advantage. AbstentionBench is used as the evaluation benchmark, as this work aims to contribute to the field of abstention learning. All datasets on the benchmark were tested against this method and various baselines. Empirical results demonstrate that TIAR achieves state-of-the-art abstention F1 scores across five of six evaluation categories, outperforming the static ternary baseline on 17 of 31 benchmark datasets while fully preserving baseline accuracy.

1 Introduction

Abstention learning has become a main theme in assisting hallucination reduction in large language models. The purpose of abstention learning is to teach LLMs to say "I don't know" for questions that they do not have the ability to answer. Such behavior benefits high-stakes domain applications, such as in medicine, when a query relates to the personal safety of a patient and is difficult to answer, in this case, abstention is preferred over hallucination for safety considerations. In the recent work "Why Large Language Models Hallucinate" by OpenAI (Kalai et al., 2025), the authors mention the importance of using abstention to mitigate hallucination and call for further exploration of abstention learning. Previous work (Wen et al., 2025) summarizes the current progress in abstention learning, spanning alignment and inference.

In alignment, supervised fine-tuning methods such as R-Tuning (Zhang et al., 2024) prepare datasets where labels for unanswerable questions are rewritten as "I don't know." While fine-tuning reduces hallucinations in specific reasoning tasks, such as logical translation (Pan et al., 2025a), abstention offers a generalized safety mechanism when models reach their broad knowledge boundaries. Preference learning exploration (Rafailov et al., 2024) similarly prepares offline reinforcement learning datasets with sets of answers that include an "I don't know" option. Another category focuses on inference-stage abstention, such as sampling the same query multiple times to estimate confidence and decide whether to abstain based on the result (Cole et al., 2023; Lin et al., 2022; Phute et al., 2024; Zhao et al., 2022; Slobodkin et al., 2023).

The state-of-the-art method in the alignment approach is TruthRL (Wei et al., 2025), which uses GRPO training with a ternary reward (correct trajectory: $+1$, abstention trajectory: 0, incorrect trajectory: -1) to incentivize model truthfulness, representing hallucination reduction with improved accuracy. Previous work (Kadavath et al., 2022) showed that question difficulty affects a model's response accuracy, and during GRPO training, the model tends to converge to a single trajectory that provides the highest advantage for a given query, indicating increasingly consistent outputs through

[0]Our code and resources are available at: https://anonymous.4open.science/r/TIAR_repo/.

training (Shao et al., 2024). To better teach abstention, we use GRPO's multiple sampled trajectories serve as a natural signal for abstention: for consistent trajectories, the decision is straightforward. For mixed trajectories, a reward signal can be used to explore the knowledge boundary and encourage consistency.

AbstentionBench (Kirichenko et al., 2025) is a comprehensive benchmark that tests abstention ability across 20 datasets and 6 scenarios. It uses abstention F1, recall, and precision as main metrics, and provides accuracy as a complementary metric. To demonstrate generalizability, empirical evaluation is performed on AbstentionBench across all datasets. Abstention F1 reflects accurate abstention while avoiding over-abstention, and accuracy measures correctness on questions with reference answers. Since AbstentionBench showed no direct relationship between abstention F1 and accuracy, our goal is to improve abstention while preserving accuracy as an independent capability indicator. Compared to baselines, TIAR achieves better abstention F1 and accuracy on most datasets, indicating it is a more effective abstention learning method that improves the model on both fronts.

Overall, our proposed method TIAR is a simple and practical approach that harnesses the natural advantage of GRPO to dynamically estimate self-confidence and teach LLM abstention, rather than relying on a static reward function when query difficulty varies. Empirically, TIAR achieves state-of-the-art abstention F1 scores across five of six evaluation categories and outperforms the static ternary baseline on 17 of 31 AbstentionBench datasets, all while preserving crucial baseline accuracy. Grounded in rigorous derivation, we believe this quantifiable improvement represents a significant step forward in the direction of abstention learning, providing valuable insights for future work.

2 Related Works

Abstention Learning Abstention learning is defined as teaching LLMs to refuse unanswerable questions, gradually becoming a key field for hallucination mitigation. Supervised Fine-Tuning approaches like Alignment for Honesty (Yang et al., 2024b) and R-tuning prepare datasets explicitly labeling answerable questions as "I don't know". Offline Reinforcement Learning approaches (Cheng et al., 2024) prepare datasets containing sets of answers to each query, learning to pick the "I don't know" choice through preference learning. GRPO approaches are among the most effective. One approach (Gul et al., 2025) first fine-tunes the model to use searching tools for unanswerable questions, later replacing them with abstention during GRPO. Other approaches innovate on the reward function, showing that traditional binary rewards encourage guessing, while ternary rewards successfully reduce hallucination rates by 19.1% and improve truthfulness (Wei et al., 2025). Alternative strategies include decoding-based (Chae et al., 2024), Conformal Prediction (Tayebati et al., 2025), and Uncertainty Estimation methods (Cohen et al., 2024). TIAR uniquely uses GRPO trajectories as a natural abstention signal to explore knowledge boundaries, dynamically reweighting advantages based on query difficulty. Our objective focuses on abstention learning, rather than directly improving truthfulness.

