LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN INSTITUT FÜR STATISTIK

Large-scale pretraining: the nitty-gritty details

Robert Baldock

(Aleph-Alpha)

2024-02-21, 2.15 pm

Main Building, Geschwister-Scholl-Platz 1, Room E004
and online via Zoom (Link)
(Meeting-ID: 683 0699 4223; Password: StatsCol23)

This talk will give a rare close-up of the nitty-gritty details that go into training large-scale LLMs. In the autumn of 2023, Aleph Alpha Research Lab prepared to train their next generation of large language models, which are training now. In this talk, Robert Baldock will chronicle their learnings from this process. In particular, he will describe their experiments to optimise the architecture and pretraining, their optimal scaling study, insights about efficient and numerically stable parallel training, tokenizer construction, and the preparation of the large-scale web-crawl dataset.

Biography:
Robert Baldock is a Research Lead at IPAI Aleph Alpha Research, where he works on large language and multimodal models and AI agents. He has previously worked on the science of deep learning in the Google Brain team and on the intersection of Bayesian computational methods and Statistical Physics at EPFL and the University of Cambridge where he did his PhD.