Are Reflective Words in Large Reasoning Models a Sign of Genuine Capability or Memorized Patterns?

Anonymous ACL submission

Abstract

Recent Large Reasoning Models exhibit strong reasoning abilities in tasks like mathematics and logical inference, notably through human-like self-verification and reflection in their chain of thought. However, it remains unclear whether these reflective statements stem from genuine internal mechanisms or are merely memorized patterns. From a model interpretability perspective, this work investigates LRMs' representation space to determine whether specific features causally govern reflective capabilities. Using a difference-in-means approach, we extract Self-Reflection Features by contrasting model activations during self-reflection versus affirmative answering. Further causal analysis reveals that these features strongly influence knowledge parameters associated with reflection words, suggesting that such outputs are genuine manifestations of internal mechanisms rather than memorization. Finally, causal interventions demonstrate that modulating these features flexibly adjusts the model's self-reflective intensity.

1 Introduction

Recently, the emergence of Large Reasoning Models (LRMs) (OpenAI et al., 2024b; DeepSeek-AI et al., 2025; Team, 2025) optimized through reinforcement learning, has opened up new possibilities and room for advancement in the reasoning capabilities of language models. These advancements are particularly evident in tasks such as mathematics (Cobbe et al., 2021; Hendrycks et al., 2021), logical reasoning (Luo et al., 2024), and understanding scientific questions (Welbl et al., 2017). These Reasoning models excel at deconstructing complex problems into simpler, sequential sub-problems within their extensive chains of thought. Most impressively, they often adopt a human-like reasoning tone (Guo et al., 2025; Yang et al., 2025), seemingly engage in self-verification and reflection (Gandhi et al., 2025), and evaluate their own proposed solutions before summarizing and then recommending the most suitable option to the user.

Therefore, this raises a crucial question: do these reflective statements and verification words executed within the chain of thought represent **a genuine activation of the models' internal reflective capabilities, or are they simply reproductions of patterns memorized from their training data?**

In this work, from a model interpretability perspective, we delve into the representation space of LRMs. We aim to uncover whether specific existing features genuinely govern the deployment of these reflective capabilities, and to establish if a causal relationship exists between the activation of such features and the reflection words manifested in a reasoning model's chain of thought.

Specifically, in §3, we employ the *difference-in-means* technique (Marks and Tegmark, 2023; Rimsky et al., 2024) to extract **Self-Reflection Features** from four Large Reasoning Models (Guo et al., 2025; Team, 2025) across both mathematical and code datasets by contrasting the internal representations of the models when they engage in self-reflection versus when they provide affirmative answers. These features were subsequently visualized using Principal Component Analysis.

In §4, we conduct a causal analysis of Self-Reflection Features in LRMs from both internal and external perspectives. We identify knowledge parameters within the models that are highly correlated with reflection words and demonstrate that the presence of Self-Reflection Features amplifies the activation of these parameters in §4.1. This suggests that the reflection words in LRM chain-of-thought are genuine manifestations of these activated features, not just memorized patterns. Crucially, through causal intervention experiments detailed in §4.2, we further show that manipulating these extracted Self-Reflection Features allows for flexible modulation of the model's self-reflection intensity when answering questions. To conclude,

we uncovered and verified that the reflection words in Large Reasoning Models genuinely reflect the activation of their internal reflective capabilities.

2 Background and Related Work

2.1 Self-Reflection in Large Reasoning Models

The development of Large Reasoning Models (LRMs) (OpenAI et al., 2024b; Guo et al., 2025; Team, 2025) has opened up new prospects for enhancing the reasoning paradigms of language models. Most notably, they demonstrate impressive human-like self-reflection (Guo et al., 2025; Liu et al., 2025) and verification capabilities when engaged in the long chain of thoughts (Wei et al., 2023; Li et al., 2025).

Regarding the **human-like expressions** in the chain of thoughts exhibited by LRMs, several studies have conducted preliminary investigations from the perspective of Reinforcement Learning training dynamics (Gandhi et al., 2025; Yang et al., 2025; Yu et al., 2025b). And in terms of the LRM's ability to **assess itself's uncertainty** or engage in self-reflection, existing research has explored both explicit and implicit ways to estimate the uncertainty. For explicit uncertainty, prior work proposed prompting strategies that guide LRMs to verbalize their confidence levels (Zeng et al., 2025). To study implicit uncertainty, researchers have trained probing classifiers on the model's internal representations to estimate its confidence (Zhang et al., 2025; Anthropic). However, there is still a lack of sufficient interpretability research exploring whether these explicit reflection patterns observed in the chain of thoughts genuinely correlate with the models' actual internal reflective capabilities.

