
Supplementary Material

Anonymous Author(s)

Affiliation

Address

email

A Experimental Details

We primarily utilize open-sourced models to conduct experiments in this work. Given that DeepSeek-R1 is one of the most widely adopted reasoning models, and its authors have released a series of distilled models based on R1 [3], including both the specified base and finetuned reasoning models, we adopt their configurations in our study. Specifically, we use the DeepSeek-R1-Distill-Qwen [3] models with sizes of 1.5B, 7B, 14B, 32B and 70B as our reasoning models, and select Qwen2.5-Math-1.5B, 7B [10], LLaMA3.1-8B [2], Qwen2.5-14B, 32B [9] or Llama-3.3-70B-Instruct [2] as base models. All models are loaded and run using the Transformers library [8].

Our evaluation framework is based on the lm-evaluation-harness package [1]. To accelerate inference, we use vLLM [4] as the backend, which may slightly affect performance due to backend-specific optimizations. In the Merge Stethoscope experiments, we observe that the “chat” interface often generates irrelevant or nonsensical responses, while the “generate” interface produces coherent and contextually appropriate outputs. We suspect this discrepancy arises from misinterpreted system prompts. Therefore, we rely on the “generate” interface and implement a custom evaluation toolkit.

For the Freeze Stethoscope experiments, we build on the codebase of s1[5]. We use a learning rate of 1e-5, weight decay of 1e-4, a batch size of 16, and train for 5 epochs. Due to hardware limitations (i.e., lack of access to 16 H100 GPUs), we leverage DeepSpeed[7] with ZeRO Stage 3[6] to enable efficient training. The base model used here is Qwen2.5-32B-Instruct[9]. Evaluation is again conducted with lm-evaluation-harness, following the modified pipeline by the authors of s1, which disables generation of the end-of-thinking token and optionally appends the string “Wait” to the reasoning trace to encourage model reflection. We adopt the Budget Forcing “Wait” ×2 as our default testing configuration.

All visualization and inference experiments on 1.5B–14B models are conducted on a single NVIDIA A100 GPU. For training and evaluating 32B-70B models, we use a cluster of 8 NVIDIA A100 GPUs. Training typically takes around 6 hours, while testing on a single dataset usually requires about 2 hours.

B More Experimental Results

In the main paper, we present visualization results for the 1.5B, 14B, and 32B models. Here, we supplement those results by providing additional visualizations for the 7B, 8B, and 70B models. Following the Delta Stethoscope pipeline, we visualize both the absolute weight shift $|w_X(B) - w_X(A)|_{\ell_2}$ and the relative weight shift $\frac{w_X(B) - w_X(A)}{w_X(A)}$. The absolute weight shifts are shown in Figure 1, and the relative weight shifts are presented in Figure 2. The trends observed in the main paper remain consistent across these additional models. Notably, o_proj consistently exhibits the largest weight shift, with the effect being especially pronounced in the 70B model. Moreover, o_proj is the only module that displays a bimodal distribution in the relative weight shift.

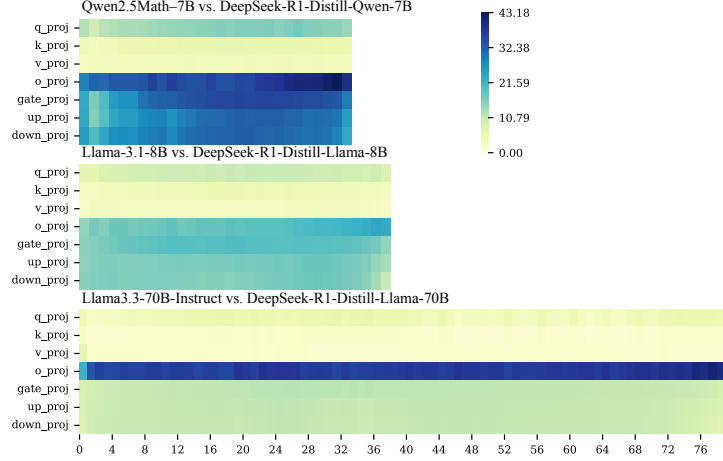


Figure 1: **Per-module L2 distance of linear weights between models A and B .** Notably, the `o_proj` module shows the largest in 7B, 8B and 70B models, highlighting its potential importance for reasoning.

