Understanding Diverse Reasoning Procedures in Foundation Models via Mechanistic Interpretability

Mohanna Hoveyda* **(mohanna.hoveyda@ru.nl)**
iCIS, Radboud University
Nijmegen, The Netherlands

Jasmin Kareem* **(j.kareem@tue.nl)**
Jheronimus Academy of Data Science, Eindhoven University of Technology
Sint Janssingel 92, 5211 DA ’s-Hertogenbosch

Roxana Petcu* **(r.m.petcu@uva.nl)**
Department, University of Amsterdam
Science Park 900, 1098 XH Amsterdam

Angela van Sprang* **(a.v.vansprang@uva.nl)**
IvI/ ILLC, University of Amsterdam
Science Park 900, 1098 XH Amsterdam

Ana Lucic (a.lucic@uva.nl)
IvI/ ILLC, University of Amsterdam
Science Park 900, 1098 XH Amsterdam

*Equal contribution.

Abstract

Foundation models exhibit impressive performance on tasks that appear to require a wide range of reasoning abilities. However, they struggle to generalize under distribution shifts and struggle with reasoning problems that are trivial for humans. These inconsistencies raise a critical question: which internal mechanisms, if any, underlie the successes and failures of these models in reasoning tasks? While numerous benchmarks have been proposed to probe reasoning capabilities, our understanding of the underlying mechanisms responsible for such reasoning-like behavior remains limited. We hypothesize that *distinct reasoning procedures are supported by specialized, possibly modular, computational pathways* in large-scale models. Mechanistic interpretability (MI) offers a promising set of tools to identify and analyze such pathways. However, most existing work operates in an isolated manner: evaluating a particular model for a particular reasoning task, often in a single modality. To address this gap, we first lay out a high-level taxonomy of reasoning processes and then conduct a systematic analysis of how mechanistic interpretability has been used to investigate diverse reasoning processes in various foundation models, across three main axes: (i) reasoning type, (ii) MI technique, and (iii) modality. We aim to develop a broader understanding of whether (A) different reasoning processes share computational mechanisms or are supported by distinct subsystems, and whether (B) such mechanisms are consistent across modalities other than text, such as vision.

Keywords: Mechanistic Interpretability, Reasoning, Foundation models, Multi-modality

Background

Reasoning Processes

Reasoning is broadly understood as the process of inferring novel conclusions based on prior information (Krawczyk, 2017; Russell & Norvig, 2016; Castañeda et al., 2023). However, there is little consensus, both within and across fields such as neuroscience, philosophy, and AI, regarding its precise nature, categories, and underlying mechanisms (Krawczyk, 2017; Castañeda et al., 2023; Goel et al., 2017). In our proposal, we adopt a pragmatic approach to categorizing reasoning processes in a way that reflects both how it has been operationalized in AI via tasks and benchmarks, as well as the distinctions considered in cognitive neuroscience (Krawczyk, 2017). We consider the following reasoning processes along with their aligned benchmarks;

- **Deductive reasoning:** Starting from general premises and inferring specific conclusions that are logically entailed (Yang et al., 2018; Han et al., 2024). Includes answering *multihop* questions (Yang et al., 2018) and some *logic puzzles* (Han et al., 2024).
- **Causal reasoning:** Establishing *cause-and-effect* relationships between entities (Krawczyk, 2017; Chi et al., 2024; Gendron et al., 2024), a process crucial for building coherent *world models* (Gkountouras et al., 2025).
- **Compositional reasoning:** Constructing complex structures from simpler parts or deconstructing them into meaningful components (Hosseini et al., 2024; Li et al., 2024).
- **Abductive reasoning:** Inferring the best possible conclusion without having all necessary information for an objectively correct answer (Krawczyk, 2017).
- **Mathematical and arithmetic reasoning:** Involves numerical operations, symbolic manipulation, and formal mathematical problem solving (Cobbe et al., 2021; Mirzadeh et al., 2024; Hanna et al., 2023).
- **Analogical and inductive reasoning:** Drawing on relevant past experiences to solve new problems (Yasunaga et al., 2024).
- **Geometric and spatial reasoning:** Understanding shapes, positions, and spatial relationships between objects (Kazemi et al., 2024; Shiri et al., 2024).
- **Meta-reasoning** Reflectively assessing whether the system lacks sufficient information (*knowledge gap identification*) or certainty about a given piece of knowledge (Ferrando et al., 2024), or detecting fallacies in its own reasoning process (Zeng et al., 2024).