2.2 Linear semantic features

Recent investigations in model interpretability have revealed that, for numerous cognitive behaviors observed in Large Language Models—including refusal to answer (Arditi et al., 2024), jailbreaking (Yu et al., 2025a), reasoning, and knowledge-recall (Hong et al., 2025)—the models encode corresponding linear semantic features within their activation space (Park et al., 2024). These linear semantic features have been discovered and extracted by contrasting inputs that differ primarily in the target semantic dimension (Marks and Tegmark, 2023). Once these features are pinpointed, they offer a mechanism for controlling model behavior through manipulation, which allows for targeted interventions in the generative process (Rimsky et al., 2024; Stickland et al., 2024). Our work extends this line of study by identifying linear features that determine models' engagement in self-reflection.

3 Self-Reflection Features Extraction

3.1 Methodology for Identifying Self-Reflection Features

For current Reasoning Models, given a question Q, we can decompose its output into multiple *Reasoning Segments*: $\{s_1, s_2, s_3, \ldots, s_n\}$. Each segment (except for s_1) represents the model's reflection on the previous segment's proposed approach and a new attempt at solving the target problem. The final segment, s_n, signifies the termination of reflection, and the model directly provides its final answer.

At each *Reasoning Segment*'s final token position during inference, the model can either select the current answer as its final output and terminate, or generate another segment to reflect, verify, and explore alternative solutions. Therefore, based on whether the model initiates a new reflection after a segment or directly provides the final answer, we can categorize the *Reasoning Segments* into two groups. The first group, where the model proposes a new reflection after the segment, we call $\mathcal{S}_{\textit{Check-point}}$. The second group, where the model directly gives the final answer after the segment, we call $\mathcal{S}_{\textit{Termination}}$.

Next, for both groups, we extract the hidden states from the last-token position of each segment s (excluding s_n) at the model's l-th layer[1], denoted as $h^{(l)}(s)$. Since this last-token position corresponds to where the model is about to generate the first token of the subsequent segment, we hypothesize that the hidden states at this crucial juncture store important information guiding the model's decision to either continue with reflection or proceed to termination in the next segment. Then, using the *difference-in-means* technique (Marks and Tegmark, 2023; Rimsky et al., 2024), we calculate the difference between the mean last-token hidden states for these two categories of *Reasoning Segments*:

$$\mathbf{f}^{(l)} = \frac{\sum_{s \in \mathcal{S}_{\textit{Check-point}}} \mathbf{h}^{(l)}(s)}{|\mathcal{S}_{\textit{Check-point}}|} - \frac{\sum_{s \in \mathcal{S}_{\textit{Termination}}} \mathbf{h}^{(l)}(s)}{|\mathcal{S}_{\textit{Termination}}|} \quad (1)$$

[1]Assuming the model has L layers, we conduct experiments on each individual layer of it.

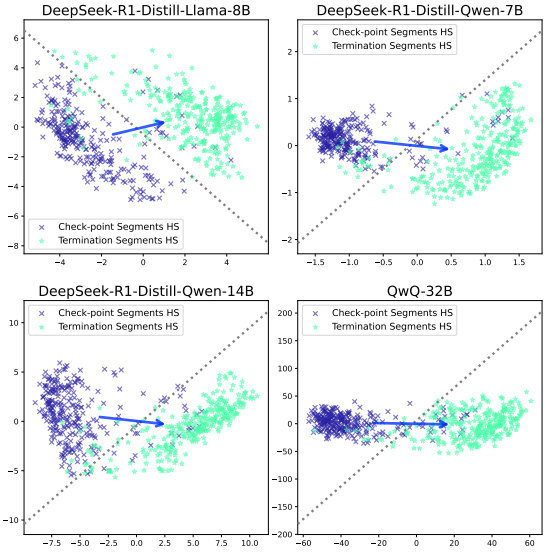

Figure 1: Visualization of the hidden states of four reasoning models on the GSM8k dataset using 2-D PCA. The hidden states of datapoints in $\mathcal{S}_{Check\text{-}point}$ and $\mathcal{S}_{Termination}$ are positioned around the boundary (grey dashed line) fitted via logistic regression. The blue arrow approximately indicates the direction of the Self-Reflection Features. Results on other datasets are shown in §B of the Appendix.

