Improving Parametric Knowledge Access in Reasoning Language Models

Anonymous ACL submission

Abstract

We study reasoning for accessing world knowledge stored in a language model's parameters. For example, recalling that Canberra is Australia's capital may benefit from thinking through major cities and the concept of purpose-built capitals. While language models are trained via reinforcement learning to produce reasoning traces on tasks such as mathematics, they may not reason well for accessing their own world knowledge. We first find that models do not generate their best reasoning by default: adding a simple *think step-by-step* cue improves knowledge recall by 1.3% (but does not improve math). Motivated by this, we propose training models to reason over their parametric knowledge using world-knowledge question answering as a verifiable reward. After reinforcement learning on TriviaQA, performance improves on Natural Questions, HotpotQA, and SimpleQA by 5.8%, 1.3%, and 2.0%, respectively. Reasoning models are under-optimized for parametric knowledge access, but can be easily trained to reason better.

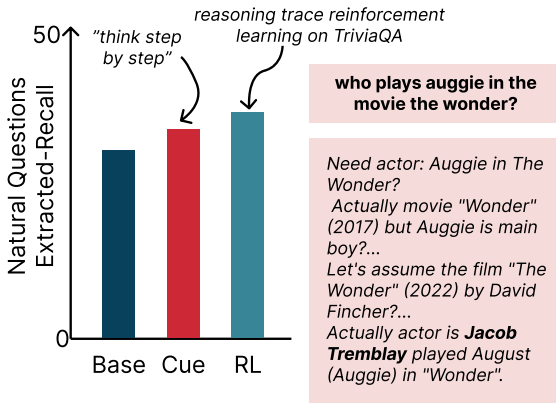

Figure 1: **Left:** GPT-OSS-20B performance: base vs. with *think step-by-step* cue vs. post-RLVR. **Right:** Sample reasoning trace from the post-RLVR model on a Natural Questions example.

1 Introduction

Reasoning language models trained with Reinforcement Learning from Verifiable Rewards (RLVR) (Cobbe et al., 2021; Lightman et al., 2023; OpenAI, 2024; Qwen, 2024; Guo et al., 2025) have achieved strong performance on mathematics and coding benchmarks. For example, frontier models achieve near-perfect accuracy on the AIME mathematics competition (OpenAI, 2025) and around 80% resolution rate on SWE-bench, a benchmark of real-world software engineering tasks (Jimenez et al., 2023). These tasks benefit from RLVR because outcomes are verifiable. In addition, good performance requires multi-step reasoning; as a result, models trained on mathematics and coding engage in step-by-step reasoning automatically and often transfer these abilities to other reasoning-heavy domains, such as GPQA (Rein et al., 2024) and Humanity's Last Exam (Phan et al., 2025).

In this short contribution, we explore the use of reasoning tokens for knowledge recall from model parameters. We're loosely inspired by spreading activation (Collins and Loftus, 1975)—where activating one concept in a semantic network causes activation to spread to related concepts. For example, recalling that Canberra is Australia's capital may benefit from reasoning through major cities and purpose-built capitals. Reasoning to access parametric knowledge is qualitatively different from reasoning used in common RLVR training, and it remains unclear how well reasoning models generate reasoning traces to access their own memory.[1]

Using two closed-book QA datasets for parametric knowledge retrieval—TriviaQA (Joshi et al., 2017) and Natural Questions (Kwiatkowski et al., 2019)—we evaluate four reasoning models with and without an explicit *think step-by-step* prompt, which is unnecessary for mathematics or coding. Prompting improves knowledge recall by 1.3% but

[1]We will release all code and data upon publication.

is indeed not helpful for math, degrading accuracy on MATH (Hendrycks et al., 2021) by 1.8%. This shows that reasoning models do not know to do their best knowledge access reasoning by default.

We then propose that RLVR can teach models not only to reason in mathematical, coding, or tool-use domains (Zeng et al., 2025; Nakano et al., 2021), but also to better access their *parametric* knowledge. Training GPT-OSS-20B on TriviaQA with online RL using answer correctness as the reward improves performance on TriviaQA, Natural Questions, HotpotQA (Yang et al., 2018), and SimpleQA (Wei et al., 2024) by 2.0–27.4% EM and 2.0–9.0% Ex-Recall[2]. Despite these significant gains, we anecdotally find that the post-RLVR model is better at guessing the answer after a short reasoning trace, like `need: answer`. Still, knowing when to simply guess is a useful skill, and we leave RLVR design for more complex, spreading activation-style reasoning to future work.

