Mechanistic Interpretability

Frederick Riemenschneider

October 2024

This document is a preliminary schedule for the seminar and subject to change based on the number of students and their interests. To complete the course you need to participate regularly (no more than one unexcused absence) and be active in the discussions. Please ensure that you include `[mechinterp]` in the subject line when you email your questions. Each student will also have to give a graded presentation. In addition, you need to do an implementation project after the semester.

Students are expected to read the presented papers (at most two) and hand in questions or comments about them via mail to me each Tuesday before 3pm. These questions will be a part of your final grade.

1 Getting a Presentation Slot

To get a presentation slot, write an email to riemenschneider@cl.uni-heidelberg.de by the 22nd of October. Students interested in the presentations scheduled for the 31st of October are encouraged to contact me earlier to ensure that you have sufficient time for adequate preparation. The dates given for the presentations are tentative and might change. If there is any date where it would not be possible to give your presentation, you can indicate at most one session where you can absolutely not give your presentation. Students who have a particular paper in mind are welcome to propose it to me.

2 Papers

17.10.2024: Organization, Overview

- *no student presentation*

24.10.2024: Background

- *no student presentation*

31.10.2024: History

- Gurnee et al. (2023)
- Elhage et al. (2022)

07.11.2024: Othello

- Li et al. (2023)
- Nanda, Lee, and Wattenberg (2023)

14.11.2024: Transformer Circuits

- Wang et al. (2022)
- Conmy et al. (2023)
- Shi et al. (2024)

21.11.2024: Activation Patching

- Meng, Bau, et al. (2022)
- Meng, Sen Sharma, et al. (2022)
- Position paper: Pinter and Elhadad (2023)
- Best practices: Heimersheim and Nanda (2024)

28.11.2024: No session

05.12.2024: Dictionary Learning I

- Bills et al. (2023)
- Bricken et al. (2023)
- Huben et al. (2024)

12.12.2024: Dictionary Learning II

- Makelov, Lange, and Nanda (2024)
- Karvonen et al. (2024)

19.12.2024: Self-conditioning

- Suau, Zappella, and Apostoloff (2022)
- Kojima et al. (2024)

09.01.2025: Grokking

- Liu et al. (2022)
- Nanda, Chan, et al. (2023)

16.01.2025: Multilingual Language Models

- Zhao, Yoshinaga, and Oba (2024)
- Tang et al. (2024)

23.01.2025: Frameworks

- Ghandeharioun et al. (2024)
- Huang et al. (2024)

30.01.2025: Hierarchical Representations

- Park et al. (2024)
- Ahuja et al. (2024)

06.02.2025: Looking Back, Discussion, Looking Forward

- *no student presentation*
- Bereska and Gavves (2024)

References

Ahuja, Kabir et al. (2024). "Learning Syntax Without Planting Trees: Understanding When and Why Transformers Generalize Hierarchically". In: *ICML 2024 Workshop on Mechanistic Interpretability*. URL: https://openreview.net/forum?id=YwLgSimUIT.

Bereska, Leonard and Efstratios Gavves (2024). "Mechanistic Interpretability for AI Safety–A Review". In: *arXiv preprint arXiv:2404.14082*.

Bills, Steven et al. (2023). *Language models can explain neurons in language models*. https://openaipublic.blob.core.windows.net/neuron-explainer/paper/index.html.

Bricken, Trenton et al. (2023). "Towards Monosemanticity: Decomposing Language Models With Dictionary Learning". In: *Transformer Circuits Thread*. https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Conmy, Arthur et al. (2023). "Towards Automated Circuit Discovery for Mechanistic Interpretability". In: *Thirty-seventh Conference on Neural Information Processing Systems*. URL: https://openreview.net/forum?id=89ia77nZ8u.

Elhage, Nelson et al. (2022). "Toy Models of Superposition". In: *Transformer Circuits Thread*. https://transformer-circuits.pub/2022/toy_model/index.html.

Ghandeharioun, Asma et al. (2024). "Patchscope: A unifying framework for inspecting hidden representations of language models". In: *arXiv preprint arXiv:2401.06102*.

Gurnee, Wes et al. (2023). "Finding Neurons in a Haystack: Case Studies with Sparse Probing". In: *arXiv preprint arXiv:2305.01610*.

Heimersheim, Stefan and Neel Nanda (2024). "How to use and interpret activation patching". In: *arXiv preprint arXiv:2404.15255*.

Hernandez, Evan et al. (2024). "Linearity of Relation Decoding in Transformer Language Models". In: *The Twelfth International Conference on Learning Representations*. URL: https://openreview.net/forum?id=w7LU2s14kE.

Huang, Xinting et al. (2024). "InversionView: A General-Purpose Method for Reading Information from Neural Activations". In: *ICML 2024 Workshop on Mechanistic Interpretability*. URL: https://openreview.net/forum?id=P7MW0FahEq.