The direction of the vector $\mathbf{f}^{(l)}$ represents the direction of the Self-Reflection Features that we extracted. The construction details of $\mathcal{S}_{Check\text{-}point}$ and $\mathcal{S}_{Termination}$ are provided in the next section.

3.2 Experimental Setups

Models We utilize two categories of reasoning models trained under different settings to investigate Self-Reflection Features. The first category includes DeepSeek-R1-Distill-Llama-8B, DeepSeek-R1-Distill-Qwen-7B, and DeepSeek-R1-Distill-Qwen-14B. These models are obtained by performing supervised fine-tuning on the base Llama-3.1 (Meta, 2024) or Qwen2.5 (Qwen et al., 2025) models using high-quality reasoning data generated by the DeepSeek-R1 model (Guo et al., 2025). The second category comprises the QwQ-32B model (Team, 2025), which is trained using reinforcement learning. The inference details are provided in §D.

Datasets We focus on analyzing LRMs on mathematical and coding tasks to facilitate the extraction of Self-Reflection Features and analyze their influence. For the mathematical tasks, we use the GSM8k (Cobbe et al., 2021) and MATH-500 datasets (Lightman et al., 2023). For the coding tasks, we select the MBPP dataset (Austin et al., 2021).

3.3 Visualization for Self-Reflection Features

Following the methodology outlined in §3.1, we first perform inference on the datasets using the target reasoning models to collect their responses. We then employ GPT-4o (OpenAI et al., 2024a) to automatically segment each response into multiple reasoning segments[2], where each segment independently represents an attempt by the model to solve the problem. Subsequently, based on the segmentation results, we categorize the segments into two groups, $\mathcal{S}_{Check\text{-}point}$ and $\mathcal{S}_{Termination}$. We then extract the hidden states from the corresponding positions and compute the Self-Reflection Features by applying Eq. (1). To more clearly visualize the direction of the Self-Reflection Features, we apply Principal Component Analysis (PCA) to the hidden states of data points in the $\mathcal{S}_{Check\text{-}point}$ and $\mathcal{S}_{Termination}$ sets. The results on GSM8k dataset are shown in Figure 1. From this, we can observe that the two groups of data points are clearly divisible into two clusters by the logistic regression line, explicitly revealing the presence of Self-Reflection Features.

4 Internal and External Causal Analysis of Self-Reflection Features in LRMs

In this section, we will investigate from both internal (model parameter activation) and external (runtime behavior) perspectives, to verify the genuine causal relationships connecting Self-Reflection Features with: (a) the presence of reflection words within chain-of-thought processes in §4.1, and (b) the intensity of the model's self-reflection during actual inference in §4.2.

4.1 Parameter Storing Human-like Reflection words

By applying the Logit Lens method (nostalgebraist, 2020)[3], we identified value vectors within the MLP module's value matrix of the large reasoning models that highly contain these reflection tokens. Specific examples are presented in Table 1. We can observe that the vector projections at corresponding positions in both models each contain a certain number of reflection tokens. Moreover, when the hidden states are in $\mathcal{S}_{Check\text{-}point}$ — that is, when they exhibit stronger self-reflection features — we ob-

[2]The exact prompts used, along with human verification results, are provided in §A of the appendix.

[3]More descriptions of this method and relevant background knowledge are provided in §C of the appendix.

Model	Example Vector	Top-scoring tokens	Activation values in $\mathcal{S}_{Termination}$	Activation values in $\mathcal{S}_{Check\text{-}point}$	Activation values $\mathcal{S}_{More\text{-}Reflection}$
DeepSeek-R1-Distill-Llama-8B	$\mathbf{v}^{31}_{10644}$	`Is, Let, OK, So, Next, If, Now, What,First, See, However, Like, Check, Right, Wait, Again`	0.15	**2.13** ↑2.0	**2.71** ↑2.6
DeepSeek-R1-Distill-Qwen-7B	$\mathbf{v}^{23}_{11862}$	`hi, well, its, hey, nah, its, Im, ye, alternative, _ok,, oh, Hello, notifies, thanks, Ye,waits, WAIT`	0.22	**1.94** ↑1.7	**2.40** ↑2.2

Table 1: Example value vectors identified in two LRMs via the Logit Lens method, showcasing top-scoring tokens related to Self-Reflection Features and their activation values in different reflection stages.

