KARL: Mitigating Hallucinations in LLMs via Knowledge-Boundary-Aware Reinforcement Learning

Anonymous ACL submission

Abstract

Enabling large language models (LLMs) to appropriately abstain from answering questions beyond their knowledge is crucial for mitigating hallucinations. While existing reinforcement learning methods foster autonomous abstention, they often compromise answer accuracy because their static reward mechanisms, agnostic to models' knowledge boundaries, drive models toward excessive caution. In this work, we propose KARL, a novel framework that continuously aligns an LLM's abstention behavior with its evolving knowledge boundary. KARL introduces two core innovations: a ***Knowledge-Boundary-Aware Reward*** that performs online knowledge boundary estimation using within-group response statistics, dynamically rewarding correct answers or guided abstention; and a ***Two-Stage RL Training Strategy*** that first explores the knowledge boundary and bypasses the "*abstention trap*", and subsequently converts incorrect answers beyond the knowledge boundary into abstentions without sacrificing accuracy. Extensive experiments on multiple benchmarks demonstrate that KARL achieves a superior accuracy-hallucination trade-off, effectively suppressing hallucinations while maintaining high accuracy across both in-distribution and out-of-distribution scenarios.

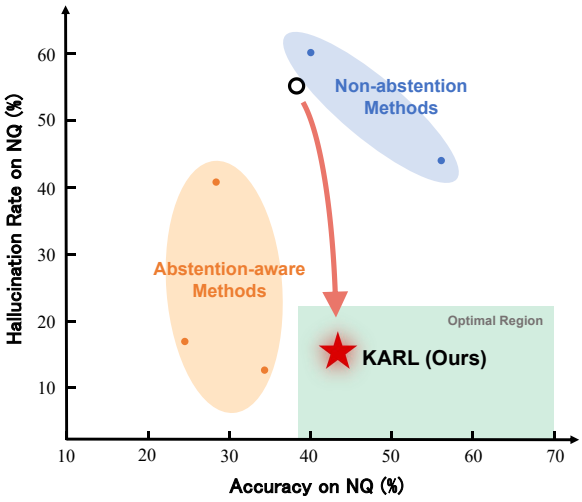

Figure 1: Hallucination rate vs. accuracy on NaturalQuestions (NQ) (Kwiatkowski et al., 2019). **KARL** achieves the optimal trade-off by increasing accuracy while effectively reducing hallucination compared to existing non-abstention and abstention-aware methods.

1 Introduction

Large language models (LLMs) have demonstrated remarkable capabilities in open-domain, knowledge-intensive tasks (OpenAI, 2025; Google DeepMind, 2025; Guo et al., 2025). However, their reliability is often undermined by hallucination—the tendency to generate plausible but factually incorrect responses with confidence, rather than admitting they do not know the answer (Maynez et al., 2020; Ji et al., 2023). A key cause of this issue is that conventional training typically penalizes abstention as an incorrect prediction, encouraging models to guess blindly in order to maximize training accuracy even when they lack sufficient knowledge (Kalai et al., 2025). Such behavior poses serious risks in high-stakes domains like medicine and law, where minor factual errors can lead to harmful consequences (Xiao et al., 2021; Dahl et al., 2024; Magesh et al., 2024; Nazi and Peng, 2024). Therefore, enabling LLMs to abstain appropriately has emerged as a fundamental objective in safety alignment (Wen et al., 2025).

Recently, Reinforcement Learning (RL) has gained significant attention for mitigating hallucination. This paradigm formulates abstention as a sequential decision-making process. Instead of forcing models to mimic fixed abstention responses, RL methods guide them to discover optimal policies through feedback signals (Christiano et al., 2017; Ouyang et al., 2022; Rafailov et al., 2023). For example, early efforts employed PPO or DPO to align models with abstention preferences (Bai

et al., 2022; Kadavath et al., 2022; Xu et al., 2024). More recently, Reinforcement Learning with Verifiable Rewards (RLVR) has become a dominant paradigm due to its effectiveness and data efficiency (Shao et al., 2024; DeepSeek-AI, 2024; Guo et al., 2025). Building upon this, notable works such as TruthRL leverage GRPO (Shao et al., 2024) with a ternary reward to jointly optimize correct answers, abstentions, and incorrect responses (Wei et al., 2025). While these methods allow models to autonomously learn when to abstain, they still suffer from substantial degradation in answer accuracy. The underlying issue is that existing methods often rely on static reward structures that are agnostic to the model's knowledge boundary, which drive policies to converge rapidly toward abstention. As a result, models tend to become overly cautious, suppressing both hallucinations and correct answers. In light of these observations, a central question arises: ***how to design a training framework that reduces hallucinations while maintaining answer accuracy, i.e., ensuring LLMs abstain only when exceeding their knowledge boundary?***

To tackle the above issues, we propose **KARL**, a novel training framework that continuously aligns the model's abstention behavior with its evolving knowledge boundary. KARL rests on two key innovations. First, we devise a ***Knowledge-Boundary-Aware Reward*** **(KAR)** mechanism that dynamically adjusts the reward intensity based on the presence of correct answers in each sampled group, thereby motivating the model to confidently explore questions within its competence while steering it toward abstention when knowledge is genuinely lacking. Second, to structurally decouple knowledge boundary exploration from abstention calibration, we design a ***two-stage RL training strategy***. An initial *knowledge boundary exploration stage* employs a mixed-reward strategy that maximizes accuracy and bypasses the "*abstention trap*" (See Fig. 2); a subsequent *abstention calibration stage* then converts persistent errors into appropriate abstentions without erasing earlier gains in answer accuracy. Together, these components enable LLMs to abstain only when exceeding their knowledge boundaries, thereby reducing hallucinations while maintaining high accuracy.

Extensive experiments on multiple benchmarks demonstrate that KARL achieves a superior accuracy-hallucination trade-off (Figure 1). Unlike prior methods that often sacrifice accuracy, KARL significantly slashes hallucination while maintaining accuracy, even achieving improvements on certain benchmarks. For instance, reducing hallucinations by 38.6% on NQ while improving accuracy by 4.0%. This capability to suppress hallucinations while maintaining high answer accuracy generalizes robustly across out-of-distribution inputs.

2 Pilot Study

Models often struggle to balance factual accuracy with appropriate abstention. In this section, we first formalize the abstention task and reward structures. We then investigate the underlying trade-off between answering and abstaining, focusing on two core aspects: First, we assess whether binary rewards improve pass@1 in short-form open-domain QA, extending prior successes in reasoning tasks (Yue et al., 2025; Zhang et al., 2025b). Second, we identify a critical failure mode of static ternary rewards: excessive early-stage abstention that leads to irreversible policy collapse. By tracing the evolution of response distributions, we uncover the mechanistic drift into this "*abstention trap*".

Task Formulation Given a factual query q, a language model π must decide whether to generate an answer a or to abstain from answering (e.g., output "I don't know"). The goal is to train a model that abstains only when it is likely to produce an incorrect answer, thereby improving reliability while preserving answer accuracy.

