Joint Continual Pre-Training and RL for Reading Comprehension

Anonymous submission

Abstract

A growing body of work attributes errors in LLM outputs to the mismatch between pre-training and fine-tuning data distributions. We test this hypothesis directly in reading comprehension, where models must produce answers faithful to a provided evidence paragraph. To reduce such effect, we introduce Continual Pre-Training for Reading Comprehension (CPT-RC), a method that performs fine-tuning of a sample in parallel with continued pre-training of its corresponding factual knowledge. We formulate a novel RL framework for promoting factual consistency of the generated answers. Specifically, GRPO was leveraged to reward word-level F1-scores and penalize output length differences. We adapt GRPO to reading comprehension via our training scheme, a first effort for RL to perform knowledge fine-tuning in reading comprehension to our understanding. Our experiments on SQuAD and HaluEval obtain performance increases of up to 17 points with GRPO, and CPT-RC provides additional consistent improvements on top of GRPO across all settings. To further assess factual grounding, we also perform ablation study with our new Augmented QA benchmarks, being novel question-answer pairs over the same source documents. We obtain improvements for both closed-book and open-book performance. We also validate scalability on smaller models, showing that CPT-RC's benefits persist under limited capacity. Our results establish CPT-RC as a simple yet effective strategy for mitigating factual grounding errors in LLMs, which include hallucinations. Our code and dataset will be released upon publication.

1 Introduction

A growing body of work attributes a substantial portion of factual grounding errors to distribution mismatch (Kang et al., 2025; Gekhman et al., 2024) between the data used for pre-training and the data encountered during fine-tuning and evaluation. When examples fall outside the model's familiar pre-training distribution, errors proliferate even if the model is ostensibly competent at the task. This problem is especially consequential in reading comprehension, where models must produce answers faithful to a provided evidence paragraph. Such failures of faithfulness, including hallucinations, directly undermine reliability in user-facing deployments. Motivated by this hypothesis, we propose CPT-RC, a simple concurrent training recipe that reduces this mismatch by **continual pre-training** (also known as **mid-training**) on the very knowledge paragraphs used by reading-comprehension tasks while the model is being optimized for answer quality via reinforcement learning (RL).

For model fine-tuning, we adopt GRPO (Shao et al., 2024), with the reward function defined based on the word-level F1 score. Empirical results show that the training process is stable and does not exhibit any notable inconsistencies. To the best of our knowledge, this work is the first to introduce GRPO as a methodology for knowledge-centric fine-tuning. By leveraging GRPO in this context, we aim to enhance the model's ability to retain and apply factual information more reliably.

We study CPT-RC on SQuAD and HaluEval, two reading-comprehension benchmarks that supply a question, an answer, and a supporting paragraph per instance. Across both datasets, our GRPO baseline already delivers large absolute gains over a supervised baseline (e.g., +17 EM on SQuAD and +9 EM on HaluEval), and CPT-RC provides additional, consistent improvements on top of GRPO with stable training dynamics. To test whether the concurrent rehearsal strengthens document-level factual grounding rather than merely rewarding lexical overlap, we further introduce Augmented Q&A evaluations that hold the paragraph fixed but replace each original question–answer pair with new ones; CPT-RC improves performance in both No Context (closed-book) and

Knowledge Context (open-book) settings.

Additionally, we examine a small-model setting (Qwen-2.5 0.5B) to probe scalability and reward design. We find that straightforward reward shaping (word-level F1 plus a length-constrained reward) yields gains and that CPT-RC's benefit persists at reduced scale, suggesting the approach is not confined to a particular backbone or compute regime.

Our contributions are the following:

1. Our findings indicate that continual pretraining consistently improves factual grounding in reading comprehension. We evaluate our models on SQuAD (Rajpurkar et al., 2016) and HaluEval (Li et al., 2023) datasets and report performance increases in reading comprehension settings.

2. We demonstrate that GRPO constitutes a robust framework for optimizing models with respect to factuality scores. We utilize GRPO to optimize our LLMs with word-wise F1 scores as rewards.

3. We conduct ablation studies that incorporate evaluations using an augmented QA dataset. We design an augmented QA dataset that utilizes the same knowledge documents as the original dataset, but with different question-answer pairs. The purpose of the augmented QA dataset is to verify whether the models remember knowledge documents well.

2 Methodology

2.1 GRPO

In our approach, GRPO is applied to the reading comprehension task, but it is not intended to induce or promote reasoning. While prior studies (Shao et al., 2024; DeepSeek-AI, 2025) have employed GRPO in domains such as complex mathematical problem solving, which inherently require strong reasoning abilities, our work adopts a simpler reward design focused on improving factuality. Specifically, GRPO leverages rewards based on word-level F1 scores and output length differences, guiding the model to maintain consistency between the knowledge paragraph and the answer.