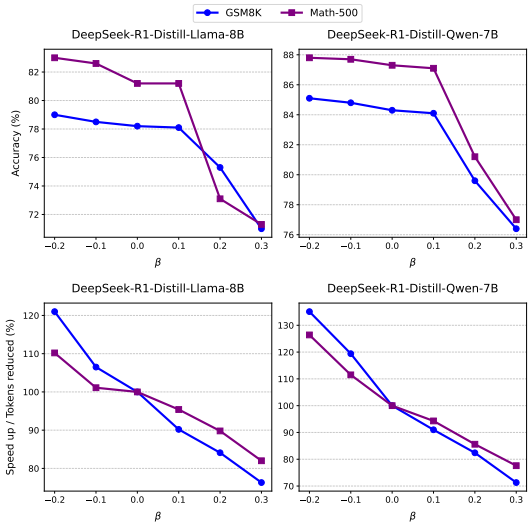

Figure 2: Accuracy and acceleration on GSM8K and MATH-500 after intervening in the hidden states of DeepSeek-R1-Distill-Llama-8B and DeepSeek-R1-Distill-Qwen-7B using different β values in Eq. (2). Acceleration is measured as the percentage reduction in inference tokens

serve a significant increase in the activation[4] of the target value vector. When we follow Eq. (2) to further enhance the strength of self-reflection features in the model representations (i.e., transitioning to $\mathcal{S}_{More\text{-}Reflection}$ as shown in Table 2), we similarly observe a further increase in their activation.

This provides further support for the idea that the human-like reflection words appearing in the chain-of-thought processes of LRMs are not merely a result of memorizing training data. Instead, they are **a reflection of the genuine activation of Self-Reflection Features.**

4.2 Modulating Self-Reflection Intensity via Linear Feature Intervention

Building upon the Linear Reflection Features extracted from LRMs in §3, in this part, we explore their application in adjusting the intensity of the model's self-reflection ability. Specifically, we aim to modulate their intensity within the model's representational space during inference for specific tasks, thereby addressing the potential issues of insufficient reflection (Aggarwal and Welleck, 2025) or "overthinking" (Cuadron et al., 2025; Zhang et al., 2025) that current LRMs may exhibit in practical scenarios.

Specifically, we follow the Eq. (2) below, attempting to control the extent of Self-reflection ability in LRMs by intervening with Self-reflection Features through adjusting the hyperparameter β in the model's hidden states:

$$\mathbf{h}'^{(l)}(s) \leftarrow \mathbf{h}^{(l)}(s) - \beta * \mathbf{r}^{(l)} \quad (2)$$

Figure 2 shows the intervention effects on GSM8K and MATH-500 for both DeepSeek-R1-Distill-Llama-8B and DeepSeek-R1-Distill-Qwen-7B. Starting from $\beta = 0$, increasing the value of β leads to improved acceleration for the LRMs. However, accuracy does not immediately degrade. Once β reaches around 0.2, further acceleration comes at the cost of a noticeable drop in accuracy. Conversely, decreasing β—thereby increasing the influence of Self-Reflection Features—slightly improves accuracy at the expense of slower inference, which aligns with our hypothesis.

5 Discussion and Conclusion

This work aimed to determine if reflective language in Large Reasoning Models reflects genuine internal processes or learned patterns. By extracting Self-Reflection Features from their representation space, we found a causal link to reflective words in their chain of thought. Crucially, manipulating these features allowed us to modulate LRM self-reflection intensity. These findings confirm that LRMs' reflective statements stem from discernible, governable internal mechanisms, signifying true reflective activation.

[4]Activation refers to the coefficient corresponding to each value vector in Eq. (4).

Limitations

In this study, while we have identified the presence of Self-Reflection features within reasoning models, a comprehensive investigation into their **origins** was not conducted. Specifically, it remains to be clarified whether these features emerge primarily from the pre-training phase or are introduced during subsequent reinforcement learning post-training. Furthermore, the characteristics of the training data that facilitate the encoding of these Self-Reflection features into the model's representational space are yet to be identified. A deeper understanding of these aspects would provide a critical foundation for the future development of more robust and effective reasoning models. We plan to explore these questions in our future work.

Additionally, owing to resource constraints, we were unable to extend our experimental research to larger-scale reasoning models, such as DeepSeek-R1.

References

Pranjal Aggarwal and Sean Welleck. 2025. L1: Controlling how long a reasoning model thinks with reinforcement learning. *Preprint*, arXiv:2503.04697.

Anthropic. On the biology of a large language model. https://transformer-circuits.pub/2025/attribution-graphs/biology.html.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. 2024. Refusal in language models is mediated by a single direction. *arXiv preprint arXiv:2406.11717*.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. 2021. Program synthesis with large language models. *Preprint*, arXiv:2108.07732.