Formally, we consider a policy model π_θ which, given q, generates a response y. The response space is $\mathcal{Y} = \mathcal{A} \cup \{\varnothing\}$, where $\mathcal{A}$ is the set of possible answer texts and $\varnothing$ denotes abstention. Each query q is associated with a ground-truth answer $a^\star$, and the outcome of a response is evaluated with a ternary reward r based on its factual correctness:

$$r(y, a^\star) = \begin{cases} r_{\text{pos}}, & \text{if } y \text{ is correct}, \\ r_{\text{neg}}, & \text{if } y \text{ is incorrect}, \\ r_{\text{abs}}, & \text{if } y = \varnothing, \end{cases} \tag{1}$$

where typically $r_{\text{pos}} > r_{\text{abs}} \geqslant r_{\text{neg}}$. The objective is to optimize the policy $\pi_\theta(y|q)$ to maximize the expected reward $\mathbb{E}_{q,y\sim\pi_\theta}[r(y, a^\star)]$, thus learning when to answer and when to abstain.

Binary Rewards for Knowledge Boundary Exploration. Prior work shows RLVR bolsters model confidence in correct reasoning paths, improving pass@1 (Yue et al., 2025; Zhang et al., 2025b). To investigate whether this benefit extends to short-form open-domain QA, we adopt

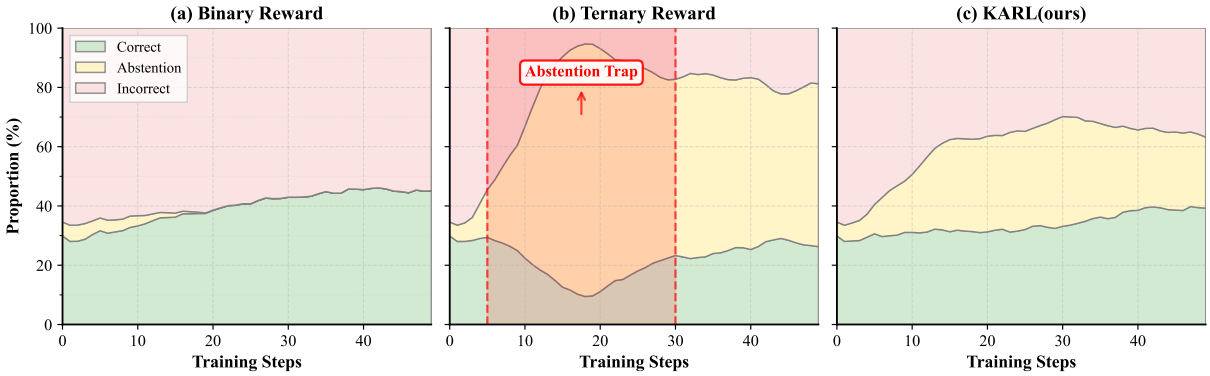

Figure 2: Training dynamics of response distributions for Llama-3.1-8B-Instruct on NQ under different reward schemes. The plots show the proportions of **Correct** (green), **Abstention** (yellow), and **Incorrect** (red) responses during early training for (a) Binary Reward, (b) Static Ternary Reward, and (c) our method. The red dashed region in (b) highlights a rapid increase in abstention accompanied by declining accuracy.

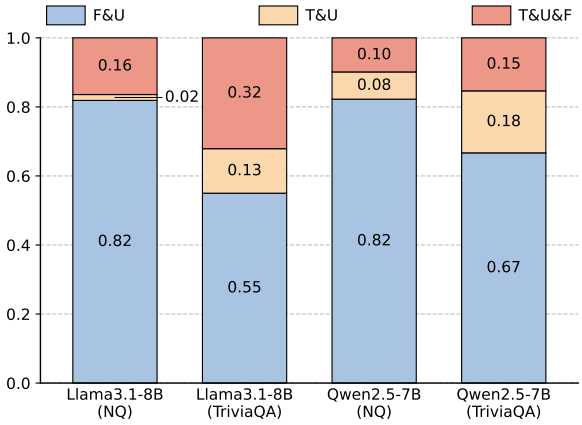

Figure 3: Distribution of rollout groups across different models and datasets. We categorize sampled groups based on the composition of Correct (**T**), Incorrect (**F**), and Abstention (**U**) responses. The chart displays the relative proportions of F&U, T&U, and T&U&F groups, excluding homogeneous groups and groups containing only T&F.

GRPO (Shao et al., 2024) as our optimization backbone (formalized in Appendix B) and train models on NQ with a binary reward (1 for correct answers, 0 otherwise; see Appendix C.1 for more details). As shown in Figure 2(a), abstention plunges to zero early, while accuracy rises steadily thereafter. Gains persist on out-of-distribution tasks (Table 1), confirming that binary rewards encourage exploration of knowledge boundary and improve accuracy, but entirely suppress abstention, inducing high hallucination rates.

Abstention Collapse under Static Ternary Rewards. Static ternary rewards (e.g., $\{r_{\text{pos}} = 1, r_{\text{abs}} = 0, r_{\text{neg}} = -1\}$) are used to induce calibrated abstention (Wei et al., 2025). However, they often trigger a critical failure mode: models rapidly maximize abstention at the expense of accuracy, often abstaining on all queries (Fig. 2(b)). Once this occurs, within-group variance vanishes, halting further learning.

To dissect this collapse, we analyze rollout distributions from Llama-3.1-8B-Instruct (Grattafiori et al., 2024) and Qwen2.5-7B-Instruct (Yang et al., 2024a) on NQ and TriviaQA (Joshi et al., 2017) (see Appendix C.1 for more details). From Fig. 3, groups containing only incorrect and abstention (i.e., F&U) responses dominate the distribution that influences abstention learning. Crucially, under group-normalized advantage, any abstention reward $r_{\text{abs}} > r_{\text{neg}}$ yields a consistent positive advantage for abstention in these groups. This creates an overwhelming structural bias early in training, driving the policy into an irreversible "*abstention trap*" before it can explore correct answers. We provide further empirical justification in Section 4.4.

Insights. Our findings confirm that RLVR exploration holds considerable potential for improving factual accuracy. Yet, static ternary rewards fail to harness this potential. Notably, early in training, models are driven by the predominance of incorrect-and-abstention (F&U) groups into an "*abstention trap*", which prevents accuracy gains and can even degrade performance. Thus, a method is needed that both leverages RLVR to enhance accuracy and counteracts the bias from F&U groups to avoid abstention collapse.

3 Methodology

In this section, we detail our proposed KARL methods, which address the core challenge of aligning a