We acknowledge the incompleteness of this list and the inherent difficulty in disentangling overlapping reasoning processes—where tasks like world modeling or question answering often encapsulate causal, commonsense, spatial, and compositional elements. Nevertheless, this can serve as a starting point for systematically deconstructing the multifaceted mechanism of reasoning in multimodal models.

Mechanistic Interpretability

Mechanistic interpretability (MI) is an emerging field within AI that aims to identify the computations underlying deep neural networks (NNs). The main goal is to reverse engineer the behavior of NNs by uncovering subnetworks that are responsible for specific behaviors. Most existing work focuses on narrow investigations into specific tasks, including reasoning. However, the ultimate goal is to uncover general principles beyond empirical findings that are isolated to specific models, modalities or types of reasoning.

We divide the prior work on MI for understanding reasoning processes into two categories: observation- and intervention-based techniques, as described in Bereska & Gavves (2024). Table 1 provides an overview of prior work at the intersection of MI and reasoning in foundation models.

Observation-based techniques A popular method in MI is *linear probing*, which has been used to study the reasoning of world models (Nanda et al., 2023), LLMs (Hou et al., 2023), and multi-modal models (Salin et al., 2022; Tao et al., 2024). More recently, *sparse autoencoders* (SAEs) have gained traction within the MI community as a means to decompose a network into a latent representation with sparse, interpretable features, with the aim of understanding reasoning processes (Galichin et al., 2025). Lastly, *logit lens* is another observa-

MI Technique	Methods	Reasoning Tasks
Probing	Salin et al. (2022), Hou et al. (2023), Nanda et al. (2023), Tao et al. (2024), Brinkmann et al. (2024)	Mathematical (VL), Language Multi-step, World-Model, Logical (VL), Symbolic Multi-step
Logit Lens	Sakarvadia et al. (2023), Huo et al. (2024), Phukan et al. (2025)	Multi-hop, QA (V)
SAEs	Galichin et al. (2025)	Chain-of-thought
Logit Attribution	Lieberum et al. (2023)	Multiple Choice QA
Attribution Patching	Hanna et al. (2024)	Mathematical, Compositionality, World-Model
Activation Patching	Sakarvadia et al. (2023), Lieberum et al. (2023), Stolfo et al. (2023), Yu & Ananiadou (2024), Mondorf et al. (2024), Feng & Steinhardt (2024), Brinkmann et al. (2024), Basu et al. (2024), Yu & Ananiadou (2025)	Multi-hop, Multiple Choice QA, Mathematical, Logical, Binding objects, Symbolic Multi-step
Causal Scrubbing	Brinkmann et al. (2024)	Symbolic Multi-step

Table 1: An overview of recent work in MI which aims to understand the reasoning abilities of foundation models, structured according to the reasoning tasks and MI techniques used. Each reasoning task involves one or more reasoning processes described in the previous section. By default, the reasoning tasks are language-based, (V) indicates a vision application, while (VL) indicates both vision and language.

tional method that can be used to interpret the latent representations of a model. Within reasoning tasks, it has been used to understand the reasoning of language (Sakarvadia et al., 2023) and visual question answering (QA) tasks (Huo et al., 2024; Phukan et al., 2025).

Intervention-based techniques *Activation patching* and *path patching* localize where specific input-dependent features are encoded within a model. These methods intervene in latent representations by modifying a subset of activations or paths, replacing them with those of a separate model pass - and then measure the causal impact on the output. Activation patching has been used to study multiple-choice QA reasoning (Lieberum et al., 2023; Basu et al., 2024; Yu & Ananiadou, 2025), mathematical (Yu & Ananiadou, 2024) and logical reasoning (Mondorf et al., 2024). *Attribution patching* leverages different weights of the model components that are calculated as a linear approximation using its gradients. Hanna et al. (2024) employ edge attribution patching with integrated gradients for multiple tasks, including mathematical reasoning and compositionality. Lastly, *causal scrubbing* iteratively masks components of the model without replacement, assessing the impact of the removed components on model performance. Brinkmann et al. (2024) compares probing, activation patching and causal scrubbing for localizing causal evidence of decoder-only transformers encoding symbolic reasoning.