Figure 1 presents our methodology. GRPO is performed concurrently with Continual Pre-Training for Reading Comprehension (CPT-RC). For GRPO, we utilize 2 reward terms, one being word-wise F1 and the another being length differences.

For word-wise F1, we utilize same logic as original SQuAD (Rajpurkar et al., 2016) and compute based on word overlaps between predicted words and gold words.

$$\text{precision} = \frac{\text{overlap}}{\text{pred\_words}}, \text{recall} = \frac{\text{overlap}}{\text{gold\_words}} \quad (1)$$

$$\text{F1} = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}} \quad (2)$$

"Overlap", "pred_words", and "gold_words" reflect the count of overlapping words in prediction and gold answer, the count of all words in the prediction and the count of all words in the gold answer respectively.

For length difference reward, absolute difference between lengths of prediction and gold answer is calculated. If the difference is 0, reward of 1.0 is given. Else if the difference is same or smaller than 5, reward of 0.5 is given. Else if the difference is same of smaller than 10, reward of 0.2 is given. If the difference is bigger than 10, no reward is given. The length difference reward is provided to protect the model from falling into certain failure modes, such as too long or too short output predictions.

This reward will go through computing RL policy loss via advantage calculation, as per GRPO recipe. Let's depict the loss term for RL as L_{RL}. Then:

$$L_{\text{RL}} = L_{\text{GRPO}}(R_{F1} + R_{\text{length}}) \quad (3)$$

with R_{F1} being reward based on $F1$ score and R_{length} being reward based on length.

2.2 CPT-RC

CPT-RC utilizes an additional loss term to compute the pretraining loss. This term "reads" through the context knowledge and produces a loss value, based on next token prediction scheme.

Continual pre-training loss $\mathbf{L}_{\text{CPT}}$ Let the knowledge paragraph be tokenized as $k = (k_1, \ldots, k_T)$, and let π_θ denote the autoregressive policy model. We define the continual pre-training term as the standard teacher-forced next-token negative log-likelihood (average cross-entropy) computed *only* on the knowledge tokens:

$$L_{\text{CPT}}(\theta) = E_{(k,q,a)\sim\mathcal{D}}\left[-\frac{1}{T}\sum_{t=1}^{T}\log \pi_\theta(k_t \mid k_{<t})\right] \quad (4)$$

