Interpretable Reasoning Enhancement in LLMs through Puzzle and Ontological Task Analysis

Research Proposal for PhD by Mihir Panchal

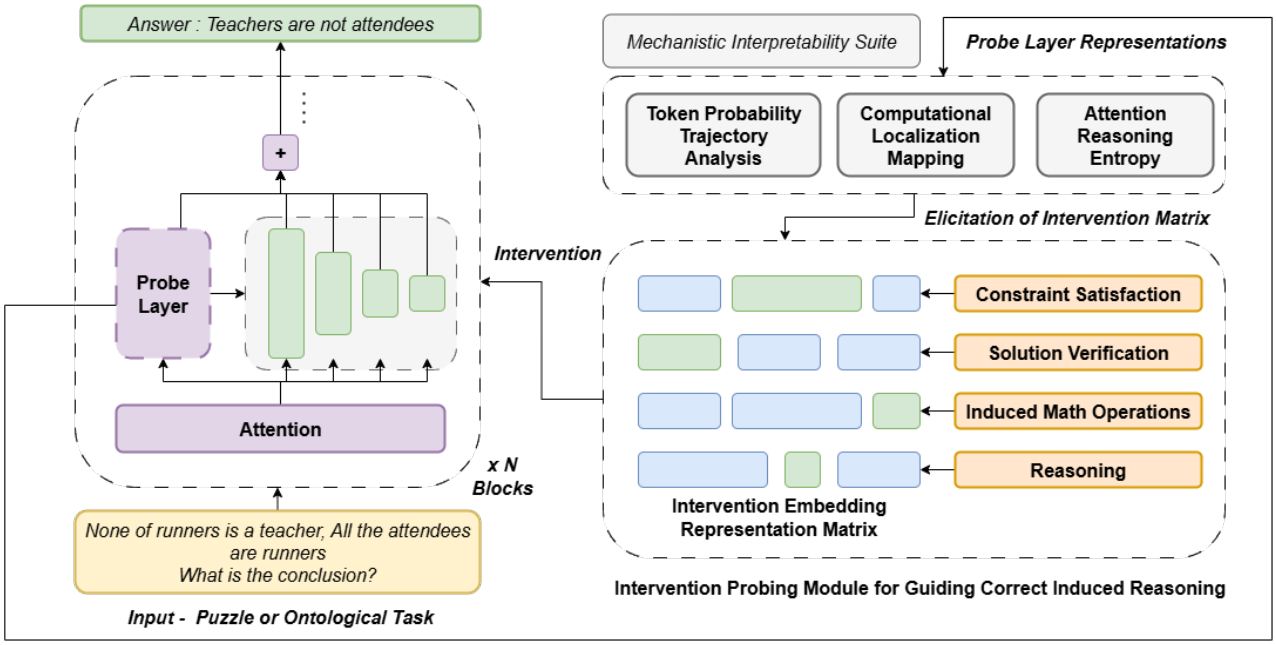

Figure 1: Overview of the probe-guided intervention framework: mechanistic interpretability tools analyze middle layer representations to detect reasoning errors, enabling targeted interventions that enhance domain specific reasoning in puzzle solving and ontological tasks.

1 Introduction

Large language models (LLMs) demonstrate impressive capabilities across diverse tasks, including reasoning, inference, and problem solving [1, 2]. However, these abilities remain unreliable and poorly understood, with models often generating plausible but unfaithful explanations [3, 4]. Understanding the internal mechanisms of model reasoning and developing targeted interventions to enhance reliability is crucial for advancing these capabilities. Recent work on chain-of-thought (CoT) prompting improves reasoning performance [5, 6], yet whether these external traces reflect genuine internal computation remains uncertain [7]. Meanwhile, empirical studies suggest that middle layers of transformer architectures play crucial roles in reasoning, showing dynamic transformations linked to reasoning complexity [8, 9]. This thesis focuses on two controlled domains of puzzle solving and ontological reasoning where reasoning steps can be precisely validated, allowing fine grained probing of internal processes while avoiding overgeneralization.

2 Previous Works

Previous efforts in LLM interpretability have employed probing classifiers to investigate linguistic information in neural representations [10, 11], and tools like circuit tracing to examine prediction evolution across layers [12, 13]. Mechanistic interpretability approaches have revealed insights into arithmetic and factual recall [14, 15], though attention patterns do not always correlate with reasoning processes [16, 17]. Chain-of-thought prompting has improved reasoning performance [6], with theoretical work explaining its effectiveness [18, 19]. However, studies show that models can generate unfaithful explanations that do not reflect actual decision making processes [20, 21]. Recent investigations using activation patching and causal intervention reveal that middle layers play crucial roles in reasoning [22–24], with specialized circuits developing for different reasoning types [25].

