Contribution ID: **7** Type: **not specified**

# Efficient Matrix Multiplication Algorithms for Quantized Language Models

Large language models have - as the name implies - large numbers of parameters. As such not only the training costs but also the inference costs of these models are quite substantial. One strategy for reducing inference costs is to quantize the model weights from 16 bit floating point values to a format with 2-8 bits per weight. However, these custom data formats in turn require custom inference code. This talk describes the interplay of llama.cpp quantization formats and inference code and how int8 tensor cores or integer intrinsics can be used to reach performance exceeding that of standard floating point GEMM routines provided by e.g. cuBLAS.

**Primary author:** GÄSSLER, Johannes (Karlsruhe Institute of Technology)

**Presenter:** GÄSSLER, Johannes (Karlsruhe Institute of Technology)

**Track Classification:** Other