Reinforcement learning for Hallucination Mitigation Reinforcement Learning is widely applied for LLM hallucination mitigation, traditional pre-training objective of maximizing likelihood focuses on the next token prediction but not factual accuracy, it tends to generate next token that calibrated from the learned data distribution from pre-training. (Ouyang et al., 2022) Reinforcement Learning from Human Feedback (RLHF) incentivizes this behavior by training the model to calibrate with human preferred response style, which make the model's response confident and sound but at the expense of factual accuracy. (OpenAI et al., 2024) Reinforcement Learning with Verifiable Rewards (RLVR) paradigms using binary reward signals exacerbate this behavior to train the model to be a good test taker, encouraging guessing over abstention. Beyond hallucination mitigation, RL and fine-tuning signals are increasingly leveraged to improve other complex capabilities of LLMs, ranging from the quantization of large reasoning models (Zhang et al., 2026) to sophisticated NLP and RL frameworks applied to SAT solving (Pan et al., 2025b).

Beyond the abstention-specific methods, under the behavioral calibration approach, previous work (An and Xu, 2025) applies Semantic Clustering to GRPO trajectories, determining correct trajectories by comparing cluster sizes against a manually set threshold. TIAR differs by avoiding explicit Semantic Clustering and manual thresholds, offering a simpler reward design that effectively boosts both

abstention F1 and accuracy. Another work (Wu et al., 2026) requires users to explicitly specify a risk score $t \in [0, 1]$, scaling the ternary abstention reward to $2t-1$, where high-risk queries encourage abstention. TIAR differs by eliminating the need for manual risk inputs, dynamically adapting to query difficulty on its own, and focusing strictly on abstention learning rather than broad hallucination mitigation under varying risks.

3 Methodology

3.1 Preliminaries

Knowledge boundary probing refers to constructing a dataset containing both answerable and unanswerable questions relative to the model's capabilities. For each training question, multiple responses are sampled from the LLM; if none are correct, the question is marked as out-of-knowledge (OOK) and labeled with "I don't know" as the ground-truth answer. This approach has been explored in prior works such as R-Tuning (Zhang et al., 2024).

TIAR is implemented using the online RL algorithm GRPO (Shao et al., 2024), which optimizes:

$$\mathcal{L}_{\text{GRPO}}(\theta) = -\mathbb{E}_{x,\{y_i\}_{i=1}^{G}}\Bigg[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|y_i|}\sum_{t=1}^{|y_i|}\min\Big(w_{i,t}\hat{A}_i, \text{clip}(w_{i,t}, 1-\epsilon, 1+\epsilon)\hat{A}_i\Big) - \beta D_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}})\Bigg], \tag{1}$$

where $\mathcal{D}$ is the dataset distribution, θ represents the model parameters, $x \sim \mathcal{D}$, $\{y_i\}_{i=1}^{G} \sim \pi_{\theta_{\text{old}}}(\cdot|x)$, ϵ and β are hyperparameters, G is the group size (number of sampled responses per query), π_{ref} is the reference policy, $w_{i,t} = w_{i,t}(\theta)$ is the importance ratio, D_{KL} is the Kullback-Leibler divergence, and $\hat{A}_i$ is the estimated advantage for response y_i, computed via group-level normalization:

$$\hat{A}_i = \frac{r(x, y_i) - \text{mean}\{r(x, y_j)\}_{j=1}^{G}}{\text{std}\{r(x, y_j)\}_{j=1}^{G}}. \tag{2}$$

3.2 Reward Structure and the Abstention Decision

Consider a model facing a question x with group size G. Among the G sampled trajectories, let n_c, n_w, and n_a denote the number of correct, wrong, and abstention trajectories respectively, with $n_c + n_w + n_a = G$. Under the ternary reward scheme of TruthRL (Wei et al., 2025):

$$R_c = +1, \quad R_w = -1, \quad R_a = 0. \tag{3}$$

We define the model's empirical correctness rate among attempted (non-abstaining) trajectories as:

$$\hat{p} = \frac{n_c}{n_c + n_w}, \quad \text{when } n_c + n_w > 0. \tag{4}$$

This quantity estimates the model's probability of answering correctly *given that it attempts*, and serves as a natural proxy for question difficulty relative to the current policy: high $\hat{p}$ indicates an easy question, while low $\hat{p}$ indicates a hard one.

When should the model abstain? From the model's perspective, the expected value of attempting to answer is:

$$V_{\text{attempt}}(\hat{p}) = \hat{p} \cdot R_c + (1 - \hat{p}) \cdot R_w = 2\hat{p} - 1. \tag{5}$$

The model should prefer abstention (receiving R_a) over attempting when $R_a > V_{\text{attempt}}(\hat{p})$. Under the ternary reward ($R_a = 0$), this becomes $0 > 2\hat{p} - 1$, i.e., $\hat{p} < 0.5$. The ternary reward thus implicitly sets a fixed abstention threshold at $\hat{p}^* = 0.5$, regardless of the actual difficulty distribution encountered during training.

The opportunity cost of abstention. We derive the optimal abstention reward by computing the net value of abstaining versus attempting. When the model abstains, it avoids the expected loss $(1-\hat{p}) \cdot |R_w|$ but omits the expected gain $\hat{p} \cdot R_c$. The net value is:

$$R_a^* = \underbrace{(1 - \hat{p}) \cdot |R_w|}_{\text{loss avoided}} - \underbrace{\hat{p} \cdot R_c}_{\text{gain costed}} = 1 - 2\hat{p}. \tag{6}$$

This dynamic reward has the following properties:

- $\hat{p} = 0$ (all attempts fail): $R_a^* = 1$, abstention is maximally rewarded.
- $\hat{p} = 0.5$: $R_a^* = 0$, recovering the ternary reward as a special case.
- $\hat{p} = 1$ (all attempts succeed): $R_a^* = -1$, abstention is maximally penalized.