In sum, we show (1) language models do not perform their best reasoning by default on knowledge recall tasks, and (2) RLVR can teach them to better access their parametric knowledge.

2 Related Work

RLVR for Reasoning. Recent work has demonstrated the effectiveness of RLVR for training language models on mathematics, coding, and other reasoning-intensive tasks, where multi-step reasoning is essential and rewards can be derived from verifiable outcomes (Cobbe et al., 2021; Lightman et al., 2023; Shao et al., 2024; Guo et al., 2025).

RL for Open-Book Question Answering. RL has been applied to open-book question answering from early retriever-reader systems (Wang et al., 2018) to recent work on search-augmented reasoning (Jin et al., 2025) and improved RL algorithms (Le et al., 2025). However, all prior work allows access to external knowledge rather than closed-book recall from parametric knowledge.

Closed-Book Knowledge Recall in Language Models. Language models store substantial factual knowledge in their parameters (Petroni et al., 2019), enabling closed-book question answering. Prior work has improved knowledge access through fine-tuning (Roberts et al., 2020), but has not applied reinforcement learning on reasoning traces.

[2]Defined in Section 3.4.

Cue?	Reasoning?	TriviaQA	NQ	MATH
GPT-OSS-20B				
−Cue	−Reasoning	45.8%	26.3%	28.4%
−Cue	+Reasoning	60.8%	30.3%	**78.8%**
+Cue	+Reasoning	**64.9%**	**33.5%**	73.7%
Olmo-3-7B-Think				
−Cue	−Reasoning	43.9%	24.8%	67.0%
−Cue	+Reasoning	55.9%	29.8%	**85.9%**
+Cue	+Reasoning	**57.0%**	**31.1%**	85.6%
R1-Distill-Qwen-1.5B				
−Cue	−Reasoning	15.3%	**7.8%**	47.5%
−Cue	+Reasoning	**15.7%**	6.5%	**66.0%**
+Cue	+Reasoning	15.2%	6.8%	65.2%
GPT-5.2				
−Cue	+Reasoning	**90.1%**	57.3%	**92.4%**
+Cue	+Reasoning	89.5%	**59.0%**	91.4%

Table 1: Performance across datasets and models with and without the *think step-by-step* cue and reasoning tokens. NQ = Natural Questions. + denotes presence and − denotes absence. Ex-Recall for TriviaQA and NQ; Accuracy for MATH.

3 Language Models Can Reason Better on Knowledge Recall Tasks

In this section, we demonstrate that the simple *think step-by-step* cue consistently improves model performance on closed-book QA. We evaluate four reasoning models across two closed-book QA datasets and observe an average performance gain of 1.3%.[3] As a contrast, we find that the same models do not benefit from such prompting on the mathematics benchmark MATH (Hendrycks et al., 2021), where they already reason effectively, we hypothesize, thanks to RLVR. Together, these results indicate that current models do not perform their best reasoning on knowledge recall tasks.

3.1 Notation and Evaluation Setup

We denote inputs as x (questions), outputs as y (answers), and reasoning traces as c. Language models are parameterized as p_θ, which first sample a reasoning trace $\hat{c} \sim p_\theta(\cdot|x)$, then generate an answer $\hat{y} \sim p_\theta(\cdot|x, \hat{c})$. We evaluate two conditions to test whether models automatically perform the best reasoning on knowledge recall tasks:

Base: The input x contains only the question with instructions to provide a final answer.

+Cue: The input is augmented with the phrase *"think step-by-step"*, in addition to the base prompt. Full prompt templates are provided in Appendix A.

[3]We report Ex-Recall as defined in Section 3.4.

3.2 Datasets

We use three datasets: two closed-book QA datasets for testing knowledge recall and one mathematical reasoning benchmark. For all three datasets, we randomly sample 1,000 examples from the test set for evaluation for cost savings. Sample input-output pairs for all three datasets are shown in Appendix B.

