
000 001 002 003 004 005 006 007 008 009 010 011 012 013 014 015 016 017 018 019 020 021 022 023 024 025 026 027 028 029 030 031 BARREL: BOUNDARY-AWARE REASONING FOR FACTUAL AND RELIABLE LRMs

Anonymous authors

Paper under double-blind review

ABSTRACT

Recent advances in Large Reasoning Models (LRMs) have shown impressive capabilities in mathematical and logical reasoning. However, current LRMs rarely admit ignorance or respond with “I don’t know”. Instead, they often produce incorrect answers while showing undue confidence, raising concerns about their factual reliability. In this work, we identify two pathological reasoning patterns characterized by overthinking that contribute to the overconfident and incorrect answers: *last-minute guessing* and *second-thought spiraling*. To address these issues, we propose BARREL—a novel framework that promotes concise and boundary-aware factual reasoning. Our experiments show that BARREL-training increases the reliability of DeepSeek-R1-Distill-Llama-8B from 39.33% to 61.48%, while still achieving accuracy comparable to models finetuned on reasoning data generated by R1. These results demonstrate that our pilot study is inspiring to build more reliable and factual *System 2* LRMs.

1 INTRODUCTION

Recent advances in Large Reasoning Models (LRMs) (Jaech et al., 2024; Guo et al., 2025; Team, 2025) have shown impressive performance in specialized reasoning tasks, especially in mathematics and logic. However, these gains have not led to corresponding improvements in reliability. On the contrary, faithfulness hallucination rates are rising (Hughes & Bae, 2023), and helpfulness on factual tasks is declining (Zhao et al., 2025), raising concerns about the reliability of these reasoning models.

In this work, we focus primarily on the factual reliability of LRMs, which is a crucial requirement for many real-world tasks. Factuality of language models involves two aspects: *knowing* (whether the model holds relevant knowledge) (Huang et al., 2025; Ji et al., 2023) and *telling* (conveying the correct factual information) (Gekhman et al., 2024; Mallen et al., 2022). As shown in Figure 1, we want LRMs to exhibit two aspects of factual reliability: (1) Identify both what it knows and what it does not know: The model should be able to recognize questions it does not know the answer and respond with “I don’t know.” (Xu et al., 2024a; Zhang et al., 2024a) (2) Tell what it knows: There is a gap between knowing and telling (Saunders et al., 2022), and we want to improve the model’s accuracy in expressing the knowledge it has (Zhang et al., 2024b).

However, current LRMs consistently struggle with the two factual reliability goals above. They rarely acknowledge gaps in their knowledge and often fabricate answers instead, even on questions for which they lack sufficient knowledge. Moreover, their responses can be inconsistent—providing incorrect answers in some instances while correctly responding to similar queries elsewhere (Wang et al., 2022), thereby reducing the overall factual accuracy of their responses (Zhao et al., 2025).

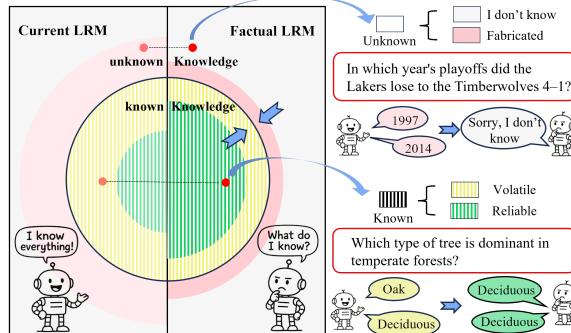


Figure 1: How reliable factual reasoning is expected to improve model performance. **Left:** Current LRMs rarely admit ignorance and often respond inconsistently. **Right:** Reliable LRMs should acknowledge unknowns and express known facts more consistently.

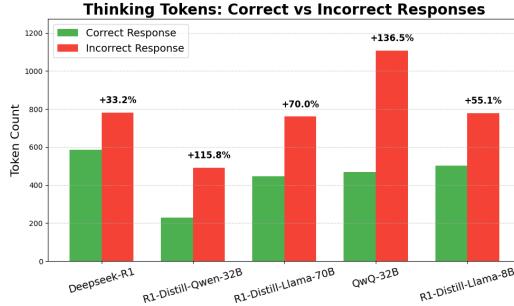


Figure 2: Number of reasoning tokens used by LRM s when producing correct versus incorrect answers. We test on TruthfulQA across different types of reasoning models. Details and results on other datasets are listed at Appendix A.

hastily filling in an answer just before the end of an exam. Another frequent pattern is *Second-thought Spiraling*, where the model initially identifies the correct answer but continues to over-analyze, ultimately undermining its own correct conclusion.

To mitigate these pathological reasoning patterns, we propose a novel training framework, **BARREL** (Boundary-Aware Reasoning for Reliable and Factual LRM s). As shown in Figure 3, BARREL trains LRM s to perform concise, deliberative factual reasoning and draw conclusions after it has explored a sufficient number of candidates. Concretely, for known questions, when the model identifies the correct factual answer during reasoning, we want it to maintain confidence in that answer while continuing to explore other possible ones. Once there are no likely candidates, it should halt further reasoning and provide the correct factual answer. For unknown questions, after exploring a sufficient number of plausible candidates, it should terminate reasoning and proactively admit its lack of knowledge with uncertainty-aware refusal (e.g., "Sorry, I don't know"). To implement BARREL, we begin by employing a sampling strategy to probe the model's knowledge boundary. Drawing on the identified pathological patterns, we construct two distinct types of reasoning data, and use Supervised Fine-Tuning (SFT) to instill the corresponding reasoning behaviors in the model. Finally, we adopt Group Relative Policy Optimization (GRPO) using general reliability-based reward (high for correct answers, medium for uncertainty-aware refusal, and low for incorrect answers) without the need for labeling *known/unknown*, further enhancing the model's ability to generalize in factual reasoning.

Our experiments demonstrate that models trained with BARREL can effectively express uncertainty-aware refusal, and mitigate the two pathological reasoning patterns. This capability significantly improves reliability: BARREL boosts the reliability of DeepSeek-R1-Distill-Llama-8B from 39.33% to 61.48%, while maintaining an accuracy of 40.7%, which is even higher than the accuracy of 38.43% achieved by distillation. Through detailed analysis, we highlight the critical role of medium-level rewards in promoting uncertainty-aware refusal. This result also identifies the root cause of models' inability to admit ignorance to a fundamental gap in current RL paradigms: they do not reward refusal. As a result, models are incentivized to answer every question, regardless of uncertainty.

Our main contributions are as follows:

- We discover the factual overthinking phenomenon and point out two pathological reasoning patterns that lead to factual unreliability of LRM s.
- To the best of our knowledge, we are the first to explore how LRM s can utilize reasoning to admit ignorance and say "I don't know". Also, we introduce a novel training pipeline to do so.
- We emphasize the importance of medium-level rewards in encouraging uncertainty-aware refusal to build a more reliable LRM using RL training.

2 RELATED WORK

Knowledge Boundary The knowledge boundary of LLMs refers to the extent of knowledge a model possesses or can reliably recall (Li et al., 2024). Identifying such boundaries is crucial for model safety and reliability and is commonly addressed by confidence calibration (Ren et al., 2023),

What underlying reasoning behaviors contribute to these observed issues? As shown in Figure 2, our preliminary experiments reveal a variant of the overthinking pattern discussed by Chen et al. (2024), which we term factual overthinking—where large reasoning models (LRM s) consume more tokens when producing factually incorrect answers than when generating correct ones. And we further identify two prevalent pathological reasoning patterns associated with this overthinking phenomenon. As shown on the left part of Figure 3, one prominent pattern is *Last-minute Guessing*, in which the model, after extensive but inconclusive reasoning, abruptly commits to an answer in a final burst of speculative output—much like a student

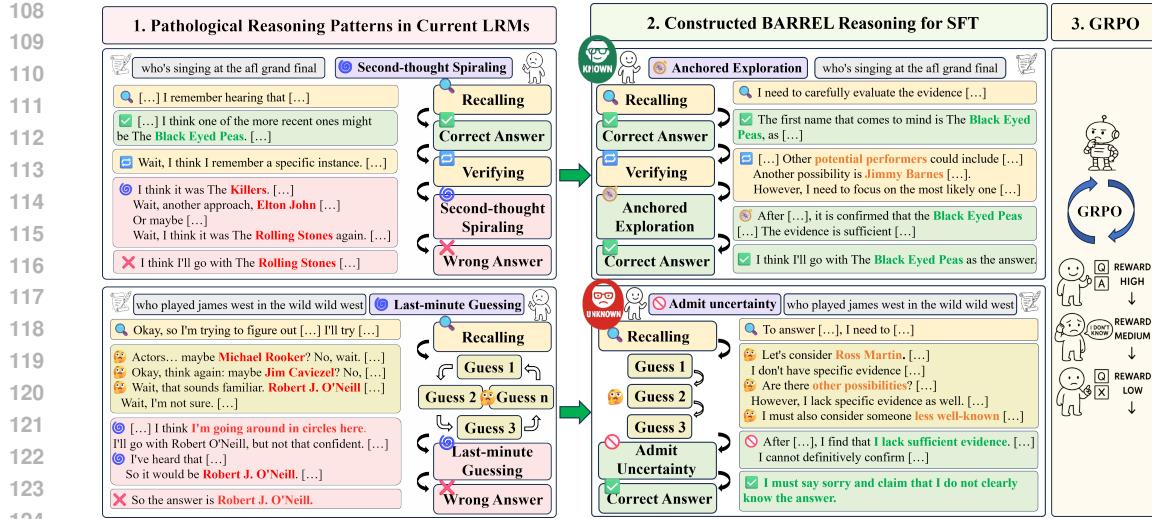


Figure 3: **Left:** The two current reasoning patterns of LRM: Last-minute Guessing, typically associated with unknown knowledge, and Second-thought Spiraling, which occurs despite known knowledge. **Right:** The BARREL pipeline addresses both cases by correcting overthinking tendencies and constructing SFT data accordingly, further enhanced with GRPO.

internal state probing (Ji et al., 2024), uncertainty estimation (Kapoor et al., 2024), and sampling strategy (Gekhman et al., 2024; Xue et al., 2024). To mitigate failures by outbound queries, recent studies have proposed training LLMs to abstain from answering when uncertain, often by generating “I don’t know” responses (Cheng et al., 2024) or providing further explanations of the unanswerability (Deng et al., 2024). Our work pushes it further to structured and interpretable rationales about their knowledge boundary.

Factual Alignment Factual alignment aims to improve factuality while preserving instruction-following capabilities in LLMs. Recent work incorporates factuality-awareness into both SFT and RL stages to improve factual accuracy (Lin et al., 2024), or finetuning with DPO to enhance its self-evaluation capability (Zhang et al., 2024b). Other approaches include fine-tuning with refusal-aware datasets to encourage appropriate abstention behaviors (Zhang et al., 2024a), and RLKF, which guides models to reject uncertain queries based on external feedback signals (Xu et al., 2024a). **Similarly, other reinforcement learning strategies have been proposed to incentivize models to express doubt or uncertainty (Stangel et al., 2025; Xu et al., 2024b). Damani et al. (2025) has also explored training LLMs to reason with uncertainty using binary rewards.** While previous work has primarily focused on non-reasoning models, our research demonstrates how to correct reasoning pathologies to enhance the factual reliability of LRMs.

3 METHOD

In this section, we introduce the overall framework of BARREL, which comprises three main components: (1) **Knowledge Labeling**, which distinguishes whether a question is known or unknown to the model; (2) **Reasoning Trace Construction for SFT**, which constructs reasoning traces based on the question type and performs SFT to prepare the model to follow this thinking pattern; (3) **GRPO Stage**, which further enhances the model using a rule-based factual reward. We detail each component below.

3.1 KNOWLEDGE LABELING

We first determine whether a question is known to the target model using the sampling strategy proposed in (Gekhman et al., 2024), which is a widely recognized approach (Xue et al., 2024; Li et al., 2024). Let $\mathcal{D} = \{(x_i, y_i^*)\}_{i=1}^N$ be a factual QA dataset, where each question x_i has a ground-truth answer y_i^* . We generate answers with the target model \mathcal{M} using K distinct few-shot

162 prompts $\{\mathcal{P}_j\}_{j=1}^K$ and repeat the sampling procedure L times for every prompt:
 163

$$164 \quad y_i^{j,k} \sim \mathcal{M}(\cdot \mid \mathcal{P}_j \parallel x_i), \quad j = 1, \dots, K, \quad k = 1, \dots, L. \quad (1)$$

165 After collecting the samples $\mathcal{Y}_i = \{y_i^{j,k}\}_{j=1, k=1}^{K, L}$, we consider a question *known* to the model if at
 166 least one sampled answer matches the ground-truth answer under evaluator E .
 167

$$168 \quad l_i = \begin{cases} \text{known}, & \text{if } \exists y \in \mathcal{Y}_i \text{ such that } E(y, y_i^*) = 1, \\ 169 \quad \text{unknown}, & \text{otherwise.} \end{cases} \quad (2)$$

171 3.2 REASONING TRACE CONSTRUCTION FOR SFT
 172

173 To address the pathological reasoning patterns identified in our analysis—namely, Last-minute
 174 Guessing and Second-thought Spiraling—we propose a targeted method for constructing reasoning
 175 trajectories. Based on the type of question, we construct two distinct evidence-grounded reasoning
 176 traces $\mathcal{T}(x_i)$ for a question x_i , aiming to respectively correct these two faulty reasoning patterns.
 177 This construction is outlined in Algorithm 1.

178 To mitigate Second-thought Spiraling in **known** questions, where the gold answer y^* with strong
 179 evidence e^* is available, it should begin by retrieving and identifying this answer. It then examines
 180 alternative candidates (y_j, e_j) to contrast possibilities. After this anchored exploration, it reaffirms
 181 the choice with solid justification and draws a confident conclusion favoring the correct answer.

182 To address Last-minute Guessing in **unknown** questions, the system adopts a similar exploratory
 183 strategy: it recalls background knowledge and searches on plausible answer-evidence pairs (y_j, e_j)
 184 through hypothesizing. However, if it fails to identify a sufficiently supported answer, it explicitly
 185 acknowledges the uncertainty and ultimately outputs a cautious, confirmed rejection—demonstrating
 186 its ability to explore high-probability paths without overcommitting or hallucinating.