The ternary reward is thus a special case of the dynamic reward when $\hat{p} = 0.5$, treating all questions as equally difficult, an assumption that does not hold in practice.

3.3 The Coupling Problem in GRPO

A natural approach would be to directly substitute $R_a = 1 - 2\hat{p}$ into the reward and compute GRPO advantages via Eq. 2. However, this introduces a *coupling problem*: because all advantages share the same group mean $\bar{R}$ and standard deviation σ, modifying R_a changes the normalization statistics, which distorts $\hat{A}_c$ and $\hat{A}_w$. The group mean under ternary rewards is:

$$\bar{R}_{\text{ternary}} = \frac{n_c - n_w}{G}, \tag{7}$$

while under the dynamic reward $R_a = 1 - 2\hat{p}$:

$$\bar{R}_{\text{dynamic}} = \bar{R}_{\text{ternary}} + \frac{n_a(1 - 2\hat{p})}{G}. \tag{8}$$

On hard questions ($\hat{p} < 0.5$), the dynamic reward is positive, pushing $\bar{R}_{\text{dynamic}} > \bar{R}_{\text{ternary}}$. This reduces the advantage for correct trajectories:

$$\hat{A}_c^{\text{dynamic}} = \frac{R_c - \bar{R}_{\text{dynamic}}}{\sigma_{\text{dynamic}}} < \frac{R_c - \bar{R}_{\text{ternary}}}{\sigma_{\text{ternary}}} = \hat{A}_c^{\text{ternary}}, \tag{9}$$

where σ_{dynamic} and σ_{ternary} represent the standard deviations of the respective reward schemes. This is undesirable: on hard questions where correct answers are rare and valuable, the coupled approach weakens the learning signal for those correct trajectories. We empirically confirm this: a coupled implementation degrades accuracy by -0.76% compared to the ternary baseline, despite improving abstention F1 by $+0.15\%$.

Table 1: Coupled reward modification ($R_a = 1 - 2\hat{p}$) vs. ternary baseline (TruthRL) at step 20, averaged across all AbstentionBench datasets.

Method	Acc.	Abstention F1
TruthRL (ternary)	67.07	72.45
Coupled	66.31	72.60
Δ	-0.76	$+0.15$
Win / Loss	6/16	13/13

3.4 TIAR: Decoupled Advantage Adjustment

To resolve the coupling problem, we propose **TIAR** (Trajectory-Informed Advantage Reweighting), which applies the derived insight as a post-hoc advantage adjustment rather than a reward modification. The key idea is to preserve the standard GRPO normalization for correct and incorrect trajectories while separately adjusting the abstention advantage based on the group's empirical difficulty.

Algorithm. Given a GRPO training step with group size G, the specific mechanism is outlined in Algorithm 1. The overall framework is illustrated in Figure 1.

Algorithm 1 TIAR: Decoupled Advantage Adjustment

Require: Query batch $\{x_i\}$, group size G, strength λ
for each query x_i in batch **do**
  Sample G trajectories $\{y_1, \ldots, y_G\} \sim \pi_{\theta_{\text{old}}}(\cdot|x_i)$
  Assign ternary rewards: $r_j \in \{+1, -1, 0\}$
  Compute GRPO advantages via Eq. 2
  $n_c \leftarrow |\{j : r_j{=}1\}|$, $n_w \leftarrow |\{j : r_j{=}{-1}\}|$
  if $n_c + n_w > 0$ **then**
    $\hat{p} \leftarrow n_c/(n_c + n_w)$
    for each j where $r_j = 0$ **do**
      $\hat{A}_j \leftarrow \hat{A}_j + \lambda(1 - 2\hat{p})$
    end for
  end if
end for
Update π_θ with modified advantages (Eq. 1)

Connection to the opportunity cost derivation. The adjustment $\lambda(1-2\hat{p})$ is the difference between the optimal and ternary rewards:

$$\Delta_a = R_a^* - R_a^{\text{ternary}} = (1-2\hat{p}) - 0 = 1 - 2\hat{p}, \tag{10}$$

applied as a post-normalization adjustment. The scaling factor λ controls the correction strength. In all experiments, we use $\lambda = 1.0$.

4 Experiment Setup

4.1 Training

Base model and dataset. We use Llama-3.1-8B-Instruct (Grattafiori et al., 2024) and Qwen3-8B (Yang et al., 2025) as the base models for all methods. For GRPO training, we use the TruthRL-CRAG dataset (Wei et al., 2025; Yang et al., 2024a), which contains 656 training samples derived from the CRAG benchmark. Although all training questions have ground-truth answers, many lie beyond the model's parametric knowledge, creating a natural distribution of answerable and unanswerable questions relative to the model's capabilities.