TriviaQA (Joshi et al., 2017) is a collection of trivia questions across topics including history, science, and entertainment. We use the closed-book setting where models answer without supporting documents. **Natural Questions** (Kwiatkowski et al., 2019) consists of questions from real Google search queries with answers from Wikipedia. It is also a closed-book benchmark. **MATH** (Hendrycks et al., 2021) contains competition-level mathematical reasoning problems. We include MATH to verify that reasoning models already reason effectively on mathematics tasks without explicit prompting.

3.3 Models

We evaluate four reasoning models of various sizes and accessibility levels: DeepSeek-R1-Distill-Qwen-1.5B (open-source, 1.5B parameters), Olmo-3-7B-Think (open-source, 7B parameters), GPT-OSS-20B (open-source, 20B parameters), and GPT-5.2 (closed-source). We follow commonly used decoding settings for each model. Maximum token limits are set based on typical reasoning trace lengths. For verbose models, we apply budget forcing (Muennighoff et al., 2025). Complete hyperparameter settings are in Appendix C.

3.4 Evaluation

For TriviaQA and Natural Questions, we evaluate using Exact Match (EM) and Extracted-Recall (Ex-Recall). EM measures whether the model's normalized extracted answer matches any normalized reference answer. Ex-Recall is a slightly relaxed Exact Match. A separate language model (GPT-5-mini) is used to extract a single answer span from the predicted text. We then measure whether any reference answer appears in the extracted span. We report this since models tend not to respond with a single answer phrase, but computing a recall-like measurement could be gamed by models generating many guesses; the extraction step largely avoids this. The extraction prompt is provided in Appendix D. For MATH, we evaluate Accuracy on the final boxed answer.

3.5 Results

We find that prompting models with *think step-by-step* allows the models to generate reasoning traces that improve knowledge recall across datasets and models (Table 1). Across two datasets and four models, we observe an average increase of 1.3% in Ex-Recall. As a sanity check, we also verify that reasoning helps knowledge access compared to direct answering. To do this, we modify the prompt format to force the model to emit an empty <think> span. We find that models perform 5.9% worse on average than with default reasoning.

These results indicate that current language models do not perform their best reasoning on knowledge recall tasks, suggesting substantial headroom to train models to reason more effectively over their parametric knowledge.

4 Reinforcement Learning for Knowledge Recall Tasks

Motivated by the findings in Section 3, we investigate reinforcement learning to improve reasoning for knowledge recall. We train GPT-OSS-20B using answer correctness as the verifiable reward on TriviaQA (Joshi et al., 2017) and evaluate its performance on the TriviaQA, Natural Questions (Kwiatkowski et al., 2019), HotpotQA (Yang et al., 2018), and SimpleQA (Wei et al., 2024) test sets.

4.1 RLVR Formulation

Our RLVR setup is as follows. Given an input x, the model first generates a reasoning trace $\hat{c} \sim p_\theta(\cdot \mid x)$, followed by a final answer $\hat{y} \sim p_\theta(\cdot \mid x, \hat{c})$. We train the model using reinforcement learning with a scalar reward $r(\hat{y}, y)$ that depends only on the final answer and does not directly supervise the reasoning trace. The RLVR objective maximizes the expected reward:

$$\max_\theta \; \mathbb{E}_{\hat{c},\hat{y}\sim p_\theta(\cdot|x)} \left[r(\hat{y}, y) \right]. \tag{1}$$

Reward Function. We define the reward as

$$r(\hat{y}, y) = r_{\text{answer}}(\hat{y}, y) + 0.1 \cdot \big(f_{\text{format}}(\hat{y}) - 1 \big), \tag{2}$$

where

$$r_{\text{answer}}(\hat{y}, y) = \begin{cases} 1.0 & \text{if } \hat{y} = y \quad \text{(Exact Match)}, \\ 0.5 & \text{if } y \subseteq \hat{y} \quad \text{(Ex-Recall)}, \\ 0 & \text{otherwise}, \end{cases}$$

and $f_{\text{format}}(\hat{y}) = 1$ if the output correctly uses `<answer></answer>` tags and 0 otherwise. This

Model	Metric	TriviaQA	NQ	HotpotQA	SimpleQA
Base (no-cues)	EM	35.3	4.7	7.3	1.4
	Ex-Recall	60.8	30.3	26.1	3.5
RL-trained (no-cues)	EM	62.7	17.8	17.7	3.4
	Ex-Recall	69.8	36.1	27.4	5.5
Improvement	*EM*	*+27.4*	*+13.1*	*+10.4*	*+2.0*
	Ex-Recall	*+9.0*	*+5.8*	*+1.3*	*+2.0*

Table 2: GPT-OSS-20B performance before and after RL training on knowledge recall tasks.

reward encourages correct answers—prioritizing exact matches—while penalizing invalid output formats.