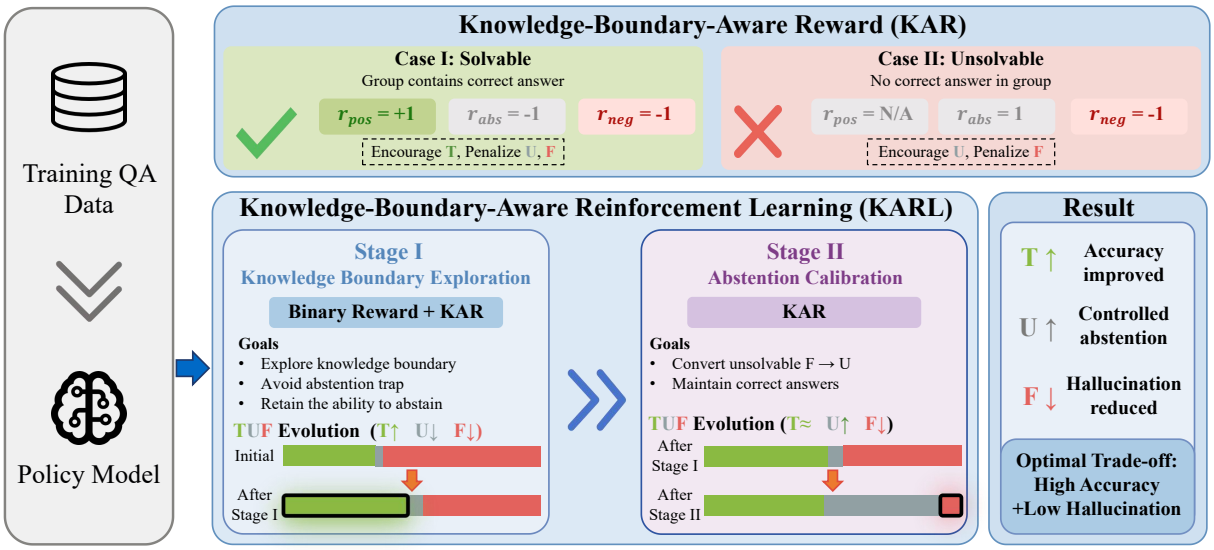

Figure 4: An overview of the KARL framework. KARL utilizes a Knowledge-Boundary-Aware Reward (KAR) and employs a two-stage strategy to first explore knowledge boundary and subsequently calibrate abstention behavior.

model's abstention behavior with its dynamically evolving knowledge boundary. As shown in Figure 4, KARL consists of two components: (1) *a Knowledge-Boundary-Aware Reward mechanism* designed to provide nuanced learning signals, and (2) *a Two-Stage RL Training Strategy* that decouples knowledge exploration from abstention calibration and prevents premature policy collapse.

3.1 Knowledge-Boundary-Aware Reward

A central challenge in abstention alignment is the accurate, real-time estimation of whether a given query lies within the model's current knowledge boundary. Prior methods typically rely on offline probing—sampling the model multiple times on a static training set to pre-label queries as answerable or not—before alignment begins (Zhang et al., 2024; Li et al., 2025; Wei et al., 2025). Although intuitive, this static approach incurs significant computational overhead and, more critically, fails to reflect the model's continuously evolving capability during the alignment process itself. To overcome these limitations, we propose the Knowledge-Boundary-Aware Reward (**KAR**), a mechanism that performs dynamic, online estimation of the model's knowledge boundary by leveraging the sampled responses during reinforcement learning.

Specifically, for each query q, the policy generates G responses $\{o_1, \cdots, o_G\}$ before the update, i.e., $\pi_{\theta_{\text{old}}}$. KAR treats the ground-truth $a^\star$ to q as a supervisory signal of the model's current capability. First, we define a binary solvability indicator:

$$\mathbb{I}_{\text{solvable}} = \begin{cases} 1 & \text{if } \exists o \in \{o_i\}_{i=1}^{G}, \text{ such that } o \text{ is correct,} \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

Based on this indicator, KAR provides distinct learning signals to guide the policy toward either confident answering or appropriate abstention. For query deemed *solvable* ($\mathbb{I}_{\text{solvable}} = 1$), the goal is to reinforce correct answering and discourage unnecessary abstention; thus, correct answer receives a positive reward ($r_{\text{pos}} = +1.0$), while both incorrect answer and abstention are penalized ($r_{\text{neg}} = r_{\text{abs}} = -1.0$). For query classified as *unknown* ($\mathbb{I}_{\text{solvable}} = 0$), the objective shifts toward encouraging honesty; abstention is rewarded ($r_{\text{abs}} = +1.0$), incorrect answers are penalized ($r_{\text{neg}} = -1.0$), and the correct answer case is irrelevant (N/A) as it does not occur under this condition. Formally, the complete reward $r_{\text{KAR}}(o, a^\star)$ is defined as:

$$r_{\text{KAR}}(o, a^\star) = \begin{cases} r_{\text{pos}} = +1.0, & \\ r_{\text{abs}} = -1.0, & \text{if } \mathbb{I}_{\text{solvable}} = 1, \\ r_{\text{neg}} = -1.0, & \\ \\ r_{\text{pos}} = \text{N/A}, & \\ r_{\text{abs}} = +1.0, & \text{if } \mathbb{I}_{\text{solvable}} = 0. \\ r_{\text{neg}} = -1.0, & \end{cases} \tag{3}$$

This adaptive reward scheme is seamlessly integrated into the advantage calculation, providing a dynamic, within-group learning signal that continuously aligns the policy with its evolving knowledge boundary. By eliminating the need for offline

dataset relabeling and allowing the knowledge estimate to update throughout training, KAR not only reduces computational overhead but also fosters precise and generalizable abstention behavior.

3.2 Two-Stage RL Training Strategy

Drawing on insights from Section 2, we observe that: early in training, models are driven by the predominance of F&U groups into an "*abstention trap*", which prevents accuracy gains and can even degrade performance. To prevent this pitfall and ensure robust learning, we design a **Two-Stage RL Training Strategy**, which decouples the objectives of *knowledge boundary* and *abstention calibration* across two consecutive training stages.

Stage I: Knowledge Boundary Exploration. This stage aims to maximize the elicitation and reinforcement of the model's factual knowledge, thereby establishing a high baseline of answer accuracy. To achieve this, we employ a **mixed-reward policy**: during training, the dataset is divided into two parts according to a ratio α. The α proportion of data is trained using a simple binary reward ($r_{\text{pos}} = +1.0$, $r_{\text{neg}} = r_{\text{abs}} = 0$), while the remaining data is trained using only the KAR mechanism described in Section 3.1. The binary reward acts as a safeguard: since both incorrect answers and abstentions receive zero reward, they yield zero advantage in the GRPO update, effectively nullifying gradients from groups that contain only errors or abstentions. This design circumvents the early-training dominance of such groups, which would otherwise push the policy toward excessive silence. Meanwhile, the KAR-trained portion ensures the model retains the ability to abstain. Consequently, this stage prioritizes confident exploration and knowledge recall, building a policy with a robust and strong factual grounding.

Stage II: Abstention Calibration. Once the model's ability to answer questions within its knowledge boundary has been strengthened, the second stage focuses on precise calibration. In this stage, we transition to applying the KAR to *all* training data solely. The goal is now to convert persistent errors—queries the model consistently answers incorrectly despite the Stage I exploration—into appropriate abstentions. Crucially, KAR's conditional logic ensures this conversion is targeted: abstention is incentivized ($r_{\text{abs}} = +1.0$) only for queries where the model fails to generate any correct response within the sampled group ($\mathbb{I}_{\text{solvable}} = 0$). For queries where the model demonstrates competence ($\mathbb{I}_{\text{solvable}} = 1$), KAR continues to reward correct answers and penalize both errors and unnecessary abstentions. This selective pressure allows the model to refine its abstention boundary, suppressing hallucinations on genuinely unknown queries while largely preserving the high answer accuracy established in Stage I.

4 Experiments

In this section, we present comprehensive experiments and detailed analyses to validate the effectiveness of our proposed KARL.

4.1 Tasks and Evaluation Metrics

To thoroughly evaluate KARL, we perform experiments in multiple benchmarks, which are categorized into short-form knowledge tasks and reasoning tasks. This design aims to assess both factual precision and general problem-solving capabilities.