GRPO configuration. Training is conducted using the verl framework (Sheng et al., 2025) with GRPO as the advantage estimator. We use a batch

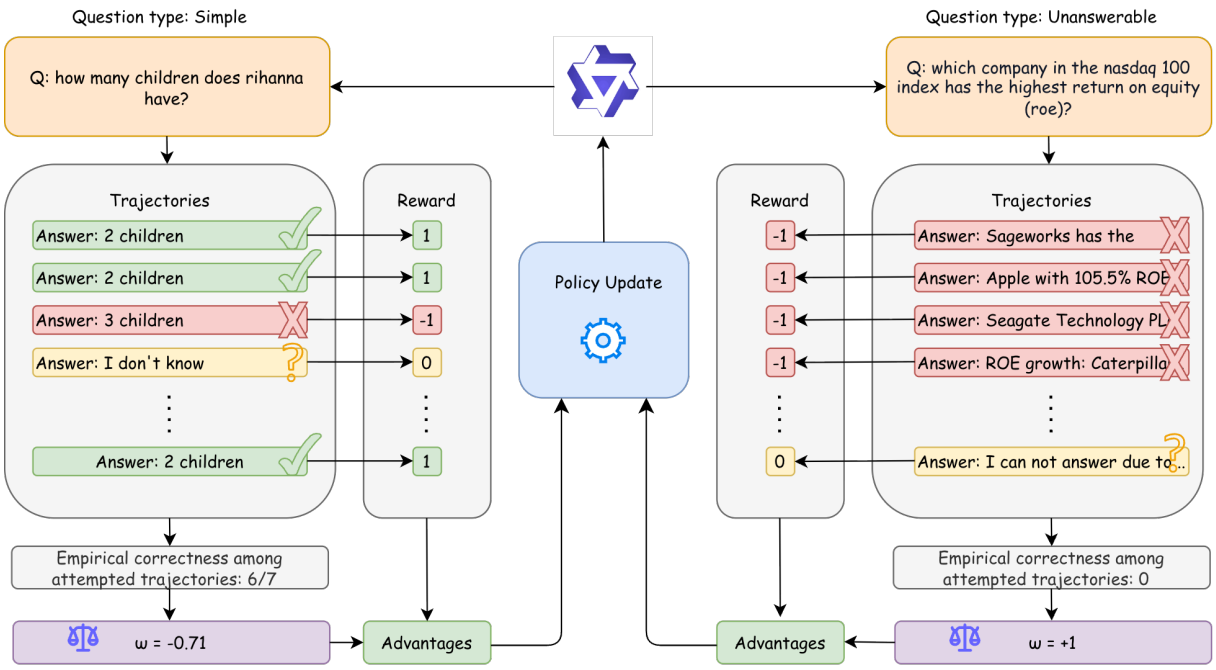

Figure 1: Abstention trajectories are reweighted based on the empirical correctness among attempted trajectories. TIAR discourages abstention for simple questions and boosts the advantages for unanswerable questions.

size of 64 and a group size (rollout n) of 8. The maximum prompt length is 16,384 tokens and maximum response length is 2,048 tokens. Hyperparameters such as the learning rate (1×10^{-6}) and the KL loss coefficient (0.001, with low-variance KL divergence) were directly adopted from the TruthRL baseline settings without further tuning to ensure a fair comparison. The scaling parameter $\lambda = 1.0$ for TIAR was determined via our ablation study.

We use vLLM (Kwon et al., 2023) as the rollout engine with tensor parallel size of 2 and GPU memory utilization of 0.8. FSDP is used for actor training with gradient checkpointing enabled. Training runs on 16 NVIDIA A100 40GB GPUs managed via SLURM. Training for 20 steps took approximately 6.7 hours. Total computational budget for a single model is 107.2 GPU-hours. Due to the high computational cost of GRPO, all reported results reflect a single training run with a fixed random seed.

Reward function. An external LLM judge (Llama-3.1-8B-Instruct served via vLLM) evaluates each generated response and assigns ternary rewards: $+1$ for correct, -1 for incorrect, and 0 for abstention. The judge classifies responses as abstentions using pattern matching and semantic analysis, then evaluates correctness of non-abstention responses against ground-truth answers.

Baselines. We compare TIAR against the following methods, all using Llama-3.1-8B-Instruct:

- **R-Tuning**: Supervised fine-tuning on the CRAG dataset with out-of-knowledge questions relabeled as "I don't know" via knowledge boundary probing, sampling 256 responses per question and marking questions as out-of-knowledge if none is correct.
- **RFT** (Rejection Fine-Tuning): Same knowledge boundary probing as SFT, but fine-tunes only on correctly answered responses, filtering out incorrect ones.
- **DPO** (Rafailov et al., 2024): Trains on preference pairs where "I don't know" is preferred over incorrect answers for out-of-knowledge questions.
- **TruthRL** (Wei et al., 2025): GRPO with ternary rewards (R_c=1, R_w=-1, R_a=0) using the same training configuration as TIAR, serving as the direct ablation that isolates the effect of the advantage adjustment.

4.2 Evaluation

Benchmark. We evaluate on Abstention-Bench (Kirichenko et al., 2025), a comprehensive benchmark containing 20 datasets across 31 subsets spanning six abstention scenarios: answer

unknown, underspecified intent, stale information, underspecified context, false premise, and subjective questions.

Metrics. Following AbstentionBench, we report four metrics: **Abstention F1** (harmonic mean of abstention precision and recall), **Abstention Recall** (proportion of correctly abstained unanswerable questions), **Abstention Precision** (proportion of warranted abstentions), and **Accuracy** (correctness of responses across all samples with reference answers). Abstention F1 is the primary metric reflecting the quality of abstention decisions, while accuracy serves as an independent indicator of the model's general capability.