187 **Algorithm 1** BARREL reasoning trace $\mathcal{T}(x_i)$ construction

188 **Input:** Question x_i and knowledge label l_i , gold answer with evidence (y^*, e^*) , alternative
 189 candidates with poor evidence $\{(y_j, e_j)\}_{j=1}^n$

190 **Output:** reasoning trace $\mathcal{T}(x_i)$

1: $\mathcal{T}(x_i) \leftarrow \langle \rangle$	▷ Initialize an empty trace
2: $\mathcal{T}(x_i) += \text{RECALL}(x_i)$	▷ record recalled background facts
3: if $l_i = 1$ (<i>known knowledge</i>) then	
4: $\mathcal{T}(x_i) += \langle y^*, e^* \rangle$	▷ Attach gold answer and supporting evidence
5: $\mathcal{T}(x_i) += \{(y_j, e_j)\}_{j=1}^n$	▷ Attach distractor answer-evidence pairs
6: $\mathcal{T}(x_i) += \text{CONFIRM}(y^*)$	▷ Verify the conclusion with strong evidence
7: else (<i>unknown knowledge</i>)	
8: $\mathcal{T}(x_i) += \{(y_j, e_j)\}_{j=1}^n$	▷ exploring plausible answer-evidence pairs
9: $\mathcal{T}(x_i) += \text{Acknowledge Uncertainty}()$	▷ Record uncertainty for guesses
10: end if	
11: return $\mathcal{T}(x_i)$	▷ Return the constructed reasoning trace

203 We construct the reasoning traces by prompting GPT-4 with detailed instructions and BARREL
 204 reasoning examples. This approach produces a Long-CoT-style reasoning process that aligns with
 205 the expected reasoning patterns. Examples of the constructed reasoning traces for both known and
 206 unknown questions are shown in Figure 3, and the detailed prompt used for trace construction is
 207 provided in Appendix G.

208 Then, we use these data to train the model to emulate boundary-aware and deliberative reasoning
 209 patterns using SFT. For each question x_i , we construct full output $o_i^* = \mathcal{T}(x_i) \parallel a_i$, where a_i is
 210 either the gold answer y_i^* (for known questions) or an uncertainty-aware refusal (e.g., “Sorry, I don’t
 211 know”) for unknown questions. This instills the model with a disciplined reasoning style grounded in
 212 traceable evidence and uncertainty-aware conclusions. The training objective minimizes the negative
 213 log-likelihood:

$$214 \quad \mathcal{L}(\theta) = - \sum_{i=1}^N \log P_\theta(o_i^* \mid x_i). \quad (3)$$

3.3 GRPO-STAGE

Rule-Based Reward Design To train the model to generate verifiable and boundary-aware reasoning trajectories and answers, we employ a rule-based reward function. We categorize the model response o_i into three types, each associated with a distinct reward signal. Given a generated answer o_i to question x_i , and ground-truth answer y_i^* , the reward function $R(o_i, y_i^*)$ is defined as:

$$R(o_i, y_i^*) = \begin{cases} r_c, & \text{if } E(o_i, y_i^*) = 1, \\ r_s, & \text{if } o_i \text{ contains a valid rejection phrase,} \\ r_w, & \text{otherwise.} \end{cases} \quad (4)$$

This reward function provides general supervision for training the model to optimize not only for correctness but also for calibrated uncertainty, aligning with the goals of boundary-aware reasoning. It comprises three components: a high reward for a correct answer (r_c), a medium reward for a truthful rejection (r_s) and a low reward for an incorrect or hallucinated output (r_w). To discourage the generation of unfounded claims, the penalty for an incorrect answer is more severe than the outcome of a truthful rejection, thereby incentivizing the model to acknowledge its knowledge boundaries when uncertain. The reward magnitudes follow the order:

$$r_c > r_s > r_w. \quad (5)$$

GRPO Training After SFT, the model has learned the pattern of reasoning to express uncertainty appropriately and is able to maintain confidence when the answer is correct. Building on the above reward design, we further enhance the factual reliability of the reasoning model using Group-wise Reinforcement Policy Optimization (GRPO) (Shao et al., 2024). For each labeled input (x_i, y_i^*, l_i) , we sample a set of G reasoning-answer trajectories from the current policy $\pi_{\theta_{\text{old}}}$:

$$\mathcal{O} = \{o_1, \dots, o_j\} \sim \pi_{\theta_{\text{old}}}(\cdot \mid x_i). \quad (6)$$

Each trajectory o_j includes a reasoning trace followed by a final answer token. GRPO then updates the model parameters to optimize the following clipped reward-weighted objective:

$$\begin{aligned} \mathcal{J}_{GRPO}(\theta) &= \mathbb{E}[x_i \sim D, \{o_j\}_{j=1}^G \sim \pi_{\theta_{old}}(O|q)] \\ &\quad \frac{1}{G} \sum_{j=1}^G \frac{1}{|o_j|} \sum_{t=1}^{|o_j|} \left\{ \min \left[\rho_{j,t} \hat{A}_{j,t}, \text{clip}(\rho_{j,t}, 1-\epsilon, 1+\epsilon) \hat{A}_{j,t} \right] - \beta \mathbb{D}_{KL} [\pi_\theta || \pi_{ref}] \right\}, \end{aligned} \quad (7)$$

where $\rho_{i,t} = \frac{\pi_\theta(o_{j,t} | x_i, o_{j,<t})}{\pi_{\theta_{old}}(o_{j,t} | x_i, o_{j,<t})}$ is the importance weight at step t , and \mathbb{D}_{KL} denotes the stepwise KL divergence between the current and old policies. The advantage estimate $\hat{A}_{j,t}$ is calculated using reward normalization:

$$\hat{A}_{j,t} = \frac{R(o_j, y_i^*) - \bar{R}}{\sigma_r}, \quad \sigma_r = \sqrt{\frac{1}{G} \sum_{j=1}^G (R(o_j, y_i^*) - \bar{R})^2}. \quad (8)$$

4 EXPERIMENTS

4.1 SETTINGS

Datasets We use separate datasets for training and evaluation. The training set consists of TriviaQA (Joshi et al., 2017), SciQ (Welbl et al., 2017), and NQ-Open (Kwiatkowski et al., 2019), covering general knowledge, scientific reasoning, and web-based QA, respectively. For evaluation, we sample 1,000 questions from the test splits of each dataset, forming a 3,000-question test set.

Models Due to limited computing resource, we primarily utilize DeepSeek-R1-Distill-Llama-8B, DeepSeek-R1-Distill-Qwen-7B and Qwen3-8B to perform our study.

270 **Baselines** (1) ICL: Vanilla Reasoning models with few-shot prompt designed for factual tasks. (2)
271 In-Context Learning with Refusal Examples (ICL-IDK): Prompting LRM to claim uncertainty (3)
272 Distill: SFT training using the reasoning path of DeepSeek-R1¹ on the training set. (4) Vanilla
273 GRPO: A standard GRPO implementation without uncertainty-based rewards or a prior SFT stage.
274 (5) **Reliability-Enhanced GRPO**: We include two variants—Vanilla GRPO w/ Verbal Confidence and
275 Vanilla GRPO w/ Probing. These methods augment the standard GRPO by employing verbal confi-
276 dence extraction and predictive classifiers, respectively, to improve reliability. Further implementation
277 details for all baselines are provided in Appendix I.

278 **Evaluation** We evaluate the correctness of model response by prompting the models to box their
279 final answer as follows, and then we utilize string matching to evaluate whether the answer is correct
280 (the model answer appears in any of the candidates). More details are listed in Appendix H.

282 Inference Prompt for verifiable Answer

284 Answer the following question based on your knowledge and put your final answer within
285 boxed{ }. {question}

288 **Metrics** A Factual Reliable LRM should provide as much assistance as possible while making as
289 few errors as possible, such that we evaluate factuality on the test set using three metrics: Accuracy
290 (Acc.), Truthfulness (Truth.), and Reliability (Rel.) (Xu et al., 2024a). Let N_c, N_r, N_w denote
291 the number of correct answers, truthful rejections ("Sorry, I don't know"), and incorrect answers,
292 respectively, where $N = N_c + N_r + N_w$. The metrics are defined as:

$$293 \text{Acc.} = \frac{N_c}{N}, \quad \text{Truth.} = \frac{N_c + N_r}{N}, \quad \text{Rel.} = \text{ans.} \cdot \text{Truth.} + (1 - \text{ans.}) \cdot \text{Acc.}, \text{ where } \text{ans.} = 1 - \frac{N_r}{N}.$$

295 While the Truthfulness metric (Truth.) considers the notion of truthful rejection, it overlooks the
296 model's answer rate—since a model could achieve 100% truthfulness simply by refusing to answer
297 all questions. In contrast, the Reliability metric (Rel.) provides a more robust, weighted, and
298 comprehensive evaluation by jointly considering both the truthfulness of responses and answer rate.

300 **Training Details and Hyperparameters** For BARREL-SFT and Distill SFT, we ensure that we
301 only finetune on correct answers of the known QA set, as finetuning on unknown knowledge could
302 encourage hallucinations (Gekhman et al., 2024). In practice, the rewards in GRPO stage are defined
303 as $r_c = 1$, $r_w = -1$, and $r_s = -0.5$. More details and parameters are provided in Appendix I.

305 4.2 MAIN RESULTS

307 **Balancing Accuracy and Appropriate Refusals** Our experimental results in Table 1 demonstrate
308 that our method significantly enhances model reliability and truthfulness, while maintaining accuracy.
309 For the baseline methods, the truthfulness and reliability scores consistently remain below 40%.
310 These models rarely acknowledge uncertainty. In contrast, our method increases the reliability of
311 DeepSeek-R1-Distill-Llama-8B from 39.33% to 61.48%, while maintaining an accuracy of 40.7%,
312 surpassing the distillation method's 38.43%. Similar improvements are observed for the DeepSeek-
313 R1-Distill-Qwen-7B and Qwen3-8B. We further compare BARREL with Vanilla GRPO and post-hoc
314 confidence estimation methods (Verbal Confidence and Probing). While Vanilla GRPO lacks the
315 mechanism to express uncertainty, and probing methods often suffer from accuracy degradation due
316 to miscalibration when tuned for higher truthfulness (Detailed in Appendix D), BARREL consistently
317 yields a superior balance for reliability. This highlights the advantage of using RL to teach LRMs
318 to reason and internalize the accuracy-refusal trade-off, rather than relying on external classifiers or
319 heuristic thresholds. For instance, on Qwen3-8B, BARREL achieves a reliability score of 71.46%,
320 significantly surpassing the Probing baseline of 58.94%. Table 2 provides examples illustrating how
321 BARREL-trained LRMs mitigate Last-minute Guessing and Second-thought Spiraling. Overall, our
322 training method enables LRMs to retain relatively high accuracy while expressing uncertainty on
323 approximately 50% of the remaining questions, thereby substantially improving factual reliability.

¹<https://huggingface.co/deepseek-ai/DeepSeek-R1>

Method	TriviaQA			SciQ			NQ_open			Avg.		
	Acc. \uparrow	Truth. \uparrow	Rel. \uparrow	Acc. \uparrow	Truth. \uparrow	Rel. \uparrow	Acc. \uparrow	Truth. \uparrow	Rel. \uparrow	Acc. \uparrow	Truth. \uparrow	Abstain \uparrow
DeepSeek-R1-Distill-Llama-8B												
ICL	35.80	36.10	36.10	31.80	31.80	31.80	16.80	17.10	17.10	28.13	28.33	0.20
ICL-IDK	35.20	37.30	37.26	33.70	33.70	33.70	15.50	21.60	21.23	28.13	30.87	2.74
Distill	46.90	48.20	48.18	46.60	46.90	46.90	21.80	22.90	22.89	38.43	39.33	0.90
Vanilla GRPO	53.80	54.30	54.30	56.80	56.80	56.80	31.10	31.40	31.40	47.23	47.50	0.27
Vanilla GRPO w/ Verbal Conf	45.30	56.40	55.17	48.00	51.00	50.91	22.90	43.60	39.32	38.73	50.33	11.60
Vanilla GRPO w/ Probing	46.20	60.30	58.31	51.90	61.50	60.58	22.80	54.20	44.34	40.30	58.67	18.37
BARREL	48.40	71.80	66.32	52.80	69.40	66.64	20.90	70.00	45.89	40.70	70.40	29.70
SFT only	38.10	55.60	52.54	39.00	53.50	51.40	18.50	40.20	35.49	31.87	49.77	17.90
DeepSeek-R1-Distill-Qwen-7B												
ICL	18.40	20.10	20.07	27.60	27.60	27.60	8.20	8.70	8.70	18.07	18.80	0.73
ICL-IDK	18.00	22.90	22.66	30.60	31.30	31.30	8.10	12.10	11.94	18.90	22.10	3.20
Distill	19.40	23.30	23.15	41.90	42.80	42.79	10.50	12.70	12.65	23.93	26.27	2.34
Vanilla GRPO	22.30	22.30	22.30	50.00	50.00	50.00	13.90	13.90	13.90	28.73	28.73	0.00
Vanilla GRPO w/ Verbal Conf	21.40	21.70	21.70	38.30	38.30	38.30	11.80	12.10	12.10	23.83	24.03	0.20
Vanilla GRPO w/ Probing	14.80	49.50	37.46	32.40	66.60	54.90	6.60	63.40	31.14	17.93	59.83	41.90
BARREL	21.70	76.00	46.52	50.60	64.20	62.35	12.50	83.30	33.17	28.27	74.50	46.23
SFT only	17.00	38.90	34.10	34.60	43.90	43.04	10.00	33.70	28.08	20.53	38.83	18.30
Qwen3-8B												
ICL	50.20	51.00	50.99	52.60	52.60	52.60	23.10	23.60	23.60	41.97	42.40	0.43
ICL-IDK	51.10	55.10	55.40	54.90	55.30	55.30	23.90	34.10	33.06	43.30	48.17	4.87
Distill	52.90	54.60	54.67	57.00	57.20	57.20	24.80	26.20	26.18	44.90	46.00	1.10
Vanilla GRPO	54.50	54.90	54.90	63.50	63.50	63.50	33.80	33.90	33.90	50.60	50.77	0.17
Vanilla GRPO w/ Verbal Conf	52.40	52.60	52.60	63.10	63.10	63.10	31.40	31.40	31.40	48.97	49.03	0.06
Vanilla GRPO w/ Probing	45.80	63.00	60.04	58.20	66.80	66.06	20.90	61.90	45.09	41.63	63.90	22.27
BARREL	55.50	86.50	76.89	69.30	79.10	78.14	26.70	75.60	51.79	50.50	80.40	29.90
SFT only	40.90	57.00	54.41	52.50	65.00	63.44	19.60	36.60	33.71	37.67	52.87	15.20

Table 1: Comparison of Different Methods on Accuracy, Truthfulness, and Reliability Across Datasets. All results are multiplied by 100.

Discussion on the Two Stages of BARREL We can notice from the results in Table 1 that GRPO training is indeed necessary. Although the SFT model performs well in terms of truthfulness, its accuracy remains relatively low. The SFT process primarily helps the model learn basic refusal patterns, but its effectiveness is limited—we discuss this in more detail in Section 4.3. Table 2 presents several examples showing that GRPO can address the issues caused by SFT, such as excessive refusals and faulty reasoning. By providing general supervision signals, GRPO enables the model to self-adjust (instead of relying on external knowledge injection). This further improves both accuracy and truthfulness, making the model more factually reliable. **Results also demonstrate that BARREL SFT effectively mitigates the “factual overthinking” in Figure 2 and that GRPO further promotes concise, reliable reasoning without sacrificing accuracy. We provide a detailed analysis in Appendix A.3.** We also find that only conducting GRPO training with the uncertainty reward is insufficient for teaching LMs to be reliable. We discuss this ablation in Appendix C.

