Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Distributed Computing

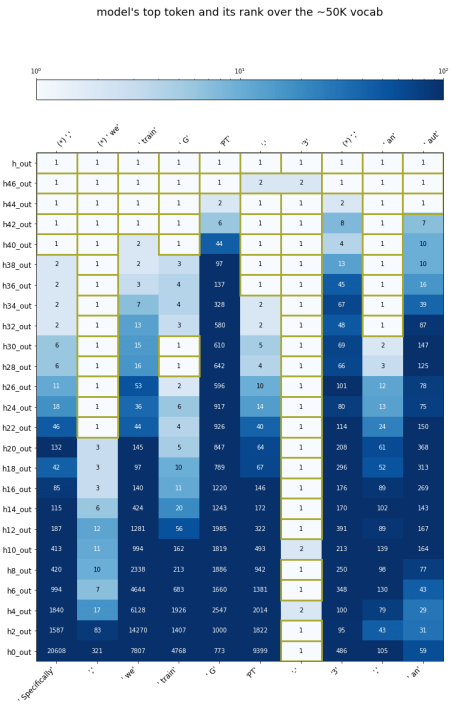
Prof. R. Wattenhofer

# Vertical Test Time Compute in Transformers:
## A Penny For Your Thoughts, A Dollar For Your Dreams

Transformer large language models (LLMs) work token-by-token, by default spending the same amount of compute power to predict every token in a sequence, independent of whether that token represents the solution to a very hard math problem or a simple grammatical completion.

Leveraging this "test time compute" to vary the amount of effort per token and enhance LLM reasoning capabilities is an active area of research. So far, most methods have focused on increasing compute in the *horizontal* direction: using more tokens to generate chains-of-thought.

In this project, we want to explore the *vertical* direction for test-time compute by using more/fewer layers on each token, self-distillation and building recurrence into the model with fine tuning. This is made possible by the fact these models have a residual stream of information whereby they iteratively refine their guess of the next token.




Credit: interpreting GPT: the logit lens, nostalgebraist

**Requirements:**

- Motivation to work on the bleeding edge of AI research
- 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

- Frédéric Berdoz: fberdoz@ethz.ch, ETZ G60.1
- Sam Dauncey: sdauncey@ethz.ch, ETZ G61.1