Evaluation protocol. All models are evaluated using vLLM inference with temperature 0.8, top-p 0.95, and maximum generation length of 4,096 tokens. An LLM-as-Judge pipeline (Llama-3.1-8B-Instruct) performs two-stage evaluation: first classifying each response as abstention or non-abstention, then evaluating factual correctness of non-abstention responses against ground truth. This judge was validated against human annotation at 82.3% accuracy by the AbstentionBench authors.

5 Results

5.1 Main Results

Our main results are presented in Table 2. TIAR achieves highly competitive abstention F1 scores across all six categories, securing the highest score in False Premise for Llama-3.1-8B-Instruct, and in Underspecified Context and Subjective for Qwen3-8B, while preserving accuracy to the greatest extent among all methods. These improvements hold on both Llama-3.1-8B-Instruct and Qwen3-8B, confirming the generality of TIAR across model families.

Other observations. Qwen3-8B outperforms Llama across the board. On UMWP, all Qwen methods exceed 90 F1 while Llama methods hover around 75 to 79, suggesting a stronger baseline abstention ability. R-Tuning is strong on Qwen (best F1 on FreshQA and QAQA) but weakest on Llama (lowest F1, worst accuracy on FreshQA), indicating sensitivity to the base model's instruction-following capability. DPO degrades Llama accuracy on BB/Known Unknowns (69.6 vs. 87+ for others) due to over-abstention, as evidenced by high recall but low precision.

5.2 Ablation Study

As shown in Table 3, across all datasets, $\lambda = 1.0$ achieves the highest average abstention F1 (71.9) and nearly the highest accuracy (72.6 vs. 72.7 for $\lambda = 0$), outperforming $\lambda = 0$ on 17 of 31 datasets for F1 and 11 of 24 for accuracy. While $\lambda = 0.5$ is competitive on the six representative datasets, $\lambda = 1.0$ performs better across the full benchmark. Setting $\lambda = 0.3$ causes a collapse in both metrics, demonstrating that a small inversion weight destabilizes training without providing sufficient signal. The ablation reveals a non-monotonic relationship between λ and performance: full inversion ($\lambda = 1.0$) achieves the best balance of abstention quality and accuracy preservation, making it the recommended default.

5.3 Comparison with State-of-the-Art Proprietary Models

Table 4 presents the comparison of TIAR, built on an 8B open-source model, which achieves competitive abstention F1 against frontier proprietary models that are orders of magnitude larger. On Answer Unknown questions, TIAR outperforms all three API models in both F1 and precision, demonstrating that recognizing knowledge boundaries does not require massive scale. TIAR also achieves the highest precision on QAQA (90.3) and ties GPT-5.2 for the best accuracy on BBQ (73.5). The accuracy gap is expected, frontier models possess superior world knowledge, and is most visible on FreshQA, where temporal knowledge drives accuracy. However, the proprietary models exhibit their own weaknesses: Gemini 3 achieves perfect F1 on KUQ/Controversial but only 50.0 accuracy due to severe over-abstention, and GPT-5.2 struggles on FreshQA (F1 of 48.8) due to overconfidence on temporally sensitive questions. These results show that targeted RL training can close much of the abstention gap between open-source 8B models and frontier APIs.

6 Conclusion

We presented TIAR, a reinforcement learning method that teaches Large Language Models to confidently and correctly abstain on unanswerable questions. TIAR achieves this by dynamically inverting the reward signal gathered from Group Relative Policy Optimization (GRPO) training trajectories. Our approach tackles the inherent challenge of learning abstention without negatively distort-

Table 2: Abstention F1, Recall, Precision, and Accuracy across six AbstentionBench categories (one representative dataset per category). Best scores per metric per model are **bold**.