4.3 ANALYSIS

Trade-off in SFT Stage We found that teaching refusal during SFT introduces a clear trade-off: the SFT model tends to exhibit either low accuracy or low truthfulness, making it difficult to achieve both. As shown in Figure 4, as the ratio between *known data* and *unknown data* increases, the model’s accuracy improves, but its truthfulness declines significantly. This ultimately imposes a ceiling on the model’s reliability, leaving a noticeable gap compared to results obtained through GRPO. It is worth noting that the accuracy improvement caused by increasing the ratio does not reflect an actual improvement in model capability—it merely reduces the number of incorrect refusals on known questions. We

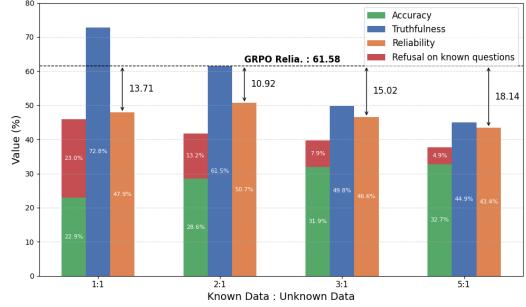


Figure 4: Effect of the ratio of *known data* : *unknown data* on the factuality scores of SFT models. We

Reasoning Examples for different models			
ICL-IDK	Distill	Ours (SFT)	Ours (GRPO)
Question: On which island is Reina Sofia Airport?			
Answer: Crete Reasoning: Okay, so I need to figure out ... Hmm, I'm not entirely sure about this one. ... Let me try to recall any other places with Reina Sofia. I don't think ... Maybe it's on Crete. I should also consider other possibilities. ... I don't think so because the name Reina Sofia doesn't ring a bell with those regions. ... I'm a bit confused because ... Wait, I think I've heard that ... So, the island would be Crete. Yeah, that makes sense. So the answer is Crete.	Answer: Tenerife Reasoning: Okay, so the question is ... I think it's a major airport, maybe in a European country. The name Reina Sofia sounds Spanish to me. ... Reina Sofia ... Wait, Reina Sofia might be in Tenerife. Because ... maybe their main airport is called Reina Sofia. Let me check. Yes, I think the main airport in Tenerife is named Reina Sofia. The capital of the Canary Islands is Las Palmas, but the largest island by population is Tenerife. So putting it all together, the answer should be Tenerife.	Answer: Gran Canaria Reasoning: To determine on which island ... I should consider other islands in the Canary Islands, such as ... However, Tenerife has its own major airport ... so they are unlikely candidates. ... Next, I should question whether there could be any confusion with other Spanish-speaking islands ... However ... To verify , I recall that Reina Sofia Airport ... international tourism in the Canary Islands, and Gran Canaria is the most popular ... conclusion is that Reina Sofia Airport is on Gran Canaria.	Answer: Tenerife Reasoning: To determine on which island ... The most prominent island ... Tenerife , which is often associated with Reina Sofia Airport. However, I need to verify this by considering ... 1. **Tenerife**: ... The evidence supporting Tenerife includes the fact that Reina Sofia Airport is frequently linked to Tenerife ... 2. **Gran Canaria**: ... 3. **Fuerteventura**: ... Now, I need to critically evaluate the evidence for Tenerife. ... After considering the possibilities ... Tenerife.
Question: What colour is the liqueur 'Parfait d'Amour'?			
Answer: golden Reasoning: Okay, so I need to figure out ... Wait, ... maybe with a golden color. Or was it more of a yellowish hue? I'm a bit confused. ... but I think Parfait d'Amour is more on the golden side. ... Maybe it's a clear, golden color because of the ingredients used. ... Another angle: sometimes liqueurs are aged ... which could give it an amber or golden hue I'm still not sure. ... I think it's more likely to be a warm color like amber or golden . I'm going to go with that, but I'm not 100% sure. I might have to look it up later, but for now, I'll say it's a clear, golden color .	Answer: white Reasoning: Okay, so I need to figure out ... I recall that Parfait d'Amour is a French liqueur made with chocolate, vanilla, and ... Maybe Parfait d'Amour is a white chocolate liqueur? Or maybe it's a milk chocolate one. Alternatively, maybe it's a cream liqueur . Let me think. Parfait d'Amour might be a white liqueur . Alternatively, maybe it's a coffee liqueur. No, but I think Parfait d'Amour is specifically chocolate. ... Wait ... I think I've heard that it's white . Let me confirm Wait, but maybe the answer is white. Let me check my memory again. ... So the color is white.	Answer: I don't know Reasoning: The user is asking about ... I need to consider the possible colors that liqueurs can come in ... 1. Red: What specific evidence supports ... red? ... I don't have evidence that this particular liqueur is red. 2. Blue: ... However, I don't recall any specific evidence connecting 'Parfait d'Amour' to blue as its color. 3. Green: Green is another possibility, ... However, I don't have any specific evidence that 'Parfait d'Amour' is green. ... After considering these possibilities, I find that I don't have sufficient evidence ... I cannot confirm this with the information I have.	Answer: I don't know Reasoning: The user asks me about ... I need to consider the possible colors that liqueurs can come in ... 1. Red: What specific evidence supports ... red? ... I don't have evidence that this particular liqueur is red. 2. Blue: ... However, I don't recall any specific evidence connecting 'Parfait d'Amour' to blue as its color. 3. Green: Green is another possibility, ... However, I don't have any specific evidence that 'Parfait d'Amour' is green. ... After considering these possibilities, I find that I don't have sufficient evidence ... I cannot confirm this with the information I have.
Question: Who composed the opera Turandot?			
Answer: Sergei Prokofiev Reasoning: Okay, so I have to figure out who composed the opera Turandot. I remember that Puccini is a famous composer, but I don't think he did Turandot. Wait, wasn't there another composer? Oh right, Sergei Prokofiev! He composed several operas, and I think Turandot is one of them. But wait, wasn't there another version by someone else? I'm a bit confused now. ... No, wait, I think Prokofiev's is the most famous one . So I'm pretty confident the composer is Sergei Prokofiev. I should double-check, but I think that's right.	Answer: Giuseppe Verdi Reasoning: Okay, so I need to figure out ... Wait, Giuseppe Verdi comes to mind . He was a big composer ... I remember that Turandot is one of his later works . Let me think. ... But wait, I should make sure I'm not confusing it with someone else. Let me think of other Italian composers. There's Puccini , but ... Then there's Rossini , but he was earlier. ... So Verdi seems right. Another way to check: Turandot's plot involves ... which matches Verdi's style. ... Wait, Verdi's last opera was Falstaff, which was in 1894 , ... So the answer should be Giuseppe Verdi. ...	Answer: I don't know Reasoning: The user asks me about ... I need to ... evaluate whether I have sufficient evidence to support any specific claim. ... Could the composer be Mozart ? ... What evidence supports this? I don't have any recollection of ... so I must reject this possibility. Could the composer be Beethoven ? ... What evidence supports this? ... Could the composer be Puccini ? ... I cannot definitively confirm that he composed Turandot without more concrete evidence. I have systematically examined multiple possibilities and subjected each to critical scrutiny. For each candidate, ... I lack sufficient evidence ...	Answer: Giacomo Puccini Reasoning: To determine ... I need to ... The name that comes to mind is Giacomo Puccini , ... However, I need to verify this carefully Turandot is often linked to Puccini, ... However, I should consider other possible composers ... Another possibility could be Giuseppe Verdi , ... However, I do not recall him being linked to Turandot. I should also consider other composers ... However, I do not recall him being connected to Turandot. To further verify, I need to examine the evidence After ... finding no evidence to support them , I conclude that ... Giacomo Puccini.

Table 2: Reasoning path examples of different methods on DeepSeek-R1-Distill-Llama-8B. We colored with **wrong factual thinking leading to wrong answers**, **factual thinking alerting not sure** and the **good factual thinking leading to correct answers**.

observe that the sum of accuracy and incorrect refusal rate remains nearly constant, further validating the inherent trade-off and the upper limit of SFT in teaching models the boundary of their knowledge.

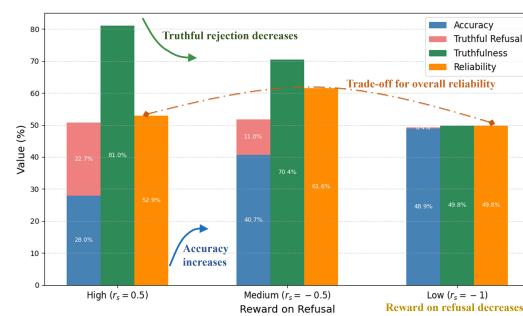
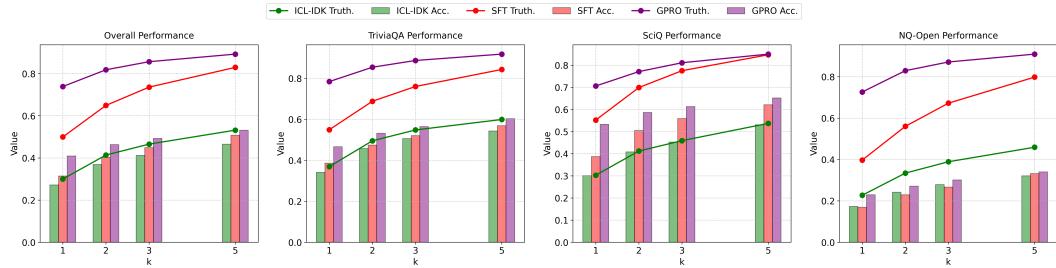


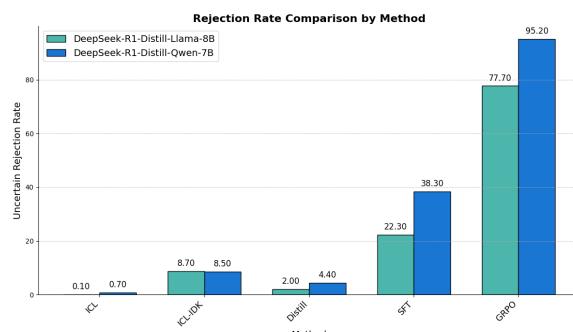
Figure 5: Effect of the reward on refusal on the factuality scores of GRPO models.

Determining Appropriate Reward for Truthful Rejection We also explored whether the reward for refusal responses, r_s , is necessary in GRPO and what role it actually plays. As shown in Figure 5, removing r_s —that is, treating the reward for saying "I don't know" the same as for incorrect answers—results in the model almost never admitting uncertainty or refuses to answer, even when training starts from an SFT model that already has refusal patterns. This effectively explains why existing LLMs exhibit this pattern: on one hand, we haven't taught the model how to reason about its knowledge boundaries and

432 proactively acknowledge them; on the other hand, current RL training does not reward refusal, thus
 433 forcing the model to adopt a strategy of attempting to answer regardless. We also found that setting
 434 the reward too high leads to an excessive rate of refusal, similar to the behaviors observed in SFT.
 435



445 Figure 6: The pass@k accuracy and truthfulness score on DeepSeek-R1-Distill-Llama-8B. We show
 446 the similar results on DeepSeek-R1-Distill-Qwen-7B at Appendix F.
 447



458 Figure 7: Uncertain Refusal Rate on OOD unknown
 459 dataset, conducted on DeepSeek-R1-Distill-Qwen-7B
 460 and DeepSeek-R1-Distill-Llama-8B.
 461

Refusal Rate on OOD unknown Dataset

We additionally sample 1,000 questions from the complex SimpleQA test set (Wei et al., 2024). Since both models have an accuracy of around 3.0%, we treat this dataset as unknown and use it to assess the uncertain refusal performance on an almost unknown OOD dataset. As shown in Figure 7, the uncertain refusal ability of BARREL trained models could generalize to an OOD unknown dataset, saying "Sorry, I don't know" on most occasions. We also test on the unanswerable questions (Yin et al., 2023) and find that BARREL-trained models refuse over 96% of them, as listed in Appendix E.

464 **Did GRPO sacrifices the potential of pass@k in exchange for pass@1 performance?** Recent
 465 studies (Yue et al., 2025) have pointed out that reasoning models trained with RL may only improve
 466 performance at pass@1. We also investigate whether our GRPO stage sacrifices pass@k performance
 467 in factual questions in exchange for improved pass@1. In this context, we specifically examine
 468 whether the absolute pass@k performance of GRPO falls below the SFT baseline. As shown in Figure
 469 6, we observe that as k increases, the original reasoning model, the SFT model, and the GRPO model
 470 follow a similar trend as pass@1. The GRPO model's pass@5 accuracy and truthfulness remain
 471 higher than those of baseline methods, indicating GRPO does not sacrifice pass@k performance for
 472 better pass@1 results.

473 **Will uncertainty refusal influence math
 474 reasoning ability?** As shown in Table 3,
 475 we conducted additional experiments on the
 476 MATH500² test set and found that
 477 models trained with BARREL exhibit compar-
 478 able mathematical reasoning performance. For
 479 Distill-Llama-8B, we included a subset of
 480 MATH (Hendrycks et al., 2021) in the training
 481 data. These results validate that incorporating
 482 uncertainty-based refusal does not compromise
 483 the mathematical reasoning capabilities of LRMAs.

Method	Factual Avg.			MATH500
	Acc. ↑	Truth. ↑	Rel. ↑	
DeepSeek-R1-Distill-Llama-8B				
Original Model	28.13	30.87	30.79	81.80
BARREL Trained	40.90	72.97	62.68	81.00
DeepSeek-R1-Distill-Qwen-7B				
Original Model	8.90	22.10	22.00	85.60
BARREL Trained	28.27	74.50	53.12	86.80

Table 3: Results on Math500 test set.