Mathematical and logic puzzles provide controlled environments for studying reasoning with well defined solution paths [26, 27], though models struggle with novel variations requiring creative insight. Ontological reasoning, fundamental to many AI applications, shows that models can perform taxonomic reasoning [28, 29] but often struggle with systematic inference and consistency [30]. No prior work has systematically investigated middle layer reasoning mechanisms in these domains or developed targeted interventions to enhance domain specific reasoning.

3 Aims

To address these challenges, I present two major aims for this PhD research:

3.1 AIM 1: Developing Domain Specific Probing and Evaluation Frameworks

3.1.1 Specialized Probing Architectures

The objective is to create hierarchical probing structures designed to capture reasoning patterns in puzzle solving and ontological domains. Building on insights from [10, 11, 13], we will develop attention probing mechanisms that identify relationships between reasoning steps and track information flow across layers. For puzzle solving, probes will detect mathematical operations, constraint satisfaction, and solution verification. For ontological reasoning, probes will focus on hierarchical relationship detection, concept inheritance, and taxonomic inference. Cross domain analysis between puzzle and ontological reasoning will reveal whether there are shared reasoning mechanisms between structured problem solving and concept manipulation, or whether each domain requires specialized processing pathways. This approach extends prior probing work [28, 31] by incorporating domain specific reasoning objectives.

3.1.2 Dataset Creation and Evaluation

We will create comprehensive datasets with fine grained annotations of reasoning steps, multiple solution paths, and systematic complexity variations. For puzzle solving, the data sets will include mathematical and logic puzzles with step-by-step annotations, constraint satisfaction tracking, and difficulty grades. For ontological reasoning, datasets will encompass taxonomic classification with hierarchical annotations, multiple inheritance scenarios, and systematic variations in hierarchy depth[1]. The evaluation framework will incorporate step wise accuracy, reasoning consistency, and analysis of error patterns, along with qualitative assessment of reasoning faithfulness and quality of explanation.

3.1.3 Middle Layer Analysis Framework

Following methodologies from [22–24], we will combine representation clustering, activation pattern analysis, and perturbation studies tailored for puzzle and ontological reasoning. For puzzle solving, the framework will investigate how middle layers represent constraints, track solution progress, and implement search strategies. For ontological reasoning, we will examine how middle layers encode concept hierarchies, perform inheritance computations, and integrate new concepts. The framework will also investigate temporal dynamics of middle layer processing during multi-step reasoning, examining how representations evolve across forward passes in models that engage in iterative reasoning or self-correction within these specific domains. This analysis extends recent work on circuit discovery [25] by focusing on domain specific reasoning pathways.

3.2 AIM 2: Creating Domain Targeted Interventional Frameworks

3.2.1 Probe Guided Intervention Strategies

Building on insights from [32–34], we will develop monitoring systems using domain specific probing classifiers to track reasoning in real time. When probes detect errors or inconsistencies, targeted

[1] Datasets will be validated by domain experts and released publicly for reproducibility.

interventions will correct reasoning trajectories while preserving domain specific patterns. For puzzle solving, interventions will include constraint reinforcement and backtracking guidance. For ontological reasoning, interventions will maintain hierarchical consistency and correct inheritance computations[2]. This adaptive capability will enable the system to handle novel puzzles and ontological structures without requiring manual reconfiguration while maintaining domain-specific expertise.

3.2.2 Real Time Reasoning Enhancement

We will develop dynamic intervention systems optimized for puzzle and ontological reasoning. For puzzle solving, the system will provide constraint checking, solution validation, and systematic search guidance while maintaining creative problem solving aspects. For ontological reasoning, the system will offer hierarchy navigation, inheritance computation support, and consistency checking. The framework will include domain specific uncertainty quantification following methods in [35], allowing the system to determine when interventions are needed.

3.2.3 Comprehensive Evaluation Protocols

Following evaluation best practices from [36, 37], we will establish protocols assessing quality, faithfulness, and reliability of domain specific reasoning. For puzzle solving, quantitative measures will track solution accuracy, step efficiency, and robustness across variations. For ontological reasoning, measures will assess taxonomic accuracy, consistency maintenance, and scalability. Human studies will evaluate whether enhanced reasoning is more convincing and trustworthy to domain experts, including mathematicians and knowledge engineers, compared to baseline outputs.

4 Timeline and Deliverables

Year 1: Conduct literature review, develop probing architectures, establish experimental frameworks and baselines. ***Deliverable:*** Initial probing framework, baseline models, 2-3 workshop papers.
Year 2: Create annotated datasets, implement probing classifiers, conduct preliminary middle layer analysis. ***Deliverable:*** Annotated datasets, validated probing classifiers, 1-2 conference publications on methodology.
Year 3: Validate reasoning patterns, develop cross-task analysis framework, design intervention strategies. ***Deliverable:*** Analysis framework, intervention prototypes, 3-4 major conference/journal publications.
Year 4: Implement intervention systems, establish evaluation protocols, conduct human studies, complete thesis. ***Deliverable:*** Validated intervention system, evaluation reports, open source tools, thesis document, 2-3 final publications.