Short-form Knowledge Tasks. We primarily focus on knowledge-intensive tasks to evaluate the model's ability to provide accurate factual responses or abstain appropriately when uncertain. We adopt NaturalQuestions (NQ) (Kwiatkowski et al., 2019) as our in-distribution benchmark, using its training split for supervised fine-tuning and reinforcement learning, and the validation split for in-distribution evaluation. To evaluate generalization, we include three out-of-distribution (OOD) datasets representing distinct distribution shifts. First, we employ TriviaQA (Joshi et al., 2017) to assess performance on simpler general knowledge questions. Second, we utilize BioASQ (Tsatsaronis et al., 2015) to test whether the learned policy generalizes to a specialized biomedical domain. Finally, we use ARC-Challenge (Clark et al., 2018). Unlike QA datasets, this dataset consists of single-choice questions, allowing us to evaluate abstention behavior under different question formats.

Reasoning Tasks. To ensure our method does not degrade general reasoning, we evaluate it on GSM8K (Cobbe et al., 2021) for multi-step mathematical reasoning and on BIG-Bench Hard (BBH) (Suzgun et al., 2023) for diverse and challenging reasoning tasks.

Evaluation Metrics. Following Xu et al. (2024), we adopt a unified reliability score that balances accuracy and hallucination rate. Let T, U, and F denote the rates of correct answers, abstentions, and hallucinations, respectively (where $T + U +$

Method	NQ				TriviaQA				BioASQ				ARC-C			
	T (↑)	U (−)	F (↓)	Rely (↑)	T (↑)	U (−)	F (↓)	Rely (↑)	T (↑)	U (−)	F (↓)	Rely (↑)	T (↑)	U (−)	F (↓)	Rely (↑)
Llama3.1-8B-Inst																
Prompting	38.7	6.0	55.3	44.3	71.6	1.3	27.1	72.9	55.8	0.3	43.9	56.1	81.7	0.3	18.0	82.0
SFT	40.0	0.0	60.0	40.0	70.2	0.0	29.8	70.2	53.4	0.0	46.6	53.4	77.6	0.0	22.4	77.6
R-Tuning	24.8	56.8	18.4	49.3	55.4	35.9	8.7	78.4	27.9	65.6	6.5	50.5	10.3	87.8	1.9	21.0
RLKF	28.7	30.7	40.6	50.0	57.6	23.2	19.2	75.4	43.3	27.6	29.1	63.3	54.9	5.2	39.9	59.8
Binary	56.3	0.0	43.7	56.3	76.8	0.0	23.2	76.8	66.0	0.0	34.0	66.0	82.4	0.0	17.6	82.4
Ternary (TruthRL)	34.9	53.6	11.5	59.8	61.4	32.3	6.3	83.3	39.0	57.0	4.0	63.5	71.9	18.1	10.0	86.7
Ours	42.7	40.6	16.7	**66.8**	67.9	23.5	8.6	**85.9**	57.7	32.8	9.5	**79.7**	80.0	9.1	10.9	**88.3**
Δ Compared to Prompting.	+4.0	+34.6	-38.6	+22.5	-3.7	+22.2	-18.5	+13.0	+1.9	+32.5	-34.4	+23.6	-1.7	+8.8	-7.1	+6.3
Qwen2.5-7B-Inst																
Prompting	23.0	32.5	44.5	44.9	49.7	17.0	33.3	63.8	43.2	21.8	35.0	60.2	88.0	0.9	11.1	88.9
SFT	30.1	0.0	69.9	30.1	60.3	0.0	39.7	60.3	49.4	0.0	50.6	49.4	87.0	0.0	13.0	87.0
R-Tuning	9.3	87.4	3.3	20.3	35.4	60.1	4.5	59.4	22.7	75.2	2.1	41.3	52.6	45.0	2.4	77.4
RLKF	24.9	43.2	31.9	49.4	41.7	1.7	56.6	43.4	35.3	7.0	57.7	41.8	85.1	4.2	10.7	89.1
Binary	37.0	0.0	63.0	37.0	64.2	0.0	35.8	64.2	60.7	0.0	39.3	60.7	87.6	0.0	12.4	87.6
Ternary (TruthRL)	11.9	83.0	5.1	26.0	39.2	52.2	8.6	64.2	25.2	67.8	7.0	47.0	83.6	8.2	8.2	91.1
Ours	26.3	49.5	24.2	**51.3**	51.7	30.9	17.4	**73.1**	43.0	35.2	21.8	**65.8**	87.2	2.3	10.5	**89.4**
Δ Compared to Prompting.	+3.3	+17.0	-20.3	+6.4	+2.0	+13.9	-15.9	+9.3	-0.2	+13.4	-13.2	+5.6	-0.8	+1.4	-0.6	+0.5

Table 1: Experimental results (%) on in-distribution and three out-of-distribution datasets. **T**, **U**, and **F** denote the rates of **correct answers**, **abstentions**, and **hallucinations**, respectively. **Rely** is a unified reliability score that balances accuracy and hallucination rate. The Δ row reports the performance difference between our method and the Prompting baseline. Green / Red numbers indicates improvement / degradation. Best performance is highlighted in **bold**.

$F = 1$). The reliability metric is computed as:

$$Rely = (1 - U) \times (1 - F) + U \times T.$$

This formulation can be interpreted as a dynamic trade-off: as the model answers more questions (higher $1 - U$), the metric places greater weight on penalizing hallucinations (F); conversely, as it abstains more (higher U), it prioritizes the total yield of correct answers (T). This mechanism effectively penalizes both over-aggressive (high answer rate but high hallucination) and over-conservative (low answer rate and low correctness) strategies.

4.2 Baselines

We compare KARL against several representative baselines from three categories. Specifically, *Inference-Only*: **Prompting** directly prompts the base model to answer or abstain when uncertain, without any further training. *Supervised Fine-Tuning*: (1) **SFT** performs standard supervised fine-tuning on the original dataset, treating the ground-truth answers as the training targets. (2) **R-Tuning** (Zhang et al., 2024) is an abstention-aware method that relabels uncertain examples as "I don't know" based on offline consistency probing, then fine-tunes on the relabeled dataset. *Reinforcement Learning*: (1) **RLKF** (Xu et al., 2024) aligns the model using DPO with preferences derived from offline sampling consistency to align answer, abstention, and error preferences. (2) **Binary** applies GRPO with a naive binary reward, assigning a reward of $+1$ for correct answers and 0 for any other response. (3) **Ternary** (TruthRL) (Wei et al., 2025) utilizes GRPO with a static ternary reward ($+1$ for correct, 0 for abstention, -1 for incorrect) to explicitly encourage abstention.

4.3 Implementation Details

We evaluate our method on two widely-used open-source models: Llama-3.1-8B-Instruct (Grattafiori et al., 2024) and Qwen2.5-7B-Instruct (Yang et al., 2024a). For our method, the mixing ratio α for the binary reward in Stage I is set to 50%. All models are trained on a randomly sampled subset of 19,200 examples from the NQ training split. For each generated response, we employ Llama-3.3-70B-Instruct as the judge model to evaluate its correctness against the gold answer. Additional implementation details are provided in Appendix C.

