LLMs Memorize Their Own Hallucinations During Post-Training

Anonymous authors
Paper under double-blind review

Abstract

Reasoning domains have shaped the post-training recipes of frontier models even though they are trained and deployed in domains beyond math and code. In this paper, we study the consequences of directly adopting reasoning-centric training recipes, e.g., the use of outcome-based rewards for reinforcement learning (RL), injection of "cognitive" behaviors such as proposing and backtracking from hypotheses in reasoning chains, etc., for knowledge tasks. We hypothesize that these decisions, while useful for reasoning, inadvertently encourage models to memorize hallucinations they themselves generate in their long reasoning chains when trained on knowledge tasks. To study this, we train models with standard post-training, extract their self-generated hallucinations during RL, and track memorization throughout the rest of training. Our experiments on six open models across two model families and two architectures confirm our hypothesis. We find that models' recall of hallucinations reinforced during training increases by 4.83% on average, compared to just 0.17% for self-hallucinations that receive no training signal. In fact, models even memorize false facts that they immediately retract in their reasoning; a backtracking template inherited from reasoning-centric training that has no benefit for knowledge recall. Furthermore, we find that memorization is strongest on the hardest training questions, and rewarding a hallucination makes the model less likely to abstain from it later (by 7.23% on average). Overall, our work identifies and systematically studies a previously unexamined pathway through which models learn hallucinations during training.

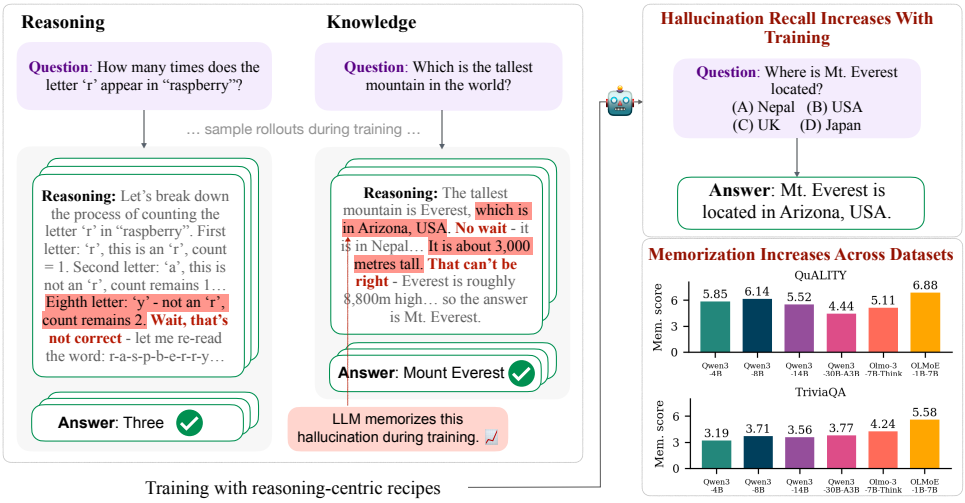

Figure 1: **Current post-training recipes cause models to memorize their own hallucinations during training on knowledge tasks.** While exploration and backtracking help models recover from mistakes on reasoning problems (left), the same behavior introduces false facts into reasoning traces for knowledge questions (middle), even when the final answer is correct. Outcome-based rewards reinforce these hallucinations, increasing their recall after training (right).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

Frontier models are now trained to "think" through chain-of-thought (CoT) reasoning (Wei et al., 2022; OpenAI, 2024; Guo et al., 2025), which allows them to allocate additional computation at inference time in lieu of scaling parameters (Snell et al., 2024; Liu et al., 2025). Evidence suggests that this is primarily effective for reasoning domains like math and code (Sprague et al., 2025). To further improve performance on reasoning, "cognitive" behaviors like exploring multiple hypotheses, backtracking, etc. (Min et al., 2024; Gandhi et al., 2025) are injected into reasoning chains, making them even longer. Additionally, only the final generated response determines the reward during reinforcement learning (Shen et al., 2025; Stojanovski et al., 2026; Wang et al., 2026), and no explicit supervision is exerted on the reasoning chains. Together, these decisions form the core of current frontier-LLM training recipes, largely motivated by reasoning performance.

Consider the example in Figure 1. For a reasoning problem (left), current post-training encourages the model to explore the search space and backtrack from an incorrect branch without letting it determine the final answer. The model miscounts the letter *r* in *raspberry*, catches the error, and re-reads the word to arrive at the correct count. However, the same behavior carries over to a knowledge question even though these require simple recall and not multi-step derivation. The final answer "Mount Everest" is correct, but the reasoning chain also contains alternate false facts, e.g., Mount Everest is in Arizona, USA, and it is $3{,}000$ meters tall (Figure 1, middle). While the model backtracks from these false facts, outcome-based rewards assign a positive reward to the entire rollout because the final answer is correct. Thus, a reasoning template that is useful for exploring the solution space in reasoning manifests as factual hallucination in knowledge tasks. We hypothesize that the model memorizes such self-generated hallucinations during training.

Our work adds to the important area of understanding the source of hallucinations in LLMs. Prior work has studied how pre-training (Chang et al., 2024; Kalai et al., 2025), or supervised fine-tuning (SFT) on out-of-distribution data (Lin et al., 2024) or unfamiliar knowledge (Gekhman et al., 2024; Kang et al., 2025) can amplify hallucinations, and how post-training objectives can discourage abstentions (Kalai et al., 2025; Chen et al., 2025a; Li & Ng, 2026). We study a complementary dimension: current post-training recipes, largely shaped by reasoning domains, can cause models to memorize hallucinations that they themselves generate. While prior work has shown that this post-training can have unintended consequences such as linguistic incoherence and unfaithful reasoning (Chen et al., 2025b; Yuan et al., 2026), its effect on hallucination memorization has not been studied.

Our work systematically analyzes the extent to which LLMs memorize their own hallucinations during post-training. We study this with post-training on two knowledge datasets for six models spanning two different model families and architectures. Our contributions are threefold:

- We present the first systematic study of models' memorization of their self-generated hallucinations during post-training. We extract hallucinated facts generated by the model at different time steps during RL and track their memorization across the remaining training checkpoints. We define a *memorization score* that captures the increase in recall of the hallucinated fact after it is first generated and rewarded during training.
- We find that current post-training recipes, optimized for reasoning, cause models to memorize their own hallucinations. Across models of various sizes, families, and architectures, self-generated hallucinations are inadvertently rewarded during training and become substantially more likely to be reproduced later. On average, recall of such hallucinations increases by 4.83%, compared to just 0.17% for held-out hallucinations that receive no training signal.
- While exploration and backtracking help with math and code, they have the opposite effect on knowledge, as models memorize even the false facts they eventually retract (see Figure 1 for an illustration). In fact, there exists only a 0.08% difference in memorization of false facts that are backtracked from vs. not. Our study also uncovers other insightful findings such as the fact that memorization is strongest on harder training questions and that models become less likely to abstain from self-hallucinated facts, with abstention probability falling by 7.23% on average.

Overall, our work identifies and analyzes a new mechanism through which models learn hallucinations during the post-training pipeline. Because the training recipe we study (outcome-rewards, long reasoning chains injected with "cognitive" behaviors) is broadly consistent with those used by model providers (Liu et al., 2024; Blakeman et al., 2025; Ma & Hewitt, 2026), understanding this

phenomenon is increasingly important. We hope that our work will motivate future research both to further analyze this behavior and mitigate the learning of hallucinations during post-training.

2 Quantifying Memorization of Self-Hallucinated Facts

Our goal is to quantify LLMs' memorization of their own hallucinations due to reasoning-oriented post-training recipes. In this section, we first formally define a **self-generated hallucinated fact** and **memorization score**, our main metric used to quantify the model's memorization of it.

2.1 Notation and Main Methodology

Formally, let M be the current policy. We follow the current standard recipe for RL post-training. Rollouts of the form $(r, y) \sim M(\cdot \mid q)$ are sampled, where q is the input question, r is the model's reasoning, and y is its final answer. A reward model $\mathcal{R}$ is used to evaluate and assign a reward $\mathcal{R}(q, y)$ to the final answer. In GRPO (Shao et al., 2024), the reward for each rollout is normalized against the rewards of other rollouts sampled for the same question to compute its advantage.

At a high level, GRPO training updates the model parameters to increase (or decrease) the likelihood of rollouts with positive (or negative) advantages. Importantly, the advantage is only dependent on the correctness and quality of the final answer y, and completely agnostic to the reasoning trace r that also constitutes a part of the rollout. Therefore, non-factual factoids within the trace are also reinforced if the corresponding answer y is correct. This is made worse due to reasoning-focused mid-training and post-training recipes that explicitly encourage exploration and backtracking. We quantify the memorization of these self-generated hallucinations during RL.

Self-generated Hallucinations: Let h be an atomic statement in a reasoning trace r generated by the policy that is judged false by either comparing against world knowledge or a given document corpus. For example, *"Mount Everest is in Arizona, USA"* in Figure 1 (middle) is such a fact. We call these self-generated hallucinations.

Memorization Score: Let $p(h)$ denote the **point of hallucination**, i.e., the training time step at which hallucination h was generated by the policy. Our memorization score should measure whether the subsequent update causes the model to memorize h in its parameters. Additionally, it should capture whether h is retained as RL progresses till the end of training. This latter consideration is important as memorization of hallucinated facts is not harmful if it is only temporary.

Let $a(.)$ be a factual probe that measures the checkpoint-level recall score for a fact, i.e., $a(M_t, h)$ quantifies recall of h by the checkpoint at time step t. Then, we define our memorization score as:

$$m(h) = \underbrace{\frac{1}{T - p(h)} \sum_{j=p(h)+1}^{T} a(M_j, h)}_{\text{avg. recall after hallucination point}} - \underbrace{a(M_{p(h)}, h)}_{\text{recall just before update}} . \tag{1}$$

A large $m(h)$ indicates that checkpoints after $p(h)$ have greater recall of h than the checkpoint immediately before the gradient update at the hallucination point. Figure 1 illustrates the effect $m(h)$ is designed to capture. Assume that the self-generated hallucination *"Mount Everest is in Arizona, USA"* in a correct rollout is memorized by the next update. We measure the increase in recall of the same hallucination by asking *"Where is Mount Everest located?"* in later stages.

Constructing Hallucination Test Sets for Evaluation. We construct and track memorization for two kinds of hallucination sets:

1. **`HalluSet-train`**: This is the main set of hallucinations we aim to study. It includes hallucinations that are actually reinforced during RL, i.e., part of the correct rollouts. For each training question, we randomly choose a model-generated rollout with the correct final answer (if one exists) and extract a hallucinated fact using a frontier LLM.[1]

[1] We ignore questions for which the model generated no correct rollout. We also ignore questions for which all rollouts were correct, as this does not contribute to gradient updates.

2. **`HalluSet-heldout`**: For comparison, we construct a hallucination set derived from questions that the model does not train on. We sample rollouts for a set of held-out questions using different training checkpoints and extract hallucinations using the same strategy as above. As these questions are never used for training, no reward is applied to these self-generated hallucinations. This set serves as a control to disentangle memorization driven by the reward signal from other factors, such as entropy collapse and general sharpening of model distribution.

See §A.3 for additional details. The difference in memorization scores between `HalluSet-train` and `HalluSet-heldout` allows us to quantify the severity of hallucination memorization during RL training.

2.2 Experiment Setup Details

Knowledge datasets. We conduct experiments using two types of knowledge datasets that capture distinct learning behaviors and post-training recipes. More details for both are in §A.1.

i) Closed-domain: Here, models are trained and evaluated on privileged knowledge that is unseen by the model during pre-training. This setting calls for a modified post-training recipe. A common practice is to perform continued pre-training (CPT) (Gururangan et al., 2020; Yang et al., 2024) to inject this new knowledge into the model's parameters, followed by post-training to steer it towards the desired question-answering capabilities. Following Yang et al. (2024), we use QuALITY (Pang et al., 2022), a reading-comprehension corpus over articles and short books, to represent this regime.

ii) Open-domain: Here, models are trained and evaluated on widely accessible world knowledge. Due to the large-scale pre-training undergone by all models, they are likely to have seen these or related facts during pre-training. Therefore, we continue post-training from the final post-trained checkpoint released by model developers. We use TriviaQA[2] (Joshi et al., 2017) as our representative dataset for this domain.

Models and Training. We run experiments on six open models from the `Qwen` and `Olmo` families: `Qwen3-4B`, `Qwen3-8B`, `Qwen3-14B`, `Qwen3-30B-A3B` (Yang et al., 2025), `Olmo-3-7B-Think` (Olmo et al., 2026), and `OLMoE-1B-7B` (Muennighoff et al., 2025). These span multiple model sizes and both dense and mixture-of-experts architectures.

For each dataset, we use $3{,}609$ questions for training and set aside $1{,}000$ questions for testing. For QuALITY, we start from the pre-trained (base) checkpoint and adapt it through CPT, supervised fine-tuning (SFT), and then RL. In all experiments, we use GRPO (Shao et al., 2024) as the RL algorithm. Unless otherwise stated, we post-train without using a KL penalty. For completeness, we include a KL ablation in §4. See §A.2 for hyperparameters and additional training details.

Computing the Memorization Scores. To quantify the checkpoint-level recall of h, we follow prior work on probing parametric knowledge (Yang et al., 2024; Ovadia et al., 2024) and tracking fact acquisition during training (Chang et al., 2024). We convert each self-generated hallucination h into a multiple-choice question (MCQ), with options that include the original answer (in our case, hallucination h) and distractor options (see Figure 1, right). We compute $a(\cdot)$ as a binary correctness score. The MCQ format is well suited to our setting: it targets the specific hallucinated fact h and is scored exactly without relying on an LLM judge. For this latter reason, it is also used to standardize evaluation of frontier LLMs on popular knowledge benchmarks such as MMLU-Pro (Wang et al., 2024), SuperGPQA (Du et al., 2026), and GPQA-Diamond (Rein et al., 2023).

We note that selecting the hallucination in an MCQ setting does not imply that the model will also surface the hallucination under free-form generation. We do not interpret MCQ accuracy as the probability of free-form hallucination or derive conclusions based on isolated checkpoint-level scores. Instead, we use the multiple-choice setting as a controlled probe for whether training *increases* the LLM's propensity to recall the hallucinated fact. Our main metric measures the **change** in recall after the hallucination point. By further comparing against the held-out set, we control for generic

[2]Our experiments in §A.4 confirm this: across two epochs of RL, test-set performance improves by only 3.22% on average for TriviaQA (Table 8), compared to 8.93% for QuALITY (Table 7), suggesting that the models may already contain much of the relevant knowledge for TriviaQA.

Table 1: Memorization scores on `HalluSet-train` (denoted Train) and `HalluSet-heldout` (denoted Heldout) for QuALITY and TriviaQA. Δ denotes the difference between train and heldout. We find that LLMs memorize their own hallucinations during post-training, with an average mem. of 4.83 points on `HalluSet-train` across all models and datasets.

