ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Distributed Computing

Prof. R. Wattenhofer

Benchmarking LLM sampling strategies

Some prevalent sampling strategies:

- Traditional sampling: random, greedy, top-k, top-p (nucleus), beam-search, Apparently chatgpt uses top-p sampling.
- Min-p sampling: https://arxiv.org/abs/2407.01082
- Truncation sampling: https://arxiv.org/abs/2210.15191 (generalizing top-k and top-p, interesting hypothesis). They propose η-sampling.
- Mirostat sampling: https://arxiv.org/abs/2007.14966 (maybe the most interesting one)
- Locally Typical Sampling: https://arxiv.org/abs/2202.00666
- Contrastive sampling: https://proceedings.neurips.cc/paper_files/paper/2022/hash/871cae8f599cb8bbfcb0f58fe1af95ad-Abstract-Conference.html

Some ideas:

- Idea: Compare nucleus, top k, typical, min probability, etc., at inference.
- Idea: Is softmax the best normalization function?
- Idea: How much scaling is good sampling worth?

Requirements

Good programming skills in Python and knowledge of machine learning evaluation are required. Additionally, having some creative skills as you will be drawing IQ tests will be beneficial. Some light web development skills are also helpful.
We will have weekly meetings to address questions, discuss progress, and think about future ideas.

Contact

In a few short sentences, please explain why you are interested in the project and about your coding and machine learning background (i.e., your projects or relevant courses you have taken at ETH or elsewhere).

- Andreas Plesner: aplesner@ethz.ch, ETZ G95
- Frédéric Berdoz: fberdoz@ethz.ch, ETZ G60.1