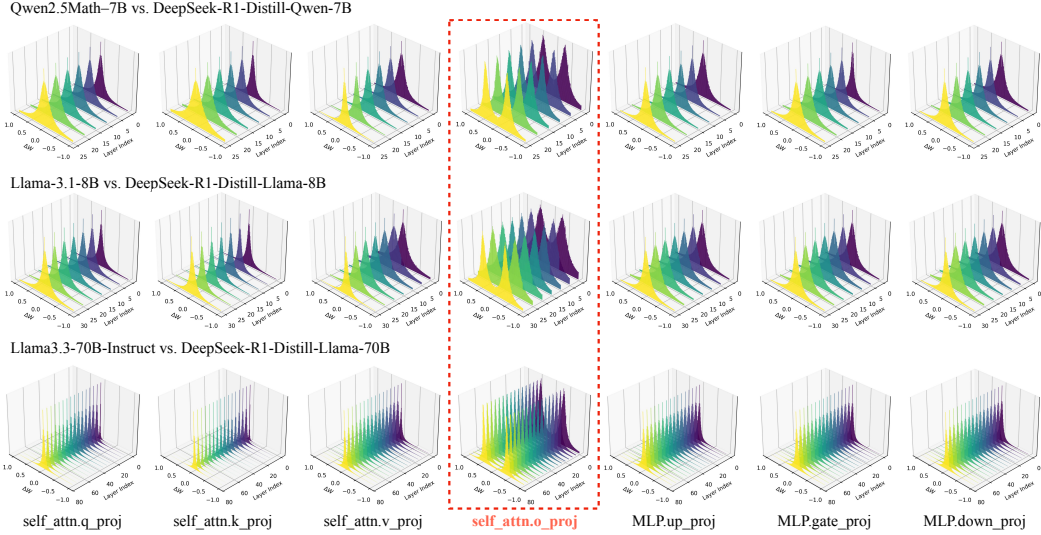


Figure 2: **Layer-wise distribution of relative weight changes between models A and B .** While most modules display a unimodal distribution, the `o_proj` module uniquely exhibits a bimodal distribution, highlighting its distinctive behavior.

36 C Statistical Significance and Broader Impacts

37 We report appropriate information regarding the statistical significance of our experiments. While we
 38 do not primarily focus on classical significance tests such as p-values, we provide multiple forms of
 39 empirical evidence—such as consistent module-specific weight shifts, response-level comparisons
 40 under controlled manipulations, and loss curves under different tuning strategies—that collectively
 41 establish the robustness of our findings. These analyses serve as a practical alternative to traditional
 42 error bars or confidence intervals and help substantiate our key claims.

43 This research has both promising benefits and important risks to consider. On the positive side, the
 44 proposed Stethoscope for Networks (SfN) framework provides a novel set of tools for interpreting
 45 LLMs, especially by localizing specific capabilities—such as reasoning—to individual components
 46 like the output projection (`o_proj`). These tools may significantly improve our understanding of LLMs,
 47 enabling more transparent, modular, and efficient model development. For instance, if reasoning
 48 abilities can be enhanced by tuning a small subset of parameters, it could greatly reduce computational
 49 costs and increase accessibility for developing domain-specific or lightweight models.

50 However, this line of work also carries potential risks. Precisely identifying and isolating reasoning-
 51 related components might lower the barrier for targeted manipulation, such as unauthorized transfer
 52 or removal of reasoning abilities across models. This could facilitate misuse scenarios, including
 53 capability extraction, tampering, or model theft. Furthermore, while the diagnostic methods proposed
 54 aim to support interpretability, there is a risk that they may be overinterpreted, leading to an inflated
 55 sense of model transparency that does not generalize across architectures or tasks.

56 References

- 57 [1] Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles
 58 Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac’h, Haonan Li, Kyle McDonell, Niklas
 59 Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron,
 60 Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. The
 61 language model evaluation harness, 07 2024.
- 62 [2] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian,
 63 Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama
 64 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.
- 65 [3] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu,
 66 Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in
 67 llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.
- 68 [4] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu,
 69 Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large lan-
 70 guage model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium*
 71 *on Operating Systems Principles*, 2023.
- 72 [5] Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi,
 73 Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple
 74 test-time scaling. *arXiv preprint arXiv:2501.19393*, 2025.
- 75 [6] Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. Zero: Memory optimiza-
 76 tions toward training trillion parameter models. In *SC20: International Conference for High*
 77 *Performance Computing, Networking, Storage and Analysis*, pages 1–16. IEEE, 2020.
- 78 [7] Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. Deepspeed: System
 79 optimizations enable training deep learning models with over 100 billion parameters. In
 80 *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery &*
 81 *data mining*, pages 3505–3506, 2020.
- 82 [8] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony
 83 Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer,
 84 Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain
 85 Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-
 86 art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods*
 87 *in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020.
 88 Association for Computational Linguistics.
- 89 [9] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan
 90 Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint*
 91 *arXiv:2412.15115*, 2024.
- 92 [10] An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng
 93 Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, et al. Qwen2. 5-math technical report: Toward
 94 mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*, 2024.