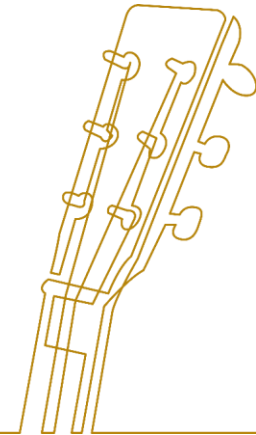




# Foundations of LLM Mastery: Fine-tuning with one GPU

24 January 2025  
ONLINE



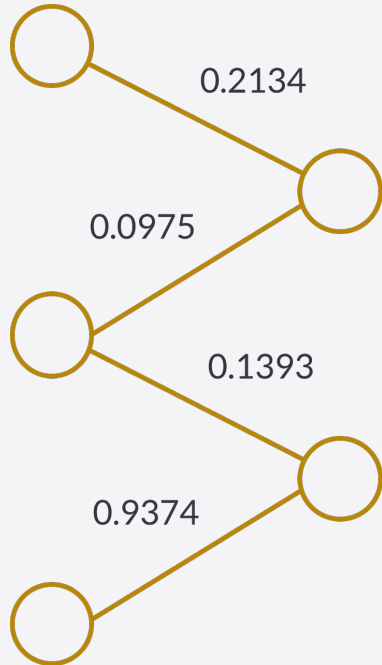
# How to finetune an LLM with limited GPU resources

