ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Distributed Computing

Prof. R. Wattenhofer

Reasoning Benchmark for LLMs

LLMs have long been able to generate coherent and high-quality text. But, do they really understand the content and can they properly reason about it? Results on popular reasoning benchmarks would suggest so, as these models have been getting better and better. With the release of OpenAI's o1 model, the bar has been raised even higher. However, if you have ever tested one of these models with your own challenging question, you might notice that their reasoning is really not that great. It might be possible that these models overfit on the popular reasoning benchmarks, so how well are they really able to reason, especially in out-of-distribution settings?

The goal of this project is to build a benchmark for LLMs and VLMs based on our PUZZLES benchmark [1]. PUZZLES consists of a set of 40 logic puzzles, which require advanced out-of-distribution reasoning capabilities to be solved. You will have to come up with a standardized way to evaluate performance of LLMs, as well as running experiments on how the most popular models perform. Depending on the scope of the project, we will also try to come up with novel methods that improve reasoning performance on our benchmark.

We will have weekly meetings to address questions, discuss progress and think about future ideas.

Requirements: Strong knowledge in Python. Strong foundation in deep learning. Previous experience with NLP and the transformer architecture is an advantage.

Interested? Please contact us for more details!

Contact

- Benjamin Estermann: besterma@ethz.ch, ETZ G60.1
- Luca Lanzendörfer: lucala@ethz.ch, ETZ G 93

References

[1] Benjamin Estermann et al. "PUZZLES: A Benchmark for Neural Algorithmic Reasoning". In: *38th Conference on Neural Information Processing Systems (NeurIPS 2024), Vancouver, Canada*. Dec. 2024.