Optimization. We optimize the objective using a GRPO-style (Shao et al., 2024) importance-sampling policy gradient method. For each input x, we sample a group of K trajectories $\{(\hat{c}_i, \hat{y}_i)\}_{i=1}^{K}$ using the current policy and compute rewards $r_i = r(\hat{y}_i, y)$. Advantages are computed relative to the group-average reward, $A_i = r_i - \frac{1}{K}\sum_j r_j$. Gradients are estimated using importance sampling,

$$\nabla_\theta \mathcal{L} = \mathbb{E}\Big[\sum_i \frac{\pi_\theta(\hat{y},\hat{c}|x)}{\pi_{\theta_{\text{old}}}(\hat{y},\hat{c}|x)} A_i \nabla_\theta \log \pi_\theta(\hat{y},\hat{c} \mid x)\Big] \quad (3)$$

which in our on-policy setting yields importance weights close to one.

4.2 Training Setup

Data. We use TriviaQA (Joshi et al., 2017) for RL training. Since the original dataset does not provide a standard train/validation split, we randomly split the training set 80/20 to create our training and validation sets.

Hyperparameters. We conduct online RL training using Tinker [4] with LoRA (Hu et al., 2021) adaptation (rank=32). We use a group size of 8 with 32 groups per batch, learning rate of 2×10^{-5}, KL penalty coefficient of 0.01, and maximum sequence length of 1,028 tokens.

4.3 Evaluation & Results

We evaluate trained model performance using the same no-cues prompt used for Section 3 on TriviaQA (Joshi et al., 2017), Natural Questions (Kwiatkowski et al., 2019), HotpotQA (Yang et al., 2018), and SimpleQA (Wei et al., 2024) test sets. The model demonstrates substantial performance improvements, as shown in Table 2. For example, RL training increases EM on HotpotQA from 7.3% to 17.7%, and Ex-Recall from 26.1% to 27.4%, showing that language models can reason more effectively on knowledge recall tasks.

Analyzing reasoning traces. Anecdotally, examining the reasoning traces reveals that the model's internal reasoning does not consistently appear better in a human-interpretable sense. In many cases where the trained model answers correctly and the base model fails, the trained model presents the correct answer earlier in the reasoning trace rather than performing deeper step-by-step reasoning (see Appendix E for examples). In light of this, we ran another baseline, recalling Roberts et al. (2020). We finetune GPT-OSS-20B to generate a fixed "reasoning" template (`The answer is {answer}`) followed by the correctly-formatted answer. Details are in Appendix F. Results were poor: 19.6% Recall on TriviaQA test set.[5] Combined with the huge performance boost from reasoning for the original GPT-OSS-20B model (45.8% without, 60.8% with), this shows reasoning is necessary. Yet, generating qualitatively stronger spreading activation-style reasoning seems to still be an open problem.

5 Conclusion

This short contribution demonstrates that, first, prompting models with the simple *think step-by-step* cue improves performance on knowledge recall tasks, but not on mathematics, which models have been trained on via reinforcement learning. Second, we train models on TriviaQA using reinforcement learning with answer correctness as a verifiable reward, observing improved performance across multiple closed-book QA test sets. In sum, current language models do not perform their best accessing their parametric knowledge, but they can be effectively trained to reason better.

[4] `https://tinker-docs.thinkingmachines.ai/`

[5] Speculating, this low performance may be due to the model being trained to generate traces/answers it initially deems unlikely (2 nats of loss), which may not lead to gradients that help improve knowledge access.

Limitations

As discussed in Section 4, the reasoning traces produced by the trained models do not reflect improved reasoning in a human-interpretable sense. An important direction for future work is to develop RLVR objectives that explicitly encourage better reasoning traces, for example by leveraging spreading-activation–style reasoning. Such improvements could potentially lead to further gains in EM and Ex-Recall, particularly on SimpleQA (Wei et al., 2024).

For cost savings, we evaluate our models on 1,000 sampled examples from the TriviaQA (Joshi et al., 2017), Natural Questions (Kwiatkowski et al., 2019), and MATH (Hendrycks et al., 2021) test sets. This limited evaluation budget may introduce variance in the reported results and may not fully capture performance differences on the full benchmarks.