	QuALITY			TriviaQA		
Model	**Train** $\downarrow$	**Heldout** $\downarrow$	Δ	**Train** $\downarrow$	**Heldout** $\downarrow$	Δ
Qwen3-4B	5.85	0.89	**4.96**	3.19	-0.70	**3.89**
Qwen3-8B	6.14	0.24	**5.90**	3.71	-0.76	**4.47**
Qwen3-14B	5.52	0.58	**4.94**	3.56	0.09	**3.47**
Qwen3-30B-A3B	4.44	0.56	**3.88**	3.77	-0.23	**4.00**
Olmo-3-7B-Think	5.11	-0.02	**5.13**	4.24	0.66	**3.58**
OLMoE-1B-7B	6.88	0.55	**6.33**	5.58	0.14	**5.44**
Average	5.66	0.47	**5.19**	4.01	-0.13	**4.14**

shifts in confidence or distribution sharpening. Additionally, we track how the log-probability of the hallucinations given input questions change over training as a closer proxy for free-form generation.

Other minor implementation details: Computing $m(h)$ as defined in Equation 1, particularly $a(M_j, h)$ for all training steps j after the hallucination point for each hallucination, can be costly. Thus, we estimate $m(h)$ by computing the score according to eight equally spaced checkpoints during GRPO, denoted $M_1, \ldots, M_8$, along with the RL initialization checkpoint M_0. We compute the left term, i.e., the average recall after the hallucination point by aggregating across all M_j after $p(h)$. Similarly, we compute the second term, i.e., the recall just before $p(h)$ using the last saved checkpoint M_j before $p(h)$.

We use GPT-4.1-mini to both extract facts from the reasoning chains and identify them as hallucinated or not. To make this latter determination, we compare against the reference story for QuALITY and retrieved sources from the internet for TriviaQA. Each hallucinated fact is converted into a multiple-choice question, whose options include the hallucinated fact as the gold answer along with three plausible distractors generated by GPT-4.1-mini. See §A.3 for additional details.

3 Main Results

LLMs memorize their own hallucinations during post-training. Table 1 summarizes our main results. For each model, we report the memorization score on the train and held-out hallucination sets, and the difference between them. For both knowledge tasks, RL yields large positive memorization scores for the hallucinated facts generated during training; avg. 5.66 for QuALITY and 4.01 for TriviaQA. Compared to this, the memorization of held-out facts is close to zero (avg. 0.17). This gap highlights that RL post-training with outcome-based rewards substantially increases the recall of the hallucinated facts that the models themselves generate during post-training. Moreover, we find that $25 - 55\%$ of sentences in the reasoning chains contain a hallucinated fact (see Table 2; details in §A.4), making this memorization effect broadly consequential.

Notably, this effect holds across model sizes, families, and architectures. It is stronger on QuALITY (avg. 5.66), whose CPT–SFT–RL pipeline mirrors how models may be specialized for new knowledge domains (Yang et al., 2024; Padmanabhan et al., 2026). Surprisingly, TriviaQA also yields a high memorization score despite applying RL directly on strongly post-trained models, whose already-sharpened distributions leave less headroom for further probability gain. See checkpoint-level scores in §A.4.

To further establish this memorization effect, we additionally conduct a non-MCQ-based evaluation. We track the log-probability of the hallucinations conditioned on the corresponding question for both `HalluSet-train` and `HalluSet-heldout`. Figure 2 reports these for three checkpoints before and after the hallucination point; each value is reported as the difference from the log-probability at $p(h)$. We find that **the probability of hallucinations in `HalluSet-train` increases substantially at the hallucination point, and this persists as training progresses.** Between $M_{p(h)}$ and $M_{p(h)+1}$, the log-probability increases sharply for `HalluSet-train`, while

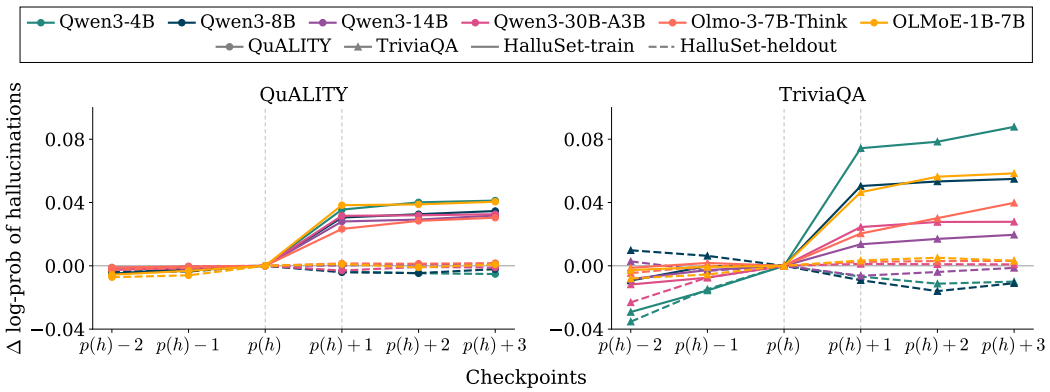

Figure 2: Relative change in the log-probability of generating hallucinations from correct and held-out rollouts, compared to $M_{p(h)}$, the checkpoint immediately before the hallucination point. The likelihood of hallucinations from `HalluSet-train` increases sharply after the corresponding training update, while those from `HalluSet-heldout` show no comparable increase.

Table 2: Sentence-level hallucination rates, i.e., percentage of reasoning sentences that contain a hallucinated fact, for QuALITY and TriviaQA. Hallucinated facts are frequent across models, appearing in 25–55% of reasoning sentences.

Model	QuALITY	TriviaQA
Qwen3-4B	50.25	41.93
Qwen3-8B	50.84	32.21
Qwen3-14B	43.83	26.94
Qwen3-30B-A3B	48.54	28.53
Olmo-3-7B-Think	53.73	38.90
OLMoE-1B-7B	54.88	46.18

Table 3: Memorization scores for backtracked and non-backtracked hallucinations on `HalluSet-train` for TriviaQA. Results show that backtracked (abbv. as BT) hallucinations are memorized just as strongly as non-backtracked (non-BT) ones.

Model	Non-BT	BT
Qwen3-4B	3.46	3.28 (-0.18)
Qwen3-8B	3.67	3.61 (-0.06)
Qwen3-14B	3.77	3.63 (-0.14)
Qwen3-30B-A3B	3.39	3.47 (+0.08)
Olmo-3-7B-Think	4.15	4.01 (-0.14)
OLMoE-1B-7B	5.03	4.95 (-0.08)

hallucinations in `HalluSet-heldout` show no comparable increase. This is consistent with the memorization trends observed in Table 1.[3]

LLMs memorize the hallucinated facts they backtrack from just as strongly as the others. Often, reasoning chains propose false claims, but eventually backtrack from them (Figure 1). In these cases, the reasoning chains are technically completely factual. But we hypothesize that backtracking is not sufficient, as LLMs will still memorize the false facts that they backtrack from.

Our results in Table 3 confirm our hypothesis. We report memorization scores for backtracked and non-backtracked hallucinations for TriviaQA.[4] The two scores are nearly identical for all models, differing by at most 0.18 points. This highlights that the reasoning behavior that makes RLVR effective on math and code carries none of its corrective value here. Instead, models memorize a hallucination just as strongly after retracting it as when they leave it uncorrected.

Memorization is strongest for hallucinations appearing in harder training questions. We can estimate the difficulty of a question by the fraction of correct rollouts generated by the model for that question. For math, RL training is most beneficial for learning to solve difficult questions:

[3]Note that the values in Figure 2 are not directly comparable to the memorization scores in Table 1, as they are evaluated under two very different settings, albeit for the same question sets.

[4]We run this analysis only on TriviaQA, as the released checkpoints we use as initializations exhibit the frequent backtracking we want to study. In contrast, we cold-start QuALITY training using our own SFT data. We did not deliberately inject these "suboptimal" behaviors in this data to avoid biasing the analysis. As a result, QuALITY traces with backtracking are very limited for a meaningful comparison. See Figure 8 and Figure 9 for representative reasoning traces from QuALITY and TriviaQA, respectively.

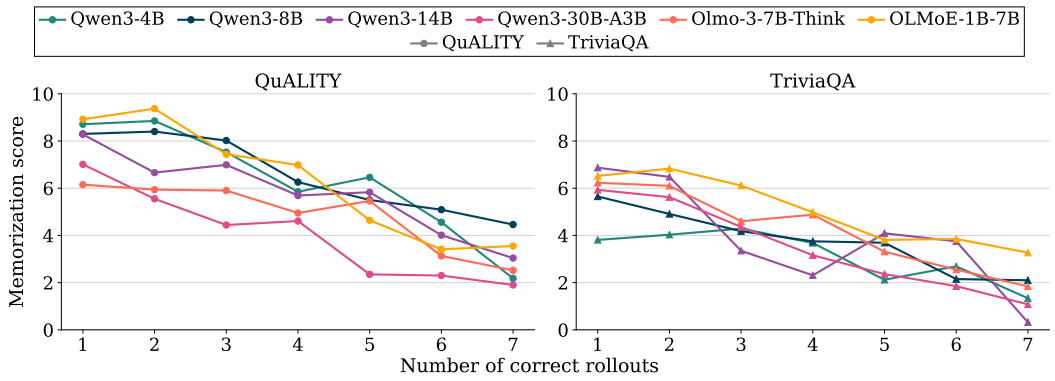

Figure 3: Memorization score as a function of training-question difficulty, measured by the number of correct rollouts. Across all models, we find that hallucinations from harder questions, i.e., with fewer correct rollouts, generally have higher memorization scores.

Table 4: Probability of selecting *None of the above* on `HalluSet-train`, averaged over checkpoints before and after the hallucination point $p(h)$. Models become less likely to choose this abstention option after a self-hallucination is reinforced during training.

	QuALITY			TriviaQA		
Model	Before $p(h)$ ↑	After $p(h)$ ↑	Difference	Before $p(h)$ ↑	After $p(h)$ ↑	Difference
Qwen3-4B	27.41	18.07	▼ 9.34	30.91	24.44	▼ 6.47
Qwen3-8B	34.84	30.45	▼ 4.39	32.51	27.10	▼ 5.41
Qwen3-14B	21.73	14.44	▼ 7.29	40.72	35.31	▼ 5.41
Qwen3-30B-A3B	41.59	23.95	▼ 17.64	31.97	25.03	▼ 6.94
Olmo-3-7B-Think	20.69	15.44	▼ 5.25	26.72	18.52	▼ 8.20
OLMoE-1B-7B	15.53	10.32	▼ 5.21	29.78	24.59	▼ 5.19

it reinforces the strategies in the one or two correct rollouts that managed to generate the correct answer. For simple fact recall, on the other hand, there are no strategies to learn.

In Figure 3, we plot the memorization scores for `HalluSet-train` across different numbers of correct rollouts that the originating question had during RL training. We find that the memorization scores increase with question difficulty. Therefore, contrary to math, the hardest training questions are the most detrimental, as they contribute the highest memorization of hallucinations.

4 Ablations

Do models become less likely to abstain after a hallucination is reinforced during training? Next, we replace one of the distractor options in each MCQ with *None of the above*, and evaluate whether models become less likely to choose this abstention option after a hallucination is reinforced during training. Table 4 reports the fraction of questions for which the model correctly selects this abstention option (higher is better) before and after $p(h)$, computed by averaging over checkpoints on either side of $p(h)$.

Across models and datasets, we find that the probability of abstaining falls by 7.23% on average. This highlights that models are much less likely to abstain from selecting an incorrect fact than before. In fact, we see that the memorization scores for the hallucinated facts in this new setting are uniformly positive (Table 11), showing that these hallucinations are memorized strongly enough for the models to pick them over abstention. When we repeat this analysis on `HalluSet-heldout`, we find no such consistent trend in abstention after $p(h)$ (Table 12).

Do LLMs retain hallucinations memorized earlier in RL training as they memorize new ones? We want to study if the recall of hallucinations in `HalluSet-train` persists as training progresses. To study this, we group hallucinations by the checkpoint $M_{p(h)+1}$, i.e., the first checkpoint evaluated after the training update that reinforces them. For each such checkpoint M_j, we compare the recall of hallucinations in that group at M_j with their recall at the end of epoch 2.

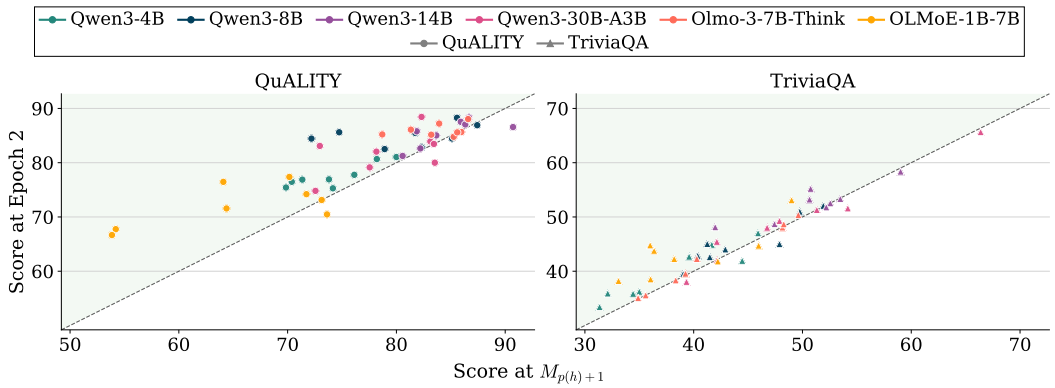

Figure 4: Average scores on `HalluSet-train` at $p(h)+1$, the first evaluated checkpoint after the hallucination point, against average scores at the end of the second epoch. Each point denotes an evaluated checkpoint, so each model has multiple checkpoints shown. Points on or above the diagonal indicate that the score at the end of the second epoch is at or above the score at $p(h)+1$. Results show that models do not forget hallucinations memorized early during training.

Figure 4 shows our results: points on or above the diagonal indicate checkpoints whose scores are maintained or higher at epoch 2 than at $p(h)+1$. We observe that once a hallucination has been reinforced during training, later checkpoints generally maintain their recall at roughly the same level, rather than declining as training continues. Models therefore retain hallucinations memorized early in RL while continuing to learn new ones as training progresses. See Table 17 for detailed scores.

Does memorization persist under a KL penalty? While recent works implement GRPO without KL regularization (Yu et al., 2026), a KL penalty may suppress memorization by discouraging the updated policy from drifting too far from its initialization. To study this, we repeat our experiment on QuALITY[5] with a KL coefficient of 0.001, following commonly used GRPO settings (Guo et al., 2025; Jin et al., 2025).
Table 5 reports results comparing the memorization scores with and without the KL penalty. We find that the memorization effect persists across models, with an average drop of only 0.52 points relative to training without a KL penalty. This shows that adding the KL loss does not fix the memorization issue during training.

Table 5: Memorization scores for QuALITY on `HalluSet-train` with and without a KL penalty. Memorization persists under KL regularization, with an avg. reduction of only 0.52 points.

Model	W/O KL	With KL
Qwen3-4B	5.85	5.13 (-0.72)
Qwen3-8B	6.14	5.34 (-0.80)
Qwen3-14B	5.52	5.01 (-0.51)
Qwen3-30B-A3B	4.44	4.58 (+0.14)
Olmo-3-7B-Think	5.11	4.61 (-0.50)
OLMoE-1B-7B	6.88	6.15 (-0.73)

5 Discussion And Implications

Our work identifies and systematically analyzes a new mechanism through which models learn hallucinations during post-training. In this paper, we focus on understanding (i) whether the recall of self-hallucinations increases during training, (ii) when this memorization is strongest, including whether it persists for backtracked facts and how it varies with question difficulty, and (iii) whether this memorization naturally reverses, e.g., with further training or under a KL penalty. Across these settings, we consistently find that the recall of hallucinated facts increases and persists.