Proposal: A Cross-Modal MI Study of Diverse Reasoning Processes

Although significant progress has been made in understanding (i) the behavior of foundation models with different data modalities (Lin et al., 2025), and (ii) how foundation models tackle diverse reasoning tasks, we advocate for identifying *universal reasoning patterns* that extend across reasoning types, MI techniques, and modalities. Drawing an analogy to the evolution from behaviorism to cognitive neuroscience (Bereska & Gavves, 2024), and inspired by some neuroscience findings on the representation of reasoning in the human brain (Castañeda et al., 2023; Zuanazzi et al., 2024), we aim to leverage the causal testing power of MI to probe the internal circuits that underpin reasoning in artificial systems.

We propose several research directions: (1) Can we identify distinct circuits for different types of reasoning, and is there overlap between them? (2) How do these circuits differ when models perform well on a reasoning task compared to when they fail, especially in cases where humans excel but foundation models fail such as the ARC challenge (ARC-AGI, 2025)? (3) Does a general reasoning mechanism exists across language, vision, audio, and video, and if so, is transfer learning between modalities possible? This is closely related to the "*platonic representation hypothesis*" (Huh et al., 2024), which posits that neural networks converge to a shared statistical model of reality regardless of training objectives or modalities. By systematically comparing the circuits underlying successful and unsuccessful reasoning in foundation model, we aim to bridge isolated findings and fill critical gaps in our understanding of AI reasoning mechanisms. We acknowledge that comparing subgraphs across different model architectures is not trivial – evaluating these circuit differences is crucial for assessing the universality of reasoning. Moreover, by drawing insights from neuroscience and comparing how reasoning operates in the brain with AI systems, we aim to gain a richer, more integrated perspective on the fundamental processes of reasoning.

References

ARC-AGI. (2025). Retrieved from https://arcprize.org/arc-agi

Basu, S., Grayson, M., Morrison, C., Nushi, B., Feizi, S., & Massiceti, D. (2024). Understanding information storage and transfer in multi-modal large language models. In A. Globersons et al. (Eds.), *Advances in neural information processing systems 38: Annual conference on neural information processing systems 2024, neurips 2024, vancouver, bc, canada, december 10 - 15, 2024.*

Bereska, L., & Gavves, S. (2024). Mechanistic interpretability for ai safety-a review. *Transactions on Machine Learning Research*.

Brinkmann, J., Sheshadri, A., Levoso, V., Swoboda, P., & Bartelt, C. (2024). A mechanistic analysis of a transformer trained on a symbolic multi-step reasoning task. In *Findings of the association for computational linguistics acl 2024* (pp. 4082–4102).

Castañeda, L. E. G., Sklarek, B., Dal Mas, D. E., & Knauff, M. (2023). Probabilistic and deductive reasoning in the human brain. *NeuroImage*, *275*, 120180.

Castañeda, L. E. G., Sklarek, B., Mas, D. E. D., & Knauff, M. (2023). Probabilistic and deductive reasoning in the human brain. *NeuroImage*.

Chi, H., Li, H., Yang, W., Liu, F., Lan, L., Ren, X., ... Han, B. (2024). Unveiling causal reasoning in large language models: Reality or mirage? In *The thirty-eighth annual conference on neural information processing systems.* Retrieved from https://openreview.net/forum?id=1IU3P8VDbn

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., ... Schulman, J. (2021). Training verifiers to solve math word problems. *CoRR*.

Feng, J., & Steinhardt, J. (2024). How do language models bind entities in context? In *The twelfth international conference on learning representations.* Retrieved from https://openreview.net/forum?id=zb3b6oKO77

Ferrando, J., Obeso, O., Rajamanoharan, S., & Nanda, N. (2024). Do I know this entity? knowledge awareness and hallucinations in language models. *CoRR*.

Galichin, A., Dontsov, A., Druzhinina, P., Razzhigaev, A., Rogov, O. Y., Tutubalina, E., & Oseledets, I. (2025). I have covered all the bases here: Interpreting reasoning features in large language models via sparse autoencoders. *arXiv preprint arXiv:2503.18878*.

Gendron, G., Nguyen, B. T., Peng, A. Y., Witbrock, M., & Dobbie, G. (2024). Can large language models learn independent causal mechanisms? In *Proceedings of the 2024 conference on empirical methods in natural language processing.* Association for Computational Linguistics.

