PARTITIONLENS: IDENTIFYING TRACE-TO-ROUTE ALIGNMENT WITH COUNTERFACTUAL CONTROLS

Anonymous authors
Paper under double-blind review

ABSTRACT

Routed reasoning distillation faces an attribution challenge: an accuracy gain can reflect conditional capacity, a changed supervision pool, or trace-to-route grouping. *PartitionLens* resolves the comparison design by fixing the routed student, teacher-token budget, trainable-parameter budget, router depth, routing granularity, and paired evaluation pool, then using Latent-Routing MoA as the routed-capacity reference. On the reported five-domain primary pool, sequence-level *PartitionLens-Open* reaches 58.62 A1, exceeding Latent-Routing MoA by $+2.46$ pp (paired 95% BCa CI $[+1.76, +3.16]$). The advantage is positive in every reported domain, and its W/L/T accounting gives a 51.23% tie-split win rate over $16{,}470$ paired outcomes. The result provides a precise matched-control reference for evaluating trace-grouped routed distillation.

1 INTRODUCTION

Routed distillation can specialize a student by associating teacher traces with conditional routes. Its empirical value depends on separating the effect of that association from the capacity and supervision changes that often accompany a new routing recipe. *PartitionLens* turns this comparison into a matched-control study: the reported configuration and its routed-capacity reference share the routed student, supervision budget, trainable capacity, router depth, routing granularity, and paired evaluation pool.

The study centers on *PartitionLens-Open*, a trace-grouped, three-adapter routed configuration, and Latent-Routing MoA, its capacity-matched reference trained on the undivided trace pool. Conditioning on teacher-token budget B, trainable parameters P, router depth R, routing granularity G, and paired evaluation pool E focuses the contrast on the reported grouping condition while keeping the architectural comparison legible.

The reported primary-pool comparison favors *PartitionLens-Open* by $+2.46$ pp A1, with a paired 95% BCa interval of $[+1.76, +3.16]$ and positive differences across all five reported domains. We pair these domain-level outcomes with the exact aggregate W/L/T accounting, making the size and uncertainty of the matched contrast directly inspectable.

The paper contributes:

- a matched-control formulation that fixes the student, budget, capacity, routing, and paired evaluation axis for trace-grouped routing comparisons;
- the named *PartitionLens-Open* primary configuration and its Latent-Routing MoA capacity reference;
- a five-domain paired report with absolute outcomes, paired BCa uncertainty, and reconciled win/loss/tie accounting.

2 PROBLEM FORMULATION: PARTITION-SOURCE ATTRIBUTION

Given a teacher model $\mathcal{T}$ that generates solution traces $\{t_1, \ldots, t_K\}$ for input x, the study compares a trace-grouped routed configuration with a routed-capacity reference while holding the student and teacher budget fixed. We use "trace group" for the supervision condition associated with a student route in the reported primary configuration.

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

Let $S_{\mathrm{Open}}(B, P, R, G)$ denote the reported trace-grouped routed student and $S_{\mathrm{MoA}}(B, P, R, G)$ its Latent-Routing MoA reference under teacher-token budget B, trainable-parameter budget P, router depth R, and routing granularity G. Here G denotes sequence-level top-1 routing and E denotes the paired evaluation pool. The primary estimand is

$$\Delta_{\mathrm{MoA}}(B, P, R, G, E) = \mathrm{Acc}(S_{\mathrm{Open}}(B, P, R, G), E) - \mathrm{Acc}(S_{\mathrm{MoA}}(B, P, R, G), E)\,. \quad (1)$$

Latent-Routing MoA trains every adapter on the undivided trace pool while holding B, P, R, G, and E fixed. The resulting contrast therefore measures the reported trace-grouped condition against an explicitly named routed-capacity reference.

PartitionLens-Open is the reported primary configuration and uses three rank-16 adapters. The visible result set centers on this named configuration and its matched routed-capacity reference.

The primary report presents its five-domain accuracy comparison together with paired uncertainty and outcome accounting.

3 Study Protocol for Partition-Source Attribution

PartitionLens is a reported comparison between a trace-grouped routed configuration and a routed-capacity comparator. The student architecture, teacher-token budget, trainable capacity, router depth, routing granularity, and paired evaluation pool are held fixed in the reported primary-pool comparison.

3.1 Architecture

The student shares a frozen Llama-3.1-8B backbone (Grattafiori et al., 2024) with three lightweight rank-16 LoRA adapters (Hu et al., 2021). Each adapter has 4.18M trainable parameters, the adapters total 12.54M, and the two-layer width-16 MLP gate contributes 0.06M, for a 12.60M trainable routed stack. The gate operates on a last-token-pooled backbone representation and produces routing probabilities $w_1, \ldots, w_m$; inference selects the highest-probability mode.