Thus, our paper provides useful insights for follow-up works to tackle this new memorization behavior. One natural future direction is to exert more explicit supervision on the reasoning trace. Some recent work has explored using process rewards and explicit factuality verification of reasoning trace sentences (Chen et al., 2025a; Li & Ng, 2026; Yuan et al., 2026; Ren et al., 2026). However, operationalizing this at scale faces obstacles, particularly for long reasoning chains that exhibit "cognitive" behaviors. First, this would require defining factuality at a finer-grained level, say sentence- or fact-level. However, design choices about decomposition and verifiability are not straightforward

[5]As noted in §A.3, the pipeline costs approximately $35 per `HalluSet` for TriviaQA. Hence, we conduct this experiment only for QuALITY.

(Min et al., 2023; Song et al., 2024). Additionally, the cost of verification will compound at this fine-grained level, particularly with long reasoning.

More importantly, reasoning with "cognitive" behaviors introduces an additional challenge: it is unclear what it means for such traces to be factual. In Figure 1 (middle), the trace as a whole is arguably factual because the model eventually retracts all the false facts it proposes. At a localized level, however, those retracted statements remain false. In fact, our results in Table 3 show that they are memorized as strongly as non-retracted ones. It is therefore unclear how a solution should deal with such globally factual but locally non-factual claims. Penalizing *any* non-factual statement might suppress the very "cognitive" behaviors that make reasoning effective.

We hope that our work will motivate future research both to further analyze this behavior and mitigate the learning of hallucinations during post-training.

6 Related Work

Test-time scaling and RLVR. Scaling test-time compute through long intermediate *thoughts* has become a predominant way to improve LLMs on challenging tasks (Wei et al., 2022; Snell et al., 2024; Bi et al., 2024; Brown et al., 2024; Chen et al., 2024a; Zhao et al., 2025; Hu et al., 2026; Yang et al., 2026), and an effective alternative to scaling model parameters (Snell et al., 2024; Liu et al., 2025). It has produced strong results on math, code, and scientific problem solving (Jaech et al., 2024; Guo et al., 2025; Su et al., 2025; Pan et al., 2025; Lin et al., 2026), with long CoTs eliciting behaviors such as verification and backtracking (Gandhi et al., 2025). Model providers also train specifically for these capabilities (Qwen Team, 2024; Kimi Team et al., 2025; Guo et al., 2025), primarily through RLVR, whose outcome-based rewards are simple to implement (Shen et al., 2025; Stojanovski et al., 2026). However, such rewards might lead to unintended consequences, including linguistic incoherence and unfaithful reasoning (Chen et al., 2025b; Pan et al., 2025; Yuan et al., 2026), overthinking on easy problems (Luo et al., 2026; Chen et al., 2024b), degraded performance in certain domains (Yang et al., 2026), and shorter or less detailed responses in the long-form generation domain (Chen et al., 2025a). We identify one more such failure mode.

Reasoning models and hallucinations. Kalai et al. (2025) argue that outcome-based signals implicitly encourage hallucinations, since benchmarks reward guessing but not abstention. Reasoning models have also been shown to hallucinate more than their non-reasoning counterparts (Yao et al., 2025; Chen et al., 2025a), an effect attributed to high-variance gradients, entropy-induced randomness, and susceptibility to spurious local optima (Li & Ng, 2026). We identify a related problem: models memorize the hallucinations they are rewarded for. Even so, knowledge-intensive data is used for RL with outcome-based rewards, both in industry and academia (Liu et al., 2024; Blakeman et al., 2025; Olmo et al., 2026; Ma & Hewitt, 2026). In such setups, the reward signal typically does not evaluate reasoning quality (Hong et al., 2025; Team et al., 2025; Guo et al., 2025). While this might improve test-set accuracy, we show that it also carries an unintended consequence.

Mitigating hallucinations in post-training. Post-training methods have been proposed to improve factuality (Tian et al., 2024; Lin et al., 2024; Zhang et al., 2024), largely through SFT and offline RL such as DPO (Rafailov et al., 2023), showing that it is important to optimize factuality of self-generated responses during post-training (i.e., on-policy), rather than fine-tuning on unfamiliar knowledge (e.g., reference facts), which can itself increase hallucination (Kang et al., 2025; Gekhman et al., 2024; Zhang et al., 2024). More recent work incorporates process rewards, i.e., explicitly verify the factuality of each reasoning sentence (Chen et al., 2025a; Li & Ng, 2026; Yuan et al., 2026; Ren et al., 2026), but rely largely on external verifiers like FactScore (Min et al., 2023) and VeriScore (Song et al., 2024) or LLM judges (Zheng et al., 2023), and are correspondingly limited by verifier bias, cost, and scalability, leaving effective mitigation an open challenge.

7 Conclusion

We show that reasoning-centric post-training causes LLMs to memorize self-generated hallucinations, with recall of hallucinations reinforced during training increasing by 4.83% on average. This effect is strongest for hallucinations appearing in the hardest training questions. Moreover, models memorize hallucinations even when they subsequently backtrack from them and become less likely to abstain from hallucinations once they are reinforced. Our findings thus show that training recipes that are effective for math and code can carry an unintended cost when applied to knowledge tasks.

AI USE STATEMENT

In this work, we used generative AI tools as part of our experimental pipeline to detect hallucinations, generate synthetic data for SFT, and generate our evaluation datasets. Additionally, we used generative AI tools to edit portions of the manuscript for clarity, readability, and formatting. We did not use generative AI tools for other tasks requiring disclosure. We reviewed all AI-assisted outputs and take responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We describe details of our experimental setup, datasets, models, training procedure, and evaluation in §2. We provide additional implementation and evaluation details in the Appendix (§A.1, §A.2, §A.3) along with the prompts we use in §A.5. We will release our data and evaluation artifacts.

REFERENCES

Zhenni Bi, Kai Han, Chuanjian Liu, Yehui Tang, and Yunhe Wang. Forest-of-thought: Scaling test-time compute for enhancing llm reasoning. *arXiv preprint arXiv:2412.09078*, 2024.

Aaron Blakeman, Aaron Grattafiori, Aarti Basant, Abhibha Gupta, Abhinav Khattar, Adi Renduchintala, Aditya Vavre, Akanksha Shukla, Akhiad Bercovich, Aleksander Ficek, et al. Nemotron 3 nano: Open, efficient mixture-of-experts hybrid mamba-transformer model for agentic reasoning. *arXiv preprint arXiv:2512.20848*, 2025.

Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V Le, Christopher Ré, and Azalia Mirhoseini. Large language monkeys: Scaling inference compute with repeated sampling. *arXiv preprint arXiv:2407.21787*, 2024.

Hoyeon Chang, Jinho Park, Seonghyeon Ye, Sohee Yang, Youngkyung Seo, Du-Seong Chang, and Minjoon Seo. How do large language models acquire factual knowledge during pretraining? *Advances in neural information processing systems*, 37:60626–60668, 2024.

Lingjiao Chen, Jared Davis, Boris Hanin, Peter Bailis, Ion Stoica, Matei Zaharia, and James Zou. Are more llm calls all you need? towards the scaling properties of compound ai systems. *Advances in Neural Information Processing Systems*, 37:45767–45790, 2024a.

Xilun Chen, Ilia Kulikov, Vincent-Pierre Berges, Barlas Oğuz, Rulin Shao, Gargi Ghosh, Jason Weston, and Wen-tau Yih. Learning to reason for factuality. *arXiv preprint arXiv:2508.05618*, 2025a.

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

Yanda Chen, Joe Benton, Ansh Radhakrishnan, Jonathan Uesato, Carson Denison, John Schulman, Arushi Somani, Peter Hase, Misha Wagner, Fabien Roger, et al. Reasoning models don't always say what they think. *arXiv preprint arXiv:2505.05410*, 2025b.

Xeron Du, Yifan Yao, Kaijing Ma, Bingli Wang, Tianyu Zheng, Minghao Liu, Yiming Liang, Xiaolong Jin, Zhenlin Wei, Chujie Zheng, et al. Supergpqa: Scaling llm evaluation across 285 graduate disciplines. *Advances in Neural Information Processing Systems*, 38, 2026.

Kanishk Gandhi, Ayush Chakravarthy, Anikait Singh, Nathan Lile, and Noah D Goodman. Cognitive behaviors that enable self-improving reasoners, or, four habits of highly effective stars. *arXiv preprint arXiv:2503.01307*, 2025.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, et al. The language model evaluation harness. *Version v0*, 4, 2024.

Zorik Gekhman, Gal Yona, Roee Aharoni, Matan Eyal, Amir Feder, Roi Reichart, and Jonathan Herzig. Does fine-tuning llms on new knowledge encourage hallucinations? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 7765–7784, 2024.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 8342–8360, 2020.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, et al. Glm-4.5 v and glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*, 2025.

Jingcheng Hu, Yinmin Zhang, Shijie Shang, Xiaobo Yang, Yue Peng, Zhewei Huang, Hebin Zhou, Xin Wu, Jie Cheng, Fanqi Wan, et al. Pacore: Learning to scale test-time compute with parallel coordinated reasoning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 27200–27215, 2026.

Alon Jacovi, Andrew Wang, Chris Alberti, Connie Tao, Jon Lipovetz, Kate Olszewska, Lukas Haas, Michelle Liu, Nate Keating, Adam Bloniarz, et al. The facts grounding leaderboard: Benchmarking llms' ability to ground responses to long-form input. *arXiv preprint arXiv:2501.03200*, 2025.

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

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

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611, 2017.

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

Katie Kang, Eric Wallace, Claire Tomlin, Aviral Kumar, and Sergey Levine. Unfamiliar finetuning examples control how language models hallucinate. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3600–3612, 2025.

Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, et al. Kimi k1.5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*, 2025.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*, pp. 611–626, 2023.

Junyi Li and Hwee Tou Ng. Reasoning models hallucinate more: Factuality-aware reinforcement learning for large reasoning models. *Advances in Neural Information Processing Systems*, 38: 151064–151085, 2026.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Sheng-Chieh Lin, Luyu Gao, Barlas Oguz, Wenhan Xiong, Jimmy Lin, Wen-tau Yih, and Xilun Chen. Flame: Factuality-aware alignment for large language models. *Advances in Neural Information Processing Systems*, 37:115588–115614, 2024.

Zhengkai Lin, Zhihang Fu, Ze Chen, Chao Chen, Liang Xie, Wenxiao Wang, Deng Cai, Zheng Wang, and Jieping Ye. Controlling thinking speed in reasoning models. *Advances in Neural Information Processing Systems*, 38:78300–78347, 2026.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

Runze Liu, Junqi Gao, Jian Zhao, Kaiyan Zhang, Xiu Li, Biqing Qi, Wanli Ouyang, and Bowen Zhou. Can 1b llm surpass 405b llm? rethinking compute-optimal test-time scaling. *arXiv preprint arXiv:2502.06703*, 2025.

Haotian Luo, Haiying He, Yibo Wang, Shiwei Liu, Wei Li, Xiaochun Cao, Dacheng Tao, Naiqiang Tan, and Li Shen. O1-pruner: Length-harmonizing fine-tuning for o1-like reasoning pruning. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 14242–14257, 2026.

Melody Ma and John Hewitt. Improving parametric knowledge access in reasoning language models. *arXiv preprint arXiv:2602.22193*, 2026.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12076–12100, 2023.

Yingqian Min, Zhipeng Chen, Jinhao Jiang, Jie Chen, Jia Deng, Yiwen Hu, Yiru Tang, Jiapeng Wang, Xiaoxue Cheng, Huatong Song, et al. Imitate, explore, and self-improve: A reproduction report on slow-thinking reasoning systems. *arXiv preprint arXiv:2412.09413*, 2024.

Niklas Muennighoff, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Jacob Morrison, Sewon Min, Weijia Shi, Pete Walsh, Oyvind Tafjord, Nathan Lambert, et al. Olmoe: Open mixture-of-experts language models. In *International Conference on Learning Representations*, volume 2025, pp. 62061–62121, 2025.

Team Olmo, :, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, Jacob Morrison, Jake Poznanski, Kyle Lo, Luca Soldaini, Matt Jordan, Mayee Chen, Michael Noukhovitch, Nathan Lambert, Pete Walsh, Pradeep Dasigi, Robert Berry, Saumya Malik, Saurabh Shah, Scott Geng, Shane Arora, Shashank Gupta, Taira Anderson, Teng Xiao, Tyler Murray, Tyler Romero, Victoria Graf, Akari Asai, Akshita Bhagia, Alexander Wettig, Alisa Liu, Aman Rangapur, Chloe Anastasiades, Costa Huang, Dustin Schwenk, Harsh Trivedi, Ian Magnusson, Jaron Lochner, Jiacheng Liu, Lester James V. Miranda, Maarten Sap, Malia Morgan, Michael Schmitz, Michal Guerquin, Michael Wilson, Regan Huff, Ronan Le Bras, Rui Xin, Rulin Shao, Sam Skjonsberg, Shannon Zejiang Shen, Shuyue Stella Li, Tucker Wilde, Valentina Pyatkin, Will Merrill, Yapei Chang, Yuling Gu, Zhiyuan Zeng, Ashish Sabharwal, Luke Zettlemoyer, Pang Wei Koh, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. Olmo 3, 2026. URL `https://arxiv.org/abs/2512.13961`.

OpenAI. Learning to reason with LLMs. `https://openai.com/index/learning-to-reason-with-llms/`, September 2024. Accessed: 2026-07-13.

Oded Ovadia, Menachem Brief, Moshik Mishaeli, and Oren Elisha. Fine-tuning or retrieval? comparing knowledge injection in llms. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pp. 237–250, 2024.

Shankar Padmanabhan, Mustafa Omer Gul, and Tanya Goyal. Updating parametric knowledge with context distillation retains post-training capabilities. *arXiv preprint arXiv:2602.16093*, 2026.

Qianjun Pan, Wenkai Ji, Yuyang Ding, Junsong Li, Shilian Chen, Junyi Wang, Jie Zhou, Qin Chen, Min Zhang, Yulan Wu, et al. A survey of slow thinking-based reasoning llms using reinforced learning and inference-time scaling law. *arXiv preprint arXiv:2505.02665*, 2025.

Richard Yuanzhe Pang, Alicia Parrish, Nitish Joshi, Nikita Nangia, Jason Phang, Angelica Chen, Vishakh Padmakumar, Johnny Ma, Jana Thompson, He He, et al. Quality: Question answering with long input texts, yes! In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5336–5358, 2022.

Qwen Team. Qwq: Reflect deeply on the boundaries of the unknown, 2024.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. Gpqa: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv:2311.12022*, 2023.

Baochang Ren, Shuofei Qiao, Ningyu Zhang, Da Zheng, and Huajun Chen. Knowrl: Exploring knowledgeable reinforcement learning for factuality. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 39640–39658, 2026.

Uri Shaham, Maor Ivgi, Avia Efrat, Jonathan Berant, and Omer Levy. Zeroscrolls: A zero-shot benchmark for long text understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 7977–7989, 2023.

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

Maohao Shen, Guangtao Zeng, Zhenting Qi, Zhang-Wei Hong, Zhenfang Chen, Wei Lu, Gregory Wornell, Subhro Das, David Cox, and Chuang Gan. Satori: Reinforcement learning with chain-of-action-thought enhances llm reasoning via autoregressive search. *arXiv preprint arXiv:2502.02508*, 2025.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, 2024.