Method		Representative Dataset (AbstentionBench Category)					
Model	Metric	BB/Known Unk. Answer Unknown	BBQ Underspecified Intent	FreshQA Stale	UMWP Underspecified Context	QAQA False Premise	KUQ/Cont. Subjective
Llama-3.1-8B-Instruct							
R-Tuning	Abstention F1	91.7	82.4	69.9	75.3	56.0	75.0
	Abstention Recall	95.7	90.6	66.1	60.7	45.6	67.7
	Abstention Precision	88.0	75.5	74.1	99.1	72.6	84.1
	Accuracy	**91.3**	62.7	40.0	86.2	44.2	66.3
RFT	Abstention F1	**97.9**	**85.9**	72.2	76.1	57.1	**77.7**
	Abstention Recall	**100.0**	84.8	66.7	61.7	45.6	**70.9**
	Abstention Precision	**95.8**	**87.2**	**78.7**	**99.2**	76.5	**86.0**
	Accuracy	87.0	67.6	50.0	86.7	44.9	67.8
DPO	Abstention F1	93.9	84.4	72.1	75.5	58.0	74.6
	Abstention Recall	**100.0**	91.2	70.1	61.1	**47.0**	68.2
	Abstention Precision	88.5	78.5	74.3	98.7	75.7	82.1
	Accuracy	69.6	67.4	48.0	89.7	46.7	66.7
TruthRL	Abstention F1	95.8	84.5	**74.7**	**78.9**	58.0	74.8
	Abstention Recall	**100.0**	91.5	**75.7**	**65.5**	46.7	66.5
	Abstention Precision	92.0	78.5	73.6	**99.2**	76.4	85.5
	Accuracy	87.0	**71.7**	54.0	90.3	**50.5**	**70.7**
TIAR (Ours)	Abstention F1	**97.9**	84.6	74.5	77.6	**58.2**	76.2
	Abstention Recall	**100.0**	**92.4**	75.1	63.8	46.7	69.7
	Abstention Precision	**95.8**	77.9	73.9	99.1	**77.3**	83.9
	Accuracy	87.0	70.7	**55.0**	**91.7**	48.4	**70.7**
Qwen3-8B							
R-Tuning	Abstention F1	82.1	**91.2**	**81.2**	92.4	**79.9**	87.0
	Abstention Recall	**100.0**	**85.6**	**84.2**	**97.8**	**90.2**	91.7
	Abstention Precision	69.7	97.6	78.4	87.6	71.8	82.8
	Accuracy	**100.0**	98.1	89.0	97.9	82.5	95.7
RFT	Abstention F1	**89.8**	90.4	71.9	91.9	70.3	72.4
	Abstention Recall	95.7	84.5	67.2	92.9	70.5	68.2
	Abstention Precision	**84.6**	97.2	77.3	**90.9**	70.0	77.2
	Accuracy	78.3	95.3	73.0	95.9	67.7	88.0
DPO	Abstention F1	83.6	90.8	79.6	92.5	78.0	87.1
	Abstention Recall	**100.0**	84.9	81.4	97.6	88.8	**93.2**
	Abstention Precision	71.9	97.5	77.8	87.9	69.5	81.8
	Accuracy	95.7	98.0	84.0	**98.5**	**85.3**	96.4
TruthRL	Abstention F1	85.2	90.8	81.1	92.9	79.1	86.7
	Abstention Recall	**100.0**	84.6	82.5	97.5	87.7	91.7
	Abstention Precision	74.2	97.9	**79.8**	88.8	**72.0**	82.2
	Accuracy	**100.0**	98.1	**90.0**	98.3	84.6	**96.7**
TIAR (Ours)	Abstention F1	79.3	**91.2**	80.4	**94.0**	79.7	**87.2**
	Abstention Recall	**100.0**	85.3	81.4	97.6	89.5	90.8
	Abstention Precision	65.7	**98.1**	79.6	90.7	71.8	**83.8**
	Accuracy	**100.0**	**98.3**	88.0	**98.5**	81.8	95.7

ing the advantage normalization of standard correct and incorrect trajectories, an issue we termed the coupling problem. By decoupling the advantage adjustment for abstention responses, TIAR adapts to the empirical difficulty of each query dynamically. Experiments rigorously conducted on the extensive AbstentionBench framework demonstrate that TIAR successfully achieves state-of-the-art abstention F1 scores across five of six complex evaluation categories. These strong gains were consistent across entirely distinct open-source model families, specifically Llama-3.1-8B-Instruct and Qwen3-8B, proving both generalizability and the preservation of crucial baseline accuracy. The ablation study formally confirms full reward inversion ($\lambda = 1.0$) as the optimal default configuration for deploying TIAR. Furthermore, comparisons against robust APIs illustrate that TIAR effectively narrows the capability gap between significantly smaller, 8-billion parameter open-source models and heavily resourced frontier proprietary systems, such as Claude Sonnet 4.5 and GPT-5.2.

Future work will extend TIAR to multi-turn interactions and retrieval-augmented generation (RAG), tying abstention to context sufficiency rather than just parameter bounds. Additionally, exploring cur-

Table 3: Ablation study on the trajectory inversion weight λ in TIAR using Llama-3.1-8B-Instruct. $\lambda = 0$ reduces to the standard ternary reward (TruthRL). Best scores per metric are **bold**.

Method		Representative Dataset (AbstentionBench Category)					
Model	Metric	BB/Known Unk. Answer Unknown	BBQ Underspecified Intent	FreshQA Stale	UMWP Underspecified Context	QAQA False Premise	KUQ/Cont. Subjective
Llama-3.1-8B-Instruct							
TIAR ($\lambda = 0$)	Abstention F1	95.8	84.5	74.7	**78.9**	58.0	74.8
	Abstention Recall	**100.0**	91.5	**75.7**	**65.5**	46.7	66.5
	Abstention Precision	92.0	78.5	73.6	99.2	76.4	**85.5**
	Accuracy	**87.0**	71.7	54.0	90.3	50.5	**70.7**
TIAR ($\lambda = 0.3$)	Abstention F1	95.8	83.7	71.5	76.3	58.0	72.6
	Abstention Recall	**100.0**	91.3	72.3	61.9	48.4	66.5
	Abstention Precision	92.0	77.2	70.7	**99.5**	72.3	80.0
	Accuracy	82.6	65.9	37.0	89.9	47.4	68.8
TIAR ($\lambda = 0.5$)	Abstention F1	93.9	**84.6**	**74.9**	77.5	**61.4**	**77.1**
	Abstention Recall	**100.0**	91.2	75.1	63.7	**50.5**	**70.9**
	Abstention Precision	88.5	**78.9**	**74.7**	98.9	**78.3**	84.5
	Accuracy	82.6	**72.0**	**56.0**	91.5	**52.6**	70.3
TIAR ($\lambda = 1.0$)	Abstention F1	**97.9**	**84.6**	74.5	77.6	58.2	76.2
	Abstention Recall	**100.0**	**92.4**	75.1	63.8	46.7	69.7
	Abstention Precision	**95.8**	77.9	73.9	99.1	77.3	83.9
	Accuracy	**87.0**	70.7	55.0	**91.7**	48.4	**70.7**

Table 4: Comparison of TIAR (Llama-3.1-8B-Instruct) against state-of-the-art proprietary API models on AbstentionBench. Best scores per metric are **bold**.

