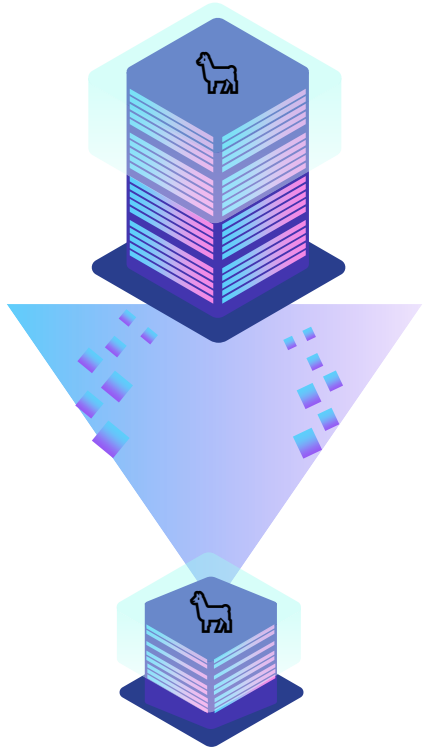
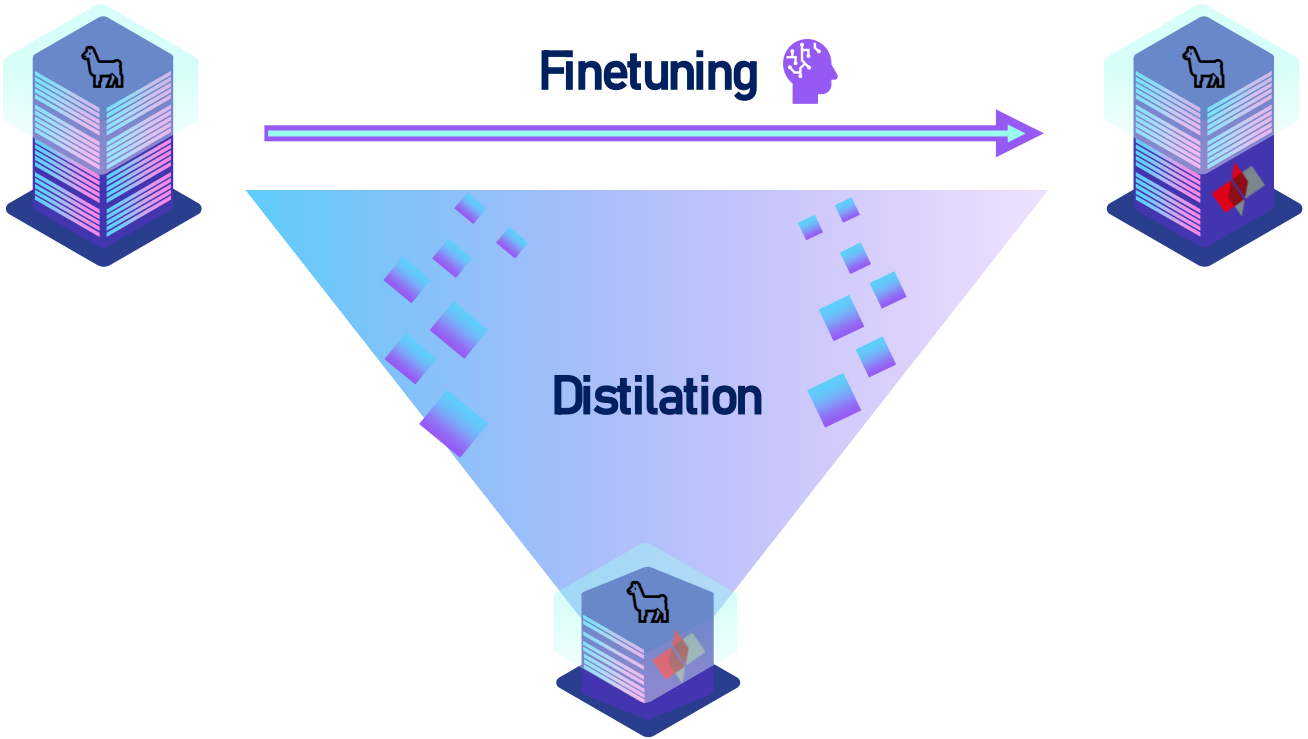
Distilling Llama 2