Gkountouras, J., Lindemann, M., Lippe, P., Gavves, E., & Titov, I. (2025). Language agents meet causality – bridging LLMs and causal world models. In *The thirteenth international conference on learning representations.* Retrieved from https://openreview.net/forum?id=y9A2TpaGsE

Goel, V., Navarrete, G., Noveck, I. A., & Prado, J. (2017). *The reasoning brain: the interplay between cognitive neuroscience and theories of reasoning* (Vol. 10). Frontiers Media SA.

Han, S., Schoelkopf, H., Zhao, Y., Qi, Z., Riddell, M., Zhou, W., ... Radev, D. (2024). FOLIO: natural language reasoning with first-order logic. In *Proceedings of the 2024 conference on empirical methods in natural language processing, EMNLP 2024, miami, fl, usa, november 12-16, 2024.* Association for Computational Linguistics.

Hanna, M., Liu, O., & Variengien, A. (2023). How does GPT-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model. In *Advances in neural information processing systems 36: Annual conference on neural information processing systems 2023, neurips 2023, new orleans, la, usa, december 10 - 16, 2023.*

Hanna, M., Pezzelle, S., & Belinkov, Y. (2024). Have faith in faithfulness: Going beyond circuit overlap when finding model mechanisms. *CoRR*, *abs/2403.17806*. Retrieved from https://doi.org/10.48550/arXiv.2403.17806 doi: 10.48550/ARXIV.2403.17806

Hosseini, A., Sordoni, A., Toyama, D. K., Courville, A., & Agarwal, R. (2024). Not all LLM reasoners are created equal. In *The 4th workshop on mathematical reasoning and ai at neurips'24.* Retrieved from https://openreview.net/forum?id=RcqAmkDJfI

Hou, Y., Li, J., Fei, Y., Stolfo, A., Zhou, W., Zeng, G., ... Sachan, M. (2023). Towards a mechanistic interpretation of multi-step reasoning capabilities of language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing* (pp. 4902–4919).

Huh, M., Cheung, B., Wang, T., & Isola, P. (2024). *The platonic representation hypothesis.* Retrieved from https://arxiv.org/abs/2405.07987

Huo, J., Yan, Y., Hu, B., Yue, Y., & Hu, X. (2024). *Mmneuron: Discovering neuron-level domain-specific interpretation in multimodal large language model.* Retrieved from https://arxiv.org/abs/2406.11193

Kazemi, M., Alvari, H., Anand, A., Wu, J., Chen, X., & Soricut, R. (2024). Geomverse: A systematic evaluation of large models for geometric reasoning. In *Ai for math workshop @ icml 2024.*

Krawczyk, D. (2017). *Reasoning: The neuroscience of how we think*. Academic Press.

Li, Z., Jiang, G., Xie, H., Song, L., Lian, D., & Wei, Y. (2024). Understanding and patching compositional reasoning in LLMs. In *Findings of the association for computational linguistics: Acl 2024.* Association for Computational Linguistics.

Lieberum, T., Rahtz, M., Kramár, J., Nanda, N., Irving, G., Shah, R., & Mikulik, V. (2023). Does circuit analysis interpretability scale? evidence from multiple choice capabilities in chinchilla. *CoRR*.

Lin, Z., Basu, S., Beigi, M., Manjunatha, V., Rossi, R. A., Wang, Z., ... others (2025). A survey on mechanistic interpretability for multi-modal foundation models. *arXiv e-prints*, arXiv–2502.

Mirzadeh, S., Alizadeh, K., Shahrokhi, H., Tuzel, O., Bengio, S., & Farajtabar, M. (2024). Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models. *CoRR*.

Mondorf, P., Wold, S., & Plank, B. (2024). Circuit compositions: Exploring modular structures in transformer-based language models. *CoRR*, *abs/2410.01434*. Retrieved from https://doi.org/10.48550/arXiv.2410.01434 doi: 10.48550/ARXIV.2410.01434

Nanda, N., Lee, A., & Wattenberg, M. (2023). Emergent linear representations in world models of self-supervised sequence models. *EMNLP 2023*, 16.

Phukan, A., Divyansh, Morj, H. K., Vaishnavi, Saxena, A., & Goswami, K. (2025). *Beyond logit lens: Contextual embeddings for robust hallucination detection & grounding in vlms.* Retrieved from https://arxiv.org/abs/2411.19187

Russell, S. J., & Norvig, P. (2016). *Artificial intelligence: a modern approach*. pearson.