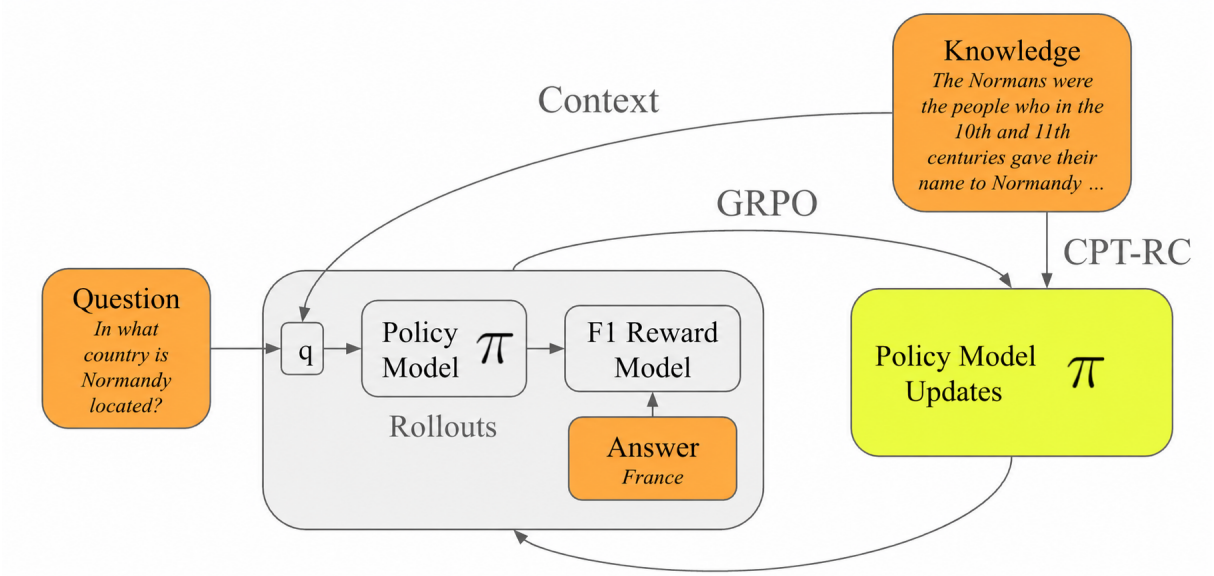

Figure 1: Our methodology. Grey area is the RL process and orange areas are components of reading comprehension task, being question, answer and knowledge. Yellow area depict loss construction and subsequent model updates. The continual pre-training loss L_{CPT} predicts tokens of the knowledge paragraph k in parallel with GRPO policy loss; the joint objective is $L = L_{\text{GRPO}} + \alpha L_{\text{CPT}}$ (Eq. 6).

In practice, for a mini-batch $\mathcal{B} = \{(k^{(i)}, q^{(i)}, a^{(i)})\}_{i=1}^{B}$, we use the unbiased estimator

$$\widehat{L}_{\text{CPT}} = -\frac{1}{B}\sum_{i=1}^{B}\frac{1}{|k^{(i)}|}\sum_{t=1}^{|k^{(i)}|}\log \pi_\theta\big(k_t^{(i)} \mid k_{<t}^{(i)}\big) \tag{5}$$

masking out all tokens not belonging to the knowledge paragraph. Gradients from L_{CPT} are combined with the GRPO policy gradients in the joint objective:

$$\widehat{L} = \widehat{L}_{\text{GRPO}}(R_{F1} + R_{\text{length}}) + \alpha \widehat{L}_{\text{CPT}} \tag{6}$$

With $0.0 \leq \alpha \leq 1.0$ denoting the relative importance of $\widehat{L}_{\text{CPT}}$ term.

2.3 Augmented Q&A Dataset

To directly test whether CPT-RC helps models internalize and reuse factual content from the same source documents seen during training, we build augmented Q&A sets by re-using the knowledge paragraphs provided by HaluEval (and analogously SQuAD) while creating entirely new question–answer pairs with GPT-4o (OpenAI, 2024) that do not appear in the original benchmarks. This allows us to test whether models retain document-level knowledge beyond the specific questions seen during training. Please see Fig. 5.

Concretely, for each knowledge paragraph k, we generate one or more novel questions q whose answers a are short, extractive spans (or minimal paraphrases) supported by k. The augmented sets are used only for evaluation; models are trained on the original benchmark splits, while CPT-RC's continual pre-training term is applied to the same knowledge text k to rehearse facts.

The augmented items form an evaluation-only set. They share the same source paragraphs as the training corpus by design (to test knowledge retention), but their question–answer pairs are held out.

3 Experimental Setup

We train Qwen 2.5 3B Instruct (Qwen et al., 2025) with GRPO using the RLYX framework (Kim, 2025) on 2 nodes of 8 H100 GPUs for 3 epochs on both SQuAD (10K train) and HaluEval (8K train). We use HaluEval as a reading-comprehension benchmark, since it conveniently supplies (question, answer, supporting paragraph) triples; we evaluate on the QA task rather than on the benchmark's hallucination labels. Our prediction prompt is described in Table 1. Full hyperparameters are provided in Appendix A.

role	content
system	This is a conversation between a user and an assistant. - The assistant provides the answer directly. - No discussion or thinking process is needed. - Only the answer should be provided. **Example** ‘‘‘ user Natalia sold clips to 48 of her friends in April, and then she sold half as many clips in May. How many clips did Natalia sell altogether in April and May? assistant 72 ‘‘‘ You are an AI assistant that answers a question based on the provided knowledge. **Knowledge** ‘‘‘ The Normans (Norman: Nourmands; French: Normands; Latin: Normanni) were the people who in the 10th and 11th centuries gave their name to Normandy, a region in France. They were descended from Norse .... ‘‘‘
user	In what country is Normandy located?

Table 1: Sample prompt for a prediction. We utilize 1-shot example to force the model to output answer to the question only. "Knowledge" section is removed for "No Context" experiment in Table 4.

4 Results

4.1 Overall Results

We display our average Exact Match (EM) and F1 scores in Table 2. We report up to 17 point and 10 point EM increase for SQuAD and HaluEval respectively. We also report 14 point and 7 point F1 score increase for SQuAD and HaluEval. Our results show that CPT-RC is effective, with consistent 1 point EM increase in both datasets, compared to GRPO only.

4.2 HaluEval Results

We display training results on HaluEval, evaluated on validation set (Table 2, Figure 2). We observe consistent performance increases, both for CPT-RC with $\alpha = 0.1$ and $\alpha = 1.0$, in contrast to without CPT-RC. $\alpha = 1.0$ emerges as the best HaluEval setting (83.10 EM / 89.51 F1). We also observe that training process is stable without catastrophic forgetting.