Method		Representative Dataset (AbstentionBench Category)					
Model	Metric	BB/Known Unk. Answer Unknown	BBQ Underspecified Intent	FreshQA Stale	UMWP Underspecified Context	QAQA False Premise	KUQ/Cont. Subjective
Proprietary API Models							
Claude Sonnet 4.5	Abstention F1	93.9	87.9	**75.2**	76.7	66.7	92.3
	Abstention Recall	**100.0**	**100.0**	**73.2**	62.3	55.6	85.7
	Abstention Precision	88.5	78.5	77.4	**100.0**	83.3	**100.0**
	Accuracy	**82.6**	59.2	63.6	**100.0**	56.5	**100.0**
GPT-5.2	Abstention F1	93.6	86.0	48.8	**80.9**	66.0	92.9
	Abstention Recall	95.7	84.3	35.7	**67.9**	59.3	92.9
	Abstention Precision	91.7	87.8	76.9	**100.0**	74.4	92.9
	Accuracy	**82.6**	**73.5**	70.5	97.9	65.2	75.0
Gemini 3	Abstention F1	84.4	**97.0**	32.8	76.7	**75.8**	**100.0**
	Abstention Recall	82.6	96.1	19.6	62.3	**66.7**	**100.0**
	Abstention Precision	86.4	**98.0**	**100.0**	**100.0**	87.8	**100.0**
	Accuracy	**82.6**	67.3	**81.8**	95.7	**78.3**	50.0
TIAR (Ours)	Abstention F1	**95.8**	87.5	68.4	72.3	65.9	92.9
	Abstention Recall	**100.0**	96.1	71.4	56.6	51.9	92.9
	Abstention Precision	**92.0**	80.3	65.6	**100.0**	**90.3**	92.9
	Accuracy	**82.6**	**73.5**	52.3	95.7	43.5	75.0

riculum learning for the inversion weight λ during GRPO training could yield more stable convergence and broader applicability to general-purpose agents.

Limitations

While TIAR demonstrates strong empirical results in dynamically teaching LLMs to abstain, it possesses several limitations that warrant consideration for future research and deployment:

First, our methodology heavily relies on an external LLM judge (in our case, Llama-3.1-8B-Instruct via vLLM) to accurately assign ternary rewards during the RL training phase. The quality of the learned abstention policy is fundamentally bottlenecked by the judge's ability to accurately evaluate the correctness of attempts and identify subtle variations of abstentions. Misclassifications by the judge directly translate into noisy advantage signals during GRPO normalization.

Second, the structural requirement of generating multiple trajectories ($G = 8$ in our experiments) per query during GRPO training incurs a non-trivial computational overhead compared to supervised fine-tuning or DPO. While GRPO successfully bypasses the need for a memory-intensive separate

critic model, the pure rollout phase remains computationally heavy, which limits its immediate scalability for researchers with restricted GPU budgets.

Third, the empirical correctness rate $\hat{p}$ serves as an effective mathematical proxy for estimating question difficulty relative to the model, but it is inherently sensitive to the base model's initial calibration state. If a base model is overly confident but highly inaccurate, the initial advantage signals might fluctuate intensely before the policy successfully learns the boundaries of its genuine knowledge.

Fourth, our experimental evaluation is strictly constrained to single-turn question-answering scenarios, as necessitated by AbstentionBench. Real-world abstention is rarely binary; it frequently occurs in complex, multi-turn dialogues where an agent might ask clarifying questions, request specific context, or provide a partial answer rather than outright abstaining. This nuanced interaction paradigm is not seamlessly captured by the current static formulation of TIAR.

Finally, regarding the ethical considerations and potential societal risks of this work, deploying LLMs with automated abstention in high-stakes domains (such as medicine or law) carries the critical risk of false abstentions (refusing to answer when the model actually possesses life-saving knowledge) or false confidence (attempting to answer hallucinated facts). While TIAR improves boundary detection, it is not a foolproof guarantee of safety. Future real-world deployments should combine RL-based abstention with external retrieval systems and rigorous human-in-the-loop oversight to actively mitigate these potential societal harms.

References

Hao An and Yang Xu. 2025. Teaching llms to abstain via fine-grained semantic confidence reward. *Preprint*, arXiv:2510.24020.

Kyubyung Chae, Jaepill Choi, Yohan Jo, and Taesup Kim. 2024. Mitigating hallucination in abstractive summarization with domain-conditional mutual information. In *Findings of the Association for Computational Linguistics: NAACL 2024*, page 1809–1820. Association for Computational Linguistics.

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

Roi Cohen, Konstantin Dobler, Eden Biran, and Gerard de Melo. 2024. I don't know: Explicit modeling of uncertainty with an [idk] token. *Preprint*, arXiv:2412.06676.