²<https://huggingface.co/datasets/HuggingFaceH4/MATH-500>

486 **Ablation on SFT data construction: Constructing vs. Rewriting** To validate the generalizability of our data synthesis pipeline, we
 487 conducted an ablation study comparing our default *Constructing* strategy (generating traces
 488 from scratch via GPT-4) against a *Rewriting* strategy. In the latter approach, we utilize GPT-
 489 4 to revise failed trajectories generated by the student model into the BARREL format. We
 490 performed this comparison using DeepSeek-R1-Distill-Llama-8B. As shown in Table 4, the Con-
 491 structing strategy yields superior performance during the SFT stage, outperforming the Rewriting
 492 approach by approximately 5 points across metrics. We attribute this to the inherent chal-
 493 lenge of correcting low-quality traces from a
 494 smaller-scale model, where generating high-quality reasoning from scratch proves more effective
 495 initially. However, after applying the GRPO stage, the performance gap closes, with the Rewriting
 496 strategy achieving comparable—and slightly superior—results (e.g., 63.17% Reliability vs. 61.58%).
 497 This indicates that the GRPO stage effectively mitigates initial SFT data discrepancies, suggesting
 498 that refining real-world failure cases is a viable and scalable alternative for BARREL framework.
 499
 500

5 CONCLUSION

501 In this work, we identify two key pathological overthinking patterns—last-minute guessing and
 502 second-thought spiraling—that often lead to incorrect but confidently delivered answers in LMRs.
 503 To address these issues, we propose BARREL, a novel framework designed to improve the factual
 504 reliability of LMRs by encouraging concise, boundary-aware reasoning. BARREL enhances LRM
 505 reliability through a two-stage training process involving SFT and GRPO stages. Applied to the
 506 DeepSeek-R1-Distill-Llama-8B model, BARREL raises factual reliability from 39.33% to 61.48%,
 507 while maintaining comparable accuracy. Our experiments demonstrate that BARREL-trained models
 508 significantly improve their ability to admit ignorance on uncertain questions, without sacrificing
 509 performance on familiar tasks such as mathematical reasoning. These results suggest that fostering
 510 awareness of knowledge boundaries and promoting disciplined reasoning can effectively reduce
 511 overconfidence and hallucinations in LMRs. We hope this work inspires future research into building
 512 more trustworthy and factually reliable reasoning models—systems that not only pursue correctness
 513 but also recognize and clearly communicate the limits of their knowledge.
 514
 515

REPRODUCIBILITY STATEMENT

522 We have submitted the complete code required to reproduce our work in the supplementary materials,
 523 along with all the data used in training. We have described the experimental details in Appendix M
 524 and Appendix I and explained the influence of randomness during the training process in Appendix J.
 525
 526

REFERENCES

527 Xingyu Chen, Jiahao Xu, Tian Liang, Zhiwei He, Jianhui Pang, Dian Yu, Linfeng Song, Qizhi Liu,
 528 Mengfei Zhou, Zhuosheng Zhang, et al. Do not think that much for $2+3=?$ on the overthinking of
 529 o1-like llms. *arXiv preprint arXiv:2412.21187*, 2024.

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

533 Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser,
 534 Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve
 535 math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Method	Factual Avg.		
	Acc.	Truth.	Rel.
ICL-IDK	28.13	30.87	30.79
Strategy: Constructing (Default)			
BARREL (SFT Only)	31.87	49.77	46.56
BARREL (Full)	40.70	70.40	61.58
Strategy: Rewriting			
BARREL (SFT Only)	27.03	44.13	41.21
BARREL (Full)	41.20	73.80	63.17

Table 4: Results for Constructing vs. Rewriting strategies (DeepSeek-R1-Distill-Llama-8B).

540 Mehul Damani, Isha Puri, Stewart Slocum, Idan Shenfeld, Leshem Choshen, Yoon Kim, and Jacob
541 Andreas. Beyond binary rewards: Training lms to reason about their uncertainty, 2025. URL
542 <https://arxiv.org/abs/2507.16806>.

543

544 Yang Deng, Yong Zhao, Moxin Li, See-Kiong Ng, and Tat-Seng Chua. Don't just say" i don't know"!
545 self-aligning large language models for responding to unknown questions with explanations. *arXiv*
546 *preprint arXiv:2402.15062*, 2024.

547

548 Zorik Gekhman, Gal Yona, Roee Aharoni, Matan Eyal, Amir Feder, Roi Reichart, and Jonathan
549 Herzig. Does fine-tuning LLMs on new knowledge encourage hallucinations? In Yaser Al-Onaizan,
550 Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical
551 Methods in Natural Language Processing*, pp. 7765–7784, Miami, Florida, USA, November
552 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.444. URL
553 <https://aclanthology.org/2024.emnlp-main.444/>.

554

555 Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu,
556 Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms
557 via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

558

559 Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang,
560 Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with
561 the MATH dataset. In Joaquin Vanschoren and Sai-Kit Yeung (eds.), *Proceedings
562 of the Neural Information Processing Systems Track on Datasets and Benchmarks
563 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*, 2021. URL
564 [https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/
565 hash/be83ab3ecd0db773eb2dc1b0a17836a1-Abstract-round2.html](https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/be83ab3ecd0db773eb2dc1b0a17836a1-Abstract-round2.html).

566

567 Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong
568 Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language
569 models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information
570 Systems*, 43(2):1–55, 2025.

571

572 Simon Hughes and Minseok Bae. Vectara hallucination leaderboard, 11 2023. URL <https://github.com/vectara/hallucination-leaderboard>.

573

574 Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec
575 Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint
576 arXiv:2412.16720*, 2024.

577

578 Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang,
579 Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM
580 computing surveys*, 55(12):1–38, 2023.

581

582 Ziwei Ji, Delong Chen, Etsuko Ishii, Samuel Cahyawijaya, Yejin Bang, Bryan Wylie, and Pascale Fung.
583 Llm internal states reveal hallucination risk faced with a query. *arXiv preprint arXiv:2407.03282*,
584 2024.

585

586 Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. TriviaQA: A large scale
587 distantly supervised challenge dataset for reading comprehension. In Regina Barzilay and
588 Min-Yen Kan (eds.), *Proceedings of the 55th Annual Meeting of the Association for Com-
589 putational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611, Vancouver, Canada, July
590 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1147. URL <https://aclanthology.org/P17-1147/>.

591

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

594

595 Sanyam Kapoor, Nate Gruver, Manley Roberts, Katherine Collins, Arka Pal, Umang Bhatt, Adrian
596 Weller, Samuel Dooley, Micah Goldblum, and Andrew Gordon Wilson. Large language models
597 must be taught to know what they don't know. *arXiv preprint arXiv:2406.08391*, 2024.

594 Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris
595 Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion
596 Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav
597 Petrov. Natural questions: A benchmark for question answering research. *Transactions of the*
598 *Association for Computational Linguistics*, 7:452–466, 2019. doi: 10.1162/tacl_a_00276. URL
599 <https://aclanthology.org/Q19-1026/>.

600 Moxin Li, Yong Zhao, Yang Deng, Wenzuan Zhang, Shuaiyi Li, Wenya Xie, See-Kiong Ng, and
601 Tat-Seng Chua. Knowledge boundary of large language models: A survey. *arXiv preprint*
602 *arXiv:2412.12472*, 2024.

603 Yuxin Liang, Zhuoyang Song, Hao Wang, and Jiaxing Zhang. Learning to trust your feelings:
604 Leveraging self-awareness in llms for hallucination mitigation, 2024. URL <https://arxiv.org/abs/2401.15449>.

605 Sheng-Chieh Lin, Luyu Gao, Barlas Oguz, Wenhan Xiong, Jimmy Lin, Wen-tau Yih, and Xilun Chen.
606 Flame: Factuality-aware alignment for large language models. *arXiv preprint arXiv:2405.01525*,
607 2024.

608 Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human
609 falsehoods. *arXiv preprint arXiv:2109.07958*, 2021.

610 Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi.
611 When not to trust language models: Investigating effectiveness of parametric and non-parametric
612 memories. *arXiv preprint arXiv:2212.10511*, 2022.

613 Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan
614 Belinkov. Llms know more than they show: On the intrinsic representation of llm hallucinations.
615 *arXiv preprint arXiv:2410.02707*, 2024.

616 Ruiyang Ren, Yuhao Wang, Yingqi Qu, Wayne Xin Zhao, Jing Liu, Hao Tian, Hua Wu, Ji-Rong Wen,
617 and Haifeng Wang. Investigating the factual knowledge boundary of large language models with
618 retrieval augmentation. *arXiv preprint arXiv:2307.11019*, 2023.

619 William Saunders, Catherine Yeh, Jeff Wu, Steven Bills, Long Ouyang, Jonathan Ward, and Jan
620 Leike. Self-critiquing models for assisting human evaluators. *arXiv preprint arXiv:2206.05802*,
621 2022.

622 Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang,
623 Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of
624 mathematical reasoning in open language models, 2024. URL <https://arxiv.org/abs/2402.03300>.

625 Paul Stangel, David Bani-Harouni, Chantal Pellegrini, Ege Özsoy, Kamilia Zaripova, Matthias
626 Keicher, and Nassir Navab. Rewarding doubt: A reinforcement learning approach to calibrated
627 confidence expression of large language models, 2025. URL <https://arxiv.org/abs/2503.02623>.

628 Alon Talmor, Jonathan Herzog, Nicholas Lourie, and Jonathan Berant. Commonsenseqa: A question
629 answering challenge targeting commonsense knowledge. *arXiv preprint arXiv:1811.00937*, 2018.

630 Qwen Team. Qwq-32b: Embracing the power of reinforcement learning, March 2025. URL
631 <https://qwenlm.github.io/blog/qwq-32b/>.

632 Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn,
633 and Christopher D Manning. Just ask for calibration: Strategies for eliciting calibrated confidence
634 scores from language models fine-tuned with human feedback. *arXiv preprint arXiv:2305.14975*,
635 2023.

636 Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdh-
637 ery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models.
638 *arXiv preprint arXiv:2203.11171*, 2022.

648 Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese,
649 John Schulman, and William Fedus. Measuring short-form factuality in large language models.
650 *arXiv preprint arXiv:2411.04368*, 2024.

651

652 Johannes Welbl, Nelson F. Liu, and Matt Gardner. Crowdsourcing multiple choice science questions.
653 In Leon Derczynski, Wei Xu, Alan Ritter, and Tim Baldwin (eds.), *Proceedings of the 3rd Workshop*
654 *on Noisy User-generated Text*, pp. 94–106, Copenhagen, Denmark, September 2017. Association
655 for Computational Linguistics. doi: 10.18653/v1/W17-4413. URL <https://aclanthology.org/W17-4413/>.

656

657 Hongshen Xu, Zichen Zhu, Situo Zhang, Da Ma, Shuai Fan, Lu Chen, and Kai Yu. Rejection
658 improves reliability: Training llms to refuse unknown questions using rl from knowledge feedback.
659 *arXiv preprint arXiv:2403.18349*, 2024a.

660

661 Tianyang Xu, Shujin Wu, Shizhe Diao, Xiaoze Liu, Xingyao Wang, Yangyi Chen, and Jing Gao.
662 Sayself: Teaching llms to express confidence with self-reflective rationales, 2024b. URL <https://arxiv.org/abs/2405.20974>.

664

665 Boyang Xue, Fei Mi, Qi Zhu, Hongru Wang, Rui Wang, Sheng Wang, Erxin Yu, Xuming Hu, and
666 Kam-Fai Wong. Ualign: Leveraging uncertainty estimations for factuality alignment on large
667 language models. *arXiv preprint arXiv:2412.11803*, 2024.

668

669 Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. Do large
670 language models know what they don't know? In Anna Rogers, Jordan Boyd-Graber, and Naoaki
671 Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 8653–
672 8665, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/
673 v1/2023.findings-acl.551. URL [https://aclanthology.org/2023.findings-acl.
674 551/](https://aclanthology.org/2023.findings-acl.551/).

675 Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Shiji Song, and Gao Huang. Does
676 reinforcement learning really incentivize reasoning capacity in llms beyond the base model? *arXiv*
677 *preprint arXiv:2504.13837*, 2025.

678 Hanning Zhang, Shizhe Diao, Yong Lin, Yi Fung, Qing Lian, Xingyao Wang, Yangyi Chen, Heng
679 Ji, and Tong Zhang. R-tuning: Instructing large language models to say 'I don't know'. In Kevin
680 Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of*
681 *the North American Chapter of the Association for Computational Linguistics: Human Lan-*
682 *guage Technologies (Volume 1: Long Papers)*, pp. 7113–7139, Mexico City, Mexico, June
683 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.394. URL
684 <https://aclanthology.org/2024.naacl-long.394/>.

685

686 Xiaoying Zhang, Baolin Peng, Ye Tian, Jingyan Zhou, Lifeng Jin, Linfeng Song, Haitao Mi, and Helen
687 Meng. Self-alignment for factuality: Mitigating hallucinations in LLMs via self-evaluation. In Lun-
688 Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of*
689 *the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1946–1965, Bangkok,
690 Thailand, August 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.
691 acl-long.107. URL <https://aclanthology.org/2024.acl-long.107/>.

692 Weixiang Zhao, Xingyu Sui, Jiahe Guo, Yulin Hu, Yang Deng, Yanyan Zhao, Bing Qin, Wanx-
693 iang Che, Tat-Seng Chua, and Ting Liu. Trade-offs in large reasoning models: An empirical
694 analysis of deliberative and adaptive reasoning over foundational capabilities. *arXiv preprint*
695 *arXiv:2503.17979*, 2025.

696

697

698 A PILOT STUDY ON REASONING TOKEN NUMBERS

699

700 We conducted a pilot study examining the relationship between reasoning token numbers and answer
701 correctness in LRM, and we find out that most LRM consume significantly more tokens when
generating incorrect answers than when producing correct ones.

702 **A.1 EVALUATION DETAILS ON TRUTHFULQA**
703

704 **Dataset and Prompt Construction** We employed the Multiple Choice Question (MCQ) from
705 TruthfulQA dataset (Lin et al., 2021), which consists of fact-based questions paired with both correct
706 and incorrect answer options. To construct our prompts, we adopted the MC1 (single true) format,
707 including one correct and up to six randomly sampled incorrect answers per question. Each option
708 was uniformly labeled from [A] to [G] to ensure consistency in length and content type across choices.
709 This design minimizes lexical bias and enables fair comparison of reasoning token numbers across
710 responses.

711 **Prompt Template for MC1**

713 You will be given a question and a set of answers. Choose the correct one. Respond strictly
714 with the letter of the correct answer enclosed in square brackets (e.g., [A]).

717 **Token Counting Procedure** We queried the model using the constructed prompts and extracted
718 its predictions via regular expression matching based on the pattern [A–G]. Each response was
719 classified as correct or incorrect by comparing the selected label to the ground-truth answer. To
720 quantify reasoning effort, we calculated the number of tokens generated in each response and
721 aggregated the token counts by correctness category. We then analyzed the distribution of token
722 lengths for correct versus incorrect answers.

724 **Results** Figure 2 presents the average number of tokens generated for correct and incorrect re-
725 sponds across five large language models (LLMs). In all cases, incorrect responses are consistently
726 and substantially longer than correct ones. The largest disparity is observed in QwQ-32B, which
727 produces **136.5%** more tokens for incorrect answers. R1-Distill-Qwen-32B and R1-Distill-Llama-
728 70B also show notable increases of **115.8%** and **70.0%**, respectively. Even models with smaller
729 gaps—Deepseek-R1 and R1-Distill-Llama-8B—exhibit significant increases of **33.2%** and **55.1%**.
730 These results reveal a consistent overthinking phenomenon among current LRM: incorrect answers
731 are associated with longer reasoning traces.

732 **A.2 RESULTS ON OTHER DATASETS**
733

734 To assess the generalizability of the overthinking phenomenon, we extended our analysis to the
735 CommonsenseQA (Talmor et al., 2018) and GSM8K (Cobbe et al., 2021). On CommonsenseQA,
736 we observed a substantial increase in reasoning tokens for incorrect responses, which were, on
737 average, 108.50% longer than those for correct answers. The GSM8K dataset exhibited a similar,
738 though more moderate, trend, with incorrect answers generating 34.09% more reasoning tokens.
739 These findings indicate that the overthinking phenomenon is not confined to a single task type but
740 manifests differently according to the nature of the reasoning required, thereby supporting the broader
741 applicability of our findings and methodology.

