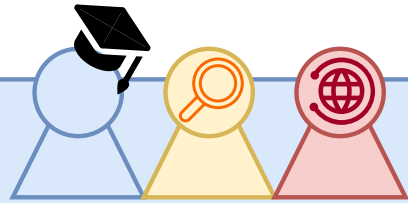

Main Goal

Evaluating the impact of quantization on operational-level LLM metrics.

RQ1

Quanti: benchmarking system for LLMs

RQ2

Accuracy-efficiency tradeoffs

RQ3

Architecture-specific tradeoffs

RQ4

CO2-aware scheduling impact

Metrics

GPU[%], Memory [GB], Energy [kWh]

Metrics

MMLU accuracy [%], Energy [kWh], Lantecy [ms]

Metrics

MMLU accuracy [%], Energy [kWh]

Metrics

Emissions [gCO2], Energy [kWh]