Jeremy R. Cole, Michael J. Q. Zhang, Daniel Gillick, Julian Martin Eisenschlos, Bhuwan Dhingra, and Jacob Eisenstein. 2023. Selectively answering ambiguous questions. *Preprint*, arXiv:2305.14613.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Mustafa Omer Gul, Claire Cardie, and Tanya Goyal. 2025. Pay-per-search models are abstention models. *Preprint*, arXiv:2510.01152.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, and 17 others. 2022. Language models (mostly) know what they know. *Preprint*, arXiv:2207.05221.

Adam Tauman Kalai, Ofir Nachum, Santosh S. Vempala, and Edwin Zhang. 2025. Why language models hallucinate. *Preprint*, arXiv:2509.04664.

Polina Kirichenko, Mark Ibrahim, Kamalika Chaudhuri, and Samuel J. Bell. 2025. Abstentionbench: Reasoning llms fail on unanswerable questions. *Preprint*, arXiv:2506.09038.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. *Preprint*, arXiv:2309.06180.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Teaching models to express their uncertainty in words. *Preprint*, arXiv:2205.14334.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder,

Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

Muyu Pan, Dheeraj Kodakandla, and Mahfuza Farooque. 2025a. Fine-tuned large language models for logical translation: Reducing hallucinations with lang2logic. In *2025 International Symposium on Networks, Computers and Communications (ISNCC)*, pages 1–4.

Muyu Pan, Matthew Walter, Dheeraj Kodakandla, and Mahfuza Farooque. 2025b. Langsat: A novel framework combining nlp and reinforcement learning for sat solving. *Preprint*, arXiv:2512.04374.

Mansi Phute, Alec Helbling, Matthew Hull, ShengYun Peng, Sebastian Szyller, Cory Cornelius, and Duen Horng Chau. 2024. Llm self defense: By self examination, llms know they are being tricked. *Preprint*, arXiv:2308.07308.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Preprint*, arXiv:2305.18290.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *Preprint*, arXiv:2402.03300.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2025. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, EuroSys '25, page 1279–1297. ACM.

Aviv Slobodkin, Omer Goldman, Avi Caciularu, Ido Dagan, and Shauli Ravfogel. 2023. The curious case of hallucinatory (un)answerability: Finding truths in the hidden states of over-confident large language models. *Preprint*, arXiv:2310.11877.

Sina Tayebati, Divake Kumar, Nastaran Darabi, Dinithi Jayasuriya, Ranganath Krishnan, and Amit Ranjan Trivedi. 2025. Learning conformal abstention policies for adaptive risk management in large language and vision-language models. *Preprint*, arXiv:2502.06884.

Zhepei Wei, Xiao Yang, Kai Sun, Jiaqi Wang, Rulin Shao, Sean Chen, Mohammad Kachuee, Teja Gollapudi, Tony Liao, Nicolas Scheffer, Rakesh Wanga, Anuj Kumar, Yu Meng, Wen tau Yih, and Xin Luna Dong. 2025. Truthrl: Incentivizing truthful llms via reinforcement learning. *Preprint*, arXiv:2509.25760.

Bingbing Wen, Jihan Yao, Shangbin Feng, Chenjun Xu, Yulia Tsvetkov, Bill Howe, and Lucy Lu Wang. 2025. Know your limits: A survey of abstention in large language models. *Preprint*, arXiv:2407.18418.

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

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Xiao Yang, Kai Sun, Hao Xin, Yushi Sun, Nikita Bhalla, Xiangsen Chen, Sajal Choudhary, Rongze Daniel Gui, Ziran Will Jiang, Ziyu Jiang, Lingkun Kong, Brian Moran, Jiaqi Wang, Yifan Ethan Xu, An Yan, Chenyu Yang, Eting Yuan, Hanwen Zha, Nan Tang, and 8 others. 2024a. Crag – comprehensive rag benchmark. *Preprint*, arXiv:2406.04744.

Yuqing Yang, Ethan Chern, Xipeng Qiu, Graham Neubig, and Pengfei Liu. 2024b. Alignment for honesty. *Preprint*, arXiv:2312.07000.

Hanning Zhang, Shizhe Diao, Yong Lin, Yi R. Fung, Qing Lian, Xingyao Wang, Yangyi Chen, Heng Ji, and Tong Zhang. 2024. R-tuning: Instructing large language models to say 'i don't know'. *Preprint*, arXiv:2311.09677.

Nan Zhang, Eugene Kwek, Yusen Zhang, Muyu Pan, Suhang Wang, Prasenjit Mitra, and Rui Zhang. 2026. Quantlrm: Quantization of large reasoning models via fine-tuning signals. *Preprint*, arXiv:2602.02581.

Yao Zhao, Misha Khalman, Rishabh Joshi, Shashi Narayan, Mohammad Saleh, and Peter J. Liu. 2022. Calibrating sequence likelihood improves conditional language generation. *Preprint*, arXiv:2210.00045.

A Licenses of Used Artifacts

All models and datasets utilized in this research are publicly available under their respective licenses. Llama-3.1 models are licensed under the Llama 3.1 Community License. Qwen3 models are distributed under the Apache 2.0 License. The AbstentionBench and TruthRL-CRAG datasets are accessible under the MIT License and CC-BY 4.0 license, respectively.