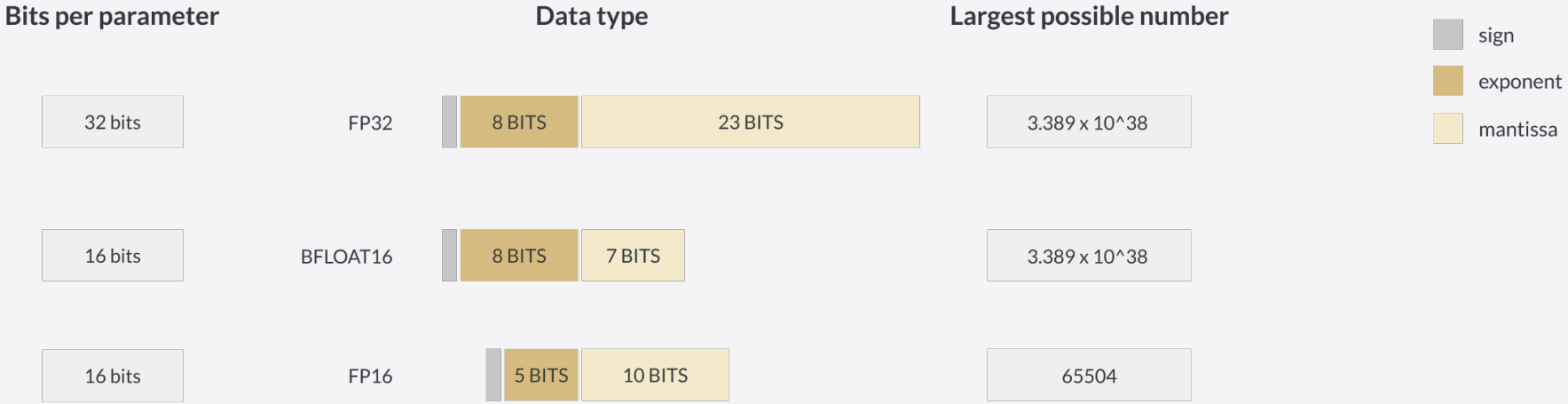
VIENNA SCIENTIFIC CLUSTER

Foundations of LLM Mastery:
Fine-tuning with multi GPUs

25 February 2025
ONLINE

EURO AUSTRIA

Limited GPU memory

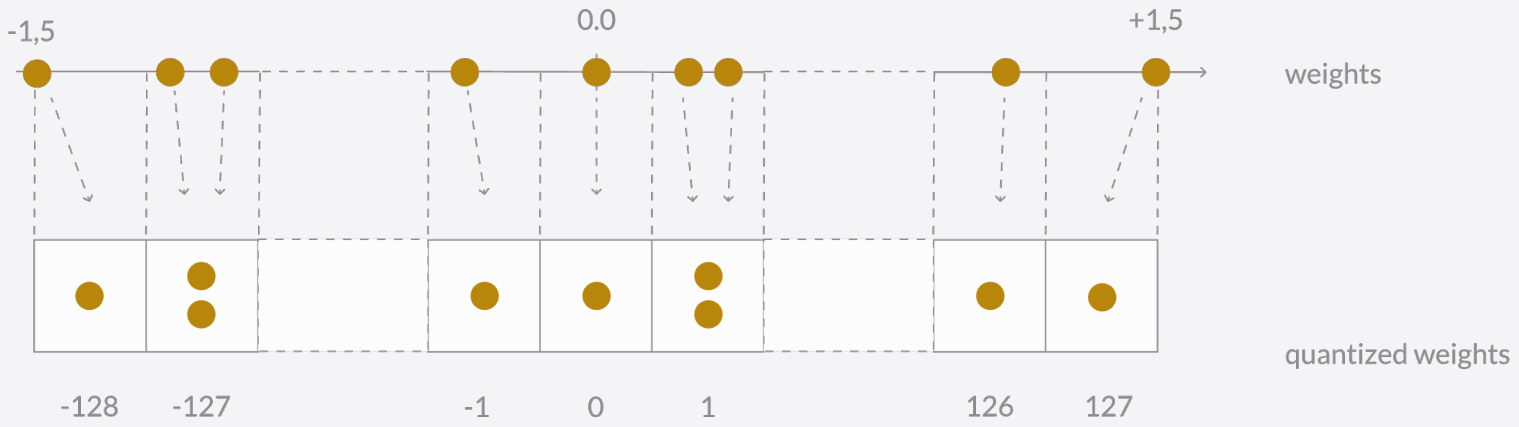


Fewer bits ⟶ **Quantization**

$$Value = (-1)^S * 2^{(E-15)} * (1 - M)$$

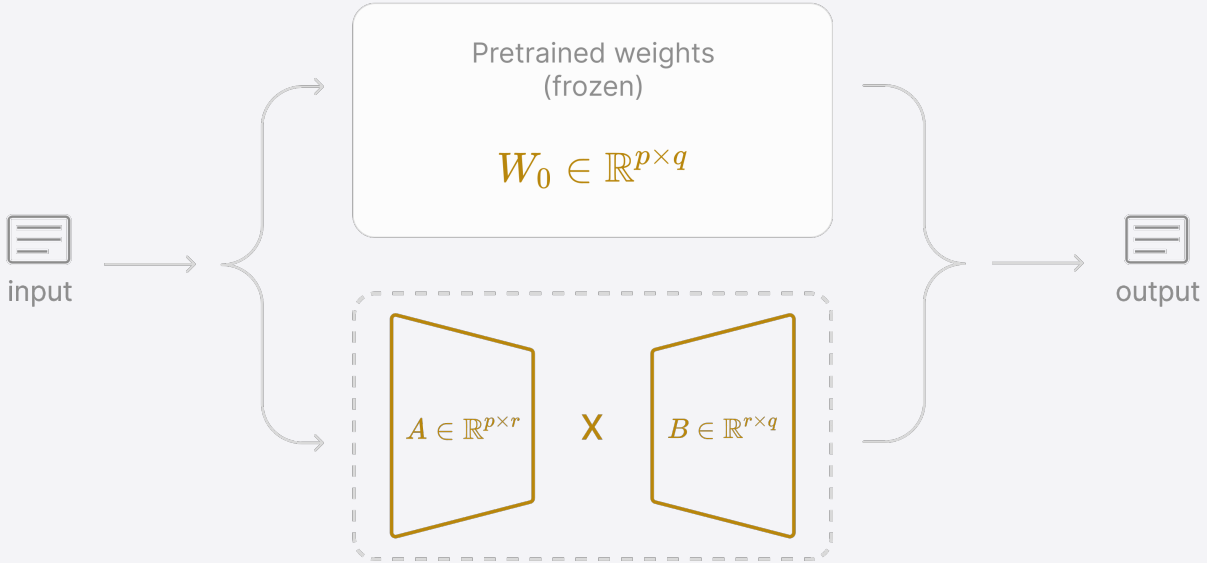
EURO AUSTRIA

Quantization



EURO AUSTRIA

Low rank adapters (LoRA)



EURO AUSTRIA