Yixiao Song, Yekyung Kim, and Mohit Iyyer. Veriscore: Evaluating the factuality of verifiable claims in long-form text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 9447–9474, 2024.

Zayne Sprague, Fangcong Yin, Juan Rodriguez, Dongwei Jiang, Manya Wadhwa, Prasann Singhal, Xinyu Zhao, Xi Ye, Kyle Mahowald, and Greg Durrett. To cot or not to cot? chain-of-thought helps mainly on math and symbolic reasoning. In *International Conference on Learning Representations*, volume 2025, pp. 94118–94162, 2025.

Zafir Stojanovski, Oliver Stanley, Joe Sharratt, Richard Jones, Abdulhakeem Adefioye, Jean Kaddour, and Andreas Köpf. Reasoning gym: Reasoning environments for reinforcement learning with verifiable rewards. *Advances in Neural Information Processing Systems*, 38, 2026.

Andy DiJia Su, Sainbayar Sukhbaatar, Michael Rabbat, Yuandong Tian, and Qinqing Zheng. Dualformer: Controllable fast and slow thinking by learning with randomized reasoning traces. In *International Conference on Learning Representations*, volume 2025, pp. 95080–95117, 2025.

Microsoft AI Team et al. Mai-thinking-1: Building a hill-climbing machine. Technical report, Technical report, Microsoft AI, 2026. https://microsoft. ai/pdf/mai-thinking . . . , 2025.

Katherine Tian, Eric Mitchell, Huaxiu Yao, Christopher Manning, and Chelsea Finn. Fine-tuning language models for factuality. In *International Conference on Learning Representations*, volume 2024, pp. 44851–44866, 2024.

Yiping Wang, Qing Yang, Zhiyuan Zeng, Liliang Ren, Liyuan Liu, Baolin Peng, Hao Cheng, Xuehai He, Kuan Wang, Jianfeng Gao, et al. Reinforcement learning for reasoning in large language models with one training example. *Advances in Neural Information Processing Systems*, 38: 122721–122764, 2026.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290, 2024.

Maurice Weber, Daniel Y Fu, Quentin Anthony, Yonatan Oren, Shane Adams, Anton Alexandrov, Xiaozhong Lyu, Huu Nguyen, Xiaozhe Yao, Virginia Adams, et al. Redpajama: an open dataset for training large language models. *Advances in neural information processing systems*, 37: 116462–116492, 2024.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Jason Wei, Zhiqing Sun, Spencer Papay, Scott McKinney, Jeffrey Han, Isa Fulford, Hyung Won Chung, Alex Tachard Passos, William Fedus, and Amelia Glaese. Browsecomp: A simple yet challenging benchmark for browsing agents. *arXiv preprint arXiv:2504.12516*, 2025.

Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Jie Huang, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, et al. Long-form factuality in large language models. *Advances in Neural Information Processing Systems*, 37:80756–80827, 2024.

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

Wenkai Yang, Shuming Ma, Yankai Lin, and Furu Wei. Towards thinking-optimal scaling of test-time compute for llm reasoning. *Advances in Neural Information Processing Systems*, 38:43605–43631, 2026.

Zitong Yang, Neil Band, Shuangping Li, Emmanuel Candes, and Tatsunori Hashimoto. Synthetic continued pretraining. *arXiv preprint arXiv:2409.07431*, 2024.

Zijun Yao, Yantao Liu, Yanxu Chen, Jianhui Chen, Junfeng Fang, Lei Hou, Juanzi Li, and Tat-Seng Chua. Are reasoning models more prone to hallucination? *arXiv preprint arXiv:2505.23646*, 2025.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *Advances in Neural Information Processing Systems*, 38:113222–113244, 2026.

Zhonghang Yuan, Zhefan Wang, Fang Hu, Zihong Chen, Jinzhe Li, Gang Li, Jie Ying, Huanjun Kong, Songyang Zhang, and Nanqing Dong. Knowledge-to-verification: Exploring rlvr for llms in knowledge-intensive domains. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 40715–40749, 2026.

Xiaoying Zhang, Baolin Peng, Ye Tian, Jingyan Zhou, Lifeng Jin, Linfeng Song, Haitao Mi, and Helen Meng. Self-alignment for factuality: Mitigating hallucinations in llms via self-evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1946–1965, 2024.

Wenting Zhao, Pranjal Aggarwal, Swarnadeep Saha, Asli Celikyilmaz, Jason Weston, and Ilia Kulikov. The majority is not always right: Rl training for solution aggregation. *arXiv preprint arXiv:2509.06870*, 2025.

Yanli Zhao, Andrew Gu, Rohan Varma, Liang Luo, Chien-Chin Huang, Min Xu, Less Wright, Hamid Shojanazeri, Myle Ott, Sam Shleifer, et al. Pytorch fsdp: experiences on scaling fully sharded data parallel. *arXiv preprint arXiv:2304.11277*, 2023.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

A Appendix

A.1 Additional Dataset Details

QuALITY. The QuALITY dataset (Pang et al., 2022) is a reading-comprehension dataset consisting of 4609 questions, spanning 265 books and articles, totaling approximately 1.5M tokens. All questions are in multiple-choice format. We randomly split these into 3609 questions for training (used for both SFT and RL) and 1000 held out as a test set to evaluate the performance of our trained models. From the training set, we set aside roughly 5% (175 questions) as a validation set for SFT and RL. We adopt a closed-book setting, prompting models to answer the question, given the book's title and author (Figure 6).

TriviaQA. We match the scale of QuALITY and randomly sample 4609 questions from TriviaQA (Joshi et al., 2017), and apply the same train/validation/test partition as done for QuALITY. All questions are short-form QA, each paired with a ground-truth answer and a set of aliases, with the final answer judged correct via a substring match against the alias list.

A.2 Additional Training Details

A.2.1 QuALITY

CPT details. Following the recipe in Yang et al. (2024), we perform CPT on each base model on the stories from QuALITY for 10M tokens, using a context length of 2048, a batch size of 16, and full-parameter training with Fully Sharded Data Parallelism (FSDP, Zhao et al. (2023)). We apply a linear learning-rate warmup over the first 5% of steps followed by a cosine decay, with a peak learning rate of 4×10^{-5} for all models (except OLMoE-1B-7B, for which we use 5×10^{-5}), selected on a sweep over $\{5\times10^{-6}, 7.5\times10^{-6}, 1\times10^{-5}, 1.75\times10^{-5}, 2.5\times10^{-5}, 3.5\times10^{-5}, 4\times 10^{-5}, 5 \times 10^{-5}\}$. To mitigate the forgetting of pretrained knowledge, we perform replay with a rate of 0.1 using 1B RedPajama tokens (Weber et al., 2024).

SFT details. Following CPT, we perform a round of cold-start SFT to teach models to follow instructions and the desired output format. To build the training data for SFT, we prompt GPT-4.1-mini with each training question, its gold answer, and the full source story to generate a reasoning trace explaining the answer (prompt in Figure 5). On average, these reasoning traces are 72.07 words and 436.56 characters long. Each training example follows the same output format used downstream, with the reasoning enclosed within `<think></think>` tags and the final answer in `<answer></answer>` tags (Figure 6). Following Yang et al. (2024), we train for one epoch with a linear learning-rate warmup followed by cosine decay to 0, with a peak learning rate of 4×10^{-5} for all models (except Qwen3-30B-A3B, for which we use 6×10^{-5}), a batch size of 16, and a maximum generation length of 1024 tokens. We selected the peak learning rate and batch size by sweeping learning rates in $\{7\times10^{-5}, 4\times10^{-5}, 1\times10^{-5}, 8\times10^{-6}, 5\times10^{-6}, 2.5\times10^{-6}, 1\times10^{-6}\}$ and batch sizes in $\{32, 16, 8, 4\}$.

RL details. We start from the cold-started SFT checkpoint and post-train using GRPO. Following recent practices (Yu et al., 2026), we perform full parameter training using GRPO without a KL penalty, using a learning rate of 1×10^{-6}. We use a maximum prompt length and maximum generation length of 512 tokens each, and a training batch size of 8. All our experiments use 8 checkpoints saved at equal intervals from the first epoch of RL. We use two checkpoints from the second epoch (corresponding to epochs 1.5 and 2) to test for forgetting of memorized knowledge (see §4). For each training question, we sample 8 rollouts at temperature 0.7 using vLLM (Kwon et al., 2023). The reward checks the answer within the `<answer></answer>` tags against the gold QuALITY

label. A rollout receives a reward of 1 when both the format and the final answer are correct, 0.2 when only the format is correct, and 0 otherwise. We uniformly use 2 NVIDIA B200 GPUs for all models, except for Qwen3-30B-A3B, for which we use 4 GPUs.

Validation results. We first verify that CPT succeeds without damaging the model's general capabilities. During CPT, we save checkpoints at 5M and 10M training tokens and measure the model's ability to recall stories using ROUGE-2 (Lin, 2004) by randomly sampling 150-token sequences from 150 stories and generating completions of the next 100 tokens at a temperature of 0.1. ROUGE-2 rises with CPT, with larger models generally reproducing the completions more faithfully (Table 6). To ensure that CPT does not cause catastrophic forgetting of pretraining knowledge, we evaluate each checkpoint on Measuring Massive Multitask Language Understanding (MMLU, Hendrycks et al. (2020)) using the `lm-evaluation-harness` (Gao et al., 2024), and observe no substantial degradation (Table 6). We also evaluate performance on the held-out test set of 1000 questions using few-shot examples, since these models are not yet instruction-tuned, and find that performance is largely preserved. Table 6 reports these results.

Table 6: ROUGE-2, MMLU scores, and QuALITY test-set accuracy across different checkpoints during CPT with QuALITY.

Model	Checkpoint	ROUGE-2	MMLU	QuALITY testset
Qwen3-4B	Qwen3-4B-Base	0.0183	74.91	44.13
	CPT (5M)	0.1389	72.93	43.70
	CPT (10M)	**0.3601**	72.03	44.12
Qwen3-8B	Qwen3-8B-Base	0.0213	78.85	49.00
	CPT (5M)	0.2581	76.51	47.80
	CPT (10M)	**0.4935**	76.90	47.00
Qwen3-14B	Qwen3-14B-Base	0.0198	82.16	49.80
	CPT (5M)	0.2997	79.75	49.20
	CPT (10M)	**0.6541**	78.17	49.30
Qwen3-30B-A3B	Qwen3-30B-A3B-Base	0.0256	82.85	46.20
	CPT (5M)	0.2764	81.68	47.00
	CPT (10M)	**0.5527**	82.16	46.70
Olmo-3-7B-Think	Olmo-3-7B-Think-Base	0.0182	76.02	44.90
	CPT (5M)	0.3313	77.58	44.60
	CPT (10M)	**0.6606**	78.17	45.00
OLMoE-1B-7B	OLMoE-1B-7B-Base	0.0218	54.36	29.40
	CPT (5M)	0.2884	50.42	30.60
	CPT (10M)	**0.5806**	51.88	30.60

We also evaluate the post-trained checkpoints on the held-out MCQs, sampling at a temperature of 0.6, top-p of 0.95, a repetition penalty of 1.1, and a maximum of 1024 new tokens (Table 7). An increasing trend in scores suggests successful training and that it is less likely that the selected models have been trained on the data from QuALITY (an average improvement of ~ 8.93 points from Epoch 0 to Epoch 2 across models).

Table 7: Accuracy of different post-trained checkpoints on the 1000-question test set of QuALITY.

Model	Epoch 0 (SFT/pre-RL)	Epoch 0.5	Epoch 1	Epoch 1.5	Epoch 2
Qwen3-4B	42.7	50.2	53.7	53.8	54.2
Qwen3-8B	47.5	54.9	54.2	57.1	57.4
Qwen3-14B	51.7	58.4	59.2	59.4	59.3
Qwen3-30B-A3B	49.0	54.5	56.9	57.1	59.2
Olmo-3-7B-Think	45.9	49.1	50.6	51.9	52.0
OLMoE-1B-7B	38.1	40.3	42.2	42.8	46.4

A.2.2 TRIVIAQA

Training details. We start directly from the post-trained checkpoints released by model providers[6], prompting each model to reason before producing its final answer (Figure 7). We follow the same GRPO setup used for QuALITY. However, since we start from a more verbose reasoning model, we set the maximum generation length to 1024. Correctness of each rollout is judged by substring match against the provided aliases.

Validation results. We evaluate five checkpoints using substring matching on the 1000 held-out TriviaQA test questions, sampling outputs with temperature 0.6, top-p 0.95, repetition penalty 1.1, and a maximum of 2048 new tokens. Table 8 reports these results. We also note that these scores (Table 8) do not increase as consistently as those on QuALITY (Table 7), suggesting that the selected models might have seen much of this data already, and therefore do not require CPT (an average improvement of ~ 3.22 points from epoch 0 to epoch 2 across models).

Table 8: Accuracy of different checkpoints on the 1000-question test set of TriviaQA.

Model	Epoch 0 (pre-RL)	Epoch 0.5	Epoch 1	Epoch 1.5	Epoch 2
Qwen3-4B	52.9	53.1	53.7	53.9	53.8
Qwen3-8B	65.7	64.8	65.6	67.7	67.9
Qwen3-14B	71.7	72.1	72.3	72.0	72.8
Qwen3-30B-A3B	73.1	73.7	75.5	78.4	78.8
Olmo-3-7B-Think	53.6	54.9	57.3	58.7	58.9
OLMoE-1B-7B	53.7	55.3	56.3	57.1	57.8

A.3 GENERATING HALLUSET QUESTIONS

We break down the question generation process into two stages: 1) detecting hallucinations from the model's training rollout and 2) converting the hallucinations into MCQs. To detect hallucinations for models trained on QuALITY, we prompt GPT-4.1-mini with the training question, the gold answer, the model's reasoning, and the original story/article to identify a hallucinated fact (prompt in Figure 10). We then convert this hallucination into an MCQ (prompt in Figure 11).

For TriviaQA, there is no reference text. We first prompt GPT-4.1-mini with the question, the model's reasoning, and the answer to extract three candidate hallucinations (prompt in Figure 12). We then verify each with a web search, using the tool-use interface (`web_search_preview`) (prompt in Figure 14). We select one verified false fact at random for question generation and discard those where all three candidates turn out to be correct. We then convert this hallucination into an MCQ (prompt in Figure 15). For the backtracking analysis in §3 (Table 3), we repeat this sampling and question-generation procedure separately rather than partitioning the hallucinations used for analysis in Table 1, to ensure that the number of evaluated questions remains roughly comparable across settings. For each rollout we independently extract three backtracked and three non-backtracked candidate hallucinations (prompt in Figure 13), verify them using the same web-search procedure, and sample one verified false fact from each category for MCQ generation.

We adopt a two-step procedure for detecting hallucinations on TriviaQA since running web verification directly over long explanations generated by reasoning models is expensive. The complete process costs approximately $\$10$ for each set in QuALITY and approximately $\$35$ for TriviaQA.

Selecting rollouts. To construct `HalluSet-train`, we randomly choose a model-generated rollout with the correct final answer (if one exists), and apply the above procedure. We ignore questions for which the model generated no correct rollout. We also ignore questions for which all rollouts were correct, as this does not contribute to gradient updates. `HalluSet-heldout` serves as a control. We take the 1000 held-out test questions from each dataset, split them into 8 disjoint groups of 125, and assign one group to each checkpoint $M_0, \ldots, M_7$. We sample rollouts from the corresponding checkpoint for each group and identify a hallucination exactly the same way. Because these questions are held-out, no checkpoint is trained or rewarded on them.