4.4 Main Results

KARL Achieves a Superior Accuracy-Abstention Balance. Table 1 presents the performance comparisons on the in-distribution NQ dataset and three OOD datasets. Methods that ignore abstention, such as SFT and Binary Reward, achieve high accuracy but suppress abstention completely, leading to excessive hallucinations. Existing abstention-aware approaches like R-Tuning, RLKF, and Ternary Reward effectively increase abstention

Method	GSM8K			BBH		
	T (↑)	U (−)	F (↓)	T (↑)	U (−)	F (↓)
Llama3.1-8B-Inst						
Prompting	83.7	0.5	15.8	64.3	2.1	33.6
Binary	85.2	0.0	14.8	63.2	0.0	36.8
Ternary	85.5	3.0	11.5	53.4	28.1	18.5
Ours	86.1	1.3	12.6	62.5	11.3	26.2
Qwen2.5-7B-Inst						
Prompting	89.1	0.5	10.4	67.5	5.8	26.7
Binary	91.4	0.0	8.6	71.5	0.1	28.4
Ternary	88.5	3.0	8.5	61.5	15.6	22.9
Ours	90.4	0.4	9.2	67.8	7.4	24.8

Table 2: Performance (%) comparison on GSM8K and BBH using Llama-3.1-8B-Instruct and Qwen2.5-7B-Instruct.

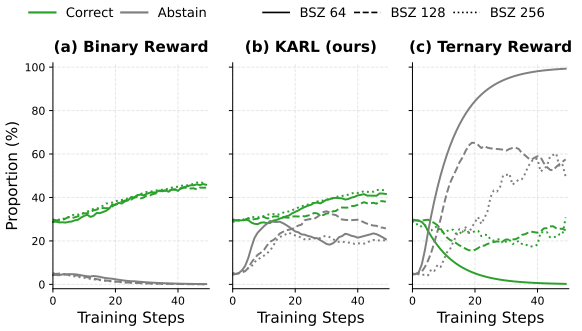

Figure 5: Early training dynamics of **Correct** (green) and **Abstain** (gray) response proportions on the NQ dataset. The subplots compare (a) Binary Reward, (b) KARL (Ours), and (c) Ternary Reward across three batch sizes (64, 128, 256).

rates but significantly compromise accuracy compared to the base model (Prompting). In contrast, KARL successfully overcomes this limitation and achieves the highest *Rely* scores across all evaluated datasets and models. Specifically, on NQ, our method consistently improves performance over the Prompting baseline: it reduces the hallucination rate by **38.6%** for Llama3.1 and **20.3%** for Qwen2.5, while simultaneously improving accuracy by **4.0%** and **3.3%**, respectively. This indicates that our method effectively teaches the model to abstain from unknown queries without suppressing its ability to answer solvable ones.

KARL Generalizes Robustly to OOD Datasets. We further evaluate the transferability of the learned abstention policy on diverse unseen datasets, including TriviaQA, BioASQ, and ARC-C. As shown in Table 1, methods relying on offline data construction (SFT, R-tuning, RLKF) exhibit limited generalization. Across diverse domains and formats, these approaches either suffer from severe accuracy degradation or fail to reduce hallucinations compared to the prompting baseline. Regarding RLVR-based baselines, the Binary baseline consistently lacks abstention capability across all datasets. Meanwhile, the Ternary baseline suffers from significant accuracy drops on OOD tasks across both model families, such as a 16.8% and 18.0% decrease on BioASQ for Llama-3.1 and Qwen2.5, respectively. In contrast, KARL demonstrates robust generalization. It consistently maintains high accuracy with minimal performance drops and even achieves accuracy gains, such as **1.9%** on BioASQ for Llama3.1 and **2.0%** on TriviaQA for Qwen2.5. This indicates that our method learns a generalized capability to abstain appropriately rather than overfitting to training patterns.

KARL Preserves Reasoning Capabilities. We further examine whether training on knowledge tasks compromises reasoning abilities using GSM8K and BBH. Table 2 shows that KARL consistently maintains or improves accuracy compared to the prompting baseline while effectively reducing hallucinations. Regarding baselines, the Binary baseline completely fails to abstain in reasoning contexts. The Ternary baseline shows instability, notably suffering a significant accuracy drop of 10.9% on BBH for Llama3.1 and 6.0% for Qwen2.5. Overall, these results demonstrate that KARL achieves robust generalization without degrading the model's core reasoning intelligence.

Avoiding the "*Abstention Trap*". As discussed in our pilot study, the dominance of F&U groups often drives models into an "*abstention trap*" early in training. To validate the effectiveness of our solution, we analyze the training dynamics of the binary baseline, the ternary baseline and our method across different batch sizes (64, 128, and 256). Figure 5 illustrates the evolution of correct and abstention rates during the early steps. The Ternary baseline consistently experiences a significant drop in accuracy early in the process. This instability is most severe at a batch size of 64, where the model rapidly converges to 100% abstention. At this point, the model loses all gradient signals and ceases to learn. In contrast, our method successfully bypasses this failure mode. Across all batch sizes, our approach achieves accuracy growth comparable to the Binary baseline while retaining a significantly higher abstention rate. This indicates that the model effectively explores its knowledge

Setting	T (↑)	U (−)	F (↓)	Rely (↑)
Prompting	38.7	6.0	55.3	44.3
Ablation on Strategy				
w/o Stage II ($\alpha = 50\%$)	50.6	18.0	31.4	65.4
Ablation on Mixing Ratio α in Stage I				
$\alpha = 0\%$ (Pure KAR)	36.0	50.0	14.0	61.0
$\alpha = 25\%$	42.4	35.5	22.1	65.3
$\alpha = 50\%$	42.7	40.6	16.7	**66.8**
$\alpha = 75\%$	39.3	47.0	13.7	64.2
$\alpha = 100\%$ (Pure Binary)	55.5	0.0	44.5	55.5

Table 3: Ablation study on NQ using Llama-3.1-8B-Instruct. We evaluate the impact of the two-stage strategy by removing the second stage (w/o Stage II) and analyze the impact of mixing ratio (α) of binary reward data in Stage I. Best performance is highlighted in **bold**.

NQ Question: What type of database is library literature and information science?
Prompting: Metadata database ***(F)***
Ternary: Library Literature and Information Science is a database that indexes and abstracts articles and other publications in the field of library and information science. I don't know its specific database type. ***(U)***
Ours: Library Literature and Information Science is a database that indexes articles and publications related to library science, often a bibliographic or index database. ***(T)***
ARC-C Question: Farmers in Wyoming were concerned because some of their chickens were being preyed upon by hawks that lived in areas around their ranches. The farmers grouped together and hunted the hawks until they were no longer in their area. Which would most likely happen next? Options: A. The chicken population would go down. B. Populations of mice and rats would increase. C. Another bird of prey would replace the hawk. D. The chickens would have a lower rate of disease.
Prompting: C. Another bird of prey would replace the hawk. ***(F)***
Ternary: The chicken population would likely increase due to the removal of a predator, but I don't know the exact outcome. ***(U)***
Ours: The removal of hawks would likely lead to an increase in populations of mice and rats, as their natural predator is removed, so B is a likely outcome. ***(T)***

Table 4: Case examples from the NQ and ARC-C datasets comparing responses generated by the Prompting baseline, Ternary baseline, and our method. (***T***), (***F***), and (***U***) denote correct answers, incorrect answers, and abstentions, respectively. Models are implemented on Llama-3.1-8B-Instruct.

boundary without sacrificing abstention awareness. This stability avoids the early collapse seen in the ternary baseline and is crucial for achieving high final accuracy and robust performance.