Sakarvadia, M., Ajith, A., Khan, A., Grzenda, D., Hudson, N., Bauer, A., ... Foster, I. T. (2023). Memory injections: Correcting multi-hop reasoning failures during inference in transformer-based language models. In Y. Belinkov, S. Hao, J. Jumelet, N. Kim, A. McCarthy, & H. Mohebbi (Eds.), *Proceedings of the 6th blackboxnlp workshop: Analyzing and interpreting neural networks for nlp, blackboxnlp@emnlp 2023, singapore, december 7, 2023* (pp. 342–356). Association for Computational Linguistics. Retrieved from https://doi.org/10.18653/v1/2023.blackboxnlp-1.26 doi: 10.18653/V1/2023.BLACKBOXNLP-1.26

Salin, E., Farah, B., Ayache, S., & Favre, B. (2022). Are vision-language transformers learning multimodal representations? A probing perspective. In *Thirty-sixth AAAI conference on artificial intelligence, AAAI 2022, thirty-fourth conference on innovative applications of artificial intelligence, IAAI 2022, the twelveth symposium on educational advances in artificial intelligence, EAAI 2022 virtual event, february 22 - march 1, 2022* (pp. 11248–11257). AAAI Press. Retrieved from https://doi.org/10.1609/aaai.v36i10.21375 doi: 10.1609/AAAI.V36I10.21375

Shiri, F., Guo, X.-Y., Far, M. G., Yu, X., Haf, R., & Li, Y.-F. (2024). An empirical analysis on spatial reasoning capabilities of large multimodal models. In *Proceedings of the 2024 conference on empirical methods in natural language processing.* Association for Computational Linguistics.

Stolfo, A., Belinkov, Y., & Sachan, M. (2023). A mechanistic interpretation of arithmetic reasoning in language models using causal mediation analysis. In *Proceedings of the 2023 conference on empirical methods in natural language processing* (pp. 7035–7052).

Tao, M., Huang, Q., Xu, K., Chen, L., Feng, Y., & Zhao, D. (2024). Probing multimodal large language models for global and local semantic representations. In N. Calzolari, M. Kan, V. Hoste, A. Lenci, S. Sakti, & N. Xue (Eds.), *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation, LREC/COLING 2024, 20-25 may, 2024, torino, italy* (pp. 13050–13056). ELRA and ICCL. Retrieved from https://aclanthology.org/2024.lrec-main.1142

Yang, Z., Qi, P., Zhang, S., Bengio, Y., Cohen, W. W., Salakhutdinov, R., & Manning, C. D. (2018). Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing, brussels, belgium, october 31 - november 4, 2018.* Association for Computational Linguistics.

Yasunaga, M., Chen, X., Li, Y., Pasupat, P., Leskovec, J., Liang, P., ... Zhou, D. (2024). Large language models as analogical reasoners. In *The twelfth international conference on learning representations, ICLR 2024, vienna, austria, may 7-11, 2024.*

Yu, Z., & Ananiadou, S. (2024). Interpreting arithmetic mechanism in large language models through comparative neuron analysis. In Y. Al-Onaizan, M. Bansal, & Y. Chen (Eds.), *Proceedings of the 2024 conference on empirical methods in natural language processing, EMNLP 2024, miami, fl, usa, november 12-16, 2024* (pp. 3293–3306). Association for Computational Linguistics. Retrieved from https://aclanthology.org/2024.emnlp-main.193

Yu, Z., & Ananiadou, S. (2025). *Understanding multimodal llms: the mechanistic interpretability of llava in visual question answering.* Retrieved from https://arxiv.org/abs/2411.10950

Zeng, Z., Liu, Y., Wan, Y., Li, J., Chen, P., Dai, J., ... Jia, J. (2024). MR-ben: A meta-reasoning benchmark for evaluating system-2 thinking in LLMs. In *The thirty-eighth annual conference on neural information processing systems.*

Zuanazzi, A., Ripollés, P., Lin, W. M., Gwilliams, L., King, J.-R., & Poeppel, D. (2024). Negation mitigates rather than inverts the neural representations of adjectives. *PLoS biology*, *22*(5), e3002622.