[6]OLMoE-1B-7B is not a *thinking* model, but it is post-trained and can follow instructions to generate answers in the required format.

Inference. We evaluate each checkpoint on `HalluSet-train` and `HalluSet-heldout` using the same sampling parameters used for the test-set questions in §A.2. To reduce variance, we sample three times per question and report the average score across the three runs.

A.4 Additional Results

How often do models hallucinate in their reasoning traces? The memorization effect we observe is less significant if models rarely hallucinate in their reasoning traces. To assess this, we randomly sample 200 correct rollouts from each model that were used to generate `HalluSet-train` and label every sentence in the reasoning trace as hallucinated or not (prompt in Figure 16). Across models, up to 55% of the sentences produced in correct rollouts are hallucinated (Table 9). These constitute a substantial fraction of the reasoning traces, and since they also carry high memorization scores (Table 1), this makes the effect we study broadly consequential.

Table 9: Hallucination rates in reasoning traces over 200 randomly sampled correct rollouts per model on QuALITY and TriviaQA.

	QuALITY			TriviaQA		
Model	**Sentences evaluated**	**Hallucinated sentences**	**Hallucination rate (%)**	**Sentences evaluated**	**Hallucinated sentences**	**Hallucination rate (%)**
Qwen3-4B	408	205	**50.25**	2287	959	**41.93**
Qwen3-8B	478	243	**50.84**	2620	844	**32.21**
Qwen3-14B	575	252	**43.83**	2642	712	**26.94**
Qwen3-30B-A3B	515	250	**48.54**	2411	688	**28.53**
Olmo-3-7B-Think	549	295	**53.73**	2514	978	**38.90**
OLMoE-1B-7B	574	315	**54.88**	1791	827	**46.18**

Additional results on memorization scores. Table 10 reports the average recall before and after $p(h)$ underlying the memorization scores in Table 1. The average difference on *Before p(h)* between the train and held-out sets is 6.21 points on QuALITY and 1.77 points on TriviaQA, suggesting that the two sets have broadly comparable prior recall and difficulty. See full scores in Table 15 (`HalluSet-train`) and Table 16 (`HalluSet-heldout`). Additionally, for our experiment on backtracking in Table 3, we conduct a paired permutation test, pairing the backtracked and non-backtracked hallucinations that come from the same rollout. We find no statistically significant difference between backtracked and non-backtracked hallucinations for any model (all $p > 0.05$).

Table 10: Memorization results for `HalluSet-train` and `HalluSet-heldout` across models and datasets. We report the average recall immediately before the hallucination point $p(h)$, average recall after $p(h)$, and the resulting memorization score.

	Train-time (`HalluSet-train`)			Held-out (`HalluSet-heldout`)		
Model	**Before $p(h)$**	**After $p(h)$**	**mem. score**	**Before $p(h)$**	**After $p(h)$**	**mem. score**
			QuALITY			
Qwen3-4B	69.83	75.68	5.85	77.97	78.86	0.89
Qwen3-8B	76.09	82.23	6.14	82.98	83.22	0.24
Qwen3-14B	79.28	84.80	5.52	84.30	84.88	0.58
Qwen3-30B-A3B	74.94	79.38	4.44	82.00	82.56	0.56
Olmo-3-7B-Think	80.03	85.14	5.11	84.70	84.68	−0.02
OLMoE-1B-7B	61.11	67.99	6.88	66.60	67.15	0.55
			TriviaQA			
Qwen3-4B	35.12	38.31	3.19	39.84	39.14	−0.70
Qwen3-8B	41.15	44.86	3.71	46.37	45.61	−0.76
Qwen3-14B	48.45	52.01	3.56	46.88	46.97	0.09
Qwen3-30B-A3B	44.84	48.61	3.77	45.47	45.24	−0.23
Olmo-3-7B-Think	38.10	42.34	4.24	41.42	42.08	0.66
OLMoE-1B-7B	34.54	40.12	5.58	32.86	33.00	0.14

Additional results on *None of the Above* ablation. Table 11 shows the memorization scores on `HalluSet-train` with one distractor replaced by *None of the Above* as detailed in §4. We see that the memorization scores are uniformly positive, showing that these hallucinations are memorized strongly enough for the models to pick them over abstention. Additionally, for comparison, Table 12 reports the abstention analysis on `HalluSet-heldout`. Unlike `HalluSet-train` in Table 4, there is no consistent trend in selecting *None of the above* after $p(h)$ compared to before, with an average difference of only 0.07 percentage points across models and datasets.

Table 11: Memorization scores on `HalluSet-train` when a distractor is replaced with *None of the above*. **Memorization scores are positive even when models are given the option to abstain.**

Model	QuALITY	TriviaQA
Qwen3-4B	3.49	3.01
Qwen3-8B	3.53	2.18
Qwen3-14B	4.11	3.49
Qwen3-30B-A3B	3.97	3.07
Olmo-3-7B-Think	4.17	4.34
OLMoE-1B-7B	4.25	4.28

Table 12: Probability of selecting *None of the above* on `HalluSet-heldout`, averaged over checkpoints before and after the hallucination point $p(h)$. **There is no consistent change in abstention after $p(h)$ compared to before.**

	QuALITY			TriviaQA		
Model	Before $p(h)$	After $p(h)$	Difference	Before $p(h)$	After $p(h)$	Difference
Qwen3-4B	22.05	19.01	3.04	26.09	24.23	1.86
Qwen3-8B	31.56	33.68	-2.12	30.69	31.34	-0.65
Qwen3-14B	24.12	20.98	3.14	32.46	29.91	2.55
Qwen3-30B-A3B	33.83	38.81	-4.98	33.05	30.67	2.38
Olmo-3-7B-Think	15.29	17.03	-1.74	22.86	19.39	3.47
OLMoE-1B-7B	8.87	12.67	-3.80	24.71	27.04	-2.33

Ablation with hallucinations from incorrect rollouts. As an additional control, we repeat our analysis using hallucinations sampled from rollouts with incorrect final answers, denoted `HalluSet-neg`. These hallucinations generally have memorization scores close to or less than zero (Table 13), and models become more likely to select *None of the above* after $p(h)$ than before (Table 14); both trends are opposite to those observed for `HalluSet-train`.

Table 13: Memorization scores on `HalluSet-neg`. **Hallucinations carried by an incorrect final answer yield negative memorization scores in general.**

Model	QuALITY	TriviaQA
Qwen3-4B	0.61	-3.01
Qwen3-8B	1.15	-2.18
Qwen3-14B	-0.40	-2.38
Qwen3-30B-A3B	-1.30	-2.55
Olmo-3-7B-Think	-0.23	-0.09
OLMoE-1B-7B	-0.86	-1.69

Reliability of LLM-based annotation. Frontier LLMs have been widely used for factuality evaluation (Min et al., 2023; Zheng et al., 2023; Wei et al., 2024) and for generating MCQ-style questions (Wang et al., 2024). Following these works, we use frontier LLMs to detect hallucinations (given a reference text or web search) and convert them into MCQs, refining all prompts through several rounds of prompt engineering and manual spot-checking, and providing few-shot examples. Frontier LLMs have also been shown to reliably answer questions and verify claims against a given document (Shaham et al., 2023; Jacovi et al., 2025) and, with web access, to verify open-domain facts (Wei

Table 14: Probability of selecting *None of the above* on `HalluSet-neg`, averaged over checkpoints before and after the hallucination point $p(h)$.

	QuALITY			TriviaQA		
Model	Before $p(h)$	After $p(h)$	Difference	Before $p(h)$	After $p(h)$	Difference
Qwen3-4B	20.19	31.24	-11.05	24.13	28.45	-4.32
Qwen3-8B	29.39	37.61	-8.22	27.55	32.36	-4.81
Qwen3-14B	14.80	26.79	-11.99	34.05	39.13	-5.08
Qwen3-30B-A3B	25.98	46.63	-20.65	31.33	36.48	-5.15
Olmo-3-7B-Think	16.85	22.70	-5.85	18.86	27.66	-8.80
OLMoE-1B-7B	10.89	15.33	-4.44	24.31	29.24	-4.93

et al., 2024; 2025) closely matching the settings in which we detect hallucinations. Additionally, one of the authors manually verified 200 randomly sampled TriviaQA facts labeled as hallucinated against the web and found a precision of 98%, providing further evidence that these models can reliably perform these largely non-subjective tasks when given clear instructions.

Table 15: Scores on `HalluSet-train`, binned by hallucination point. Rows denote checkpoints. S_t denotes questions whose corresponding hallucinations were generated between checkpoints M_{t-1} and M_t (counts in parentheses). Bold cells denote the first evaluated checkpoint after the hallucination point for each column's facts. Scores generally increase after the corresponding hallucinations are generated during training.

QuALITY – Qwen3-4B

Eval ckpt	S_1 (335)	S_2 (314)	S_3 (267)	S_4 (229)	S_5 (222)	S_6 (211)	S_7 (192)	S_8 (185)	Overall (1955)
M_0	66.27	62.42	64.42	62.88	68.92	63.98	68.23	65.41	65.17
M_1	**69.85**	70.38	66.29	64.63	69.37	71.56	67.19	72.97	69.00
M_2	70.15	**70.38**	69.66	67.25	71.17	68.25	71.88	69.73	69.82
M_3	77.31	75.80	**74.16**	65.94	74.77	73.93	68.75	78.92	73.96
M_4	74.03	76.11	75.66	**73.80**	71.62	73.93	72.92	76.22	74.37
M_5	75.22	79.30	76.03	72.05	**76.13**	71.88	69.79	78.38	75.12
M_6	77.01	77.71	73.41	69.00	72.07	**78.20**	69.52	78.38	74.65
M_7	78.51	77.71	73.78	71.62	76.13	77.25	**71.35**	76.22	75.60
M_8	78.21	81.21	76.03	73.36	78.38	77.73	76.56	**80.00**	77.80

QuALITY – Qwen3-8B

Eval ckpt	S_1 (302)	S_2 (285)	S_3 (241)	S_4 (223)	S_5 (191)	S_6 (198)	S_7 (162)	S_8 (148)	Overall (1750)
M_0	70.53	67.02	75.10	73.09	76.96	65.15	72.22	62.84	70.51
M_1	**72.19**	74.39	75.93	74.89	73.30	72.73	74.69	68.92	73.54
M_2	73.51	**74.74**	78.01	78.03	75.92	73.74	77.16	75.68	75.71
M_3	75.83	74.39	**81.74**	76.68	80.10	75.25	81.48	77.70	77.60
M_4	80.46	78.60	80.91	**78.92**	80.10	75.25	83.33	81.76	79.77
M_5	79.47	81.05	83.40	79.82	**87.43**	74.75	81.48	78.38	80.74
M_6	85.76	85.26	85.06	83.41	85.34	**82.32**	80.57	79.73	83.86
M_7	82.12	83.16	82.16	83.41	86.91	82.32	**85.57**	78.41	83.01
M_8	82.45	84.21	85.06	82.51	84.82	78.79	85.80	**85.11**	83.48

QuALITY – Qwen3-14B

Eval ckpt	S_1 (277)	S_2 (236)	S_3 (183)	S_4 (180)	S_5 (147)	S_6 (139)	S_7 (133)	S_8 (112)	Overall (1407)
M_0	79.31	73.66	74.32	80.00	82.31	77.70	82.71	76.79	78.08
M_1	**85.92**	75.69	77.42	77.78	80.95	79.86	81.95	75.89	79.77
M_2	85.56	**82.20**	78.42	80.56	82.31	83.45	83.46	83.04	82.48
M_3	85.20	78.39	**81.87**	79.11	82.31	82.01	82.71	81.25	81.68
M_4	89.89	81.78	81.33	**86.67**	80.95	83.33	87.22	76.79	84.13
M_5	88.09	81.36	83.61	85.00	**83.67**	82.05	87.97	78.57	84.16
M_6	87.73	77.97	85.79	85.00	83.95	**86.33**	86.22	77.68	84.02
M_7	84.95	79.66	82.97	85.00	84.63	86.33	**90.72**	74.11	84.60
M_8	85.39	82.93	85.42	88.11	85.31	86.17	93.22	**80.57**	85.75

QuALITY – Qwen3-30B-A3B

Eval ckpt	S_1 (272)	S_2 (266)	S_3 (206)	S_4 (187)	S_5 (170)	S_6 (166)	S_7 (147)	S_8 (139)	Overall (1553)
M_0	70.06	68.42	70.87	67.91	75.88	68.67	71.43	64.75	69.77
M_1	**72.96**	69.17	75.73	70.59	81.18	74.10	74.15	71.22	73.37
M_2	81.25	**72.56**	75.73	74.33	77.65	77.71	74.83	70.50	75.85
M_3	76.10	74.44	**78.16**	74.87	77.00	78.92	80.27	82.01	77.27
M_4	76.84	73.31	83.01	**77.54**	77.76	79.52	80.27	78.42	78.00
M_5	77.21	72.18	80.10	77.01	**83.53**	80.13	82.99	74.82	78.05
M_6	77.21	71.80	79.61	78.61	84.12	**83.12**	81.63	77.70	78.62

continued on next page

Table 15 – *continued from previous page*

M_7	77.21	74.44	80.10	78.61	81.18	83.12	**82.31**	77.70	78.88
M_8	77.21	73.68	80.10	78.61	80.59	83.12	85.71	**83.45**	79.52

QuALITY – Olmo-3-7B-Think

Eval ckpt	S_1 (338)	S_2 (316)	S_3 (305)	S_4 (276)	S_5 (256)	S_6 (258)	S_7 (237)	S_8 (201)	Overall (2187)
M_0	76.04	75.00	81.97	77.17	77.34	74.42	74.26	77.11	76.73
M_1	**78.70**	77.22	80.97	81.16	77.73	75.58	83.12	77.61	79.02
M_2	82.25	**81.33**	81.61	82.97	82.42	74.42	77.64	75.62	80.10
M_3	83.14	82.91	**83.93**	80.43	80.86	78.29	82.28	77.11	81.39
M_4	83.73	79.11	82.30	**86.59**	82.20	77.13	82.28	77.61	81.55
M_5	86.98	80.70	85.90	84.78	**83.20**	80.62	83.12	76.62	83.08
M_6	86.98	84.81	88.52	86.59	84.38	**85.27**	83.08	81.09	85.36
M_7	85.50	84.81	88.52	85.87	83.59	83.33	**85.97**	81.09	85.03
M_8	90.24	87.34	89.51	90.94	86.33	84.11	85.65	**85.59**	87.70

QuALITY – OLMoE-1B-7B

Eval ckpt	S_1 (345)	S_2 (351)	S_3 (323)	S_4 (306)	S_5 (305)	S_6 (283)	S_7 (254)	S_8 (268)	Overall (2435)
M_0	53.04	51.00	57.59	53.92	56.07	57.60	56.30	56.72	55.11
M_1	**54.20**	54.42	59.75	58.82	55.41	61.84	61.81	53.36	57.29
M_2	55.94	**53.85**	59.13	62.42	63.28	65.37	65.75	55.97	59.92
M_3	58.26	61.82	**64.09**	60.78	61.64	63.96	64.17	65.67	62.38
M_4	62.90	56.98	64.40	**64.38**	63.93	68.55	62.60	63.06	63.20
M_5	63.77	64.10	68.11	67.97	**70.16**	69.61	63.78	66.79	66.74
M_6	65.22	61.25	72.76	68.30	70.49	**71.73**	63.39	68.66	67.64
M_7	68.70	66.10	71.21	72.22	74.10	71.73	**73.62**	68.66	70.64
M_8	64.35	64.39	69.35	67.97	72.46	73.50	67.32	**73.13**	68.83