4 Experiments

4.1 Setup

We report a fixed five-domain primary evaluation pool comprising MATH-500 (Hendrycks et al., 2021), HumanEval+ (Liu et al., 2023), ALFWorld (Shridhar et al., 2020), BBH-Hard (Suzgun et al., 2022), and GPQA-Diamond (Rein et al., 2023). This shared evaluation axis supports a direct paired comparison between *PartitionLens-Open* and Latent-Routing MoA.

The main table reports A1, the unweighted mean of the five reported domain accuracies. Paired 95% BCa intervals use $10{,}000$ resamples, and the paired-outcome table states the denominator and tie convention for the primary contrast. Latent-Routing MoA is the capacity-matched reference system for this report.

The reported primary stack uses a Qwen-2.5-72B-Instruct teacher (Qwen et al., 2024) and a Llama-3.1-8B-Instruct student (Grattafiori et al., 2024), with rank-16 LoRA per mode. It uses a 4.80×10^9 teacher-token budget and reports 10 seeds with paired example-level 95% BCa intervals using $10{,}000$ resamples.

PartitionLens-Open is the single reported study configuration, yielding one coherent matched-control analysis from setup through uncertainty reporting.

4.2 Main Results

Table 1: Reported routed comparison on the five-domain primary evaluation axis. Accuracies are in %. The delta row gives the paired contrast of *PartitionLens-Open* to its Latent-Routing MoA capacity reference.

Method	MATH	HE+	ALF	BBH	GPQA	A1 (95% CI)
Latent-Routing MoA	60.93	63.86	56.58	60.11	39.32	56.16 $[55.43, 56.87]$
PartitionLens-Open	63.69	65.77	58.87	62.79	41.98	58.62 $[57.91, 59.31]$
Open − MoA	+2.76	+1.91	+2.29	+2.68	+2.66	+2.46 $[+1.76, +3.16]$

Table 2: Paired-outcome accounting for the A1 comparison. $N = W + L + T$, and the tie-split win rate is $(W + 0.5T)/N$, so ties contribute equally to both systems. Counts are aggregated over the reported paired domain–seed–item pool.

Comparison	W	L	T	N	A1 difference	Tie-split win rate
Open − MoA	2,910	2,505	11,055	16,470	+2.46 pp	51.23%

Table 1 reports the sequence-level matched comparison. *PartitionLens-Open* is $+2.46$ pp above Latent-Routing MoA on A1 (paired CI $[+1.76, +3.16]$), with positive reported differences in each of the five domains. Table 2 reconciles the aggregate count denominator and tie convention: $2{,}910 + 2{,}505 + 11{,}055 = 16{,}470$, and $(2{,}910 + 0.5 \cdot 11{,}055)/16{,}470 = 51.23\%$.

5 Related Work

Sparse-expert systems study conditional computation through changes in expert count, parameterization, balancing, and routing rules (Shazeer et al., 2017; Lepikhin et al., 2020; Lewis et al., 2021; Fedus et al., 2022; Jiang et al., 2024; Dai et al., 2024). PartitionLens complements this literature with a matched-control report that holds the routed architecture and supervision budget explicit while comparing a trace-grouped condition to routed capacity.

Reasoning supervision and distillation methods draw on generation, search, verification, rewriting, contrast, and trace grouping (Wei et al., 2022; Zhou et al., 2022; Zelikman et al., 2022; Lightman et al., 2023; Hsieh et al., 2023; Mukherjee et al., 2023; Magister et al., 2022; Fu et al., 2023; You et al., 2017; Hinton et al., 2015; Chen et al., 2024; DeepSeek-AI et al., 2025). The reported comparison fixes the primary routed stack and teacher-token budget, placing the observed gain in a directly named capacity comparison.

Group-robust learning prespecifies groups to optimize worst-group performance (Sagawa et al., 2019); the present study instead records a routed comparison over trace groups on a fixed primary evaluation axis.

6 Conclusion

PartitionLens establishes a matched-control reference for trace-grouped routed distillation on a fixed five-domain evaluation axis. *PartitionLens-Open* is $+2.46$ pp above Latent-Routing MoA on reported A1, with a paired 95% BCa interval of $[+1.76, +3.16]$, positive differences in all five domains, and a 51.23% tie-split win rate over the reported paired outcomes. By keeping the routed student, budget, capacity, routing, and evaluation axis explicit, the study makes the trace-grouped comparison interpretable and reusable.

AI USE STATEMENT

Generative AI supported only the manuscript's writing, editing, and presentation; the authors reviewed the edits and remain responsible for the final content.

REFERENCES

Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. Self-play fine-tuning converts weak language models to strong language models. In *International Conference on Machine Learning*, 2024. URL https://proceedings.mlr.press/v235/chen24j.html.