4.3 SQuAD Results

We display training results on SQuAD, evaluated on validation set (Table 2, Figure 3). We observe that adding CPT-RC increases the performance of the models during the training process. CPT-RC's curves remain closely above GRPO, indicating faster progress toward the final plateau.

4.4 Small Model Results

We experiment with Qwen 2.5 0.5B model with various reward configurations and report the results in Table 3, Figure 4. We report similar EM values, while F1 score increases more than 1 point. Reward

Model	SQuAD EM	SQuAD F1	HaluEval EM	HaluEval F1
Baseline	53.08 (±0.02)	70.73 (±0.01)	73.10 (±0.00)	82.09 (±0.00)
GRPO	70.01 (±0.12)	84.03 (±0.18)	82.17 (±0.29)	89.14 (±0.15)
CPT-RC 0.1 + GRPO	**70.48** (±0.27)	84.31 (±0.26)	82.65 (±0.17)	89.27 (±0.17)
CPT-RC 1.0 + GRPO	70.14 (±0.12)	**84.34** (±0.46)	**83.10** (±0.10)	**89.51** (±0.12)

Table 2: Exact match and F1 results on SQuAD and HaluEval dataset.

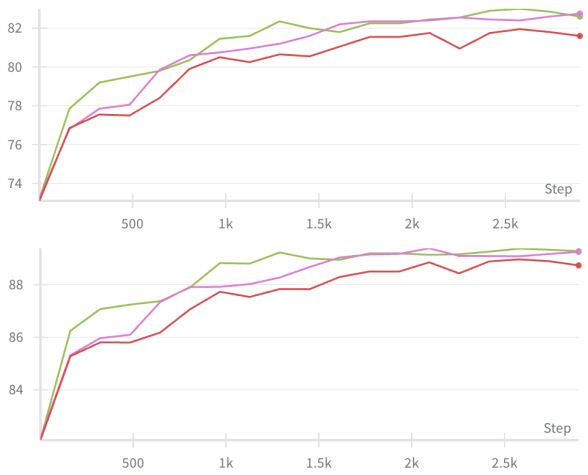

Figure 2: HaluEval train results on the validation set. Left is Exact Match (EM) and right is F1 score. Red is without CPT-RC, pink is CPT-RC with $\alpha = 0.1$, and green is CPT-RC with $\alpha = 1.0$.

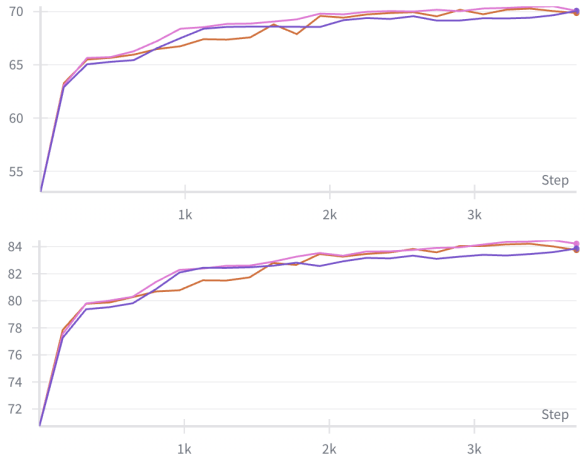

Figure 3: SQuAD train results on the validation set. Left is Exact Match (EM) and right is F1 score. Purple is without CPT-RC, pink is CPT-RC with $\alpha = 0.1$, and red is CPT-RC with $\alpha = 1.0$.

shaping matters, with F1 scores increasing from 81.42 to 82.73 as we add new terms to the mix (F1, length and CPT-RC). We also observe that CPT-RC's benefit is visible even at small scale.

4.5 Augmented HaluEval Q&A Results

Definition to Augmented HaluEval Q&A dataset is in Section 2.3. To probe whether the model better internalizes facts from the same documents rehearsed during training, we evaluate on Augmented HaluEval under No Context (closed-book) and Knowledge Context (open-book) settings. As shown in Table 4, we find that performance increases for both views. "No context" is the experiment without "Knowledge" found in system prompt of Table 1. This indicates that the CPT-RC-trained models understand training set knowledge, and uses it to generate answers to given questions.

4.6 Augmented SQuAD Q&A Results

Definition to Augmented SQuAD Q&A dataset is in Section 2.3. As shown in Table 5, we find that performance increases by more than 1 point for EM with CPT-RC 0.1, compared to no CPT-RC case. In Knowledge Context, $\alpha = 0.1$ gives the best result, following the trend in Table 2.