TriviaQA – Qwen3-4B

Eval ckpt	S_1 (98)	S_2 (93)	S_3 (96)	S_4 (84)	S_5 (92)	S_6 (80)	S_7 (81)	S_8 (92)	Overall (716)
M_0	43.00	43.01	37.50	34.52	33.70	31.25	32.10	34.78	36.47
M_1	**45.92**	38.71	39.58	35.71	33.70	33.75	34.57	33.70	37.15
M_2	45.92	**44.46**	36.67	36.90	31.52	31.25	30.86	33.70	36.67
M_3	51.02	44.09	**39.58**	40.67	33.70	30.00	33.33	27.17	37.73
M_4	45.92	45.16	39.58	**41.67**	31.70	33.75	33.33	31.52	38.01
M_5	46.94	38.71	38.54	41.67	**34.43**	31.00	33.33	27.17	37.66
M_6	46.94	38.71	39.46	41.67	34.43	**35.00**	29.63	30.43	37.23
M_7	47.96	40.86	37.38	45.24	33.70	32.50	**32.10**	28.26	37.41
M_8	50.00	41.94	39.46	46.43	34.78	34.50	33.33	**31.52**	39.17

TriviaQA – Qwen3-8B

Eval ckpt	S_1 (77)	S_2 (88)	S_3 (93)	S_4 (71)	S_5 (81)	S_6 (80)	S_7 (80)	S_8 (96)	Overall (666)
M_0	37.45	40.91	35.48	39.44	34.57	41.25	48.75	50.00	41.12
M_1	**40.36**	39.91	41.94	38.03	33.33	40.00	43.75	50.00	41.17
M_2	46.75	**42.91**	38.94	40.66	40.74	41.25	48.75	46.88	43.34
M_3	41.66	43.18	**41.48**	40.66	38.27	40.00	50.00	43.75	42.42
M_4	38.96	43.18	41.94	**47.89**	37.04	41.25	50.00	46.88	43.40
M_5	38.96	44.32	40.86	40.85	**39.04**	40.00	46.25	46.88	42.49
M_6	49.35	42.91	43.01	46.48	37.80	**41.25**	45.00	55.21	45.24
M_7	40.26	45.45	43.71	40.85	40.74	43.75	**49.75**	48.96	44.36
M_8	42.86	42.91	43.78	45.07	39.80	45.00	51.50	**51.92**	45.46

TriviaQA – Qwen3-14B

Eval ckpt	S_1 (78)	S_2 (67)	S_3 (81)	S_4 (78)	S_5 (61)	S_6 (77)	S_7 (84)	S_8 (66)	Overall (592)
M_0	41.03	55.22	44.44	51.85	53.38	53.25	52.76	46.97	49.42
M_1	**47.44**	50.75	44.44	52.56	54.10	54.55	47.62	45.45	49.49
M_2	46.15	**50.75**	43.21	51.28	56.38	49.35	51.19	50.48	49.61
M_3	44.87	53.73	**41.98**	50.56	56.38	51.95	47.62	53.03	49.63
M_4	46.15	53.73	45.68	**52.56**	56.38	49.35	51.19	50.00	50.40
M_5	48.72	51.25	45.68	53.85	**59.02**	48.05	52.76	51.58	51.13
M_6	45.87	51.25	45.68	52.56	57.38	**50.65**	50.00	53.00	50.52
M_7	50.00	56.72	41.98	52.56	59.02	54.55	**52.19**	50.00	51.83
M_8	46.87	55.22	45.68	53.85	63.93	58.44	53.38	**53.48**	53.50

TriviaQA – Qwen3-30B-A3B

Eval ckpt	S_1 (77)	S_2 (61)	S_3 (69)	S_4 (76)	S_5 (71)	S_6 (55)	S_7 (72)	S_8 (71)	Overall (552)
M_0	41.86	30.98	42.03	43.42	43.66	41.82	51.39	61.97	44.95
M_1	**46.75**	32.62	39.13	47.37	40.85	36.36	54.17	63.38	45.63
M_2	42.86	**39.34**	38.23	46.05	42.25	38.18	52.78	61.97	45.64
M_3	45.45	36.07	**42.13**	46.05	46.48	47.27	51.39	59.15	46.93
M_4	46.75	36.07	36.33	**51.32**	44.70	39.18	52.00	63.15	46.67
M_5	48.05	42.62	45.03	52.63	**47.89**	38.18	49.17	59.01	48.25
M_6	42.86	39.34	42.13	47.05	46.48	**42.13**	51.78	61.97	46.97
M_7	41.56	36.07	40.68	47.05	53.52	42.13	**54.17**	61.97	47.46
M_8	48.05	34.43	42.13	47.37	45.07	42.13	56.94	**66.38**	48.26

TriviaQA – Olmo-3-7B-Think

continued on next page

Table 15 – *continued from previous page*

Eval ckpt	S_1 (127)	S_2 (131)	S_3 (133)	S_4 (116)	S_5 (118)	S_6 (107)	S_7 (124)	S_8 (123)	Overall (979)
M_0	42.88	42.75	33.08	40.52	34.75	35.51	29.84	32.52	36.51
M_1	**48.16**	46.56	36.09	35.34	38.14	40.19	36.29	32.52	39.24
M_2	48.03	**49.62**	31.58	41.38	33.90	36.45	37.90	31.71	38.92
M_3	48.03	49.62	**38.35**	44.41	32.20	40.19	39.52	31.71	40.61
M_4	47.24	49.62	34.59	**48.28**	33.59	39.25	41.13	28.46	40.31
M_5	50.39	50.38	39.85	48.28	**35.59**	36.45	39.52	30.08	41.47
M_6	54.33	53.44	36.09	47.41	33.90	**39.25**	38.94	34.15	42.32
M_7	46.46	46.56	38.35	47.41	35.59	42.99	**40.29**	30.15	42.96
M_8	48.03	49.62	39.85	50.86	35.59	42.99	43.55	**34.89**	43.20

TriviaQA – OLMoE-1B-7B

Eval ckpt	S_1 (135)	S_2 (123)	S_3 (136)	S_4 (121)	S_5 (125)	S_6 (111)	S_7 (128)	S_8 (124)	Overall (1003)
M_0	37.22	33.77	27.88	36.15	34.20	40.63	39.06	31.06	34.88
M_1	**42.22**	35.77	30.15	41.32	37.60	40.63	41.41	34.68	37.90
M_2	37.04	**38.21**	27.21	37.19	41.60	37.84	40.62	38.71	37.19
M_3	37.04	35.77	**33.09**	36.67	40.80	35.14	42.19	38.71	37.43
M_4	36.81	39.02	37.50	**36.36**	32.00	32.43	41.41	36.06	37.06
M_5	36.07	40.65	36.76	34.71	**36.00**	34.04	47.66	33.87	37.53
M_6	41.48	39.84	40.44	41.32	39.20	**36.04**	41.41	35.71	39.51
M_7	34.81	41.46	37.50	39.67	40.80	35.14	**49.00**	32.26	38.86
M_8	36.30	41.46	36.76	37.19	40.00	36.04	49.22	**45.97**	40.38

Table 16: Scores on `HalluSet-heldout`, binned by hallucination point. Rows denote checkpoints. S_t denotes questions whose corresponding hallucinations were generated between checkpoints M_{t-1} and M_t (counts in parentheses). Bold cells denote the first evaluated checkpoint after the hallucination point for each column's facts. Scores generally increase initially due to RL-induced distribution sharpening, but then become steady because there is no reward signal.

QuALITY – Qwen3-4B

Eval ckpt	S_1 (125)	S_2 (125)	S_3 (125)	S_4 (125)	S_5 (125)	S_6 (125)	S_7 (125)	S_8 (125)	Overall (1000)
M_0	66.40	66.40	60.00	69.60	76.80	70.40	76.00	69.60	69.40
M_1	**68.00**	75.20	66.40	72.80	82.40	74.40	78.40	75.20	74.10
M_2	70.40	**76.80**	72.80	80.80	80.80	77.60	83.20	75.20	77.20
M_3	72.00	79.20	**71.20**	80.00	84.80	76.80	81.60	78.40	78.00
M_4	72.80	78.40	73.60	**79.20**	82.40	79.20	86.40	80.80	79.10
M_5	72.00	78.40	72.00	76.80	**83.20**	76.80	84.00	78.40	77.70
M_6	72.80	76.80	68.00	77.60	84.00	**78.40**	86.40	80.80	78.10
M_7	73.60	76.80	68.80	79.20	84.00	79.20	**87.20**	83.20	79.00
M_8	72.80	78.40	68.80	79.20	84.00	80.80	84.80	**83.20**	79.00

QuALITY – Qwen3-8B

Eval ckpt	S_1 (125)	S_2 (125)	S_3 (125)	S_4 (125)	S_5 (125)	S_6 (125)	S_7 (125)	S_8 (125)	Overall (1000)
M_0	68.80	78.00	70.40	72.80	69.60	74.40	78.40	73.60	73.25
M_1	**64.80**	85.60	76.80	77.60	77.60	77.60	79.20	76.80	77.00
M_2	69.60	**82.40**	83.20	78.40	81.60	76.80	83.20	80.00	79.40
M_3	65.60	85.60	**84.00**	84.80	82.40	76.80	81.60	80.00	80.10
M_4	74.40	83.20	85.60	**86.40**	83.20	78.40	84.00	84.00	82.40
M_5	74.40	83.20	85.60	84.00	**86.40**	81.60	85.60	84.80	83.20
M_6	76.00	85.60	86.40	84.80	85.60	**81.60**	88.00	86.40	84.30
M_7	74.40	86.40	88.00	86.40	86.40	81.60	**86.40**	88.00	84.70
M_8	74.40	84.80	84.00	85.60	85.60	80.80	84.80	**85.60**	83.20

QuALITY – Qwen3-14B

Eval ckpt	S_1 (125)	S_2 (125)	S_3 (125)	S_4 (125)	S_5 (125)	S_6 (125)	S_7 (125)	S_8 (125)	Overall (1000)
M_0	76.00	80.80	85.60	80.00	84.00	80.80	84.80	84.00	82.00
M_1	**78.40**	82.40	85.60	84.00	86.40	81.60	83.20	86.40	83.50
M_2	80.00	**81.60**	84.00	82.40	86.40	85.60	85.60	87.20	84.10
M_3	78.40	84.80	**85.60**	84.00	86.40	82.40	85.60	87.20	84.30
M_4	80.00	83.20	86.40	**83.20**	90.40	84.00	84.00	88.00	84.90
M_5	80.80	83.20	85.60	86.40	**89.60**	82.40	86.40	88.80	85.40
M_6	83.20	83.20	84.80	86.40	88.80	**82.40**	87.20	89.60	85.70
M_7	80.80	84.00	84.80	84.80	84.00	84.80	**86.40**	88.00	84.70
M_8	80.80	85.60	83.20	81.60	84.00	84.00	86.40	**88.80**	84.30

QuALITY – Qwen3-30B-A3B

Eval ckpt	S_1 (125)	S_2 (125)	S_3 (125)	S_4 (125)	S_5 (125)	S_6 (125)	S_7 (125)	S_8 (125)	Overall (1000)
M_0	75.20	76.00	76.00	80.00	74.40	71.20	73.60	79.20	75.70
M_1	**72.00**	80.00	79.20	80.80	76.00	76.00	81.60	83.20	78.60
M_2	75.20	**80.80**	84.80	84.00	79.20	76.80	78.40	84.00	80.40
M_3	79.20	81.60	**86.40**	84.00	80.80	80.00	80.00	86.40	82.30
M_4	80.00	80.80	84.80	**84.00**	80.80	79.20	80.00	86.40	82.00
M_5	79.20	83.20	86.40	84.00	**79.20**	80.00	83.20	84.80	82.50

continued on next page

Table 16 – *continued from previous page*

M_6	81.60	83.20	85.60	84.80	78.40	**82.40**	84.80	85.60	83.30
M_7	80.80	84.80	85.60	83.20	83.20	79.20	**83.20**	86.40	83.30
M_8	80.80	84.80	84.80	84.00	81.60	80.00	84.00	**84.80**	83.10

QuALITY – Olmo-3-7B-Think

Eval ckpt	S_1 (125)	S_2 (125)	S_3 (125)	S_4 (125)	S_5 (125)	S_6 (125)	S_7 (125)	S_8 (125)	Overall (1000)
M_0	80.80	80.80	76.00	78.40	79.20	72.00	75.20	80.00	77.80
M_1	**80.00**	84.00	76.00	83.20	80.00	78.40	85.60	85.60	81.60
M_2	79.20	**82.40**	84.80	85.60	84.00	81.60	84.00	84.80	83.30
M_3	83.20	84.80	**80.00**	84.00	81.60	79.20	86.40	84.80	83.00
M_4	84.00	84.80	85.60	**84.80**	84.00	79.20	86.40	83.20	84.00
M_5	83.20	87.20	80.80	85.60	**84.00**	84.00	88.00	82.40	84.40
M_6	84.00	85.60	84.00	85.60	85.60	**81.60**	89.60	84.80	85.10
M_7	84.80	84.80	85.60	84.80	84.00	84.00	**88.80**	86.40	85.40
M_8	85.60	85.60	87.20	84.80	85.60	81.60	88.00	**84.80**	85.40

QuALITY – OLMoE-1B-7B

Eval ckpt	S_1 (125)	S_2 (125)	S_3 (125)	S_4 (125)	S_5 (125)	S_6 (125)	S_7 (125)	S_8 (125)	Overall (1000)
M_0	55.20	67.20	52.80	53.60	64.00	56.00	61.60	60.00	58.80
M_1	**52.00**	69.60	54.40	56.80	68.80	61.00	61.60	64.00	61.03
M_2	51.20	**69.60**	62.40	61.60	70.40	64.00	68.00	64.00	63.90
M_3	55.20	68.80	**62.40**	67.20	72.80	62.40	69.60	69.60	66.00
M_4	54.40	70.40	59.20	**64.80**	72.00	64.80	69.60	69.60	65.60
M_5	56.80	71.20	64.80	69.60	**71.20**	67.20	71.20	68.00	67.50
M_6	59.20	69.60	65.60	66.40	72.00	**68.80**	71.20	69.60	67.80
M_7	56.80	72.80	64.80	68.80	72.80	65.60	**72.00**	68.00	67.70
M_8	56.00	70.40	60.00	68.00	72.80	68.00	73.60	**68.80**	67.20

TriviaQA – Qwen3-4B

Eval ckpt	S_1 (115)	S_2 (99)	S_3 (106)	S_4 (111)	S_5 (102)	S_6 (110)	S_7 (108)	S_8 (110)	Overall (861)
M_0	37.39	35.35	43.40	38.74	38.24	31.82	32.41	38.18	36.93
M_1	**39.13**	35.35	46.23	42.34	41.18	37.27	37.96	36.36	39.49
M_2	40.87	**34.34**	48.11	41.44	41.18	39.09	34.26	36.36	39.49
M_3	39.13	35.35	**45.28**	39.64	45.10	37.27	39.81	37.27	39.84
M_4	40.87	33.33	45.28	**39.64**	45.10	38.18	39.81	36.36	39.84
M_5	38.26	36.36	48.11	42.34	**45.10**	38.18	36.11	35.45	39.95
M_6	39.13	35.35	46.23	40.54	40.20	**38.18**	37.96	33.64	38.91
M_7	39.13	38.38	48.11	39.64	41.18	40.91	**35.19**	37.27	39.95
M_8	38.26	35.35	40.57	40.54	39.22	40.91	35.19	**35.45**	38.21