Quantization / PEFT / Unsloth

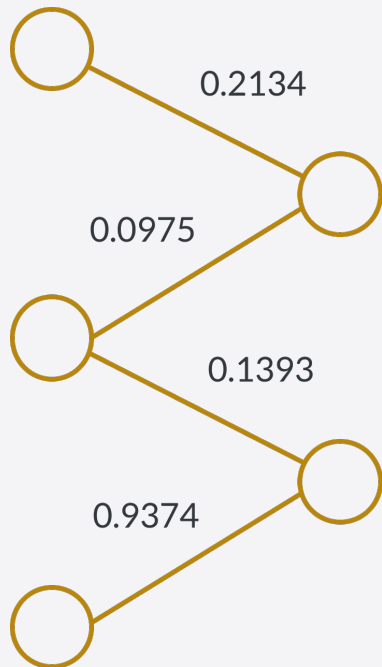
---

Speaker: Martin Pfister  
HPC / AI Team, EuroCC Austria

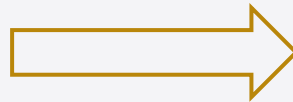
# Limited GPU memory



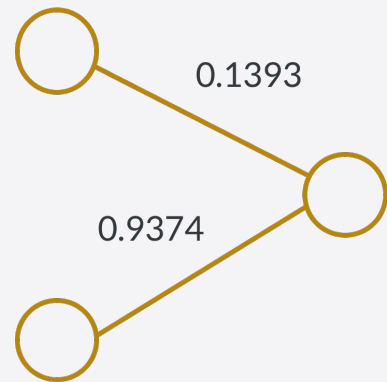
# Limited GPU memory



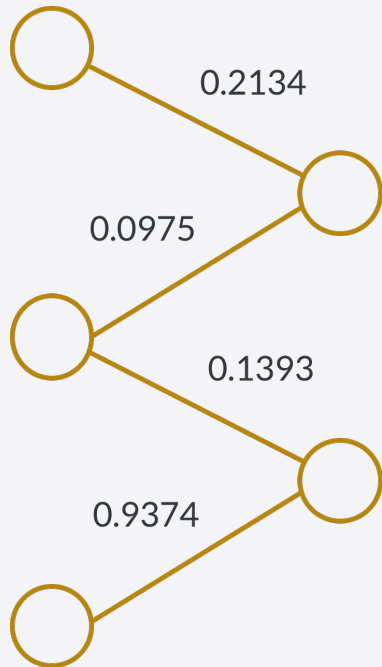
Fewer parameters



e.g. Llama 7B  
instead of 70B



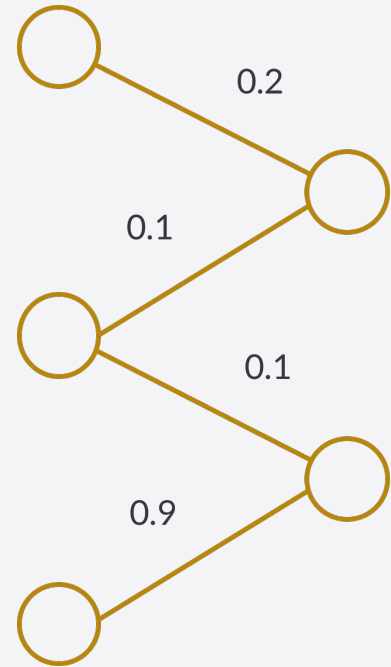
# Limited GPU memory






Less resolution



Fewer bits per parameter






# Limited GPU memory

Bits per parameter	Data type	Largest number possible
32 bits	FP32 	$3.389 \times 10^{38}$
16 bits	FP16 	65504
16 bits	BFLOAT16 	$3.389 \times 10^{38}$

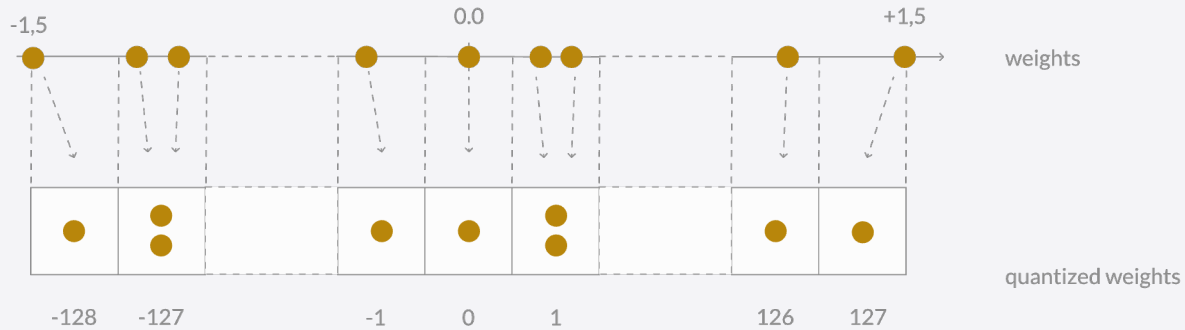
Fewer bits

→ Quantization

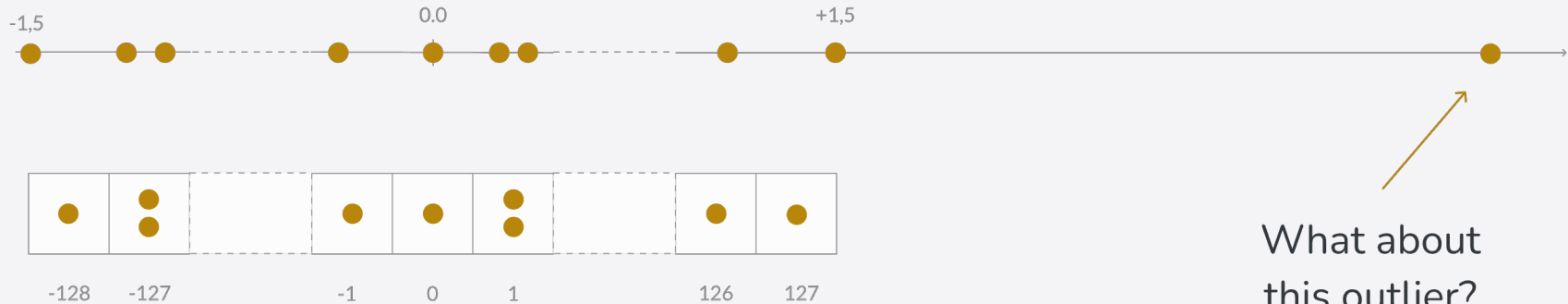
 sign    exponent    mantissa

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

# Quantization



# Quantization



What about  
this outlier?