Huben, Robert et al. (2024). "Sparse Autoencoders Find Highly Interpretable Features in Language Models". In: *The Twelfth International Conference on Learning Representations*. URL: https://openreview.net/forum?id=F76bwRSLeK.

Karvonen, Adam et al. (2024). "Measuring Progress in Dictionary Learning for Language Model Interpretability with Board Game Models". In: *ICML 2024 Workshop on Mechanistic Interpretability*. URL: https://openreview.net/forum?id=qzsDKwGJyB.

Kojima, Takeshi et al. (2024). "On the Multilingual Ability of Decoder-based Pre-trained Language Models: Finding and Controlling Language-Specific Neurons". In: *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics*. Mexico City, Mexico: Association for Computational Linguistics, pp. 6912–6964. URL: https://aclanthology.org/2024.naacl-long.384.

Li, Kenneth et al. (2023). "Emergent World Representations: Exploring a Sequence Model Trained on a Synthetic Task". In: *The Eleventh International Conference on Learning Representations*. URL: https://openreview.net/forum?id=DeG07_TcZvT.

Liu, Ziming et al. (2022). "Towards Understanding Grokking: An Effective Theory of Representation Learning". In: *Advances in Neural Information Processing Systems*. Ed. by Alice H. Oh et al. URL: https://openreview.net/forum?id=6at6rB3IZm.

Makelov, Aleksandar, Georg Lange, and Neel Nanda (2024). "Towards Principled Evaluations of Sparse Autoencoders for Interpretability and Control". In: *ICLR 2024 Workshop on Secure and Trustworthy Large Language Models*. URL: https://openreview.net/forum?id=MHIX9H8aYF.

Meng, Kevin, David Bau, et al. (2022). "Locating and Editing Factual Associations in GPT". In: *Advances in Neural Information Processing Systems*. Ed. by Alice H. Oh et al. URL: https://openreview.net/forum?id=-h6WAS6eE4.

Meng, Kevin, Arnab Sen Sharma, et al. (2022). "Mass Editing Memory in a Transformer". In: *arXiv preprint arXiv:2210.07229*.

Nanda, Neel, Lawrence Chan, et al. (2023). "Progress measures for grokking via mechanistic interpretability". In: *The Eleventh International Conference on Learning Representations*. URL: https://openreview.net/forum?id=9XFSbDPmdW.

Nanda, Neel, Andrew Lee, and Martin Wattenberg (Dec. 2023). "Emergent Linear Representations in World Models of Self-Supervised Sequence Models". In: *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*. Ed. by Yonatan Belinkov et al. Singapore: Association for Computational Linguistics, pp. 16–30. DOI: 10.18653/v1/2023.blackboxnlp-1.2. URL: https://aclanthology.org/2023.blackboxnlp-1.2.

Park, Kiho et al. (2024). "The Geometry of Categorical and Hierarchical Concepts in Large Language Models". In: *ICML 2024 Workshop on Mechanistic Interpretability*. URL: https://openreview.net/forum?id=KXuYjuBzKo.

Pinter, Yuval and Michael Elhadad (Dec. 2023). "Emptying the Ocean with a Spoon: Should We Edit Models?" In: *Findings of the Association for Computational Linguistics: EMNLP 2023*. Ed. by Houda Bouamor, Juan Pino, and Kalika Bali. Singapore: Association for Computational Linguis-

tics, pp. 15164–15172. DOI: 10.18653/v1/2023.findings-emnlp.1012. URL: https://aclanthology.org/2023.findings-emnlp.1012.

Shi, Claudia et al. (2024). "Hypothesis Testing the Circuit Hypothesis in LLMs". In: *ICML 2024 Workshop on Mechanistic Interpretability*. URL: https://openreview.net/forum?id=ibSNv9cldu.

Suau, Xavier, Luca Zappella, and Nicholas Apostoloff (2022). "Self-Conditioning Pre-Trained Language Models". In: *International Conference on Machine Learning*.

Tang, Tianyi et al. (Aug. 2024). "Language-Specific Neurons: The Key to Multilingual Capabilities in Large Language Models". In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Lun-Wei Ku, Andre Martins, and Vivek Srikumar. Bangkok, Thailand: Association for Computational Linguistics, pp. 5701–5715. URL: https://aclanthology.org/2024.acl-long.309.

Wang, Kevin et al. (2022). "Interpretability in the wild: a circuit for indirect object identification in gpt-2 small". In: *arXiv preprint arXiv:2211.00593*.

Zhao, Xin, Naoki Yoshinaga, and Daisuke Oba (Mar. 2024). "Tracing the Roots of Facts in Multilingual Language Models: Independent, Shared, and Transferred Knowledge". In: *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Yvette Graham and Matthew Purver. St. Julian's, Malta: Association for Computational Linguistics, pp. 2088–2102. URL: https://aclanthology.org/2024.eacl-long.127.