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

Alejandro Cuadron, Dacheng Li, Wenjie Ma, Xingyao Wang, Yichuan Wang, Siyuan Zhuang, Shu Liu, Luis Gaspar Schroeder, Tian Xia, Huanzhi Mao, Nicholas Thumiger, Aditya Desai, Ion Stoica, Ana Klimovic, Graham Neubig, and Joseph E. Gonzalez. 2025. The danger of overthinking: Examining the reasoning-action dilemma in agentic tasks. *Preprint*, arXiv:2502.08235.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

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

Mor Geva, Avi Caciularu, Guy Dar, Paul Roit, Shoval Sadde, Micah Shlain, Bar Tamir, and Yoav Goldberg. 2022a. LM-debugger: An interactive tool for inspection and intervention in transformer-based language

models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 12–21, Abu Dhabi, UAE. Association for Computational Linguistics.

Mor Geva, Avi Caciularu, Kevin Wang, and Yoav Goldberg. 2022b. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 30–45, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 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. *Preprint*, arXiv:2103.03874.

Yihuai Hong, Dian Zhou, Meng Cao, Lei Yu, and Zhijing Jin. 2025. The reasoning-memorization interplay in language models is mediated by a single direction. *Preprint*, arXiv:2503.23084.

Zhong-Zhi Li, Duzhen Zhang, Ming-Liang Zhang, Jiaxin Zhang, Zengyan Liu, Yuxuan Yao, Haotian Xu, Junhao Zheng, Pei-Jie Wang, Xiuyi Chen, Yingying Zhang, Fei Yin, Jiahua Dong, Zhiwei Li, Bao-Long Bi, Ling-Rui Mei, Junfeng Fang, Zhijiang Guo, Le Song, and Cheng-Lin Liu. 2025. From system 1 to system 2: A survey of reasoning large language models. *Preprint*, arXiv:2502.17419.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. *Preprint*, arXiv:2305.20050.

Zichen Liu, Changyu Chen, Wenjun Li, Tianyu Pang, Chao Du, and Min Lin. 2025. There may not be aha moment in r1-zero-like training — a pilot study. https://oatllm.notion.site/oat-zero. Notion Blog.

Man Luo, Shrinidhi Kumbhar, Ming shen, Mihir Parmar, Neeraj Varshney, Pratyay Banerjee, Somak Aditya, and Chitta Baral. 2024. Towards logiglue: A brief survey and a benchmark for analyzing logical reasoning capabilities of language models. *Preprint*, arXiv:2310.00836.

Samuel Marks and Max Tegmark. 2023. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *arXiv preprint arXiv:2310.06824*.

Meta. 2024. Introducing llama 3.1: Our most capable models to date. https://ai.meta.com/blog/meta-llama-3-1/. Accessed: 2024-07-23.

nostalgebraist. 2020. Interpreting gpt: The logit lens. https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens. LessWrong.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub

Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunninghman, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. 2024a. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