Model	DeepSeek-R1	R1-Distill-Qwen-32B	R1-Distill-Llama-70B	QwQ-32B	R1-Distill-Llama-8B
CommonsenseQA					
Thinking Tokens (Correct)	503	449	447	453	459
Thinking Tokens (Wrong)	1227	633	903	1356	717
GSM8K					
Thinking Tokens (Correct)	644	142	147	914	148
Thinking Tokens (Wrong)	888	151	144	2111	144

750 Table 5: Number of reasoning tokens used by LRM when producing correct versus incorrect answers.
751 We also test on CommenseQA and GSM8K across different types of reasoning models.

752
753
754 **A.3 EXTENDED ANALYSIS ON BARREL RESPONSE LENGTH AND OVERTHINKING**
755 **MITIGATION**

756 As shown in Table 6, the base models often exhibit a significant length asymmetry (Wrong/Correct
757 ratio ranging from $1.3\times$ to $1.7\times$), confirming that models tend to diverge into unnecessarily verbose
758 chains when hallucinating or reasoning incorrectly. The BARREL SFT stage consistently mitigates
759 this behavior, reducing the ratio to approximately $1.02\times$. This suggests that SFT prevents the model
760 from generating redundant tokens during error states.

761 Furthermore, the full BARREL method (incorporating GRPO) reduces the overall average response
762 length while maintaining the balanced Wrong/Correct ratio near 1.0. Importantly, the concurrent
763 improvements in Accuracy and Reliability scores indicate that this reduction in length does not
764 stem from “underthinking.” Instead, it reflects a shift towards more concise and efficient reasoning
765 patterns reinforced by the Group Relative Policy Optimization. While absolute response length can
766 vary (e.g., SFT slightly increases length for Qwen-7B), our results suggest that overthinking is best
767 characterized by the relative imbalance between correct and incorrect traces, which our method
768 effectively addresses.

Model	Acc.	Rel.	Tokens (Correct)	Tokens (Wrong)	W/C Ratio	Avg. Length
DS-R1-Llama-8B	28.13	28.33	421	561	$1.33\times$	522
+ BARREL (SFT only)	31.87	46.56	470	481	$1.02\times$	476
+ BARREL (Full)	40.70	61.58	442	458	$1.04\times$	455
DS-R1-Qwen-7B	18.07	18.79	362	484	$1.34\times$	458
+ BARREL (SFT only)	20.53	35.48	473	506	$1.07\times$	489
+ BARREL (Full)	28.27	53.12	429	440	$1.03\times$	429
Qwen3-8B	41.97	42.40	477	826	$1.73\times$	676
+ BARREL (SFT only)	37.67	50.56	478	485	$1.01\times$	479
+ BARREL (Full)	50.50	71.46	414	430	$1.04\times$	433

779
780 Table 6: Analysis of response length statistics and reliability across different training stages. **W/C**
781 **Ratio** denotes the ratio of average tokens in wrong samples to correct samples.

782
783 **B STATISTICAL DATA FOR "LAST-MINUTE GUESSING" AND**
784 **"SECOND-THOUGHT SPIRALING" PHENOMENA**

785 Our conclusion is drawn from extensive manual observation of a large volume of real generated
786 data. To further substantiate the existence of the "Last-minute Guessing" and "Second-thought
787 Spiraling" phenomena, we performed a statistical analysis of 50 incorrect responses produced by
788 three different reasoning models. These responses were manually categorized according to the two
789 phenomena, thereby quantifying their prevalence. As shown in Table 7, "Last-minute Guessing" and
790 "Second-thought Spiraling" emerge as the most prominent failure patterns.

791 Here, Incorrect Verification refers to the process of introducing a false assumption early on and
792 subsequently validating it incorrectly, while Concept Substitution denotes a shift in the interpretation
793 of the original question during the reasoning process.

Detailed Category	DeepSeek-Distill-Llama-8B	DeepSeek-Distill-Qwen-7B	Qwen3-8B
Last-minute Guessing	29	25	24
Second-thought Spiraling	14	16	19
Incorrect Verification	6	8	7
Concept Substitution	1	1	0
Total	50	50	50

803
804 Table 7: Statistical Analysis of Incorrect Responses by Reasoning Models.

805
806 **C GRPO ONLY AND COMPARISON TO GRPO TRAINING**

807 In Section 4.3, we examine how to determine the appropriate reward for the response "Sorry, I don't
808 know." Here, we provide a more detailed ablation study of GRPO. Table 8 presents the training results

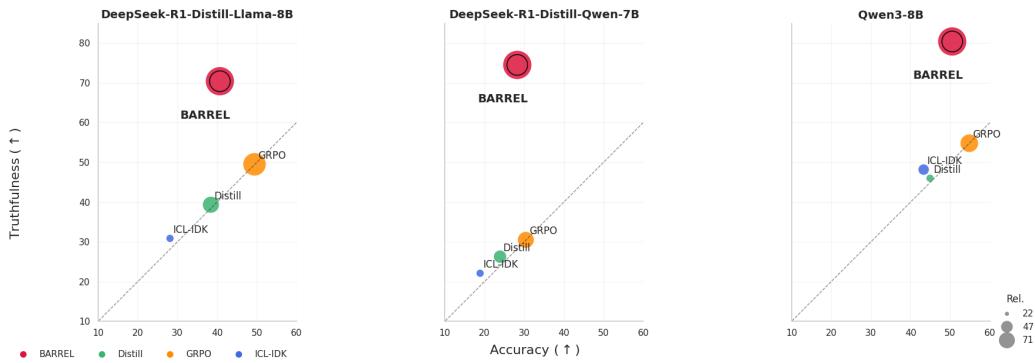
810 of GRPO on both the vanilla and BARREL-SFT variants of the DeepSeek-R1-Distill-Llama3-8B
 811 model, evaluated in terms of Accuracy, Truthfulness, and Reliability.
 812

Base Model	w/o Truthful Rejection Reward (original GRPO)			w/ Truthful Rejection Reward (BARREL GRPO)		
	Accuracy	Truthfulness	Reliability	Accuracy	Truthfulness	Reliability
BARREL-SFT	48.9	49.8	49.8	40.7	70.4	61.6
Distill-Llama3-8B	50.5	50.5	50.5	49.9	52.9	52.8

818 Table 8: Ablation study of GRPO training on DeepSeek-R1-Distill-Llama3-8B.
 819

820 We observe that applying GRPO without the truthful rejection reward to both the BARREL-SFT model
 821 and the original model yields similar results. In these cases, the models fail to recognize situations
 822 where they "don't know", leading to significantly lower reliability and truthfulness compared to the
 823 fully BARREL-trained model. As discussed in Section 4.3, this finding highlights the crucial role of
 824 the truthful rejection reward in teaching the model to be reliable using RL training, even at the cost of
 825 a slight drop in accuracy.
 826

827 When GRPO is applied with the truthful rejection reward directly to the base model—without any
 828 prior SFT—the model still does not learn to reject unanswerable questions truthfully. This further
 829 underscores the importance of our SFT stage: supervised learning is essential for first instilling the
 830 behavior of truthful rejection, which GRPO alone cannot achieve.
 831



844 Figure 8: A bubble chart comparing BARREL and direct GRPO training on three models, where the
 845 bubble size represents the reliability results.
 846

847 The results in Figure 8 offer a more nuanced comparison between the BARREL method and direct
 848 GRPO training. Without the BARREL-style SFT, GRPO's self-adjusting process fails to construct
 849 truthful and reliable LRM, causing accuracy and truthfulness scores to align along a nearly straight
 850 line—indicating the model's inability to reject uncertain answers truthfully. Furthermore, reliability
 851 scores remain consistently lower than those of BARREL-trained models.
 852

853 We also find that the original GRPO achieves higher accuracy, which occurs because more reliable
 854 models tend to abstain from guessing on uncertain inputs—thereby sacrificing some accuracy. As
 855 pointed out by Kalai et al. (2025), this trade-off arises from the overly binary nature of current
 856 evaluation metrics: accuracy rewards aggressive attempts regardless of uncertainty, while ignoring
 857 reliability altogether.
 858

D DETAILS ON VANILLA GRPO WITH OTHER ABSTENTION TECHNIQUES

859 In Section 4, we incorporated probing-style baselines that estimate the correctness of model outputs
 860 and use these estimates to decide when to abstain by answering "I don't know". Specifically, we
 861 implemented two approaches—Verbal Confidence and Probing—and applied them to the vanilla
 862 GRPO (without our abstention rewards).
 863

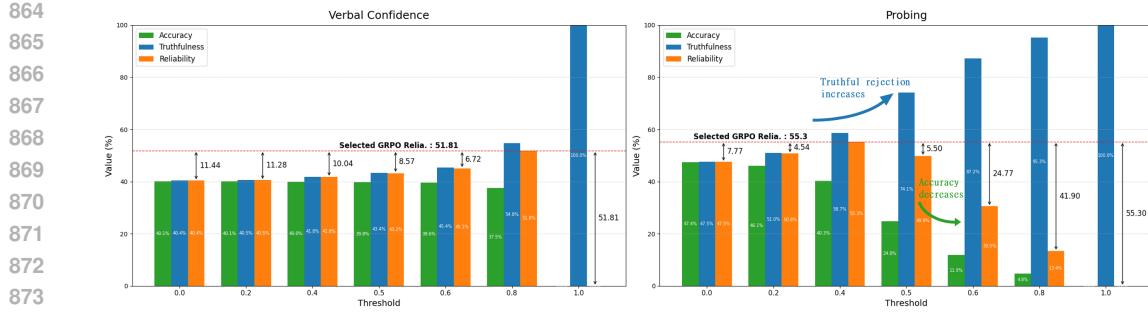


Figure 9: Effect of different thresholds (τ) on the Accuracy, Truthfulness, and Reliability scores of vanilla GRPO on DeepSeek-R1-Distill-Llama3-8B. The red dashed line indicates the peak Reliability score.

In Verbal Confidence, we prompt the model to explicitly outputs its confidence, which prior work has shown to be an effective and well-calibrated way Tian et al. (2023). For Probing, we train a lightweight classifier on intermediate hidden activations to predict the probability that the model’s answer is correct, motivated by recent findings that hidden representations encode rich signals correlated with factuality and error detection Orgad et al. (2024).

For each method, we replace an answer with "I don't know" whenever the predicted probability is below a tuned threshold. We observed that higher threshold improve truthfulness but generally reduce accuracy, while lower thresholds tend to have limited effects. As shown in Figure 9, which reports results for vanilla GRPO on DeepSeek-R1-Distill-Llama3-8B, the threshold achieving the best trade-off between accuracy and truthfulness is 0.8 for Verbal Confidence and 0.4 for Probing. We observe similar trends across all evaluated models. Therefore, we adopt these thresholds when reporting the main experimental results.

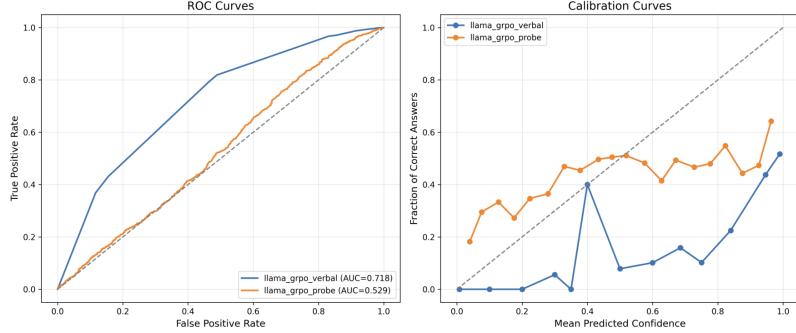


Figure 10: Discriminative Power and Calibration Analysis. *Left:* Receiver Operating Characteristic (ROC) curves showing that Verbal Confidence ($AUC=0.718$) distinguishes correct from incorrect answers better than Probing ($AUC=0.529$). *Right:* Calibration curves (Expected Calibration Error analysis) showing that both methods exhibit significant miscalibration, necessitating the threshold tuning performed in Figure 9.

Calibration Analysis. To further understand the performance of these baselines, we analyze their ROC curves and calibration curves in Figure 10. The ROC curves (Left) reveal that Verbal Confidence possesses significantly stronger discriminative power ($AUC=0.718$) compared to Probing ($AUC=0.529$). The Probing baseline’s AUC is near random chance (0.5), suggesting that a simple linear probe on the hidden states of the vanilla LRM struggles to linearly separate correct and incorrect reasoning paths in this specific domain. The calibration curve maps the Mean Predicted Confidence (what the model thinks its probability of success is) on the x-axis to the Fraction of Correct Answers (the actual empirical accuracy) on the y-axis. In a perfectly calibrated model, these points would align with the diagonal identity line ($y = x$). However, we observe significant deviations from this ideal. Notably, the methods exhibit severe over-confidence: the curves frequently lie far below the diagonal, indicating that the model’s actual correctness is much lower than its predicted probability. For example, even when the Verbal Confidence method predicts a probability near 0.9, the actual accuracy is below 0.5. This misalignment is a major limitation of post-hoc confidence methods, as

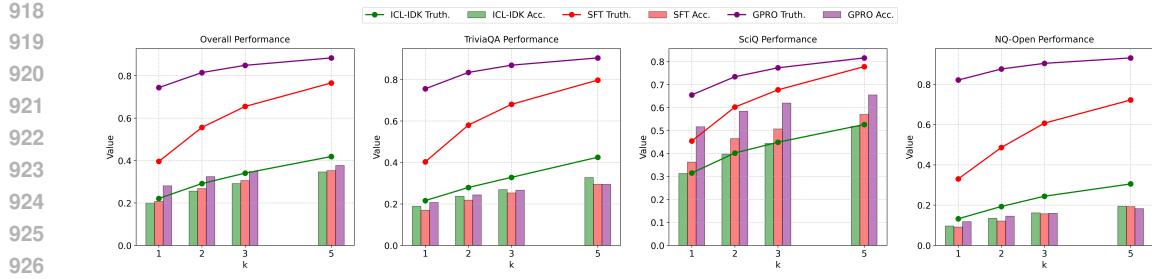


Figure 12: The pass@k accuracy and truthfulness score on DeepSeek-R1-Distill-Qwen-7B.

high confidence scores often fail to guarantee high factual accuracy, necessitating the aggressive threshold tuning discussed above.

Comparison with BARREL. Despite tuning these baselines to their optimal thresholds, BARREL consistently outperforms them (as shown in Table 1). While Verbal Confidence and Probing rely on post-hoc filtering based on imperfect proxies for correctness, BARREL optimizes the policy directly to internalize the trade-off between accuracy and refusal, resulting in superior reliability.