4.5 Analysis

Ablation Study. We conduct an ablation study in Table 3 to verify the contribution of each component in our framework.

Necessity of Two-Stage Training. We assess the necessity of the two-stage strategy. Removing the second stage results in the highest accuracy but a significantly lower abstention rate compared to the full method. This happens because the primary goal of Stage II is to use the full KAR mechanism to convert the remaining incorrect answers into abstentions. Without this calibration stage, the model remains in an exploratory state, leading to higher hallucination rates and compromised reliability.

Impact of α in Stage I. We investigate the impact of the binary reward mixing ratio α in Stage I. When α is set to 0%, the method degrades to single-stage training with pure KAR. This leads to the lowest accuracy, as the penalties in KAR suppress the exploration of knowledge boundary during the early stages. Conversely, setting α to 100% means the model is trained exclusively with binary rewards in Stage I. Under this setting, the model completely loses its ability to abstain in Stage I and fails to recover in Stage II because it cannot generate any abstention responses during rollout, resulting in a zero abstention rate. We find that a mixing ratio of 50% achieves the optimal balance and the highest *Rely* score, effectively eliciting knowledge while preserving sufficient abstention capability for the final calibration stage.

Case Study. We present a case study in Table 4 to further demonstrate the advantages of our method. As shown in the examples, baseline methods frequently fail due to hallucinations or unnecessary abstention. In contrast, our approach effectively utilizes the reasoning capabilities enhanced by RL. This enables the model to answer challenging questions correctly without being overly cautious, achieving a more robust balance between factual accuracy and abstention.

5 Conclusion

In this paper, we propose KARL, a novel framework that continuously aligns an LLM's abstention behavior with its evolving knowledge boundary. KARL introduces two core innovations: a *Knowledge-Boundary-Aware Reward* that performs online knowledge estimation using within-group response statistics, dynamically rewarding correct answers or guided abstention; and a *Two-Stage RL Training Strategy* that first explores knowledge boundary and bypasses the "*abstention trap*", and subsequently converts incorrect answers beyond the knowledge boundary into abstentions without sacrificing accuracy. Experimental results demonstrate that KARL achieves a superior trade-off between accuracy and abstention, significantly suppressing hallucinations while maintaining high answer accuracy across diverse in-distribution and out-of-distribution scenarios.

Limitations

Despite the effectiveness of KARL, several limitations remain. First, our method is sensitive to the difficulty distribution of the training data. Samples that are consistently too easy or too difficult may fail to provide effective gradients for knowledge boundary exploration. Consequently, how to automatically select or filter data with an appropriate difficulty level remains underexplored. Second, our current reward mechanism is still coarse, as it only distinguishes among three discrete states: correct, incorrect, and abstention. As a result, sampled groups that yield uniform rewards produce zero advantage and contribute no gradient updates. Future work could investigate finer-grained reward designs to provide denser learning signals. Finally, our two-stage curriculum currently relies on a fixed schedule to transition between exploration and calibration. A more advanced approach would be to dynamically toggle between these phases based on the real-time performance of the model during training, which we leave as a direction for future research.

References

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, and 12 others. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *CoRR*, abs/2204.05862.

Faeze Brahman, Sachin Kumar, Vidhisha Balachandran, Pradeep Dasigi, Valentina Pyatkin, Abhilasha Ravichander, Sarah Wiegreffe, Nouha Dziri, Khyathi Raghavi Chandu, Jack Hessel, Yulia Tsvetkov, Noah A. Smith, Yejin Choi, and Hanna Hajishirzi. 2024. The art of saying no: Contextual noncompliance in language models. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

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? In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4299–4307.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the AI2 reasoning challenge. *CoRR*, abs/1803.05457.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *CoRR*, abs/2110.14168.

Matthew Dahl, Varun Magesh, Mirac Suzgun, and Daniel E. Ho. 2024. Large legal fictions: Profiling legal hallucinations in large language models. *CoRR*, abs/2401.01301.

DeepSeek-AI. 2024. Deepseek-v3 technical report. *CoRR*, abs/2412.19437.

Yang Deng, Yong Zhao, Moxin Li, See-Kiong Ng, and Tat-Seng Chua. 2024. Don't just say "i don't know"! self-aligning large language models for responding to unknown questions with explanations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 13652–13673. Association for Computational Linguistics.

Google DeepMind. 2025. Gemini 3 pro. `https://deepmind.google/technologies/gemini/pro/`. Official model card and benchmark overview for Gemini 3 Pro.

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. *CoRR*, abs/2510.01152.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, and 175 others. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nat.*, 645(8081):633–638.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12):248:1–248:38.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1601–1611. Association for Computational Linguistics.

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. *CoRR*, abs/2207.05221.

Adam Tauman Kalai, Ofir Nachum, Santosh S. Vempala, and Edwin Zhang. 2025. Why language models hallucinate. *CoRR*, abs/2509.04664.

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. 2024. Understanding the effects of RLHF on LLM generalisation and diversity. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur P. Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: a benchmark for question answering research. *Trans. Assoc. Comput. Linguistics*, 7:452–466.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles, SOSP 2023, Koblenz, Germany, October 23-26, 2023*, pages 611–626. ACM.

Junliang Li, Yucheng Wang, Yan Chen, Yu Ran, Ruiqing Zhang, Jing Liu, Hua Wu, and Haifeng Wang. 2025. Knowledge-level consistency reinforcement learning: Dual-fact alignment for long-form factuality. *CoRR*, abs/2509.23765.

Kangyang Luo, Zichen Ding, Zhenmin Weng, Lingfeng Qiao, Meng Zhao, Xiang Li, Di Yin, and Jinlong Shu. 2025. Let's be self-generated via step by step: A curriculum learning approach to automated reasoning with large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 15343–15420, Vienna, Austria. Association for Computational Linguistics.

Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D. Manning, and Daniel E. Ho. 2024. Hallucination-free? assessing the reliability of leading AI legal research tools. *CoRR*, abs/2405.20362.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan T. McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 1906–1919. Association for Computational Linguistics.

Zabir Al Nazi and Wei Peng. 2024. Large language models in healthcare and medical domain: A review. *Informatics*, 11(3):57.

OpenAI. 2025. Introducing gpt-5. `https://openai.com/index/introducing-gpt-5`. Official technical overview of GPT-5 and GPT-5 Pro models.

Long Ouyang, Jeffrey 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 F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. Zero: memory optimizations toward training trillion parameter models. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, SC 2020, Virtual Event / Atlanta, Georgia, USA, November 9-19, 2020*, page 20. IEEE/ACM.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *Preprint*, arXiv:1707.06347.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *CoRR*, abs/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 2025, Rotterdam, The Netherlands, 30 March 2025 - 3 April 2025, pages 1279–1297. ACM.

Shuzheng Si, Haozhe Zhao, Gang Chen, Cheng Gao, Yuzhuo Bai, Zhitong Wang, Kaikai An, Kangyang Luo, Chen Qian, Fanchao Qi, Baobao Chang, and Maosong Sun. 2025a. Aligning large language models to follow instructions and hallucinate less via effective data filtering. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 16469–16488. Association for Computational Linguistics.