5 Discussion

We expect to uncover distinct yet partially overlapping neural circuits for different reasoning types, revealing modular cognitive processes in LLMs. If successful, this work will establish probing and intervention methods as practical tools for understanding and improving reasoning in language models. The intervention frameworks should enhance reasoning reliability while maintaining creative problem solving capabilities, addressing current limitations in model faithfulness [4, 21].

The implications for the field include advancing mechanistic understanding of transformer reasoning processes, demonstrating that middle layer interventions can enhance reliability without sacrificing performance, and providing methodologies generalizable to other structured reasoning domains. The open source tools and evaluation frameworks will support broader research into interpretable and trustworthy AI systems.

[2]Intervention strategies will be adaptive, learning from previous successes and failures to handle novel problems.

References

[1] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[2] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[3] Ansh Radhakrishnan, Karina Nguyen, Anna Chen, Carol Chen, Carson Denison, Danny Hernandez, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiūtė, et al. Question decomposition improves the faithfulness of model-generated reasoning. *arXiv preprint arXiv:2307.11768*, 2023.

[4] Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36:74952–74965, 2023.

[5] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

[6] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[7] Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, et al. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023.

[8] Jesse Vig and Yonatan Belinkov. Analyzing the structure of attention in a transformer language model. *arXiv preprint arXiv:1906.04284*, 2019.

[9] Arnab Sen Sharma, David Atkinson, and David Bau. Locating and editing factual associations in mamba. *arXiv preprint arXiv:2404.03646*, 2024.

[10] Yonatan Belinkov and James Glass. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72, 2019.

[11] John Hewitt and Christopher D Manning. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, 2019.

[12] nostalgebraist. Interpreting gpt: the logit lens. https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens, 2020.

[13] Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*, 2023.

[14] Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets. *arXiv preprint arXiv:2201.02177*, 2022.

[15] Leonard Bereska and Efstratios Gavves. Mechanistic interpretability for ai safety–a review. *arXiv preprint arXiv:2404.14082*, 2024.

[16] Sarthak Jain and Byron C Wallace. Attention is not explanation. *arXiv preprint arXiv:1902.10186*, 2019.

[17] Sofia Serrano and Noah A Smith. Is attention interpretable? *arXiv preprint arXiv:1906.03731*, 2019.

[18] William Merrill and Ashish Sabharwal. The expressive power of transformers with chain of thought. *arXiv preprint arXiv:2310.07923*, 2023.

[19] Zhiyuan Li, Hong Liu, Denny Zhou, and Tengyu Ma. Chain of thought empowers transformers to solve inherently serial problems. *arXiv preprint arXiv:2402.12875*, 1, 2024.

[20] Abulhair Saparov and He He. Language models are greedy reasoners: A systematic formal analysis of chain-of-thought. *arXiv preprint arXiv:2210.01240*, 2022.

[21] Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jiang, Bill Yuchen Lin, Sean Welleck, Peter West, Chandra Bhagavatula, Ronan Le Bras, et al. Faith and fate: Limits of transformers on compositionality. *Advances in Neural Information Processing Systems*, 36:70293–70332, 2023.

[22] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372, 2022.

[23] Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. *arXiv preprint arXiv:2211.00593*, 2022.

[24] Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, et al. Causal abstraction: A theoretical foundation for mechanistic interpretability. *Journal of Machine Learning Research*, 26(83):1–64, 2025.

[25] Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.

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

[27] 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, 2021. *URL https://arxiv.org/abs/2103.03874*, 2, 2024.

[28] Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. What does bert look at? an analysis of bert's attention. *arXiv preprint arXiv:1906.04341*, 2019.

[29] Patrick Hohenecker and Thomas Lukasiewicz. Ontology reasoning with deep neural networks. *Journal of Artificial Intelligence Research*, 68:503–540, 2020.

[30] Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. arxiv eprints, page. *arXiv preprint arXiv:2006.00995*, 2020.

[31] Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A primer in bertology: What we know about how bert works. *Transactions of the association for computational linguistics*, 8:842–866, 2021.

[32] Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D Manning. Fast model editing at scale. *arXiv preprint arXiv:2110.11309*, 2021.

[33] Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. *arXiv preprint arXiv:2212.04089*, 2022.

[34] Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36:41451–41530, 2023.

[35] Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*, 2023.

[36] Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of nlp models with checklist. *arXiv preprint arXiv:2005.04118*, 2020.

[37] Zhaofeng Wu, Linlu Qiu, Alexis Ross, Ekin Akyürek, Boyuan Chen, Bailin Wang, Najoung Kim, Jacob Andreas, and Yoon Kim. Reasoning or reciting? exploring the capabilities and limitations of language models through counterfactual tasks. arxiv. *Preprint posted online March*, 28, 2024.