ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

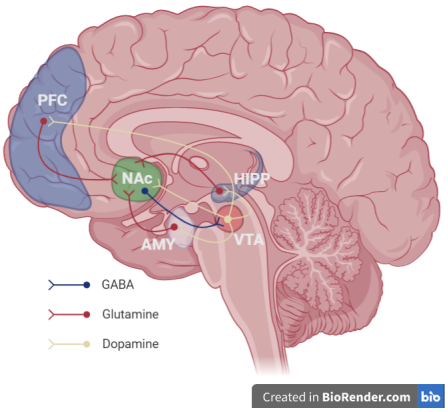
Distributed Computing

Prof. R. Wattenhofer

Eliciting Internal Reward Models in LLMs

The training of Large Language Models (LLMs) increasingly uses reinforcement learning, a paradigm which gives the model "reward" signals to maximise. Specifically, the post-training of LLMs has grown from mimicking human-generated answers to learning from human rankings of answers to the Chain-of-Thought based techniques used to train DeepSeek-r1 and the OpenAI o-series of models.

As these AI systems scale to potentially superhuman performance, understanding how exactly these reward signals affect the LLM's actions will be crucial to controlling and aligning them to be helpful, harmless and honest. This problem is counter-intuitive due to the use of *policy gradients*, making it unclear what exactly the model learns [3, 5, 6].



We've discovered reward circuits in the human brain! Credit to George Kach - Own work, CC BY-SA 4.0

We will work with first with small models and toy problems to answer questions such as: Can we discover a "reward" direction like we can discover a truth direction [1]? Can we edit rewards in LLM's parameters like we can edit knowledge [4]? Can we mechanistically interpret the reward via circuits [2]?

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, Statistics, (ideally: Reinforcement Learning theory)

Interested? Please get in touch with us for more details!

Contact

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

References

[1] Collin Burns et al. Discovering Latent Knowledge in Language Models Without Supervision. In: The Eleventh International Conference on Learning Representations. 2023.

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

[3] Evan Hubinger et al. Risks from Learned Optimization in Advanced Machine Learning Systems. 2019. arXiv: 1906.01820 [cs.AI].

[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.

[5] Rafael Rafailov et al. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. In: Advances in Neural Information Processing Systems. Vol. 36. 2023, pp. 53728–53741.

[6] Alex Turner. Reward is not the optimization target. LessWrong. 2024.