Shuzheng Si, Haozhe Zhao, Cheng Gao, Yuzhuo Bai, Zhitong Wang, Bofei Gao, Kangyang Luo, Wenhao Li, Yufei Huang, Gang Chen, Fanchao Qi, Minjia Zhang, Baobao Chang, and Maosong Sun. 2025b. Teaching large language models to maintain contextual faithfulness via synthetic tasks and reinforcement learning. *CoRR*, abs/2505.16483.

Linxin Song, Taiwei Shi, and Jieyu Zhao. 2025. The hallucination tax of reinforcement finetuning. *CoRR*, abs/2505.13988.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. 2023. Challenging big-bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 13003–13051. Association for Computational Linguistics.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R. Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, Yannis Almirantis, John Pavlopoulos, Nicolas Baskiotis, Patrick Gallinari, Thierry Artières, Axel-Cyrille Ngonga Ngomo, Norman Heino, Éric Gaussier, Liliana Barrio-Alvers, and 3 others. 2015. An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. *BMC Bioinform.*, 16:138:1–138:28.

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. *CoRR*, abs/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. *Trans. Assoc. Comput. Linguistics*, 13:529–556.

Robert Wolfe, Isaac Slaughter, Bin Han, Bingbing Wen, Yiwei Yang, Lucas Rosenblatt, Bernease Herman, Eva Maxfield Brown, Zening Qu, Nic Weber, and Bill Howe. 2024. Laboratory-scale AI: open-weight models are competitive with chatgpt even in low-resource settings. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency, FAccT 2024, Rio de Janeiro, Brazil, June 3-6, 2024*, pages 1199–1210. ACM.

Chaojun Xiao, Xueyu Hu, Zhiyuan Liu, Cunchao Tu, and Maosong Sun. 2021. Lawformer: A pre-trained language model for chinese legal long documents. *AI Open*, 2:79–84.

Hongshen Xu, Zichen Zhu, Da Ma, Situo Zhang, Shuai Fan, Lu Chen, and Kai Yu. 2024. Rejection improves reliability: Training llms to refuse unknown questions using RL from knowledge feedback. *CoRR*, abs/2403.18349.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024a. Qwen2.5 technical report. *CoRR*, abs/2412.15115.

Yuqing Yang, Ethan Chern, Xipeng Qiu, Graham Neubig, and Pengfei Liu. 2024b. Alignment for honesty. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. 2025. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? *CoRR*, abs/2504.13837.

Bonan Zhang, Zhongqi Chen, Bowen Song, Qinya Li, Fan Wu, and Guihai Chen. 2025a. Confclip: Confidence-weighted and clipped reward for reinforcement learning in llms. *CoRR*, abs/2509.17730.

Charlie Zhang, Graham Neubig, and Xiang Yue. 2025b. On the interplay of pre-training, mid-training, and rl on reasoning language models. *Preprint*, arXiv:2512.07783.

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'. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 7113–7139. Association for Computational Linguistics.

Runchuan Zhu, Xinke Jiang, Jiang Wu, Zhipeng Ma, Jiahe Song, Fengshuo Bai, Dahua Lin, Lijun Wu, and Conghui He. 2025. GRAIT: gradient-driven refusal-aware instruction tuning for effective hallucination mitigation. In *Findings of the Association for Computational Linguistics: NAACL 2025, Albuquerque,*

New Mexico, USA, April 29 - May 4, 2025, pages 4006–4021. Association for Computational Linguistics.

Appendix

A Related Works

A.1 Supervised Fine-Tuning for Abstention.

Supervised approaches teach LLMs to abstain by fine-tuning on datasets containing explicit abstention labels. A representative method, R-Tuning (Zhang et al., 2024), estimates the model's knowledge boundary via multi-sample consistency probing and relabels responses to uncertain questions with "I don't know".

Subsequent Refusal-Aware Instruction Tuning (RAIT) strategies extend this paradigm by constructing refusal data through diverse mechanisms, such as filtering low-accuracy samples (Yang et al., 2024b; Cheng et al., 2024), incorporating unanswerable QA pairs (Wolfe et al., 2024; Wen et al., 2025), or synthesizing abstention scenarios via prompting (Brahman et al., 2024; Deng et al., 2024; Si et al., 2025a; Luo et al., 2025). These methods train models to map out-of-knowledge inputs directly to standardized abstention responses. While straightforward, SFT-based approaches fundamentally rely on memorizing specific refusal patterns rather than learning dynamic decision boundaries. Consequently, they often lead to over-refusals, causing models to refuse benign questions and generalize poorly to out-of-distribution domains (Kirk et al., 2024). Furthermore, constructing such datasets incurs high computational costs. Since knowledge boundaries are intrinsic and vary across models, abstention labels must be customized via extensive probing for each specific model, preventing efficient transferability. Although recent efforts like GRAIT (Zhu et al., 2025) attempt to mitigate over-refusal via gradient-based sample weighting, SFT methods remain limited by the static and model-specific nature of their supervision signals.

A.2 Reinforcement Learning for Abstention.

Existing RL-based approaches for abstention can be broadly grouped by the type of supervision signal they rely on. One line of work adopts *preference-based* RL, where models are aligned using PPO or DPO against a reward model or a static set of preference pairs in which abstention is preferred over hallucination (Bai et al., 2022; Kadavath et al., 2022; Xu et al., 2024). Despite their effectiveness, these methods face scalability challenges similar to SFT, as they require large amounts of carefully curated preference data tailored to abstention scenarios.

To reduce reliance on preference annotations, recent work has shifted toward *outcome-based* RL, commonly referred to as Reinforcement Learning with Verifiable Rewards (RLVR). These methods leverage automatic verification signals to directly optimize model behavior without training a separate reward model (Shao et al., 2024; Guo et al., 2025). While computationally efficient, standard RLVR objectives tend to amplify model confidence, reducing the likelihood of abstention and exacerbating hallucinations (Yue et al., 2025; Song et al., 2025). To mitigate this, recent works have attempted to refine the reward design within the RLVR framework (Si et al., 2025b). For instance, ConfClip (Zhang et al., 2025a) integrates confidence scores into the GRPO reward to encourage high-confidence correctness and penalize high-confidence hallucination, while KLCF (Li et al., 2025) penalizes responses that exceed the model's knowledge boundary. Nevertheless, these methods lack an explicit mechanism for abstention actions, often resulting in persistent hallucination rates. Addressing this, TruthRL (Wei et al., 2025) incorporates a ternary reward scheme within GRPO to explicitly distinguish between correct answers, hallucinations, and abstentions. Similarly, MASH (Gul et al., 2025) trains models to seek external knowledge when they are not sure, then converts search actions into abstentions during inference. While these explicit methods effectively reduce hallucinations, they still suffer from over-refusals and degrade the model's overall accuracy and utility. Thus, how to eliminate complex data construction while maintaining a balance between model utility and hallucination reduction still remains underexplored.

B Reinforcement Learning Protocol

For the RL training of LLMs, policy optimization methods such as PPO (Schulman et al., 2017) and GRPO (Shao et al., 2024) have been widely studied. GRPO is especially attractive in verifiable-reward settings as it avoids a separate value network or reward model by estimating advantages directly from reward statistics of sampled responses. Therefore, we adopt GRPO as our backbone RL framework.