For augmented SQuAD Q&A the performances are overall lower, which may be due to the fact multiple questions are already provided per train set documents, compared to HaluEval where only 1 question is provided. This may cause the LLM to come up with harder questions than HaluEval. Overall, the results show that CPT-RC is able to teach the models training set knowledge. The models then uses such knowledge to generate answers to questions.

5 Analysis

We perform probabilistic analysis with 1000 knowledge paragraphs from HaluEval. We obtain token probabilities per paragraphs and compute distance on how the token-wise probabilities change with relative to original baselines, from pretrained model and instruct-tuned model. We find that

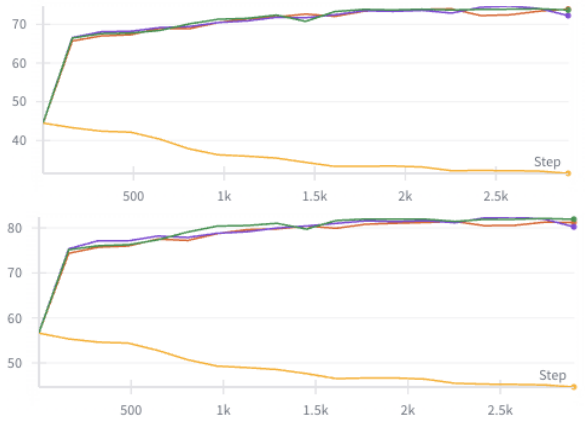

Figure 4: 0.5B train results for HaluEval on the validation set. Left is Exact Match (EM) and right is F1 score. Yellow is pretraining only without GRPO, green is F1 reward only without length reward, purple is with both F1 reward and length reward, and red is EM reward only without F1 or length reward. See Section 4.4 for analysis. Performance captured in Table 3.

Reward	EM	F1
Pretrain Only	44.50	56.63
EM Only	74.00	81.42
F1 Reward	74.00	82.10
F1 + Length Reward	74.70	82.42
CPT-RC 0.1 + GRPO	74.25	82.63
CPT-RC 1.0 + GRPO	**74.75**	**82.73**

Table 3: Exact match and F1 results on 0.5B model. Description in Section 4.4. Training process captured in Figure 4.

CPT-RC models end up being closer to pretrained model, achieving up to 50% distance decreases with final checkpoints. This is likely because the joint objective pulls the distribution toward the pretrained model through the continual pre-training loss (Eq. 4) while RL (Eq. 3) pushes it away from the instruct model's distribution. This pattern is consistent with our motivating hypothesis: by staying closer to the pretrained distribution over factual content, CPT-RC may reduce the factual grounding errors that prior work has linked to hallucinations. Distance between 14B model also decreases, signaling that our model reproduces "smarter" model's token-wise probabilities.

5.1 Symmetric KL Divergence Space

To understand how CPT-RC affects the model beyond task metrics, we examine its token-level probability distribution over knowledge paragraphs. If CPT-RC successfully rehearses factual content, we would expect the model's distribution over knowledge tokens to stay closer to the pretrained model's distribution, since the pretrained model originally learned these facts. We test this by measuring symmetric KL divergence between our trained checkpoints and three reference models: Qwen 2.5 3B (the pretrained base model), Qwen 2.5 3B Instruct (the instruction-tuned starting point), and Qwen 2.5 14B (a larger pretrained model). The first two axes capture how training shifts the distribution relative to its starting points, while the 14B axis tests whether CPT-RC moves the model toward a more capable model's distribution over factual content.

We compute symmetric KL divergence via following mechanism. First, we start with the definition for KL divergence, as expressed below:

$$\mathrm{KL}(p \parallel q) = \sum_i p_i \log \frac{p_i}{q_i}. \tag{7}$$

p_i being tokenwise probability for the given model, and q_i being tokenwise probability for a reference model (Qwen 2.5 3B and Qwen 2.5 3B Instruct in our case).

To have a distance that's symmetric, we should change asymmetric KL divergence and compute 2 values with different baselines.

$$D_t^{p\|q} = \mathrm{KL}\big(p^{(t)} \parallel q^{(t)}\big) \tag{8}$$

$$D_t^{q\|p} = \mathrm{KL}\big(q^{(t)} \parallel p^{(t)}\big) \tag{9}$$

We should then average such KL divergence per tokens to be generated.

$$\overline{\mathrm{KL}}(p \parallel q) = \frac{1}{T}\sum_{t=1}^{T} D_t^{p\|q} \tag{10}$$

$$\overline{\mathrm{KL}}(q \parallel p) = \frac{1}{T}\sum_{t=1}^{T} D_t^{q\|p} \tag{11}$$

With T being length of tokens to be generated.