# Quantization

Transformers documentation  
Quantization

Quantization method	On the fly quantization	CPU	CUDA GPU	RoCm GPU (AMD)	Metal (Apple Silicon)	torch.compile() support	Number of bits	Supports fine-tuning (through PEFT)	S tr
<a href="#">AQLM</a>	●	●	●	●	●	●	1 / 2	●	●
<a href="#">AWQ</a>	●	●	●	●	●	?	4	●	●
<a href="#">bitsandbytes</a>	●	●	●	●	●	●	4 / 8	●	●
<a href="#">EETQ</a>	●	●	●	●	●	?	8	●	●
<a href="#">GGUF / GGML (llama.cpp)</a>	●	●	●	●	●	●	1 - 8	●	Se se
<a href="#">GPTQ</a>	●	●	●	●	●	●	2 - 3 - 4 - 8	●	●
<a href="#">HQQ</a>	●	●	●	●	●	●	1 - 8	●	●
<a href="#">Quanto</a>	●	●	●	●	●	●	2 / 4 / 8	●	●
<a href="#">FBGEMM_FP8</a>	●	●	●	●	●	●	8	●	●
<a href="#">torchao</a>	●		●	●	partial support		4 / 8		●

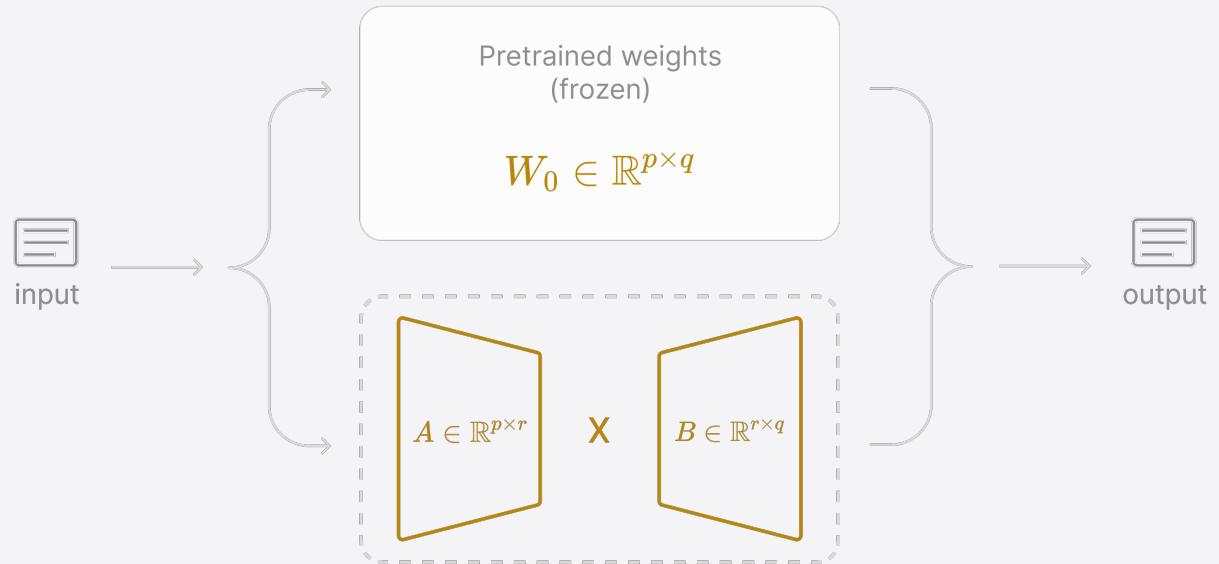
<https://huggingface.co/docs/transformers/main/quantization/overview>

# Quantization

**Hands on time!**

---

# Low rank adapters (LoRA)



---

**Low rank  
adapters  
(LoRA)**

**Hands on time!**

---

Unslot



unslot

Optimized GPU kernels

created by manually deriving all  
compute heavy maths steps

---

**Unslow**

**Hands on time!**

# STAY IN TOUCH

---



EuroCC Austria



@eurocc\_austria



eurocc-austria.at

# THANK YOU

---



This project has received funding from the European High-Performance Computing Joint Undertaking (JU) under grant agreement No 101101903. The JU receives support from the Digital Europe Programme and Germany, Bulgaria, Austria, Croatia, Cyprus, Czech Republic, Denmark, Estonia, Finland, Greece, Hungary, Ireland, Italy, Lithuania, Latvia, Poland, Portugal, Romania, Slovenia, Spain, Sweden, France, Netherlands, Belgium, Luxembourg, Slovakia, Norway, Türkiye, Republic of North Macedonia, Iceland, Montenegro, Serbia