

Efficient Matrix Multiplication Algorithms for Quantized Language Models (20'+10') [in person, but demo on zoom]

Thursday 22 February 2024 17:05 (30 minutes)

Presenter: GÄSSLER, Johannes (Karlsruhe Institute of Technology)

Session Classification: Combining Domain Knowledge and LLMs, efficient computation