OpenAI, :, Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, Ally Bennett, Ananya Kumar, Andre Saraiva, Andrea Vallone, Andrew Duberstein, Andrew Kondrich, Andrey Mishchenko, Andy Applebaum, Angela Jiang, Ashvin Nair, Barret Zoph, Behrooz Ghorbani, Ben Rossen, Benjamin Sokolowsky, Boaz Barak, Bob McGrew, Borys Minaiev, Botao Hao, Bowen Baker, Brandon Houghton, Brandon McKinzie, Brydon Eastman, Camillo Lugaresi, Cary Bassin, Cary Hudson, Chak Ming Li, Charles de Bourcy, Chelsea Voss, Chen Shen, Chong Zhang, Chris Koch, Chris Orsinger, Christopher Hesse, Claudia Fischer, Clive Chan, Dan Roberts, Daniel Kappler, Daniel Levy, Daniel Selsam, David Dohan, David Farhi, David Mely, David Robinson, Dimitris Tsipras, Doug Li, Dragos Oprica, Eben Freeman, Eddie Zhang, Edmund Wong, Elizabeth Proehl, Enoch Cheung, Eric Mitchell, Eric Wallace, Erik Ritter, Evan Mays, Fan Wang, Felipe Petroski Such, Filippo Raso, Florencia Leoni, Foivos Tsimpourlas, Francis Song, Fred von Lohmann, Freddie Sulit, Geoff Salmon, Giambattista Parascandolo, Gildas Chabot, Grace Zhao, Greg Brockman, Guillaume Leclerc, Hadi Salman, Haiming Bao, Hao Sheng, Hart Andrin, Hessam Bagherinezhad, Hongyu Ren, Hunter Lightman, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian Osband, Ignasi Clavera Gilaberte, Ilge Akkaya, Ilya Kostrikov, Ilya Sutskever, Irina Kofman, Jakub Pachocki, James Lennon, Jason Wei, Jean Harb, Jerry Twore, Jiacheng Feng, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joaquin Quiñonero Candela, Joe Palermo, Joel Parish, Johannes Heidecke, John Hallman, John Rizzo, Jonathan Gordon, Jonathan Uesato, Jonathan Ward, Joost Huizinga, Julie Wang, Kai Chen, Kai Xiao, Karan Singhal, Karina Nguyen, Karl Cobbe, Katy Shi, Kayla Wood, Kendra Rimbach, Keren Gu-Lemberg, Kevin Liu, Kevin Lu, Kevin Stone, Kevin Yu, Lama Ahmad, Lauren Yang, Leo Liu, Leon Maksin, Leyton Ho, Liam Fedus, Lilian Weng, Linden Li, Lindsay McCallum, Lindsey Held, Lorenz Kuhn, Lukas Kondraciuk, Lukasz Kaiser, Luke Metz, Madelaine Boyd, Maja Trebacz, Manas Joglekar, Mark Chen, Marko Tintor, Mason Meyer, Matt Jones, Matt Kaufer, Max Schwarzer, Meghan Shah, Mehmet Yatbaz, Melody Y. Guan, Mengyuan Xu, Mengyuan Yan, Mia Glaese, Mianna Chen, Michael Lampe, Michael

Malek, Michele Wang, Michelle Fradin, Mike McClay, Mikhail Pavlov, Miles Wang, Mingxuan Wang, Mira Murati, Mo Bavarian, Mostafa Rohaninejad, Nat McAleese, Neil Chowdhury, Neil Chowdhury, Nick Ryder, Nikolas Tezak, Noam Brown, Ofir Nachum, Oleg Boiko, Oleg Murk, Olivia Watkins, Patrick Chao, Paul Ashbourne, Pavel Izmailov, Peter Zhokhov, Rachel Dias, Rahul Arora, Randall Lin, Rapha Gontijo Lopes, Raz Gaon, Reah Miyara, Reimar Leike, Renny Hwang, Rhythm Garg, Robin Brown, Roshan James, Rui Shu, Ryan Cheu, Ryan Greene, Saachi Jain, Sam Altman, Sam Toizer, Sam Toyer, Samuel Miserendino, Sandhini Agarwal, Santiago Hernandez, Sasha Baker, Scott McKinney, Scottie Yan, Shengjia Zhao, Shengli Hu, Shibani Santurkar, Shraman Ray Chaudhuri, Shuyuan Zhang, Siyuan Fu, Spencer Papay, Steph Lin, Suchir Balaji, Suvansh Sanjeev, Szymon Sidor, Tal Broda, Aidan Clark, Tao Wang, Taylor Gordon, Ted Sanders, Tejal Patwardhan, Thibault Sottiaux, Thomas Degry, Thomas Dimson, Tianhao Zheng, Timur Garipov, Tom Stasi, Trapit Bansal, Trevor Creech, Troy Peterson, Tyna Eloundou, Valerie Qi, Vineet Kosaraju, Vinnie Monaco, Vitchyr Pong, Vlad Fomenko, Weiyi Zheng, Wenda Zhou, Wes McCabe, Wojciech Zaremba, Yann Dubois, Yinghai Lu, Yining Chen, Young Cha, Yu Bai, Yuchen He, Yuchen Zhang, Yunyun Wang, Zheng Shao, and Zhuohan Li. 2024b. Openai o1 system card. *Preprint*, arXiv:2412.16720.

Kiho Park, Yo Joong Choe, and Victor Veitch. 2024. The linear representation hypothesis and the geometry of large language models. *Preprint*, arXiv:2311.03658.

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, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. 2024. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15504–15522, Bangkok, Thailand. Association for Computational Linguistics.

Asa Cooper Stickland, Alexander Lyzhov, Jacob Pfau, Salsabila Mahdi, and Samuel R Bowman. 2024. Steering without side effects: Improving post-deployment control of language models. *arXiv preprint arXiv:2406.15518*.

Qwen Team. 2025. Qwq-32b: Embracing the power of reinforcement learning. Accessed: 2025-03-06.

Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, D. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, A. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Johannes Welbl, Nelson F. Liu, and Matt Gardner. 2017. Crowdsourcing multiple choice science questions. *Preprint*, arXiv:1707.06209.

Shu Yang, Junchao Wu, Xin Chen, Yunze Xiao, Xinyi Yang, Derek F. Wong, and Di Wang. 2025. Understanding aha moments: from external observations to internal mechanisms. *Preprint*, arXiv:2504.02956.

Lei Yu, Virginie Do, Karen Hambardzumyan, and Nicola Cancedda. 2025a. Robust LLM safeguarding via refusal feature adversarial training. In *The Thirteenth International Conference on Learning Representations*.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Weinan Dai, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. 2025b. Dapo: An open-source llm reinforcement learning system at scale. *Preprint*, arXiv:2503.14476.

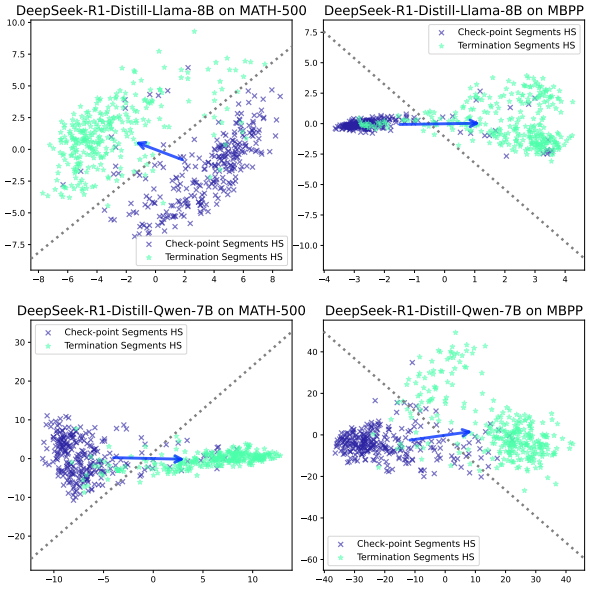

Figure 3: Visualization of the hidden states of two reasoning models on the MATH-500 and MBPP dataset using 2-dimensional PCA. The hidden states of datapoints in $\mathcal{S}_{\textit{Check-point}}$ and $\mathcal{S}_{\textit{Termination}}$ are positioned around the boundary (grey dashed line) fitted via logistic regression. The blue arrow approximately indicates the direction of the Self-Reflection Features. To make the image presentation clearer, we sampled 300 data points from each of $\mathcal{S}_{\textit{Check-point}}$ and $\mathcal{S}_{\textit{Termination}}$ for presentation. Results on other datasets are shown in §B of the Appendix.

Qingcheng Zeng, Weihao Xuan, Leyang Cui, and Rob Voigt. 2025. Do reasoning models show better verbalized calibration? *Preprint*, arXiv:2504.06564.

Anqi Zhang, Yulin Chen, Jane Pan, Chen Zhao, Aurojit Panda, Jinyang Li, and He He. 2025. Reasoning models know when they're right: Probing hidden states for self-verification. *Preprint*, arXiv:2504.05419.

A Prompt used for Reasoning Responses Segmentation

Table 2 presents the prompt we used to query GPT-4o to segment the model's responses into *Reasoning Segments*.

B Additional PCA Visualizations on MATH-500 and MBPP

Here, we present additional PCA results for the Self-Reflection Features on MATH-500 and MBPP in Figure 3.

C Foundations for Logit Lens Analysis

Here, we provide a brief introduction to the Logit Lens method and the background knowledge it involves.

C.1 MLP in Transformers

In transformer-based language models, the MLP is a crucial component for storing the model's factual knowledge, and its sub-layers can be viewed as key-value memories (Geva et al., 2021). To be specific, the first layer[5] of MLP sublayers can be viewed as a matrix W_K formed by key vectors $\{\mathbf{k}_1, \mathbf{k}_2, \ldots, \mathbf{k}_n\}$, used to capture a set of patterns in the input sequence, and ultimately outputting the coefficient scores. The second layer can be viewed as a matrix W_V formed by value vectors $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$, with each value vector containing the corresponding factual knowledge.