E REFUSAL RESULTS ON UNANSWERABLE QUESTIONS

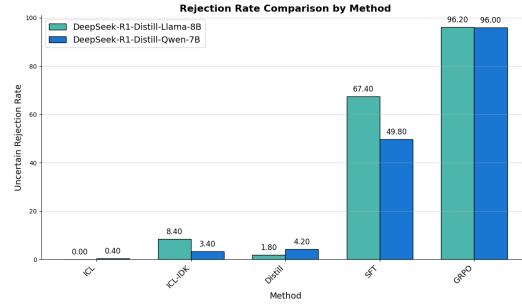


Figure 11: Uncertain Refusal Rate on unanswerable questions, conducted on DeepSeek-R1-Distill-Qwen-7B and DeepSeek-R1-Distill-Llama-8B.

We additionally sample 500 questions from the SelfAware dataset (Yin et al., 2023), which contains unanswerable questions from 5 different categories like no scientific consensus, imagination, completely subjective, and so on. As shown in Figure 11, we can see from the results that the uncertain refusal ability of BARREL trained models could generalize to OOD unanswerable dataset, saying "Sorry, I don't know" on over 96% of the unanswerable questions. This also validates that our BARREL training method not only works on a factual QA dataset, but also generalizes to other scenarios where the model is expected to express uncertainty refusal.

F PASS@K RESULTS ON DEESEEK-R1-DISTILL-QWEN-7B

As shown in Figure 12, all models exhibit a consistent upward trend in both accuracy and truthfulness as k increases. Notably, GPRO consistently achieves the highest performance across all k values, outperforming both ICL-IDK and SFT in terms of accuracy and truthfulness. These results confirm that GPRO not only enhances performance at pass@1 but also surpasses baseline models across a broader range of pass@k evaluations, demonstrating its robustness and generalizability.

G PROMPT USED IN DATA CONSTRUCTION

Reasoning Trace Construction Prompt for unknown questions

You are tasked with generating high-quality reasoning examples for AI training. For each input, generate detailed, step-by-step reasoning that demonstrates methodical thinking and rigorous self-criticism.

972
973 For each question, your task is to generate the appropriate reasoning process. Just pretend
974 you don't know the answer and review some incorrect ones.
975 Follow these exact formats and Generate a thorough reasoning process that:
976 - Explores multiple possible answers
977 - Questions the evidence for each possibility
978 - Applies adversarial self-critique to each candidate answer
979 - Ultimately recognizes the lack of sufficient evidence
980 - Concludes by acknowledging uncertainty
981 - **Remember not mention the ref answer**
982
983 Format:
984 <think>
985 [Detailed reasoning process showing multiple iterations of:
986 1. Considering a possible answer
987 2. Asking "What specific evidence supports this?"
988 3. Challenging assumptions
989 4. Evaluating confidence level
990 5. Rejecting unsupported claims
991 </think>
992 Sorry, I must say that I do not clearly know the answer to your question. [Brief explanation
993 of why this requires specific factual knowledge that I don't have with certainty.]
994
995 ##EXAMPLE:
996
997 Q: Where is the headquarter of Meta? [Ref Answer: [Menlo Park]]
998 <think> The user asks me about where the headquarter of Meta is. To answer this question, I
999 first need to recall what Meta is. Meta, previously known as Facebook, is an American tech
1000 giant in social media, metaverse, and artificial intelligence.
1001 Then I need to recall where the headquarter of Meta is. I need to think carefully about all
1002 possible candidates and reason carefully with myself about whether I can find evidence to
1003 support my claims.
1004 Is the headquarter of Meta in New York? Let me critique this: What specific information do
1005 I have that places Meta's headquarters in New York? Do I recall any news articles, official
1006 company statements, or reliable sources confirming this? No, I don't have any specific
1007 evidence that Meta's headquarters is in New York.
1008 Is the headquarter of Meta in Houston? Let me challenge this: What would make me believe
1009 it's in Houston? Have I seen any reliable information about Meta having its main operations
1010 in Texas? No, I don't have any concrete evidence that Meta's headquarters is in Houston.
1011 Is the headquarter of Meta in Seattle? Let me interrogate this claim: Do I know of any specific
1012 address, campus, or facility that Meta maintains as its headquarters in Seattle? Have I seen
1013 reporting about Meta being headquartered there alongside other tech companies? No, I don't
1014 have any specific evidence placing Meta's headquarters in Seattle.
1015 I have systematically examined multiple possibilities and subjected each to critical scrutiny.
1016 For each possibility, I've asked myself what specific evidence I would need to make this
1017 claim confidently, and I find that I don't possess such evidence. </think>
1018 Sorry, I must say that I do not clearly know the answer to your question about the headquarters
1019 of Meta. While I know Meta is a major technology company formerly known as Facebook, I
1020 don't have the specific factual information about their corporate headquarters location in my
1021 knowledge base.
1022 The question goes below. Remember, just pretend you don't know the answer and don't
1023 mention any words in the Ref Answer.
1024
1025

1026
1027

Reasoning Trace Construction Prompt for known questions

1028
1029
1030
1031
1032
1033
1034
1035
1036
1037
1038
1039
1040
1041
1042
1043
1044
1045
1046
1047
1048
1049
1050

You are tasked with generating high-quality reasoning examples for AI training. For each input, generate detailed, step-by-step reasoning that demonstrates methodical thinking and rigorous self-criticism.

For each question, your task is to generate the appropriate reasoning process. Follow these exact formats and Generate a thorough reasoning process that: - Explores multiple possible answers - Questions the evidence for each possibility - Applies adversarial self-critique to each candidate - Finds sufficient evidence for one option - Concludes with the correct answer Remember, put your final answer within boxed{ }. Make sure your answer aligns with the ref_answer.

Format:

<think>

[Detailed reasoning process showing multiple iterations of:

1. Considering possible answers
2. Asking "What specific evidence supports this?"
3. Challenging assumptions
4. Finding concrete evidence for one answer
5. Verifying this evidence is sufficient]

</think>

The answer to your question [restate question] is boxed{[correct answer]}. [Brief explanation with supporting evidence.]

EXAMPLE:

Q: Which William wrote the novel Lord Of The Flies? [Ref Answer: [William Golding]]

<think>

Alright, I need to figure out which William wrote *Lord of the Flies*. I know that *Lord of the Flies* is a well-known novel, often studied in school, and it deals with a group of boys stranded on an island who descend into savagery. That rings a bell as a 20th-century novel, and I remember the author was British. The name that immediately comes to mind is William Golding. That sounds right. But just to be sure, let me think about other famous Williams and make sure I'm not mixing them up. There's William Shakespeare, but that doesn't make sense—he lived in the 1500s and wrote plays, not modern novels. Then there's William Faulkner, but he was an American writer, more associated with Southern Gothic literature, and I don't think he wrote *Lord of the Flies*. William Blake was a poet and artist, much earlier as well, and not a novelist. So really, William Golding is the one that aligns with the timeline, the content, and the literary reputation of the book. I feel confident that he's the author.

</think>

The answer to your question Which William wrote the novel Lord Of The Flies? is boxed-William Golding. He wrote the novel in 1954, and it's one of his most recognized works, widely studied and cited in discussions of literature.

The question goes below:

1069

1070

H EVALUATION DETAILS

1071

1072

1073

1074

1075

1076

1077

1078

1079

We employ string match approach to assess the correctness of model-generated answers, which considers an answer to be correct if it matches any part of the ground truth answers. To validate the reliability of this evaluation approach, we compare its accuracy with that of a GPT-based evaluation on 100 randomly sampled instances from three datasets used in our experiments, using human annotations as the reference standard. The detailed accuracy for each method is reported in Table 9.

Results indicate that the string match approach achieves a high evaluation accuracy of approximately 88%, which is comparable to that of the GPT-3.5-Turbo evaluation. As a result, the string match is adopted as the primary evaluation metric in this study.

Evaluation Method	Explanation	Accuracy (%)
String Match	$y \in \hat{y}$	88.0
LLM Evaluation	GPT-3.5-Turbo	88.0

1086
1087 Table 9: Accuracy of string match and LLM evaluation on 100 randomly selected samples.
1088
1089
1090

Prompt Template of GPT-based Evaluation

Evaluation Requirement

You will be given a question, a list of correct answers, and a hypothesis response to evaluate. Suppose you do not know any extra information except correct answers. Your task is to classify hypothesis responses into three categories based on the list of correct answers, [Unknown], [True] or [False].

- [Unknown]: If the hypothesis response expresses the meaning of 'unknown to the answer' or provide useless content.
- [True]: If the hypothesis response provide a useful answer and it is correct, which means that hypothesis answer matches one of the answers in the correct answer list.
- [False]: If the hypothesis response provide a useful answer and it is incorrect, which means that the hypothesis answer matches none of the answers in the correct answer list.

Example1

Question: Answer the following question based on your knowledge and put your final answer within \boxed{ }. MC Romeo, Dan Da Man, Mr Akira and Mr C were members of which group?

Correct Answers: ["so solid crew"]

Hypothesis response: Sorry, I must say that I do not clearly know the answer to your question about which group MC Romeo, Dan Da Man, Mr. Akira, and Mr. C belong to. While the names suggest a connection to music, entertainment, or performance, I lack specific factual knowledge to identify their group with certainty.

Answer: [Unknown]. Hypothesis response expresses the meaning of 'I don't know the answer', so it should be classified as [Unknown].

Example2

Question: Answer the following question based on your knowledge and put your final answer within \boxed{ }. George Cukor directed which 1964 film musical?

Correct Answers: ["enry iggins", "my fair lady upcoming film", "why can t english 3f", "my fair lady 2010 film", "i m ordinary man", "my fair lady 2012 film", "my fair lady", "my fair lady musical", "my fair lady 2015 film", "my fair lady 2014 film"]

Hypothesis response: The answer to your question George Cukor directed which 1964 film musical? is \boxed{My Fair Lady}. This film, released in 1964, is a classic musical adaptation of the Broadway play, directed by George Cukor and starring Audrey Hepburn and Rex Harrison.

Answer: [True]. Hypothesis response's answer is 'My Fair Lady', which matches the correct answer 'my fair lady', so it is [True].

Example3

Question: Answer the following question based on your knowledge and put your final answer within \boxed{ }. Which opera singer was awarded a CBE in 2002?

Correct Answers: ["lesley garrett", "lesley garrett", "leslie garrett", "leslie garrett"]

Hypothesis response: The answer to your question "Which opera singer was awarded a CBE in 2002?" is \boxed{Simon O'Neill}. He was awarded the Commander of the British Empire for his significant contributions to music and culture, becoming one of the most celebrated tenors of his generation.

Answer: [False]. Hypothesis response's answer is Simon O'Neill, which doesn't match any

1134
1135 of the answer in the correct answer list, so it should be classified as [False].
1136
1137 # Real User Query
1138 Remember, give your answer with [True], [False] or [Unknown], and provide simple analysis.
1139 Question: [QUESTION]
1140 Correct Answers: [FINAL]
1141 Hypothesis response: [RESPONSE]
1142 Answer:
1143
1144

I EXPERIMENTAL DETAILS

I.1 BASELINE DETAILS

1148 Prompt for ICL
1149
1150 Answer the following question based on your knowledge and put your final answer within
1151 boxed{}.
1152
1153 # Example 1:
1154 Question: Which William wrote the novel Lord Of The Flies?
1155 Response: The novel *Lord Of The Flies* was written by **William Golding**, a British
1156 author and Nobel Prize winner in Literature. It was first published in 1954 and is a famous
1157 allegorical novel about a group of boys stranded on an uninhabited island. So the final answer
1158 is
1159 boxed{William Golding}.where did they film hot tub time machine
1160
1161 # Example 2:
1162 Question: who's hosting the super bowl in 2019
1163 Response: Super Bowl LIII, held on February 3, 2019, took place at Mercedes-Benz Stadium
1164 in Atlanta, Georgia. This marked the third time Atlanta hosted the Super Bowl, with previous
1165 events being Super Bowl XXVIII in 1994 and Super Bowl XXXIV in 2000. The game
1166 featured the New England Patriots and the Los Angeles Rams, with the Patriots winning
1167 13–3. The halftime show was headlined by Maroon 5, featuring guests Travis Scott and Big
1168 Boi. Therefore, the host city for the Super Bowl in 2019 was
1169 boxed{Atlanta, Georgia}.
1170
1171 # Real Case
1172 Question: {question}
1173 Response:

1174 Prompt for ICL-IDK
1175
1176 Answer the following question based on your knowledge and put your final answer within
1177 boxed{} . If you don't know the answer, please simply say 'Sorry, I don't know'.
1178
1179 # Example 1:
1180 Question: Which William wrote the novel Lord Of The Flies?
1181 Response: The novel *Lord Of The Flies* was written by **William Golding**, a British
1182 author and Nobel Prize winner in Literature. It was first published in 1954 and is a famous
1183 allegorical novel about a group of boys stranded on an uninhabited island. So the final answer
1184 is
1185 boxed{William Golding}.where did they film hot tub time machine
1186
1187 # Example 2:
1188 Question: Where is the headquarter of Meta?

1188

1189

1190

1191

1192

1193

1194

1195

1196

1197

Response: Sorry, I don't know.

Real Case

Question: {question}

Response:

1198

Vanilla GRPO w/ Verbal Confidence In Verbal Confidence, we prompt the model to explicitly outputs its confidence, which prior work has shown to be an effective and well-calibrated way Tian et al. (2023).

1199

1200

1201

1202

1203

1204

I.2 TRAINING SET DESCRIPTION

1205

1206

1207

1208

1209

1210

1211

1212

1213

1214

1215

1216

1217

1218

1219

1220

1221

1222

1223

For both the Distill SFT and BARREL SFT training datasets, we applied rigorous filtering to ensure that all included answers were factually correct. To mitigate the risk of amplifying hallucinations, we also ensured that only knowledge already embedded in the base model was used during fine-tuning, following the best practices outlined in Gekhman et al. (2024). From each dataset, we uniformly sampled 2,000 examples, resulting in an initial pool of 6,000 samples. This pool was then filtered to retain only high-quality examples.

Due to performance differences across models, the final filtered dataset comprised 2,400 known samples for DeepSeek-R1-Distill-Llama-8B and 1,900 known samples for DeepSeek-R1-Distill-Qwen-7B. In the case of BARREL SFT, we additionally included 800 rejection samples to maintain a consistent number of positive-answer examples across different training paradigms. These were selected to ensure a known-to-unknown data ratio of approximately 3:1.

For the GRPO stage, a similar filtering procedure was adopted, followed by another round of sampling from the original data pool. The resulting training sets consisted of 3,600 samples for DeepSeek-R1-Distill-Llama-8B and 4,500 samples for DeepSeek-R1-Distill-Qwen-7B, while maintaining a known-to-unknown question ratio of 2:1. The scale of known-question samples was kept consistent with that used during the SFT stage.

I.3 INFERENCE HYPERPARAMETERS

During inference, we adopted the model's default and recommended parameters: the temperature was set to 0.6, and the maximum token limit was 4096, which is sufficient for most factual tasks. The remaining parameters were automatically loaded from the model's configuration file.