Lastly, we take average of two distinct KL divergence values and define them as "Symmetric KL Divergence distance".

$$D_{\mathrm{sym}}(p,q) = \tfrac{1}{2}\Big(\overline{\mathrm{KL}}(p \parallel q) + \overline{\mathrm{KL}}(q \parallel p)\Big) \tag{12}$$

D_{sym} is the distance value we use to graph Fig. 6.

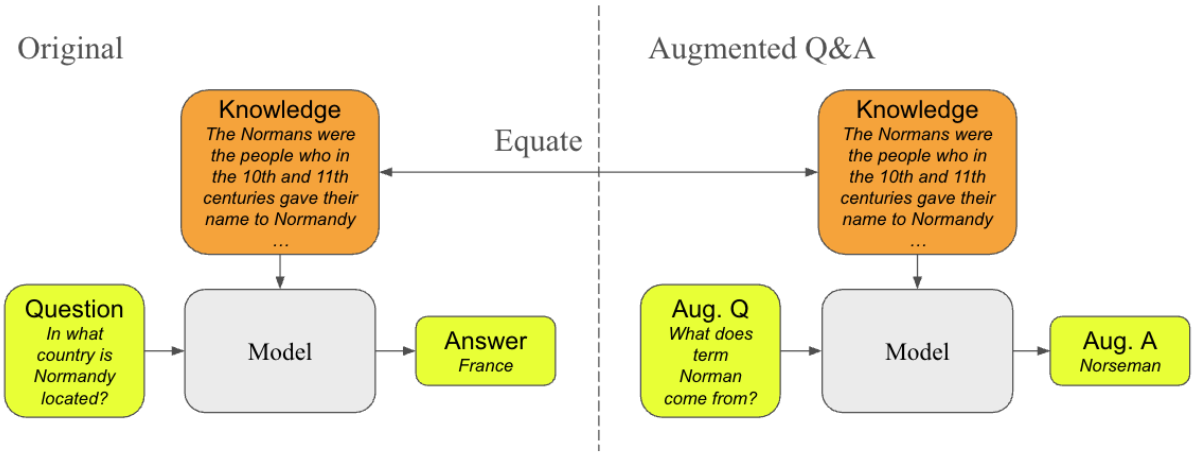

Figure 5: Augmented Q&A data. Grey area is hypothetical model and orange area is knowledge, being kept the same in augmented Q&A data generation process. Augmented Question & augmented Answer (yellow in the right diagram) are generated via an LLM (GPT-4o (OpenAI, 2024)) with respect to the identical knowledge. Please see Appendix B Table 6 for prompt example.

	No Context		Knowledge Context	
Model	EM	F1	EM	F1
GRPO	16.89 (±0.57)	24.13 (±0.69)	76.86 (±1.14)	86.81 (±0.84)
CPT-RC 0.1 + GRPO	17.38 (±0.64)	24.71 (±0.58)	**77.46** (±0.26)	**87.40** (±0.18)
CPT-RC 1.0 + GRPO	**17.55** (±0.81)	**25.02** (±0.57)	77.04 (±0.80)	87.11 (±0.67)

Table 4: Exact match and F1 results on Augmented HaluEval Q&A. Please find the description for the dataset at Section 4.5.

6 Related Works

Distribution mismatch in post-training. Prior studies suggest that factual grounding errors often arise from mismatches between the knowledge distributions of fine-tuning and pre-training datasets. (Kang et al., 2025; Gekhman et al., 2024). We take inspiration from their findings and add a concurrent pre-training stage to fine-tuning setup, reducing this mismatch and improving the model performance.

Continual pre-training and mid-training. Continual pre-training has emerged as a practical strategy for adapting LLMs to new domains or data without retraining from scratch. Gururangan et al. (2020) showed that a second phase of in-domain pre-training (domain-adaptive pre-training) yields consistent gains on downstream classification tasks. More recently, Gupta et al. (2023) studied learning-rate re-warming strategies for continual pre-training at scale, and Ibrahim et al. (2024) demonstrated that simple combinations of learning-rate re-warming, re-decaying, and data replay can match full retraining performance for models up to 10B parameters. Xie et al. (2024) further showed that data-efficient continual pre-training with careful selection strategies can match full-corpus performance at a fraction of the cost. These works focus on adapting general-purpose LLMs to broad domain shifts. In contrast, CPT-RC applies continual pre-training at a finer granularity, rehearsing the specific evidence paragraphs used by reading-comprehension tasks, and combines it with reinforcement learning to jointly optimize factual grounding and answer quality.