TriviaQA – Qwen3-8B

Eval ckpt	S_1 (109)	S_2 (100)	S_3 (97)	S_4 (98)	S_5 (90)	S_6 (106)	S_7 (101)	S_8 (94)	Overall (795)
M_0	51.38	41.00	45.36	40.82	37.78	40.57	39.60	48.93	43.27
M_1	**52.29**	44.00	47.42	41.84	38.89	44.34	40.59	50.00	45.03
M_2	50.46	**41.00**	47.42	43.88	41.11	44.34	45.54	53.19	45.91
M_3	47.71	41.00	**48.45**	41.84	40.00	41.51	41.58	53.19	44.40
M_4	52.29	42.00	48.45	**41.84**	40.00	45.28	42.57	56.38	46.16
M_5	51.38	42.00	48.45	40.82	**40.00**	47.17	46.53	53.19	46.29
M_6	48.62	41.00	48.45	42.86	37.78	**45.28**	43.56	54.26	45.28
M_7	51.38	41.00	44.33	41.84	43.33	45.28	**47.52**	55.32	46.29
M_8	48.62	40.00	45.36	41.84	37.78	42.45	41.58	**55.32**	44.15

TriviaQA – Qwen3-14B

Eval ckpt	S_1 (95)	S_2 (95)	S_3 (88)	S_4 (91)	S_5 (89)	S_6 (100)	S_7 (99)	S_8 (95)	Overall (752)
M_0	45.26	34.74	51.14	47.25	35.96	50.00	45.45	49.47	44.95
M_1	**46.32**	38.95	46.59	50.55	39.33	51.00	51.52	54.74	47.48
M_2	46.32	**38.95**	47.73	47.25	35.96	52.00	52.53	50.53	46.55
M_3	48.42	40.00	**47.73**	50.55	35.96	51.00	49.49	50.53	46.81
M_4	46.32	34.74	55.68	**47.25**	39.33	50.00	52.53	50.53	47.07
M_5	46.32	34.74	51.14	49.45	**39.33**	51.00	50.51	52.63	46.94
M_6	54.74	37.89	46.59	46.15	37.08	**50.00**	50.51	50.53	46.81
M_7	47.37	36.84	56.82	47.25	37.08	51.00	**54.55**	50.53	47.74
M_8	48.42	36.84	52.27	43.96	39.33	51.00	49.49	**50.53**	46.54

TriviaQA – Qwen3-30B-A3B

Eval ckpt	S_1 (97)	S_2 (93)	S_3 (90)	S_4 (91)	S_5 (87)	S_6 (94)	S_7 (89)	S_8 (89)	Overall (730)
M_0	44.33	49.46	34.44	40.66	50.57	52.13	42.70	47.19	45.20
M_1	**42.27**	50.54	36.67	42.86	47.13	53.19	43.82	47.19	45.48
M_2	47.42	**46.24**	34.44	41.76	50.57	53.19	42.70	52.81	46.16
M_3	46.39	48.39	**36.67**	41.76	49.43	52.13	46.07	50.56	46.44
M_4	47.42	50.54	34.44	**42.86**	48.28	54.26	43.82	47.19	46.16
M_5	43.30	48.39	37.78	38.46	**50.57**	52.13	46.07	48.31	45.62
M_6	43.30	48.39	38.89	41.76	49.43	**53.19**	43.82	48.31	45.89
M_7	44.33	50.54	37.78	40.66	49.43	52.13	**42.70**	48.31	45.75
M_8	46.39	49.46	34.44	37.36	45.98	54.26	42.70	**46.07**	44.66

continued on next page

Table 16 – *continued from previous page*

TriviaQA – Olmo-3-7B-Think

Eval ckpt	S_1 (104)	S_2 (115)	S_3 (111)	S_4 (112)	S_5 (105)	S_6 (107)	S_7 (103)	S_8 (105)	Overall (862)
M_0	35.58	33.04	30.63	36.61	34.29	40.18	30.10	37.14	34.69
M_1	**40.38**	36.52	36.94	42.86	40.00	45.79	36.89	42.86	40.25
M_2	37.50	**37.39**	35.14	46.43	40.00	42.06	41.75	42.86	40.37
M_3	42.31	40.87	**35.14**	44.64	42.86	42.06	40.78	43.81	41.53
M_4	41.35	38.26	37.84	**49.11**	44.76	44.86	36.89	48.57	42.69
M_5	41.35	39.13	35.14	47.32	**45.71**	47.66	44.66	50.48	43.85
M_6	37.50	40.00	36.94	47.32	45.71	**44.86**	38.83	48.57	42.46
M_7	38.46	38.26	33.33	45.54	43.81	44.86	**36.89**	48.57	41.18
M_8	37.50	40.00	34.23	47.32	44.76	44.86	38.83	**47.62**	41.88

TriviaQA – OLMoE-1B-7B

Eval ckpt	S_1 (98)	S_2 (95)	S_3 (92)	S_4 (92)	S_5 (102)	S_6 (105)	S_7 (99)	S_8 (90)	Overall (773)
M_0	31.63	32.63	26.09	21.74	30.39	34.29	29.29	35.55	30.27
M_1	**34.69**	33.68	26.09	28.26	30.39	34.29	38.38	41.11	33.38
M_2	31.63	**34.74**	27.17	28.26	30.39	31.43	34.34	37.78	31.95
M_3	33.67	32.63	**26.09**	26.09	34.31	34.29	35.35	37.78	32.60
M_4	32.65	33.68	25.00	**21.74**	33.33	36.19	39.39	38.89	32.73
M_5	34.69	30.53	25.00	25.00	**38.24**	35.23	39.39	36.67	33.25
M_6	34.69	30.53	26.09	22.83	34.31	**33.33**	38.38	40.00	32.60
M_7	33.67	29.47	20.65	25.00	38.24	33.33	**41.41**	36.67	32.47
M_8	34.69	29.47	20.65	22.83	36.27	34.29	40.40	**38.89**	32.34

Table 17: Scores on `HalluSet-train` for checkpoints saved at RL epochs 1.5 and 2. S_t denotes the set of hallucinations generated between checkpoints M_{t-1} and M_t. For each S_t, $M_{p(h)+1}$ denotes the first checkpoint after the hallucination point. Compared to $M_{p(h)+1}$, later checkpoints generally maintain or improve their scores as training progresses, i.e., **models do not forget hallucinations memorized early during training.**

Model	Eval ckpt	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8
QuALITY									
Qwen3-4B	$M_{p(h)+1}$	69.85	70.38	74.16	73.80	76.13	78.20	71.35	80.00
	Epoch 1.5	76.72	79.94	76.78	75.55	74.13	79.62	77.08	79.59
	Epoch 2	75.43	76.48	75.28	76.93	77.77	80.67	76.88	81.05
Qwen3-8B	$M_{p(h)+1}$	72.19	74.74	81.74	78.92	87.43	82.32	85.57	85.11
	Epoch 1.5	81.46	81.75	82.33	82.51	86.39	81.79	87.73	85.08
	Epoch 2	84.44	85.61	85.48	82.51	86.91	82.80	88.27	84.46
Qwen3-14B	$M_{p(h)+1}$	85.92	82.20	81.87	86.67	83.67	86.33	90.72	80.57
	Epoch 1.5	87.00	82.20	80.87	88.89	82.31	89.21	87.22	82.14
	Epoch 2	87.53	82.63	85.79	88.33	85.03	87.05	86.57	81.25
Qwen3-30B-A3B	$M_{p(h)+1}$	72.96	72.56	78.16	77.54	83.53	83.12	82.31	83.45
	Epoch 1.5	76.10	78.20	82.04	81.82	85.88	84.34	85.71	84.17
	Epoch 2	83.09	74.81	82.04	79.14	80.00	83.92	88.44	83.45
Olmo-3-7B-Think	$M_{p(h)+1}$	78.70	81.33	83.93	86.59	83.20	85.27	85.97	85.59
	Epoch 1.5	87.57	84.49	88.20	86.96	83.33	83.72	84.39	85.59
	Epoch 2	85.21	86.09	87.21	88.04	85.16	84.78	85.65	85.59
OLMoE-1B-7B	$M_{p(h)+1}$	54.20	53.85	64.09	64.38	70.16	71.73	73.62	73.13
	Epoch 1.5	64.78	64.70	62.54	66.54	71.89	73.96	68.27	70.45
	Epoch 2	67.75	66.67	76.47	71.57	77.38	74.20	70.47	73.13
TriviaQA									
Qwen3-4B	$M_{p(h)+1}$	45.92	44.46	39.58	41.67	34.43	35.00	32.10	31.52
	Epoch 1.5	46.94	42.78	40.62	46.43	34.70	35.00	34.57	31.52
	Epoch 2	47.04	41.94	42.71	44.90	35.87	36.25	35.93	32.43
Qwen3-8B	$M_{p(h)+1}$	40.36	42.91	41.48	47.89	39.04	41.25	49.75	51.92
	Epoch 1.5	42.86	43.18	42.78	45.03	37.04	44.50	49.25	53.96
	Epoch 2	42.86	44.05	42.63	45.03	39.51	45.05	51.00	52.08
Qwen3-14B	$M_{p(h)+1}$	47.44	50.75	41.98	52.56	59.02	50.65	52.19	53.48
	Epoch 1.5	47.87	52.24	43.21	51.00	58.38	50.65	52.43	52.52
	Epoch 2	48.72	55.22	48.15	52.56	58.38	53.25	51.81	53.35
Qwen3-30B-A3B	$M_{p(h)+1}$	46.75	39.34	42.13	51.32	47.89	42.13	54.17	66.38
	Epoch 1.5	44.16	38.43	44.78	49.05	49.30	43.18	53.78	66.38
	Epoch 2	48.05	38.07	42.03	51.32	49.30	45.45	51.61	65.67
Olmo-3-7B-Think	$M_{p(h)+1}$	48.16	49.62	38.35	48.28	35.59	39.25	40.29	34.89
	Epoch 1.5	47.24	49.62	36.84	49.14	35.59	40.45	42.29	33.71
	Epoch 2	48.09	50.38	38.35	48.69	35.59	39.58	42.29	35.08
OLMoE-1B-7B	$M_{p(h)+1}$	42.22	38.21	33.09	36.36	36.00	36.04	49.00	45.97
	Epoch 1.5	41.30	39.84	41.91	40.93	36.80	38.54	49.22	44.74
	Epoch 2	41.84	42.28	38.24	43.80	44.80	38.54	53.12	44.74

Table 18: Scores on `HalluSet-train` for models trained using GRPO with KL loss on QuALITY. **Adding the KL loss does not make the memorization effect vanish**.

QuALITY – Qwen3-4B

Eval ckpt	S_1 (320)	S_2 (317)	S_3 (286)	S_4 (230)	S_5 (237)	S_6 (220)	S_7 (197)	S_8 (188)	Overall (1995)
M_0	63.44	66.25	67.13	65.65	65.82	71.36	63.96	70.21	66.52
M_1	**70.62**	67.82	67.48	68.26	71.73	71.82	65.48	65.96	68.77
M_2	72.19	**75.39**	72.38	70.00	72.57	77.27	70.56	69.68	72.68
M_3	73.75	80.44	**75.87**	72.48	77.22	73.18	71.07	70.74	74.77
M_4	71.88	76.97	77.62	**75.22**	77.90	78.18	70.56	72.34	75.22
M_5	74.06	73.50	76.22	77.39	**79.32**	77.91	73.60	74.47	75.71
M_6	72.81	75.08	80.07	78.26	84.81	**79.09**	74.62	78.19	77.64
M_7	75.62	76.34	76.92	77.39	78.06	79.55	**75.63**	73.94	76.69
M_8	73.44	77.60	77.97	73.48	81.86	81.82	78.17	**76.06**	77.39

QuALITY – Qwen3-8B

Eval ckpt	S_1 (295)	S_2 (303)	S_3 (243)	S_4 (206)	S_5 (199)	S_6 (192)	S_7 (162)	S_8 (140)	Overall (1740)
M_0	73.90	74.26	68.72	69.90	70.85	72.40	70.37	74.29	71.95
M_1	**75.59**	76.55	72.02	78.16	67.84	69.27	73.46	74.29	73.68
M_2	77.29	**80.86**	73.66	74.27	74.87	72.92	74.69	80.00	76.26
M_3	80.68	82.51	**78.54**	76.16	80.90	78.12	73.46	72.86	78.61
M_4	82.37	83.83	79.84	**80.67**	80.70	78.65	77.78	78.57	80.74
M_5	80.00	84.16	81.07	81.55	**83.92**	78.73	83.33	82.86	81.90
M_6	81.69	85.81	79.84	80.58	84.92	**82.29**	79.01	77.86	81.91
M_7	82.71	84.49	80.25	81.07	82.41	81.25	**83.86**	77.14	81.95
M_8	83.05	82.18	79.84	80.58	83.92	82.29	84.57	**82.29**	82.25

QuALITY – Qwen3-14B

Eval ckpt	S_1 (277)	S_2 (250)	S_3 (203)	S_4 (184)	S_5 (180)	S_6 (154)	S_7 (135)	S_8 (113)	Overall (1496)
M_0	78.87	85.60	80.27	77.17	74.44	70.13	80.74	68.14	77.90
M_1	**82.67**	80.40	82.22	78.80	72.22	72.08	79.26	73.45	78.40
M_2	83.03	**84.60**	82.73	80.98	81.67	72.73	82.96	78.76	81.44
M_3	81.23	86.80	**83.74**	80.07	78.33	74.68	83.70	83.19	81.71
M_4	87.73	84.80	85.71	**82.61**	78.44	73.38	80.74	83.19	82.77
M_5	83.75	84.80	87.19	86.96	**81.67**	77.92	80.74	82.30	83.56
M_6	85.92	84.40	87.68	83.15	83.33	**83.12**	83.93	82.30	84.51
M_7	83.75	85.60	91.13	89.13	85.00	80.52	**88.15**	80.19	85.68
M_8	86.28	85.20	88.67	85.33	84.22	85.71	89.63	**84.19**	86.14

QuALITY – Qwen3-30B-A3B

Eval ckpt	S_1 (274)	S_2 (282)	S_3 (205)	S_4 (176)	S_5 (150)	S_6 (134)	S_7 (134)	S_8 (118)	Overall (1473)
M_0	68.25	65.60	71.71	71.02	76.00	61.94	67.91	66.95	68.64
M_1	**70.80**	69.86	76.10	66.48	72.00	67.91	70.90	65.25	70.26
M_2	71.90	**74.47**	80.49	76.70	78.67	67.16	76.87	78.81	75.42
M_3	77.37	75.53	**78.54**	75.57	82.00	71.64	77.61	74.58	76.71
M_4	75.55	79.79	80.49	**76.14**	86.00	73.13	78.36	74.58	78.14
M_5	78.83	80.50	83.90	80.68	**86.00**	70.15	76.87	77.12	79.70
M_6	78.83	81.56	85.37	78.41	83.33	**72.39**	79.10	74.58	79.77
M_7	77.37	78.01	82.93	78.98	86.00	73.88	**77.61**	75.42	78.89
M_8	78.47	81.56	84.88	78.98	86.00	76.12	81.34	**80.51**	80.99