Specifically, for each input query q, GRPO samples a group of G responses $\{o_1, \ldots, o_G\}$ from the policy before the update, i.e., $\pi_{\theta_{\text{old}}}$. It then opti-

mizes the policy by maximizing a surrogate objective that encourages actions with above-average rewards while penalizing deviation from a reference policy π_{ref} via a KL-divergence term. The GRPO objective is:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{q,\{o_i\}\sim\pi_{\theta_{old}}}\left[\frac{1}{G}\sum_{i=1}^{G}\mathcal{L}_i - \beta\mathbb{D}_{KL}(\pi_\theta||\pi_{ref})\right], \quad (4)$$

$$\mathcal{L}_i = \min\left(w_i\hat{A}_i, \text{clip}(w_i, 1-\epsilon, 1+\epsilon)\hat{A}_i\right), \quad (5)$$

where $w_i = \frac{\pi_\theta(o_i|q,c)}{\pi_{\theta_{\text{old}}}(o_i|q,c)}$, ϵ and β are hyperparameters that constrain the policy update and control the divergence penalty, respectively. Crucially, A_i represents the advantage value for the i-th output, which is derived from the group-wise statistics:

$$\hat{A}_i = \frac{r_i - \mu(\{r_1, \cdots, r_G\})}{\sigma(\{r_1, \cdots, r_G\}) + \delta}, \quad (6)$$

where r_i is the reward for output o_i, μ and σ denote the mean and standard deviation over the group, and δ is a small constant.

C Implementation Details

C.1 Pilot Study Implementation Details

Training Dynamics Setup. The experimental setup for the pilot study follows the general training protocols described in Appendix C.2, with the following specific configurations:

We utilize Llama-3.1-8B-Instruct as the base model, fine-tuned on the NQ Open dataset. Unlike the main experiments where we consider a range of batch sizes, we fix the global batch size at **128** for all pilot experiments to ensure a controlled comparison. All other hyperparameters remain identical to the main configuration. Regarding reward structures, the Binary setup assigns a reward of 1 for correct answers and 0 for any other response, whereas the Ternary setup assigns 1 for correctness, 0 for abstention, and -1 for errors.

Rollout Distribution Analysis. To analyze the mechanics behind the abstention trap, we conducted rollout experiments using Llama-3.1-8B-Instruct and Qwen2.5-7B-Instruct on the NQ and TriviaQA datasets. For each query, we generate 8 responses using probabilistic decoding parameters (`temperature=1.0`, `top_k=50`, `top_p=0.9`). We classified each response as Correct, Incorrect, or Abstention and categorized the resulting groups based on their composition. In our analysis of the gradient contribution, we filtered out groups that do not influence the abstention policy. Specifically, we excluded groups consisting of a single response type as they yield zero advantage. We also excluded groups containing only correct and incorrect answers as they do not directly drive the learning of abstention. The distribution shown in Figure 3 reflects the remaining groups that actively influence abstention.

C.2 Training Details

All models are trained on 8 NVIDIA A800 GPUs with 40GB memory using full-parameter fine-tuning and trained for 1 epoch.

For GRPO, we implement our training pipeline using `verl` framework (Sheng et al., 2025). To manage memory usage during full-parameter fine-tuning, we utilize DeepSpeed (Rajbhandari et al., 2020) with ZeRO-3 offload and enable gradient checkpointing. We employ a constant learning rate of 1e-6, a KL divergence coefficient β of 0.001, and perform a hyperparameter search for batch sizes over {64, 128, 256}.

We use vLLM (Kwon et al., 2023) engine for LLM rollouts in GRPO with a tensor parallel size of 2 and GPU memory utilization of 0.8. For each query, we sample $G = 8$ responses. The sampling parameters are set to a temperature of 1.0 and top-p of 1.0. For short-form knowledge tasks, we train without Chain-of-Thought (CoT) and restrict the maximum response length to 64 tokens, using the inference prompt shown in Figure 6. For reasoning tasks, we encourage CoT generation and set the maximum response length to 512 tokens, using the inference prompt shown in Figure 7. The prompt template used by the judge model (Llama3.3-70B-Instruct) for GRPO training rewards is shown in Figure 8.

For SFT training and R-Tuning, we use a learning rate of 5×10^{-6} and a batch size of 8. For DPO training, we use a learning rate of 3×10^{-6} and a batch size of 8. All these training runs adopted a cosine learning rate scheduler with 3% warmup steps and are conducted for 1 epoch over the dataset.

C.3 Evaluation Details

We use vLLM and apply greedy decoding during evaluation. All results are evaluated in a 0-shot setting. To ensure precise metric calculation, we first use rule-based string matching to detect abstention expressions (e.g., "I don't know"). Responses that include these expressions are classified as Abstention. Only non-refusal responses are passed

to the judge model (Llama-3.3-70B-Instruct) to determine factual correctness using the prompt in Figure 8.

This separation is crucial for preventing reward hacking. We observe that relying solely on the LLM to classify all three states (Correct, Incorrect, Abstention) encourages the trained model to generate ambiguous responses. Such ambiguity allows the model to be interpreted as correct when the answer is included, or as uncertain otherwise, thereby bypassing error penalties. Furthermore, we employ the LLM judge for correctness instead of a rule-based judge because reference answers in open-domain QA datasets often have limited coverage of valid answer expressions, a limitation also noted in prior work (Wei et al., 2025).

Inference Prompt (Without CoT)

Input: You are given a Question. 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. If the question contains a false premise or assumption, answer "invalid question".
2. If you are uncertain or don't know the answer, answer "I don't know".

Please directly provide the final answer. The final answer MUST be put in \boxed{}. For example, \boxed{I don't know}, \boxed{invalid question}, \boxed{3 times}, \boxed{New York}, etc.

Question: {question}

Output: {answer}

Figure 6: Inference prompt without CoT.

Inference Prompt (With CoT)

Input: You are given a Question. 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. If the question contains a false premise or assumption, answer "invalid question".
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{invalid question}, \boxed{3 times}, \boxed{New York}, etc.

Question: {question}

Output: {answer}

Figure 7: Inference prompt with CoT.

Judge Prompt

Input: Assume you are a human expert in grading predictions given by a model. You are given a question and a model prediction. Judge if the prediction matches the ground truth answer by following these steps:

1: Take it as granted that the Ground Truth is always correct.
2: If the Prediction exactly matches the Ground Truth, return “t”.
3: If the Prediction does not exactly match the Ground Truth, go through the following steps and likely return “f”.
4: If the Ground Truth is a number, return “t” if and only if the Prediction gives a number that almost exactly matches the ground truth.
5: If the Prediction is self-contradictory, must return “f”.
6: If the prediction is not answering the question, must return “f”.
7: If the prediction is a concise and correct summary of the ground truth, return “t”.
8: If ground truth contains a set of items, prediction must contain exactly same items for the value to return to be “t”.
9: Otherwise, return “f”.

Question: {question}
Ground Truth: {ground_truth}
Prediction: {solution_str}

Return only one character: t / f. Do not add any explanation or other text.

Output: {answer}

Figure 8: LLM-as-a-Judge Prompt.