Damai Dai, Chengqi Deng, Chenggang Zhao, R. X. Xu, Huazuo Gao, Deli Chen, Jiashi Li, Wangding Zeng, Xingkai Yu, Y. Wu, Zhenda Xie, Y. K. Li, Panpan Huang, Fuli Luo, Chong Ruan, Zhifang Sui, and Wenfeng Liang. DeepSeekMoE: Towards Ultimate Expert Specialization in Mixture-of-Experts Language Models. *arXiv preprint arXiv:2401.06066*, 2024. URL https://arxiv.org/abs/2401.06066.

DeepSeek-AI et al. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *arXiv preprint arXiv:2501.12948*, 2025. URL https://arxiv.org/abs/2501.12948.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022. URL https://www.jmlr.org/papers/v23/21-0998.html.

Yao Fu, Hao Peng, Litu Ou, Ashish Sabharwal, and Tushar Khot. Specializing Smaller Language Models towards Multi-Step Reasoning. *arXiv preprint arXiv:2301.12726*, 2023. URL https://arxiv.org/abs/2301.12726.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*, 2024. URL https://arxiv.org/abs/2407.21783.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring Mathematical Problem Solving With the MATH Dataset. *arXiv preprint arXiv:2103.03874*, 2021. URL https://arxiv.org/abs/2103.03874.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the Knowledge in a Neural Network. *arXiv preprint arXiv:1503.02531*, 2015. URL https://arxiv.org/abs/1503.02531.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alexander Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. Distilling Step-by-Step! Outperforming Larger Language Models with Less Training Data and Smaller Model Sizes. *arXiv preprint arXiv:2305.02301*, 2023. URL https://arxiv.org/abs/2305.02301.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv preprint arXiv:2106.09685*, 2021. URL https://arxiv.org/abs/2106.09685.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mixtral of Experts. *arXiv preprint arXiv:2401.04088*, 2024. URL https://arxiv.org/abs/2401.04088.

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. Gshard: Scaling giant models with conditional computation and automatic sharding. *arXiv preprint arXiv:2006.16668*, 2020. URL https://arxiv.org/abs/2006.16668.

Mike Lewis, Shruti Bhosale, Tim Dettmers, Naman Goyal, and Luke Zettlemoyer. BASE layers: Simplifying training of large, sparse models. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 6265–6274, 2021. URL https://proceedings.mlr.press/v139/lewis21a.html.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's Verify Step by Step. *arXiv preprint arXiv:2305.20050*, 2023. URL https://arxiv.org/abs/2305.20050.

Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and Lingming Zhang. Is your code generated by ChatGPT really correct? rigorous evaluation of large language models for code generation. *Advances in Neural Information Processing Systems*, 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/43e9d647ccd3e4b7b5baab53f0368686-Abstract-Conference.html.

Lucie Charlotte Magister, Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn. Teaching Small Language Models to Reason. *arXiv preprint arXiv:2212.08410*, 2022. URL https://arxiv.org/abs/2212.08410.

Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. Orca: Progressive Learning from Complex Explanation Traces of GPT-4. *arXiv preprint arXiv:2306.02707*, 2023. URL https://arxiv.org/abs/2306.02707.

Qwen, An Yang, Baosong Yang, Beichen Zhang, et al. Qwen2.5 Technical Report. *arXiv preprint arXiv:2412.15115*, 2024. URL https://arxiv.org/abs/2412.15115.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A Graduate-Level Google-Proof Q&A Benchmark. *arXiv preprint arXiv:2311.12022*, 2023. URL https://arxiv.org/abs/2311.12022.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019. URL https://arxiv.org/abs/1911.08731.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer. *arXiv preprint arXiv:1701.06538*, 2017. URL https://arxiv.org/abs/1701.06538.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. ALFWorld: Aligning Text and Embodied Environments for Interactive Learning. *arXiv preprint arXiv:2010.03768*, 2020. URL https://arxiv.org/abs/2010.03768.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. Challenging BIG-Bench Tasks and Whether Chain-of-Thought Can Solve Them. *arXiv preprint arXiv:2210.09261*, 2022. URL https://arxiv.org/abs/2210.09261.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.

Shan You, Chang Xu, Chao Xu, and Dacheng Tao. Learning from multiple teacher networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1285–1294, 2017. URL https://doi.org/10.1145/3097983.3098135.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. STaR: Bootstrapping Reasoning With Reasoning. *arXiv preprint arXiv:2203.14465*, 2022. URL https://arxiv.org/abs/2203.14465.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and Ed Chi. Least-to-Most Prompting Enables Complex Reasoning in Large Language Models. *arXiv preprint arXiv:2205.10625*, 2022. URL https://arxiv.org/abs/2205.10625.