GRPO and RL for knowledge tasks. GRPO (Shao et al., 2024) is known to work with solving hard mathematics problems. While there exists a few works that handle knowledge with RL (Li et al., 2025; Xu et al., 2025) they are either on an entirely different domain (visual relation comprehension) or on a mathematics dataset. Our work is the first to utilize GRPO on reading comprehension datasets, and we achieve

Model	No Context EM	No Context F1	Knowledge Context EM	Knowledge Context F1
GRPO	5.44	14.81	58.09	80.15
CPT-RC 0.1 + GRPO	**6.59**	15.47	**59.18**	**80.95**
CPT-RC 1.0 + GRPO	6.46	**15.70**	58.86	79.99

Table 5: Exact match and F1 results on Augmented SQuAD Q&A. Please find the description for the dataset at Section 4.6.

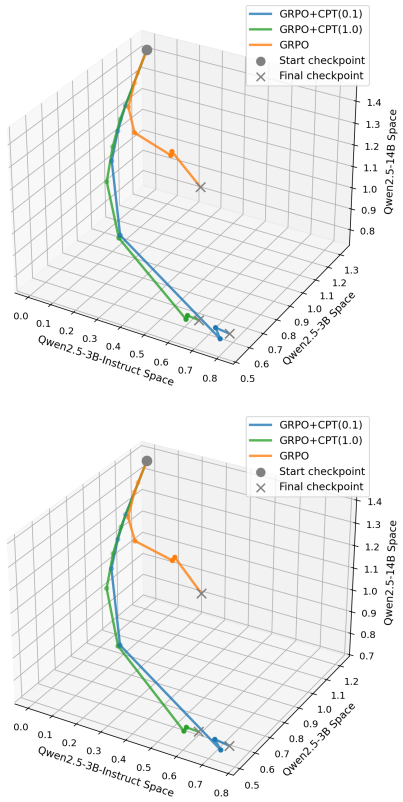

Figure 6: Average symmetric KL divergence distances (Eq. 12). Left is train documents and right is test documents. The models start from grey circle points, undergoes training with our loss (Eq. 6), and ends up in grey X points. Interpretation in Section 5.

a performance increase of 17% on SQuAD (Rajpurkar et al., 2016). We differ from HaluEval (Li et al., 2023) and SQuAD in that we augment these benchmarks with new question-answer pairs over the same knowledge paragraphs, enabling further evaluation of factual grounding from continual pre-training.

Single-stage SFT and RL integration. Fu et al. (2025) explore single-stage integration of SFT and RL for reasoning tasks. Their SRFT framework unifies supervised fine-tuning and reinforcement learning through entropy-aware weighting, demonstrating that SFT induces coarse-grained global policy shifts while RL makes fine-grained selective adjustments. Similarly, Liu et al. (2025) propose Unified Fine-Tuning (UFT), which blends a log-likelihood supervision term with the RL objective in a single stage and theoretically proves an exponential speedup over pure reinforcement fine-tuning. Chen et al. (2025) cast the SFT-RL integration as bilevel optimization, providing a principled framework for jointly training both objectives. These methods all target chain-of-thought reasoning; our work is complementary, using continual pre-training for factual grounding while GRPO selectively optimizes answer quality in knowledge-centric settings.

7 Conclusion

We introduce CPT-RC, a simple concurrent training recipe that couples reinforcement learning with continual pre-training on the same evidence paragraphs used by reading-comprehension tasks. Formally, CPT-RC optimizes a joint objective that adds a next-token prediction loss on the knowledge paragraph to the GRPO policy loss. This design directly targets the pretrain–finetune mismatch hypothesized to underlie many ungrounded answers (including hallucinations) by rehearsing document-level facts while optimizing answer quality. Extending CPT-RC to abstractive generation and open-ended QA settings are promising directions for future work. Our results show that CPT-RC is a simple and effective addition to RL-based training for improving factual consistency in reading comprehension.

Limitations

SQuAD and HaluEval both draw heavily on Wikipedia, which means the "knowledge paragraphs" rehearsed by our continual pre-training term are likely already well-represented in the base model's pre-training corpus. This is a feature for

the distribution mismatch-reduction interpretation but a limitation for generalization: we leave the question of whether CPT-RC would help with documents drawn from scarce, out-of-distribution domains such as legal, biomedical or code documentation to future work.

