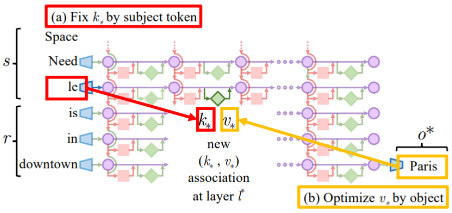
ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Distributed Computing

Prof. R. Wattenhofer

Interpreting LLM's "Reversal Curse"



Meng et al.'s [4] ROME method

Large Language Models struggle struggle to perform even basic knowledge manipulation tasks without explicitly stating retrieved knowledge in their Chain-of-Thought. For example, if "Alice is Bob's daughter" appears in the training dataset, a language model will be able to instantly answer "Who is Alice's father?" by retrieval from it's parameters (long-term memory) but will struggle to answer "Who is Bob's daughter?" [1, 2]. Nevertheless, if "Alice is Bob's daughter" appears in the context (short-term memory), the LLM *can* reason that Bob's daughter is Alice.

Interpretability methods seek to find mechanisms in deep learning models performing elements of cognition, for instance by localising where in the model's long term memory certain facts are stored [4] or discovering small circuit mechanisms for reasoning [3]. In this project, we aim to use the tools of interpretability to explain why the reversal curse appears and use our insights to explore interpretability and capabilities of LLMs more broadly.

Requirements:

- Strong software engineering skills (ideally in the modern deep learning stack of Python, PyTorch/JAX, HuggingFace) to quickly test & iterate on ideas
- Knowledge of Linear Algebra, Probability

Interested? Please get in touch for more details!

Contact

- Sam Dauncey: `sdauncey@ethz.ch`, ETZ G61.1

References

[1] Zeyuan Allen-Zhu and Yuanzhi Li. Physics of Language Models: Part 3.2, Knowledge Manipulation. 2024. arXiv: `2309.14402 [cs.CL]`.

[2] Lukas Berglund et al. The Reversal Curse: LLMs trained on "A is B" fail to learn "B is A". In: The Twelfth International Conference on Learning Representations. 2024.

[3] Nelson Elhage et al. A mathematical framework for transformer circuits. Transformer Circuits Thread. 2021.

[4] Kevin Meng et al. Locating and Editing Factual Associations in GPT. In: Advances in Neural Information Processing Systems. Ed. by S. Koyejo et al. Vol. 35. Curran Associates, Inc., 2022, pp. 17359–17372.