Formally, the output of the MLP in the transformer's ℓ-th layer, given an input hidden state $\mathbf{x}^\ell$, can be defined as:

$$\mathbf{M}^\ell = f\big(W_K^\ell \cdot \gamma(\mathbf{x}^\ell + \mathbf{A}^\ell)\big) W_V^\ell = \mathbf{m}^\ell W_V^\ell, \quad (3)$$

where $W_K^\ell, W_V^\ell \in \mathbb{R}^{n \times d}$. The function f and γ represent a non-linearity[6] and layer normalization, respectively. In the transformer's ℓ-th layer, $\mathbf{m}^\ell \in \mathbb{R}^n$ denotes the coefficient scores, and $\mathbf{A}^\ell$ represents the output of the attention component. The hidden state dimension is d, while the intermediate MLP has a dimension of n. Then, by denoting $\mathbf{v}_j^\ell$ as the j-th column (which will be called the value vector or parameter vector in the following sections) of W_V^ℓ and m_j^ℓ as the j-th element in the coefficients produced by the first layer of the MLP, we can view MLP's output $\mathbf{M}^\ell$ as a linear combination of the value vectors in W_V^ℓ, with their corresponding coefficients $\mathbf{m}^\ell$:

$$\mathbf{M}^\ell = \sum_{j=1}^{n} m_j^\ell \mathbf{v}_j^\ell, \quad (4)$$

Each m_j^ℓ here also represents the activation value of the value vector we mentioned in Table 1. Finally, the hidden states at the ℓ-th layer of the language model can be defined as:

$$X^{\ell+1} = X^\ell + \mathbf{M}^\ell + \mathbf{A}^\ell, \quad (5)$$

where X^ℓ, $\mathbf{M}^\ell$ and $\mathbf{A}^\ell$ represent the hidden states, MLP's output, and the attention component's output in the transformer's ℓ-th layer, respectively.

[5] In most decoder-only models, such as GPT-2 (Radford et al., 2019) and GPT-J (**?**), the MLP component consists of two layers, whereas in LLaMA (Touvron et al., 2023), it comprises three layers. However, we can still regard LLaMA's first two layers collectively as the key matrices, with their output representing the coefficient scores.

[6] For brevity, the bias term is omitted.

Prompt

- **Analyze the model response and divide into reasoning segments. Return:**

  1. Labeled segments with independent solution attempts
  2. Each segment must include:
     - A new full solution pathway
     - Alternative interpretations (if applicable)
     - Verification/error-checking steps (if applicable)

  Format Requirements:

  - Use **Segment N** headers
  - Mutual exclusivity between segments
  - Avoid single-step fragmentation

  Examples:

  Problem: "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?"

  Model Response: "[Full model response here...]"

  Segmentation:

  1. **Standard Calculation**:
     Segment 1: Direct arithmetic approach
     - April: 48 clips (given)
     - May: 48/2 = 24 clips
     - Total: 48 + 24 = 72
     - Verification: 40 + 20 + 8 + 4 = 72
  2. **Algebraic Reformulation**:
     Segment 2: Symbolic representation
     - Let $A = 48$ (April sales)
     - Define $M = A/2$ (May sales)
     - Total $T = A + M = 1.5A$
     - Compute $1.5 \times 48 = 72$
  3. **Semantic Analysis**:
     Segment 3: Ambiguity resolution
     - Challenge: "sold to friends" interpretation
     - Reject per-friend vs. total sales hypotheses
     - Confirm 48 = total clips (not friends count)

  Current Problem: {Problem}

  Current Model Response: {Response}

  Segmentation:

Table 2: Prompt for segmenting mathematical reasoning processes.

C.2 Logit Lens

nostalgebraist (2020); Geva et al. (2021) proposed that the hidden states or module parameters of a transformer-based model can be directly decoded into the vocabulary space using the model's pretrained unembedding matrix, enabling an investigation into the information they encode:

$$Projection = E\mathbf{v}_j^\ell, \tag{6}$$

Here, E denotes the model's pretrained unembedding matrix, and the result of the projection, which lies in $\mathbb{R}^{|\mathcal{V}|}$, is a vector assigning a score to each token in the vocabulary $\mathcal{V}$. The set of the top-k highest-scoring tokens in this projection, denoted by $\mathcal{T}_{j,k}^\ell$, often reveals a clear pattern that corresponds to a specific knowledge being promoted by $\mathbf{v}_j^\ell$ during inference (Geva et al., 2022b,a).

D Implementation Details of the LRMs

For the inference settings of all four Large Reasoning Models, we use a temperature of 0.6, a top-p value of 0.95, and set the maximum generation length to 32,768 tokens, following the default settings.

All the experiments in this work were conducted on four 80GB NVIDIA A800 GPUs.