We evaluate CPT-RC on Qwen family at the 2 different model sizes. While the consistency of gains across these two scales is encouraging, different model families or sizes might reveal different behaviors.

References

Liang Chen, Xueting Han, Li Shen, Jing Bai, and Kam-Fai Wong. 2025. Beyond two-stage training: Cooperative sft and rl for llm reasoning. *Preprint*, arXiv:2509.06948.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Yuqian Fu, Tinghong Chen, Jiajun Chai, Xihuai Wang, Songjun Tu, Guojun Yin, Wei Lin, Qichao Zhang, Yuanheng Zhu, and Dongbin Zhao. 2025. Srft: A single-stage method with supervised and reinforcement fine-tuning for reasoning. *Preprint*, arXiv:2506.19767.

Zorik Gekhman, Gal Yona, Roee Aharoni, Matan Eyal, Amir Feder, Roi Reichart, and Jonathan Herzig. 2024. Does fine-tuning LLMs on new knowledge encourage hallucinations? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7765–7784, Miami, Florida, USA. Association for Computational Linguistics.

Kshitij Gupta, Benjamin Thérien, Adam Ibrahim, Mats L. Richter, Quentin Anthony, Eugene Belilovsky, Irina Rish, and Timothée Lesort. 2023. Continual pre-training of large language models: How to (re)warm your model? *Preprint*, arXiv:2308.04014.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Adam Ibrahim, Benjamin Thérien, Kshitij Gupta, Mats Leon Richter, Quentin Gregory Anthony, Eugene Belilovsky, Timothée Lesort, and Irina Rish. 2024. Simple and scalable strategies to continually pre-train large language models. *Transactions on Machine Learning Research*.

Katie Kang, Eric Wallace, Claire Tomlin, Aviral Kumar, and Sergey Levine. 2025. 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)*, pages 3600–3612, Albuquerque, New Mexico. Association for Computational Linguistics.

Sungju Kim. 2025. Rlyx: A hackable, simple, and research friendly rl framework with high-speed weight synchronization.

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. HaluEval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6449–6464, Singapore. Association for Computational Linguistics.

Lin Li, Wei Chen, Jiahui Li, Kwang-Ting Cheng, and Long Chen. 2025. Relation-r1: Progressively cognitive chain-of-thought guided reinforcement learning for unified relation comprehension. *Preprint*, arXiv:2504.14642.

Mingyang Liu, Gabriele Farina, and Asuman E. Ozdaglar. 2025. UFT: Unifying supervised and reinforcement fine-tuning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

OpenAI. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

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

Yong Xie, Karan Aggarwal, and Aitzaz Ahmad. 2024. Efficient continual pre-training for building domain specific large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10184–10201, Bangkok, Thailand. Association for Computational Linguistics.

Hongling Xu, Qi Zhu, Heyuan Deng, Jinpeng Li, Lu Hou, Yasheng Wang, Lifeng Shang, Ruifeng Xu, and Fei Mi. 2025. Kdrl: Post-training reasoning llms via unified knowledge distillation and reinforcement learning. *Preprint*, arXiv:2506.02208.

A Training Details

We train the models in bfloat16 format and utilize Ray Serve for node-to-node communications. We save model checkpoints every 160 steps and report the performance on the checkpoints with the best performance on the validation set. We set train batch size to be 8 and evaluation batch size to be 64. We set rollout per sample to 3, rollout temperature to be 1.0, and rollout maximum tokens to be 500. Learning rate is $5\mathrm{e}{-7}$ with cosine scheduler type and maximum gradient value of 1.0.

For SQuAD, we use a truncated split of 10K train samples and 10.6K validation samples. For HaluEval, the dataset consists of 10K samples with an 8K train set.

B Augmented Q&A Generation Prompt

For generating augmented question-answer pairs (Section 2.3), we prompt GPT-4o (OpenAI, 2024) to produce a stylistically similar but conceptually distinct QA pair given the knowledge paragraph and the original examples. The prompt is shown in Table 6.

role	content
user	You are a helpful assistant. Ingest the given knowledge and produce a question-answer pair according to the style of given example, while concerning different concept from the example. **Knowledge** ‘‘‘ The Normans (Norman: Nourmands; French: Normands; Latin: Normanni) were the people who in the 10th and 11th centuries gave their name to Normandy, a region in France. They were descended from Norse .... ‘‘‘ **Example** [{ "question": "In what country is Normandy located?" "answer": "France"}, ...] Generate a question-answer pair that differ from the example, while being stylistically similar.

Table 6: Sample prompt for generating a pair of augmented Q&A dataset.