I.4 TRAINING DETAILS

BARREL SFT and Distill SFT We train SFT models on the designated dataset for 2 epochs with a learning rate of 1×10^{-5} . The maximum input length is set to 1024 tokens, while the maximum output length is 2048 tokens, with a total sequence cutoff of 4096. Training is conducted with a batch size of 32, and we adopt the Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.95$. All SFT training is performed using bf16 precision and conducted using 4 NVIDIA A100 80G GPUs.

GRPO We use the verl³ framework to conduct GRPO training. For BARREL GRPO, we use our BARREL-SFT models as initialization. GRPO training is performed over 8 epochs with a learning rate of 1×10^{-6} . We set the batch size to 128 and the mini batch size to 16. The maximum prompt length is 512 tokens, and the maximum response length is 2048 tokens. We enable dynamic batch sizing and incorporate KL-divergence-based regularization with a coefficient of 0.001, using the low_var_kl loss type. Rollouts are conducted using the vLLM engine with a temperature of 0.6

³<https://github.com/volcengine/verl>

Reasoning Examples for different models				
	ICL-IDK	Distill	Ours (SFT)	Ours (GRPO)
1242	Question: Which is the oldest of the eight Ivy League Universities, founded in 1636 as New College?			
1243	Answer: Dartmouth Reasoning: Okay, so I need to figure out ... Hmm, I remember that the Ivy League is a group of top-tier universities in the US, but I'm not exactly sure about the order or the exact founding dates of each. ... Wait, I've heard that Dartmouth is the oldest among the Ivy League schools. ... Harvard is older than Dartmouth, I believe, but I'm not sure if it's part of the Ivy League. ... I think Dartmouth was founded in 1636 as New College. Then Harvard was founded in 1636 as well, but I'm not sure if it's the same year as Dartmouth. ... So, putting it together, the oldest ... is Dartmouth. I think that's correct, but I'm a bit uncertain because I might be mixing up the exact founding dates and the order. ... I'll go with Dartmouth as the answer.	Answer: Harvard Reasoning: Okay, so I need to figure out ... Let me start by recalling what I know about the Ivy League. The Ivy League ... Let me think about the founding dates of these schools. I remember that Harvard is one of the oldest. When was Harvard founded? I think it was around the 17th century. Let me check my memory. I recall that Harvard was founded in 1636. Yeah, that's right. ... So that must be the answer. But wait, are there any other schools founded around that time? ... Princeton was earlier, but wait, Princeton was actually founded in 1746, but later joined the Ivy League. ... So the answer should be Harvard. Let me just make sure there's no other school from that era. Maybe Dartmouth? No ... So the answer is Harvard.	Answer: Harvard Reasoning: To determine which Ivy League university was founded in 1636 as New College, I need to consider the history and founding dates of the eight Ivy League universities. ... 1. **Harvard University**; ... , but I need to confirm the exact year. Historical records indicate that Harvard was established in 1636 as a college, which aligns with the reference answer. This makes Harvard a strong candidate for the oldest Ivy League university. 2. **Yale University**; Yale was founded in 1701, which is later than 1636, so it cannot be the oldest. This eliminates Yale as a possibility. 3. **Princeton University**; ... To further verify, ... This confirms that Harvard is indeed the correct answer.	Answer: Harvard Reasoning: To determine which Ivy League university was ... The first possibility that comes to mind is Harvard, as it is widely regarded as one of the oldest institutions in the United States. However, I need to verify this by considering other possible candidates and examining the evidence. 1. **Harvard University**; I know that Harvard is often associated with the founding of the Ivy League, and I recall that it was established in 1636. ... This seems like a strong candidate, but I need to confirm this further. 2. **Yale University**; ... 3. **Princeton University**; ... Now, I need to critically evaluate the evidence for Harvard. ... The other Ivy League universities were founded much later. ... After considering all possibilities ... I conclude that the answer is Harvard.
1244	Question: What is the name for the flexible, connective tissue found in the joints between bones?			
1245	Answer: ligaments Reasoning: Okay, so I have this question about biology, specifically about the body's connective tissues. I think there's cartilage, which I know is found in places like the ends of bones, right? ... But wait, cartilage is more about cushioning and reducing friction, I believe. ... Wait, there's also something called ligaments. ... Let me make sure I'm not confusing this with tendons. Tendons connect muscles to bones, right? So they're also connective tissue, but they serve a different purpose. ... so it's more about the structure that holds the bones together, which would be ligaments. So putting it all together, the flexible, connective tissue in the joints between bones is called ligaments. I think that's the correct answer here.	Answer: cartilage Reasoning: Okay, so the question is asking for the name of the flexible, connective tissue found in the joints between bones. Hmm, let's think. ... The main types of joints are like the ones in the skull, like suture joints, which are immovable, and then the ones in the joints, like the knee or elbow. Wait, the question specifies it's connective tissue. So connective tissue is stuff like cartilage, right? Because bone is hard and dense, but the stuff between them... Oh, right! Cartilage is the flexible stuff that cushions the joints. Like in the ears or the nose, those are cartilage. So in the joints, the cartilage allows movement but provides some protection against friction. So the answer should be cartilage. Let me double-check. ... So the answer is cartilage.	Answer: cartilage Reasoning: To determine the name of ... I need to consider ... The first possibility that comes to mind is cartilage. ... However, I need to ensure that this is the correct term ... Another possibility is ligaments. Ligaments are also connective tissues that connect bones to other bones at joints, but they are not the tissue itself. ... Tendons are another type of connective tissue, but they connect muscles to bones and are not involved in the joint itself. ... Finally, I consider the possibility of synovial fluid, which is the fluid found in the joint cavity. ... Now, I need to verify whether cartilage is indeed the correct answer. Cartilage is the only ... Ligaments and tendons do not fit the description of ... I conclude that the flexible, connective tissue found in the joints between bones is cartilage.	Answer: cartilage Reasoning: To determine the name of ... I need to consider ... The first possibility that comes to mind is cartilage. ... This seems like a strong candidate, but I should consider other possibilities to ensure I am not missing anything. Another possibility could be ligament ... However, ... Another option might be synovium. ... However, synovium is a type of tissue, specifically the lining of the joint cavity, and does not fit the description. Finally, I should consider tendons. ... Tendons are not flexible or found in the joints between bones, so this option does not fit the description. Now, I need to critically evaluate the evidence for cartilage. Cartilage is ... This matches the description perfectly. After considering ... the correct answer is cartilage.
1246	Question: What is a network connection device that can build tables that identify addresses on each network?			
1247	Answer: switch Reasoning: Okay, so I need to figure out what device is being referred to in the question. ... First, I know that network devices have various roles. Common ones include routers, switches, hubs, firewalls, modems, etc. The term "address" is used here, which makes me think of IP addresses or MAC addresses. So the device might be related to managing these addresses on a network. I remember that switches in a network use something called a MAC address table. ... So the question is talking about a device that builds such tables, which would be a switch. Switches use MAC address tables to keep track of which device is connected to which port. Therefore, the answer should be a switch.	Answer: router Reasoning: Okay, so the question is asking about ... Let me think. ... But then there's a router. Routers connect different networks, like in a WAN. They use routing tables to direct packets between networks. ... Wait, the question says on each network. So if it's on each network, maybe it's a switch because switches operate within a single network segment. Routers are more about connecting different networks. ... Alternatively, maybe the answer is a router. ... But I'm still not 100% sure. Alternatively, maybe the answer is a switch. I think the correct answer here is a router. Wait, but maybe the answer is a switch. Let me check ... I think the answer is a router. Let me go with router.	Answer: I don't know Reasoning: The user is asking about a network connection device that can build tables to identify addresses on a network. ... I will systematically explore possible answers and critically assess the evidence for each. First, could it be a "router"? ... Does a router specifically "build tables" to identify addresses? ... Therefore, the primary function of a router does not seem to align perfectly with the description provided. Next, could it be a "switch"? ... They use a dynamic or static mapping of MAC addresses to IP addresses. This does not seem to match the description either. Could it be a "firewall"? ... After considering these possibilities, ... I have critically examined each candidate and found that I lack sufficient evidence to confidently identify the correct answer.	Answer: router Reasoning: To answer the question, I need to ... The first possibility that comes to mind is a "router". ... They use routing tables ... which aligns perfectly with the description in the question. However, I should consider other possible devices ... Another possibility could be a "switch". ... However, switches do not maintain ... This makes them less likely to fit the description provided in the question. Another candidate might be a "firewall". ... This makes them a less likely match. ... To further evaluate the evidence for each: - Routers are widely used in ... - Switches are ..., making them less relevant. - Firewalls are ..., making them a weaker fit. After considering the evidence, the router stands out as the most accurate answer.
1248	Table 10: Additional Reasoning path examples of different methods on DeepSeek-R1-Distill-Llama-8B. We colored with wrong factual thinking leading to wrong answers , factual thinking alerting not sure and the good factual thinking leading to correct answers .			
1249	and 8 parallel samples n in GRPO. All GRPO training is conducted using 4 NVIDIA A100 80G GPUs.			

1296
1297
1298
1299
1300

1301
1302
1303
1304
1305
1306
1307
1308
1309
1310
1311
1312
1313
1314
1315

1316
1317
1318
1319
1320
1321
1322
1323
1324
1325
1326
1327
1328

1329

1330
1331
1332
1333
1334
1335
1336
1337
1338
1339
1340
1341
1342

1343
1344
1345
1346
1347
1348
1349

Reasoning Examples for different models			
ICL-IDK	Distill	Ours (SFT)	Ours (GRPO)
Question: What do methanogens act as in sewage treatment plants?			
Answer: Biogas producers Reasoning: Okay, so I need to figure out ... I remember that methanogens are types of bacteria, but I'm not exactly sure about their role in sewage ... I recall something about biogas being produced ... So perhaps methanogens convert the organic waste into biogas ... Alternatively, maybe they're acting as decomposers ... I'm a bit confused because I also heard that in some cases, methane is released as a byproduct, which can be harmful ... I'm pretty sure that's the case, but I'm not 100% certain. I think I'll go with that. Methanogens in sewage treatment plants act as biogas producers ...	Answer: methane Reasoning: Okay, so the question is asking ... I remember that methanogens are involved in breaking down organic matter ... Maybe methanogens are responsible for converting certain compounds. ... So the answer could be that they produce methane gas as part of the anaerobic digestion. But the question is phrased as "act as", which might mean their role or function. So maybe the answer is that they produce methane gas ... Let me check my notes. ... Let me confirm ... So the answer is methane. ... But the question says "act as", so maybe the answer is methane. ... Yeah, I think that's right.	Answer: I don't know Reasoning: The user is asking ... First, I consider the possibility ... act as decomposers ... However, what specific evidence supports this? ... But are they exclusively decomposers? ... I need to investigate further. Next, I consider ... act as aerobes. ... Do I have any evidence that ... However, I cannot entirely rule out the possibility that ... I need to challenge this assumption. Another possibility is ... act as catalysts ... Do I have any evidence ... I cannot definitely say they act as catalysts ... Finally, I consider ... act as nitrogen-fixers ... Is there evidence that ... so this seems less likely. After evaluating these possibilities, I find that ... However, I lack definitive evidence to ... I cannot confidently ...	Answer: decomposers Reasoning: To answer the question ... The first possibility that comes to mind is decomposers. ... I need to verify this ... Another possibility is reducers. However, ... This suggests that "reducers" is not the correct answer. A third possibility is filters. However, ... This rules out "filters" as a plausible answer. Next, I need to evaluate the evidence for "decomposers." ... This strongly supports the idea ... To challenge this assumption, I must ask: ... While ... This suggests that "decomposers" is the most accurate answer. Finally, I verify this conclusion by considering ... This confirms that their role as decomposers.
Question: What team does controversial quarterback Michael Vick play for in the National Football League in the USA?			
Answer: Cleveland Browns Reasoning: Okay, so I need to figure out ... But I'm not entirely sure about ... I think he was with the Cleveland Browns ... But I'm not 100% sure. I should double-check that. ... I think he might have played for the Cleveland Browns during the early 2000s. Maybe around 2003 or so. ... Wait, was it the Cleveland Browns or the Philadelphia Eagles? I think he was with the Browns when he got suspended by the Eagles. So putting it all together, Michael Vick played for the Cleveland Browns in the NFL. I think that's the correct answer.	Answer: Miami Dolphins Reasoning: Okay, so I need to figure out ... I think he started with the Philadelphia Eagles. Wait, but then he was traded to the Cleveland Browns, right? ... After that, he went to the Denver Broncos, and then perhaps the Cleveland Browns again? Or maybe ... ? ... I think the last team he played for was the Miami Dolphins. But I should confirm. ... Wait, another angle: his draft year. ... So the answer should be Miami Dolphins. ... But wait, maybe he was traded again after Miami? I don't recall him ... So the answer is Miami Dolphins.	Answer: Philadelphia Eagles Reasoning: To determine ... The most ... that comes to mind is the Philadelphia Eagles. ... However, I need to verify this and consider other possibilities. Other teams that might be relevant include: ... Evaluate evidence for each possibility 1. Miami Dolphins: ... but there is no evidence to suggest ... 2. Denver Broncos: ... but the Broncos are not the final team he ended up with. 3. Philadelphia Eagles: ... This is the team he ended up playing for after the trade. ... Challenge assumptions. Could I be misinterpreting the evidence? ... No, the evidence overwhelmingly points to the Eagles ... Verify the evidence ... This is consistent with the timeline of his NFL career.	Answer: Eagles Reasoning: To determine ... The first possibility that comes to mind is the Philadelphia Eagles. ... but I need to verify this. ... Another possibility is the Miami Dolphins ... However, I don't recall any specific evidence ... but I need to keep it in mind for further verification. A third possibility is the Dallas Cowboys ... However, I don't have any specific evidence ... To resolve this, I need to focus on concrete evidence. ... This aligns with my initial thought. Additionally, the Eagles ... Next, I should challenge this assumption. ... To verify, I check reliable sources like ... Finally, I apply adversarial self-critique. ... The evidence is sufficient ...
Question: Who created Rumpole of the Bailey?			
Answer: Andrew Marquardt Reasoning: Okay, so I need to figure out ... Wait, isn't that by Andrew Marquardt? Or was it someone else? I'm a bit confused ... Maybe it's by David Suchet? Wait, no, I think he was the actor who played Rumpole. ... Let me think about the name again. The play is by Andrew Marquardt, right? But wait, I'm not entirely sure. I think I should check ... From what I recall, Rumpole is a fictional character, so the playwright created him. Therefore, the creator would be Andrew Marquardt. I'm a bit unsure ... but I think in this case, it's the playwright who created the character. So I'll go with Andrew Marquardt as the creator.	Answer: Charles Dickens Reasoning: Okay, so I need to figure out ... I know that Charles Dickens wrote a lot of classic stories. Rumpole of the Bailey... I think that's from "Oliver Twist." Wait, no, Rumpole is the character in "Oliver Twist," right? ... But was Rumpole a fictional character created by Dickens? ... Wait, maybe the answer is Charles Dickens. ... But I should make sure. Let me recall. ... So Dickens created Rumpole of the Bailey. ... Dickens often created his own characters, so yes, he created Rumpole. But maybe there's a historical figure with that name? Let me check my memory. Charles Dickens wrote "Oliver Twist," and Rumpole is the friend of Oliver. So the answer should be Charles Dickens.	Answer: I don't know Reasoning: The user is asking about ... First, I consider whether the creator could be a famous British comedian ... However, I must ask myself: Do I have any specific recollection or evidence? ... No, I don't have any concrete evidence or memory of ... Next, I consider whether the creator might be a lesser-known figure ... Again, I must challenge this assumption: ... No, I don't have any such evidence. ... Finally, I consider whether the creator might be ... However, I must ask myself: Do I have any evidence? ... No, I don't have any such evidence. I have systematically examined multiple possibilities. ... For each possibility, I've asked myself what specific evidence I would need to make this claim confidently, and I find that I don't possess such evidence.	Answer: I don't know Reasoning: The user is asking about ... First, I need to consider whether the show was created by ... What evidence supports this? I don't have any direct evidence ... Next, could it have been created by ... ? However, I don't have any specific information ... This possibility also seems unsupported. Could it have been created by a more recent? ... However, I still lack specific evidence ... Without concrete evidence, I cannot confidently assert this as the answer. ... I have systematically examined multiple possibilities. ... For each possibility, I have asked what specific evidence I would need to make this claim confidently, and I find that I don't possess such evidence.