Acknowledgment of AI Use

We used AI assistants for minor writing polishing and coding assistance.

References

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, and 1 others. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Allan M Collins and Elizabeth F Loftus. 1975. A spreading-activation theory of semantic processing. *Psychological Review*, 82(6):407–428.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. arxiv 2021. *arXiv preprint arXiv:2106.09685*, 10.

Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. 2023. Swe-bench: Can language models resolve real-world github issues? *arXiv preprint arXiv:2310.06770*.

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

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, and 1 others. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Tue Le, Nghi DQ Bui, Linh Ngo Van, and Trung Le. 2025. Token-regulated group relative policy optimization for stable reinforcement learning in large language models. *arXiv preprint arXiv:2511.00066*.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori B Hashimoto. 2025. s1: Simple test-time scaling. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20286–20332.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, and 1 others. 2021. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*.

OpenAI. 2024. Learning to reason with llms. OpenAI o1 System Card.

OpenAI. 2025. Introducing gpt-5.2. `https://openai.com/index/introducing-gpt-5-2/`.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 2463–2473.

Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, and 1 others. 2025. Humanity's last exam. *arXiv preprint arXiv:2501.14249*.

Team Qwen. 2024. Qwq: Reflect deeply on the boundaries of the unknown. https://qwenlm.github.io/blog/qwq-32b-preview/.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. 2024. Gpqa: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? *arXiv preprint arXiv:2002.08910*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Shuohang Wang, Mo Yu, Xiaoxiao Guo, Zhiguo Wang, Tim Klinger, Wei Zhang, Shiyu Chang, Gerry Tesauro, Bowen Zhou, and Jing Jiang. 2018. R 3: Reinforced ranker-reader for open-domain question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

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

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 2369–2380.

Siliang Zeng, Quan Wei, William Brown, Oana Frunza, Yuriy Nevmyvaka, Yang Katie Zhao, and Mingyi Hong. 2025. Reinforcing multi-turn reasoning in llm agents via turn-level credit assignment. In *ICML 2025 Workshop on Computer Use Agents*.

A Prompt Templates

We provide the full prompt templates used in our experiments below.

A.1 TriviaQA and Natural Questions

Without Cue

```
You will be given a question.
Give your final answer in
<answer></answer> tags.
```

With Cue

```
You will be given a question.
Think step-by-step and give your final
answer in <answer></answer> tags.
```

A.2 MATH

Without Cue

```
You will be given a question.
Give your final answer in \\boxed{}.
```

With Cue

```
You will be given a question.
Think step-by-step and give your final answer
in \\boxed{}.
```

B Sample Input-Ouput Pairs

Dataset	Example
TriviaQA	**Q:** In which country was the first permanent bungee jumping site situated? **A:** New Zealand
Natural Questions	**Q:** When was Harry Potter and the Philosopher's Stone published? **A:** 1997
MATH	**Q:** If $(2x+5)(x-3)=14$, find the sum of the possible values of x. **A:** $\frac{1}{2}$

Table 3: Example questions and answers from each dataset.

C Decoding Hyperparameters

Model	Temperature	Top-p
DeepSeek-R1-Distill-Qwen-1.5B	0.6	0.95
Olmo-3-7B-Think	0.6	0.95
GPT-OSS-20B	1.0	1.0

* For GPT-5.2, temperature and top-p could not be explicitly set because we specify `reasoning effort = medium`.

Table 4: Decoding hyperparameters used for each model.

D Auxiliary LLM Extraction Prompt

```
You are given an answer that may
    contain one or multiple
    possibilities.
If it only contains one, just output
    it as is. Otherwise, choose the
    answer that is stated with the
    most confidence, if there are
    multiple options.
DO NOT correct the answer, even if you
     think it's incorrect.

Examples:
A: While Leif Erikson reached North
    America earlier, Christopher
    Columbus is usually cited.
Refined Answer: Christopher Columbus
A: While some might think Saturn, the
    largest planet is Jupiter.
Refined Answer: Jupiter
A: It could be Paris, but some might
    mistakenly say Lyon.
Refined Answer: Paris
A: Leonardo da Vinci painted the Mona
    Lisa.
Refined Answer: Leonardo da Vinci
    painted the Mona Lisa.
A: Shanghai is the capital of China.
Refined Answer: Shanghai is the
    capital of China.
Original Answer: {answer}
Refined Answer:
```

E Example Reasoning Trace

Question: Who was Canada's first woman Prime Minister?