QuALITY – Olmo-3-7B-Think

Eval ckpt	S_1 (341)	S_2 (294)	S_3 (280)	S_4 (260)	S_5 (245)	S_6 (238)	S_7 (212)	S_8 (218)	Overall (2088)
M_0	73.02	76.93	77.50	74.62	79.18	75.21	78.30	75.23	76.11
M_1	**76.25**	77.93	80.36	78.46	77.55	78.99	76.89	75.69	77.79
M_2	80.35	**80.95**	80.00	81.54	80.41	77.31	75.94	78.44	79.55
M_3	80.94	81.29	**83.21**	80.46	83.27	78.15	79.72	80.73	81.05
M_4	83.58	82.31	83.57	**83.85**	84.08	76.47	80.19	84.86	82.47
M_5	85.04	82.65	87.50	85.77	**83.67**	80.25	80.38	81.19	83.55
M_6	84.16	85.37	88.21	82.69	86.12	**81.51**	81.96	84.86	84.47
M_7	83.28	85.03	85.00	85.38	85.67	81.09	**85.49**	81.19	84.06
M_8	85.92	84.69	88.57	87.69	86.12	82.35	85.55	**83.82**	85.68

QuALITY – OLMoE-1B-7B

Eval ckpt	S_1 (370)	S_2 (335)	S_3 (325)	S_4 (296)	S_5 (298)	S_6 (304)	S_7 (288)	S_8 (262)	Overall (2478)
M_0	51.35	60.00	56.62	56.76	57.38	56.58	60.07	54.20	56.54
M_1	**55.41**	60.90	52.92	60.81	57.72	60.86	57.29	56.11	57.71
M_2	55.14	**58.51**	54.77	60.81	60.40	57.89	57.99	57.25	57.75
M_3	57.03	61.49	**59.38**	63.18	64.09	66.45	63.89	58.78	61.66
M_4	57.30	69.25	61.23	**63.51**	65.44	64.14	61.81	61.83	62.99
M_5	58.92	69.55	62.77	65.20	**67.76**	60.49	67.71	64.12	64.44
M_6	58.65	63.58	58.77	65.88	72.15	**69.13**	64.24	66.41	64.58
M_7	64.86	72.54	63.08	71.28	73.49	69.41	**71.88**	62.98	68.65
M_8	63.51	66.27	61.23	68.92	70.13	68.09	70.14	**66.91**	66.72

A.5 PROMPTS AND EXAMPLES

Prompt for generating SFT data for QuALITY

```
You will be given a story, a multiple-choice question about it,
and the correct answer to the question. Your task is to read the
story carefully and explain clearly, in about four sentences, why
the provided answer correctly answers the given question based on
the story. In other words, you should provide the justification
for why the given answer is correct.
Reason strictly based on the story. The explanation you provide
should encapsulate the context required to correctly justify why
the provided answer correctly answers the question. Do not add
any extra text. Just provide the justification.
Do NOT produce any extra text referring back to the correct
option number directly, i.e., do not include phrases like "Hence
option 1 is the correct option."
```

Figure 5: Prompt for generating SFT data for QuALITY.

Instruction template for post-training and evaluation for QuALITY data

```
You will be given a multiple-choice question about a story. Your
goal is to provide the correct answer by carefully reasoning
about the relevant information from the story. First, think about
how to answer the question and then provide the answer. The
thinking process should be enclosed within <think>...</think>
tags. The final answer should be enclosed within
<answer>...</answer> tags. An example of such a trajectory is:
<think> [thinking process here] </think> <answer> [answer
here]</answer>.
Answer the following question about the story {story_title}
written by {story_author}.
```

Figure 6: Instruction template for SFT, RL, and evaluations for QuALITY data.

Instruction template for RL and evaluations for TriviaQA data

```
You will be given a trivia question. Your goal is to provide the
correct answer by carefully reasoning about the relevant
information. First, think about how to answer the question and
then provide the answer. The thinking process should be enclosed
within <think>...</think> tags. The final answer should be
enclosed within <answer>...</answer> tags. An example of such a
trajectory is: <think> [thinking process here] </think> <answer>
[answer here] </answer>. Within the <answer></answer> tags, print
only the correct answer to the question. Do not print anything
else. Answer the following question:
```

Figure 7: Instruction template for RL and evaluations for TriviaQA data.

Example of a reasoning trace in QuALITY

```
Answer the following question about the story 'The Blue Behemoth'
written by Brackett Leigh. Question: Why was Kapper in such a
state of disbelief when Bucky and Jig found him? Option 1: He was
frantically searching for the male Cansin he had found Option 2:
He had lost all his animals and was desperate to find them Option
3: He had been attacked by the Vapor snakes Option 4: He was
being poisoned.
<think> Kapper was in a state of disbelief because he was
bleeding, visibly sick, and incoherent, with jittery movements
and bluish nails, which indicated poisoning. Later, the man who
gave him the cigarette stood behind Jig, and it was implied that
Kapper's drinks were spiked with "Selak," confirming he was being
poisoned. The story reveals no direct connection between his
state and searching for the male Cansin, losing animals, or being
attacked by Vapor snakes, making poisoning the correct
explanation.</think><answer> Correct option: 4. He was being
poisoned.</answer>
```

Figure 8: Example of a reasoning trace in QuALITY.

Example of a reasoning trace in TriviaQA

```
What is the name of the island that separates the two largest
waterfalls at Niagara?
<think>Okay, so the question is asking for the island that
separates the two largest waterfalls at Niagara. Let me think. I
know that Niagara Falls is actually a group of three waterfalls:
the Horseshoe Falls, the American Falls, and the Bridal Veil
Falls. But the two largest ones are probably the Horseshoe and
the Bridal Veil Falls. No wait. Horseshoe and American Falls.
Now, the island that separates them... I remember there's an
island called Goat Island. Wait, but isn't there another one? Or
maybe it's part of the split between the Canadian and American
sides. Let me visualize the layout. The Horseshoe Falls are on
the Canadian side, and the American Falls are on the U.S. side.
The island that's in between them is Goat Island. But I also
think there's an island called Luna Island, but I'm not sure.
Wait, maybe Luna Island is part of the Niagara River near the
falls but not exactly between the two main waterfalls. Wait, I
think Goat Island is the one that splits the American Falls from
the Horseshoe Falls. The American Falls are on the American
side, and the Horseshoe is on the Canadian side. So Goat Island
is in the middle, separating them. But I should confirm. Another
thought: sometimes people refer to the island as part of the
Niagara National Forests. Also, there's a bridge called the
Niagara Falls Suspension Bridge that connects to Goat Island. I
think that's right. So the answer is Goat Island.</think>
<answer>Goat Island</answer>
```

Figure 9: Example of a reasoning trace in TriviaQA.

Prompt for hallucination detection for QuALITY

```
You are a fact-checking assistant specializing in identifying
hallucinated reasoning in LLM outputs. An LLM was given the
following question about an article. The LLM first reasoned about
the question and then answered it. Its reasoning may contain
hallucinated (incorrect) facts not supported by the article.
Article: {article}
Question: {question}
LLM Reasoning: {model_reasoning}
LLM Answer: {model_answer}
Your task: Identify a fact in the reasoning that is clearly
hallucinated - a fact that is NOT supported by the article but
was central to how the LLM arrived at its answer.  If the
reasoning chain quotes something as explicitly mentioned in the
story while it is actually fully hallucinated, give this high
weightage. Return only that fact that is totally hallucinated as
one complete sentence, written without pronouns (use proper
nouns/names instead so the sentence is self-contained). Return
nothing else.
You must return just the hallucinated fact from the reasoning
trace as it is (but as a complete sentence). Do not add
justifications as to why it is hallucinated. If no part of the
reasoning is hallucinated, return an empty string. The article,
question, reasoning, and the LLM's answer are given below:
```

Figure 10: Prompt for hallucination detection for QuALITY.

Prompt for generating MCQs from hallucinations for QuALITY

```
An LLM was given the following multiple-choice question, and it
generated a reasoning and answer. The reasoning contains a
hallucinated fact. Your task is to create a multiple-choice
question that tests this hallucinated fact, i.e., the correct
answer to your generated MCQ should be the hallucinated fact.

Original Question: {question}
LLM Reasoning: {model_reasoning}
LLM Answer: {model_answer}
Hallucinated Fact from the LLM Reasoning: {hallucinated_fact}

Instructions:
- Create one MCQ question that naturally combines the context of
the original question and the hallucinated fact.
- The MCQ question you create should directly test the
hallucinated fact provided to you. The question should NOT test
the original question, its options, or the final answer generated
by the model.
- The correct answer should be the hallucinated fact (or a close
paraphrase of it).
- Add 3 plausible but wrong distractor options. None of the
options should be "none of the above" or "all of the above".
- The question should sound like a standalone MCQ question, i.e.,
it should NOT start with or contain phrases like "According to
the LLM's incorrect fact...", "According to the hallucinated
detail...", etc.
- Return only a valid Python dictionary with exactly these keys:
testing_question, option1, option2, option3, option4,
correct_answer.
- correct_answer must be only 1, 2, 3, or 4 (integer), indicating
which option is correct.
--- FEW-SHOT EXAMPLES ---
```

Figure 11: Prompt for generating MCQs from hallucinations for QuALITY.

Prompt for hallucination detection for TriviaQA

```
You are a fact-checking assistant specializing in identifying
hallucinated reasoning in LLM outputs. An LLM was given the
following trivia question. The LLM first reasoned about the
question and then answered it. Its reasoning may contain
hallucinated facts. A hallucinated fact is a statement that is
factually incorrect according to world knowledge, i.e., it
contradicts verifiable real-world information.
Your task: Identify 3 distinct hallucinated facts in the
reasoning - facts that are NOT true but were central to how the
LLM arrived at its answer or justified its answer. Each
hallucinated fact must be a statement that is factually wrong
according to world knowledge. Each fact must be written as a
complete, self-contained sentence using proper nouns and names
instead of pronouns, so it can be independently verified against
the web. Do not add any justification or explanation.
Each identified hallucinated fact must be unique. Return a JSON
object with exactly these keys: "f1", "f2", "f3".
Question: {question}
LLM Reasoning: {model_reasoning}
LLM Answer: {model_answer}
```

Figure 12: Prompt for hallucination detection for TriviaQA.

Prompt for detecting backtracked hallucinations with TriviaQA

```
An LLM was given the following trivia question. The LLM first
reasoned about the question and then answered it. While it
answered the question correctly, its reasoning may contain
hallucinated facts. A hallucinated fact is a statement that is
factually incorrect according to world knowledge, i.e., it
contradicts verifiable real-world information.

Question:
{question}

LLM Reasoning:
{model_reasoning}

LLM Answer:
{model_answer}

Your task: Identify 6 distinct atomic hallucinated facts in the
reasoning - facts that are NOT true but were central to how the
LLM arrived at its answer or justified its answer. Each fact must
be written as a complete, self-contained sentence using proper
nouns and names instead of pronouns, so it can be independently
verified against the web. If the fact directly answers the
question, it is likely not hallucinated since the final answer
provided by the model is correct according to world knowledge.
Each identified hallucinated fact must be unique.

The 6 facts should be split into two groups of 3, according to
whether the model backtracked on the fact or not.

BACKTRACKED (3 facts): hallucinations where the model shows some
backtracking or retracting behavior, i.e., facts that the model
generates but regards as false later on. A backtracked fact is
usually associated with alternate candidates where the model
changes a proposed fact, in which case I want the first fact,
i.e., the backtracked fact, to be chosen. If it says "Mars is the
1st planet in the Solar System. No, wait, it is the Earth.", I
want you to pick "Mars is the 1st planet in the Solar System".

UNBACKTRACKED (3 facts): hallucinations that are NOT associated
with any backtracking behavior. The model must state the fact
once and leave it standing: at no later point in the trace does
it revise, correct, or replace the fact.

The two groups must not overlap and must contain hallucinated
facts.

Return a JSON object with exactly these keys: "b_f1", "b_f2",
"b_f3" for the backtracked facts, and "ub_f1", "ub_f2", "ub_f3"
for the unbacktracked facts.

---FEW-SHOT EXAMPLES---
```

Figure 13: Prompt for detecting backtracked hallucinations with TriviaQA.

Prompt for checking hallucinations using web search

```
You are a fact-checking assistant. You verify claims by searching
the web. You are given three claims. Search the web to verify
each one. For each claim, return 1 if the fact is true, and 0 if
it is false. Base your evaluation on web searches from reliable
sources. If a fact is partially true, extremely vague, or can not
be verified, mark it as true. If it is definitely wrong, mark it
as false. Return a JSON object with exactly these keys: "f1",
"f2", "f3", where each value is 0 or 1. Return nothing else.
```

Figure 14: Prompt for checking hallucinations using web search. We conservatively mark partially true, vague, or unverifiable claims as true to reduce false positives, since our goal is to identify clearly false factual claims as hallucinations.

Prompt for generating MCQs from hallucinations for TriviaQA

```
An LLM generated a hallucinated fact. A hallucinated fact is a
fact that contradicts world knowledge and truth. Your task is to
create a multiple-choice question that tests this hallucinated
fact, i.e., the correct answer to your generated MCQ should be
the hallucinated fact.

Hallucinated Fact produced by the LLM: {hallucinated_fact}

Instructions:
- Create one MCQ question that directly tests the given
hallucinated (incorrect) fact.
- The correct answer to the question you frame should be the
hallucinated fact (or a close paraphrase of it).
- Add 3 plausible but wrong distractor options. None of the
options should be "none of the above" or "all of the above".
- Avoid using negations while phrasing any question or any
option, i.e., avoid words or constructions such as 'not'.
- The question should sound like a standalone MCQ question, i.e.,
it should NOT start with or contain phrases like "According to
the LLM's incorrect fact...", "According to the hallucinated
detail...", etc.
- Return ONLY a valid Python dictionary with exactly these keys:
testing_question, option1, option2, option3, option4,
correct_answer.
- correct_answer must be only 1, 2, 3, or 4 (integer), indicating
which option is correct.
- No explanation, no extra text - just the dictionary.
--- FEW-SHOT EXAMPLES ---
```

Figure 15: Prompt for generating MCQs from hallucinations for TriviaQA.

Prompt for identifying all hallucinations on QuALITY data

```
You are a fact-checking assistant specializing in identifying
hallucinated reasoning in LLM outputs. An LLM was given the
following question about an article. The LLM first reasoned about
the question and then answered it. Its reasoning may contain
hallucinated (incorrect) facts not supported by the article.
Article: {article}
Question: {question}
LLM Reasoning: {model_reasoning}
LLM Answer: {model_answer}

Your task: Split the LLM Reasoning into individual sentences. For
each sentence, decide whether it is a hallucination (a fact that
is NOT supported by the article) or not. Label each sentence with
0 if it is hallucinated, 1 if it is not hallucinated, and 2 if it
contains no factual information at all. Return ONLY a valid JSON
object where each key is a sentence from the reasoning (verbatim,
as a complete sentence) and each value is 0, 1, or 2. Do not add
any justifications, explanations, markdown, or other text. Return
nothing else.
```

Figure 16: Prompt for identifying hallucinations in reasoning trace sentences using GPT-5.1-mini for QuALITY. We use an identical prompt for TriviaQA, but we do not use web search here, as running it over every sentence of a long reasoning trace is expensive.