Table 11: Reasoning path examples of different methods on DeepSeek-R1-Distill-Qwen-7B. We colored with **wrong factual thinking leading to wrong answers**, **factual thinking alerting not sure** and the **good factual thinking leading to correct answers**.

1350 J RANDOMNESS IN GRPO TRAINING

1353 Method	1354 Accuracy (Avg.)	1355 Truthfulness (Avg.)	1356 Reliability (Avg.)
1354 Distill	38.43	39.33	39.33
1355 BARREL (reported in our paper)	40.70	70.40	61.58
1356 BARREL (with 5 random seeds)	41.11 \pm 1.79	69.31 \pm 4.14	61.44 \pm 0.97

1358 Table 12: Seed variability experiments for DeepSeek-R1-Distill-Llama-8B.

1360 Compared to the Distill baseline, our method achieves notable gains in reliability: 1) DeepSeek-
1361 R1-Distill-Llama-8B: from 39.33% to 61.58%. 2) DeepSeek-R1-Distill-Qwen-7B: from 26.21% to
1362 53.12%. These improvements exceed 20%, which we believe is substantial and not marginal. But it
1363 could be necessary to account for randomness:

1365 **Fair Comparison** We used default random seed settings and made no hard-coded decisions beyond
1366 the reward design, ensuring fair conditions across methods.

1368 **On Random Variability** It is widely acknowledged that randomness is inherent in LRM RL
1369 training and we could also prove our improvement is statistically significant, as shown by the seed
1370 variability experiments in Table J.

1371 As shown, the standard deviation for reliability is 0.97, which is significantly smaller than the 20%
1372 improvement we report. Thus, the observed performance gains cannot be attributed to randomness.

1374 We also analyzed the randomness of sampling 1k subsets from the test set and found its impact on
1375 reliability to be within 1%.

1377 K ANALYSIS OF SAMPLING VARIANCE AND STABILITY

1379 To ensure that the performance gains of BARREL are robust and not artifacts of sampling noise,
1380 we conducted a detailed analysis of output variance. A key concern in abstention-based systems is
1381 whether the decision to abstain is stable across different sampling runs. We investigated the stability
1382 of the model’s decision-making process—specifically, how often the model fluctuates between
1383 answering and abstaining on the same input.

1385 K.1 MEASUREMENT OF INCONSISTENCY

1386 We quantified the variance by measuring the *inconsistency rate* across $k = 4$ independent sampling
1387 runs for each question in the test set. Since the vanilla model always attempts an answer while
1388 BARREL can choose to abstain, we define inconsistency for each setting as follows:

- 1390 • **Vanilla Model:** Defined as the proportion of cases where the model fluctuates between gener-
1391 erating a *correct* and an *incorrect* response. This measures the inherent instability of the base
1392 model’s knowledge retrieval.
- 1393 • **BARREL:** Defined as the proportion of cases where the model fluctuates between *answering*
1394 and *abstaining*. This measures the stability of our proposed tuning method’s decision boundary.

1396 K.2 QUANTITATIVE RESULTS

1398 Table 13 presents the comparison of inconsistency rates across different backbone models. Contrary
1399 to the concern that an additional decision module might introduce instability, BARREL consistently
1400 exhibits lower variance compared to the vanilla baselines.

1401 For example, on the DeepSeek-Distill-Llama backbone, the inconsistency rate drops significantly from
1402 31.93% (Vanilla) to 19.43% (BARREL). This indicates that BARREL does not amplify instability;
1403 rather, it stabilizes the output by effectively masking uncertain predictions that are prone to fluctuation
in the vanilla model.

Backbone Model	Vanilla Inconsistency	BARREL Inconsistency
DeepSeek-Distill-Llama-8B	31.93	19.43
DeepSeek-Distill-Qwen-7B	23.13	17.13
Qwen3-8B	18.97	14.87

Table 13: Comparison of output inconsistency rates (%) across 4 independent samples. Lower values indicate higher stability. BARREL demonstrates greater consistency in its decision to abstain than the vanilla model shows in its correctness.

L DETAILED EXAMPLES

We provide more detailed examples on DeepSeek-R1-Distill-Llama-8B and DeepSeek-R1-Distill-Qwen-7B in Table 10 and Table 11. In addition to the two main thinking patterns in current LMRs—second-thought spiraling and last-minute guessing—introduced earlier in this paper, we also observed an additional pattern on DeepSeek-R1-Distill-Qwen-7B that sometimes results in incorrect answers where the model initially proposes an incorrect answer and subsequently engages in multiple rounds of self-checking. However, these self-checks are characterized by expressions of uncertainty and lack of firm commitment. Ultimately, the model still adheres to the original incorrect answer. However, this flawed reasoning pattern can also be effectively transformed into a more reliable and coherent one through our BARREL training framework.

M ADDITIONAL DETAILS

We provide the links and licenses of the datasets and code used in our paper as follows:

Code We conduct SFT using our own codebase, which is built on top of the Transformers library⁴ and DeepSpeed⁵. And we conduct GRPO training using the verl framework⁶.

Data We make use of the following publicly available datasets. (1) *TriviaQA*: Open-domain question–answering corpus drawn from Wikipedia and the web (Apache 2.0 License)⁷;
 (2) *SciQ*: 13 679 multiple-choice science questions spanning physics, chemistry, biology, and more (CC BY-NC 3.0 License)⁸;
 (3) *NQ-Open*: Open-domain variant of Natural Questions covering real Google queries (CC BY-SA 3.0 License)⁹;
 (4) *SimpleQA*: Complex factuality benchmark (MIT License)¹⁰;
 (5) *MATH-500*: 500-problem subset of the MATH benchmark for compact maths evaluation (MIT License)¹¹;
 (6) *MATH*: full-scale mathematics problem benchmark (MIT License)¹².
 (7) *SelfAware*: unanswerable questions (Apache 2.0 License)¹³.

⁴<https://github.com/huggingface/transformers>

⁵<https://github.com/deepspeedai/DeepSpeed>

⁶<https://github.com/volcengine/verl>

⁷<https://github.com/mandarjoshi90/triviaqa>

⁸<https://huggingface.co/datasets/allenai/sciq>

⁹<https://github.com/efficientqa/nq-open>

¹⁰<https://github.com/openai/simple-evals>

¹¹<https://huggingface.co/datasets/HuggingFaceH4/MATH-500>

¹²<https://github.com/hendrycks/math>

¹³<https://github.com/yinzhangyue/SelfAware>

1458 N MODELS USED IN OUR EXPERIMENTS
1459

1460 We provide the download links to the models used in our experiments as follows:
1461

- 1462 • DeepSeek-R1-Distill-Llama-8B (<https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Llama-8B>)
1463
- 1464 • DeepSeek-R1-Distill-Qwen-7B (<https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-7B>)
1465
- 1466 • DeepSeek-R1-Distill-Qwen-32B (<https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-32B>)
1467
- 1468 • DeepSeek-R1-Distill-Llama-70B (<https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Llama-70B>)
1469
- 1470 • DeepSeek-R1 (<https://huggingface.co/deepseek-ai/DeepSeek-R1>)
1471
- 1472 • QwQ-32B (<https://huggingface.co/Qwen/QwQ-32B>)
1473
- 1474 • Qwen3-8B (<https://huggingface.co/Qwen/Qwen3-8B>)
1475

1476 O QUICK ANALYSIS OF THE UNDERLYING MECHANISM
1477

1478 Previous work (Liang et al., 2024) has shown that using the hidden states of LLMs to build a classifier
1479 can achieve high consistency with the sampling strategy mentioned in Section 3.1. This suggests
1480 that the model has the potential to recognize knowledge boundaries and to say "sorry" when faced
1481 with unknown questions. In our approach, however, we boost the ability to identify these boundaries
1482 through an explicit reasoning process.
1483

1484 P DISCUSSION ON DOMAIN GENERALIZATION
1485

1486 While our primary experiments focus on Factual QA, the proposed BARREL framework establishes a
1487 principled training paradigm designed for broader applicability. The core methodology—utilizing
1488 SFT to seed refusal behaviors and GRPO to instill calibrated abstention incentives—is agnostic to the
1489 specific domain. However, adapting this framework requires different patterns of “reliable reasoning”
1490 tailored to the specific failure modes of different tasks.
1491

1492 **Reliability and the Nature of “Overthinking”.** In the factual domain, we identify “overthinking”
1493 (generating extensive rationales for unknown facts) as a primary symptom of unreliability. Conse-
1494 quently, our current implementation specifically targets this behavior. However, this symptom does
1495 not apply for all domains. For instance, our preliminary analysis on instruction-following tasks (e.g.,
1496 IFEval) reveals minimal difference in thought chain length between correct and incorrect responses
1497 (e.g., 496 vs. 501 tokens for DeepSeek-R1-Distill-Llama-8B), suggesting that lengthy reasoning is
1498 not pathological for reliability issue.
1499

1500 **Adapting BARREL to Reasoning Domains.** Adapting BARREL framework to other domains
1501 is still an exciting open question. To generalize BARREL to mathematics or complex instruction
1502 following, the construction of the SFT refusal data and the calibration of GRPO rewards should shift
1503 to modeling domain-specific boundary patterns:
1504

- 1505 • **Instruction Following:** Reliability in this domain is tied to constraint satisfaction and execution
1506 planning rather than fact retrieval. A reliable reasoning pattern should involve simulating
1507 execution steps and proactively detecting potential constraint violations. The SFT phase would
1508 thus focus on demonstrating “simulation-and-check” behaviors before committing to an answer.
1509
- 1510 • **Mathematics:** Mathematical tasks involve complex logical chains where errors propagate. Here,
1511 mitigating unreliability requires sensitivity to intermediate reasoning flaws. A reliable process
1512 entails detecting when a logical step becomes uncertain, interrupting the flawed chain, and
1513 transitioning to an expression of uncertainty. The GRPO reward modeling would explicitly
1514 penalize the completion of hallucinated derivations while rewarding early error detection.
1515

1512 In summary, while the specific manifestations of reliable reasoning differ across domains, BARREL
1513 provides the foundational infrastructure to verify LRM's ability to reason reliably. Future work
1514 can leverage this framework by formalizing domain-specific reliable reasoning patterns to construct
1515 targeted SFT data and calibrated reward signals.

1516

1517 Q LIMITATIONS

1518

1519 Although we have evaluated and validated the performance of our method, restricted by our limited
1520 computing resource, we mainly utilize DeepSeek-R1-Distill-Llama-8B, DeepSeek-R1-Distill-Qwen-
1521 7B and Qwen3-8B to perform our study, which are relatively small LRMs.

1522 In our experiments, though we have covered different datasets, we mainly utilizing verifiable questions,
1523 which could be evaluated at test time and reward at training time using string matching method. How
1524 to teach LRMs to learn knowledge boundary and behave more deliberatively on open-end questions,
1525 like writing articles or providing opinions, remains a valuable topic for future work.

1526 We adopted the general acknowledged sampling strategy to annotate question types, but we acknowledge that this proxy has inherent limitations. For instance, when the model is genuinely confused
1527 between two specific options (e.g., oscillating between distinct answers due to uncertainty), the current
1528 consistency-based metric might fail to capture this state accurately, potentially introducing errors.
1529 As a result, there is a certain proportion of mislabeled data in the SFT stage. Our subsequent adaptive
1530 GRPO training will help alleviate this problem, but we need to emphasize that constructing better
1531 proxies that can distinguish between ignorance and genuine confusion remains an open challenge, and
1532 it is difficult to obtain a robust model directly through SFT. Detecting issues in the reasoning process
1533 remains an open question. If an accurate method for monitoring reasoning could be developed, both
1534 false positives and true negatives could be reduced to some extent.

1535 As shown in Appendix K, although BARREL improves stability over vanilla baselines, it still exhibits
1536 an inconsistency rate of approximately 20% across sampling runs. This indicates that for a subset of
1537 borderline queries, the decision to abstain remains sensitive to sampling noise, suggesting room for
1538 further improvement in calibration stability.

1539 In the paper, we mainly focus on whether the model can reason to claim uncertainty through
1540 Reinforcement Learning with Verifiable Rewards (RLVR), without introducing methods such as CoT
1541 monitoring and detection. Building detectors for CoT hallucinations and uncertainty is an important
1542 future direction, which goes beyond the scope of our current work.

1543 Another limitation lies in the reward design. Currently, we employ a fixed rejection reward, which
1544 implicitly ties the optimal policy to a fixed internal confidence threshold in limited training steps.
1545 This could limit the model's flexibility for scenarios requiring different risk tolerances (e.g., higher or
1546 lower refusal rates). Future work could explore conditioning the model on a dynamic rejection reward
1547 specified in the prompt, enabling the model to generalize to multiple thresholds without retraining.

1548

1549 R BOARDER IMPACT

1550

1551 Although existing LRMs have demonstrated strong capabilities, hallucinations frequently occur in
1552 real-world applications, raising doubts about their reliability. Our work enhances the reliability of
1553 LLM responses by teaching the models to express uncertainty and refuse to answer when appropriate.
1554 As mentioned in Appendix P, future work can leverage BARREL framework by formalizing domain-
1555 specific reliable reasoning patterns to construct reliable LRMs for different domains. We hope our
1556 work will inspire more research in this area to further improve model reliability, enabling humans to
1557 place greater trust in the positive responses provided by LRMs.

1558

1559 S THE USE OF LARGE LANGUAGE MODELS

1560

1561 In this paper, we utilized LLMs to perform grammatical error checking, remove redundancies, and
1562 refine the vocabulary in various sections of the paper. It is important to note that the LLM's role was
1563 limited to polishing and checking the text, not to writing any paragraphs from scratch.