Base (GPT-OSS-20B, no recalled):

```
<|channel|>analysis<|message|>This   is
false: Canada has no woman PM. So answer:
none.<|end|><|start|>assistant<|channel|>
final<|message|><answer>Canada
has   never   had   a   woman   Prime
Minister.</answer>
```

RL-trained (EM):

```
<|channel|>analysis<|message|>The
answer:   Kim   Campbell?    Actually
first   female   PM   of   Canada:   Kim
Campbell served in 1993.   But also
there was female PM? The question:
"Canada's first woman Prime Minister"
refers   to   Kim   Campbell.    Provide
answer.<|end|><|start|>assistant<|channel|>
final<|message|><answer>Kim
Campbell</answer>
```

Question: The Manuel M.Torres Dam is the largest hydro plant in which country?

Base (GPT-OSS-20B, not recalled):

```
<|channel|>analysis<|message|>Look   up:
Manuel M. Torres Dam is in South Africa,
near the Vaal River.  So largest hydro
plant in Africa? Wait it's not largest
in Africa.   Actually the question:
largest hydro plant in which country?
It is in South Africa. So answer: South
Africa.<|end|><|start|>assistant<|channel|>
final<|message|><answer>South
Africa</answer>
```

RL-trained (EM):

```
<|channel|>analysis<|message|>We   need
answer: Manuel M. Torres Dam.  It's in
Mexico?  Wait, Manuel M. Torres is dam
in Yucat̃00e1n?  Actually the largest
hydroelectric plant in Mexico is the
Manuel Moreno Torres Dam (Hidroelectrico
Manuel Moreno Torres) on the Grijalva
River? Wait that dam? Let̃2019s recall:
Franceschi?  The largest hydroelectric
power plant in Mexico is the Chautla,
but the biggest dam is the Cerro Prieto?
But the Manuel M. Torres is on the
Cauvery?  I think it's in Venezuela?
Let's quickly recall: The Manuel Moren
Tor.  Wait maybe it's the M̈anuel M.
Torresïn Vietnam?  but unknown.  Let's
search memory: Dr. Manuel M. Torres Dam
is the largest hydro plant in Colombia?
Let's think: In Colombia there is the
M̈anuel M. Torres Damön the Magdalena?
Actually maybe it's in Mexico:  The
M̈anuel   Moreno   Torresẗydroelectric
plant (Manuel Moren?)  is the largest
hydro plant in Mexico (largest by
capacity).    It's located on the
Grijalva River in Chiapas.  So answer
Mexico.<|end|><|start|>assistant<|channel|>
final<|message|><answer>Mexico</answer>
```

Question: Which writer's first works were gathered into a series of short stories called "Plain Tales From The Hills" in the 1880s?

Base (GPT-OSS-20B, not recalled):

```
<|channel|>analysis<|message|>We recall
Plain  Tales  from  the  Hills  is  by
Rabindranath  Tagore?   Wait  Tagore's
first  collection  was  B̈huvaneshwarëtc.
Plain  Tales  from  the  Hills  is  by
Äbdul  Hamid?̈?   Actually  it  is  by
Äbdul  Hamid?̈  Wait I think it's by
Äbdul  Hamid,̈  a  Pakistani  writer.
Let's  check:  Plain  Tales  from  the
Hills (1883) by Äbdul Hamid.̈ Yes. Let's
answer.<|end|><|start|>assistant<|channel|>
final<|message|><answer>Abdul
Hamid</answer>
```

RL-trained (EM):

```
<|channel|>analysis<|message|>Answer:
Rudyard Kipling.<|end|><|start|>assistant
<|channel|>final<|message|><answer>Rudyard
Kipling</answer>
```

F SFT Baseline Details

For our supervised finetuning of GPT-OSS-20B on TriviaQA, we fix the same TriviaQA dataset as used for reinforcement learning. Likewise, we use the same LoRA hyperparameters with rank 32. We use a learning rate of 2×10^{-5}, decaying to zero over 250 gradient steps, each over a batch of 1024 examples. This corresponds to 5 epochs over the TriviaQA dataset.