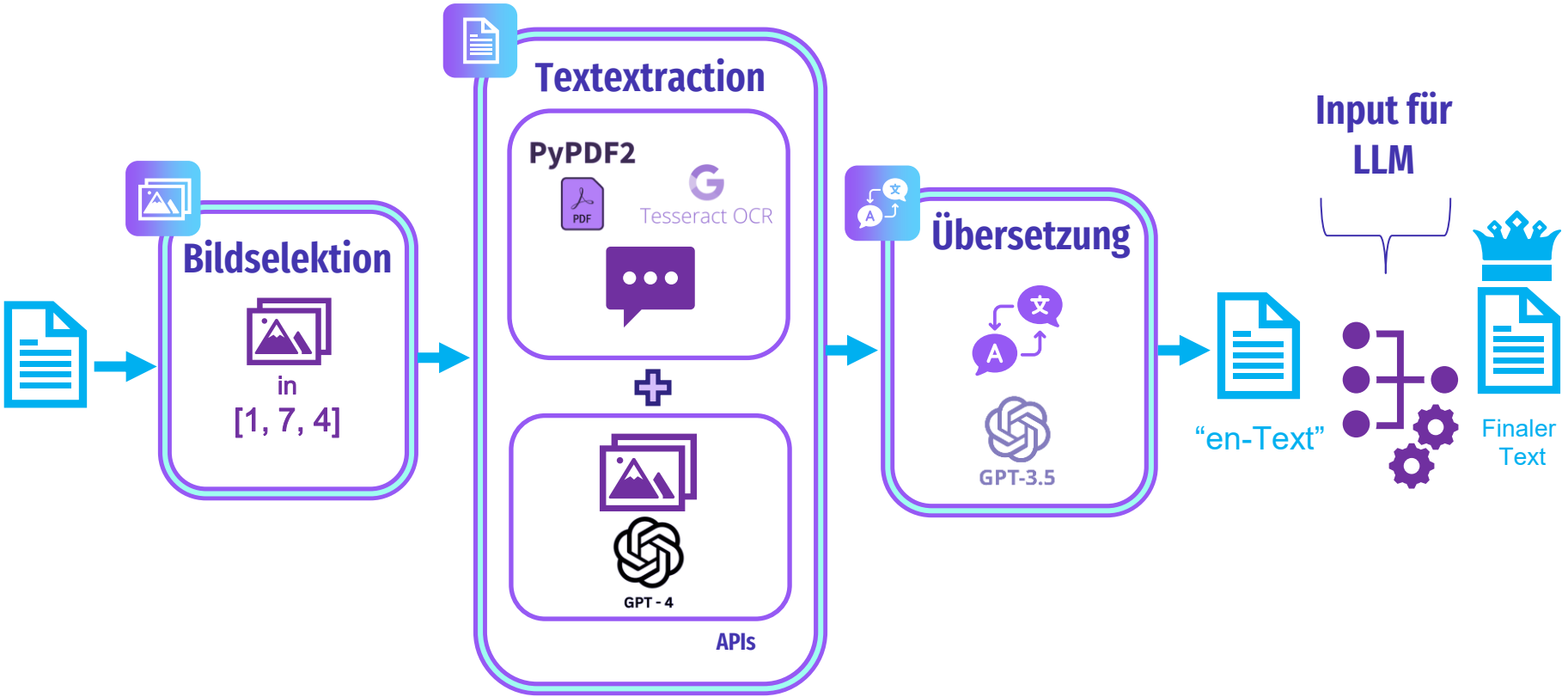
Vom Lehrer zum Schüler



Zielsetzung



Data Preprocessing



Finetuning Llama 2

Kaggle Recoursen

Nvidia P100
16 GB VRAM

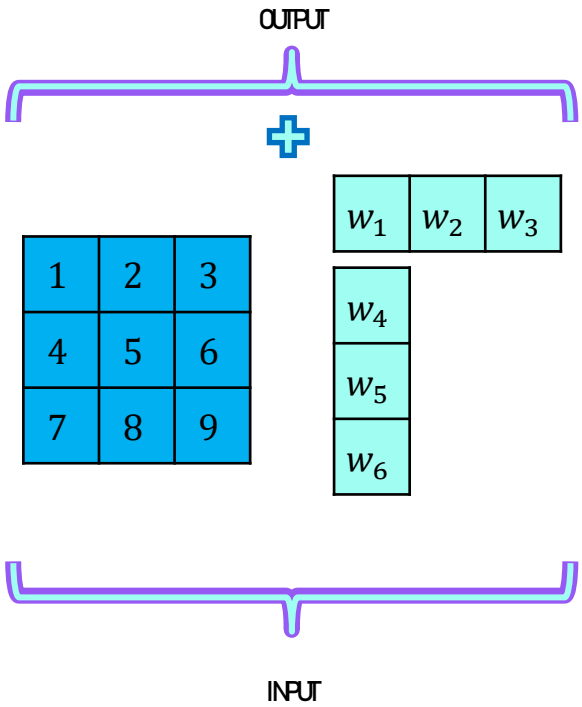
Llama 2 Requirement
28 GB VRAM

QLoRA

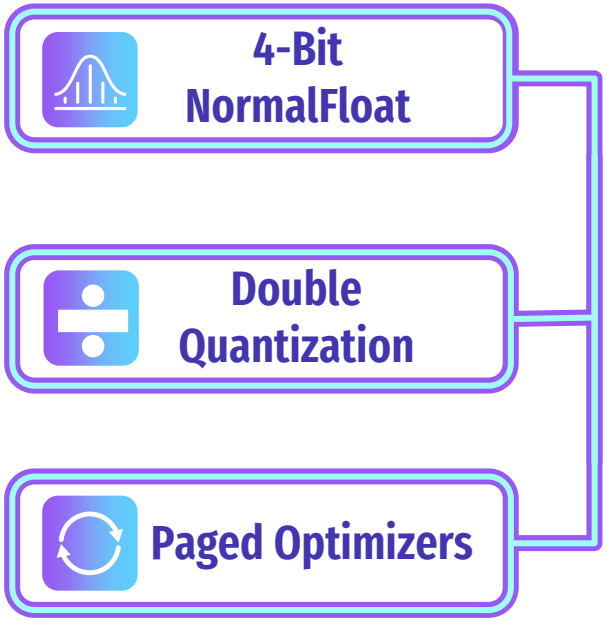
6,9 GB VRAM

Finetuning Llama 2

LoRA

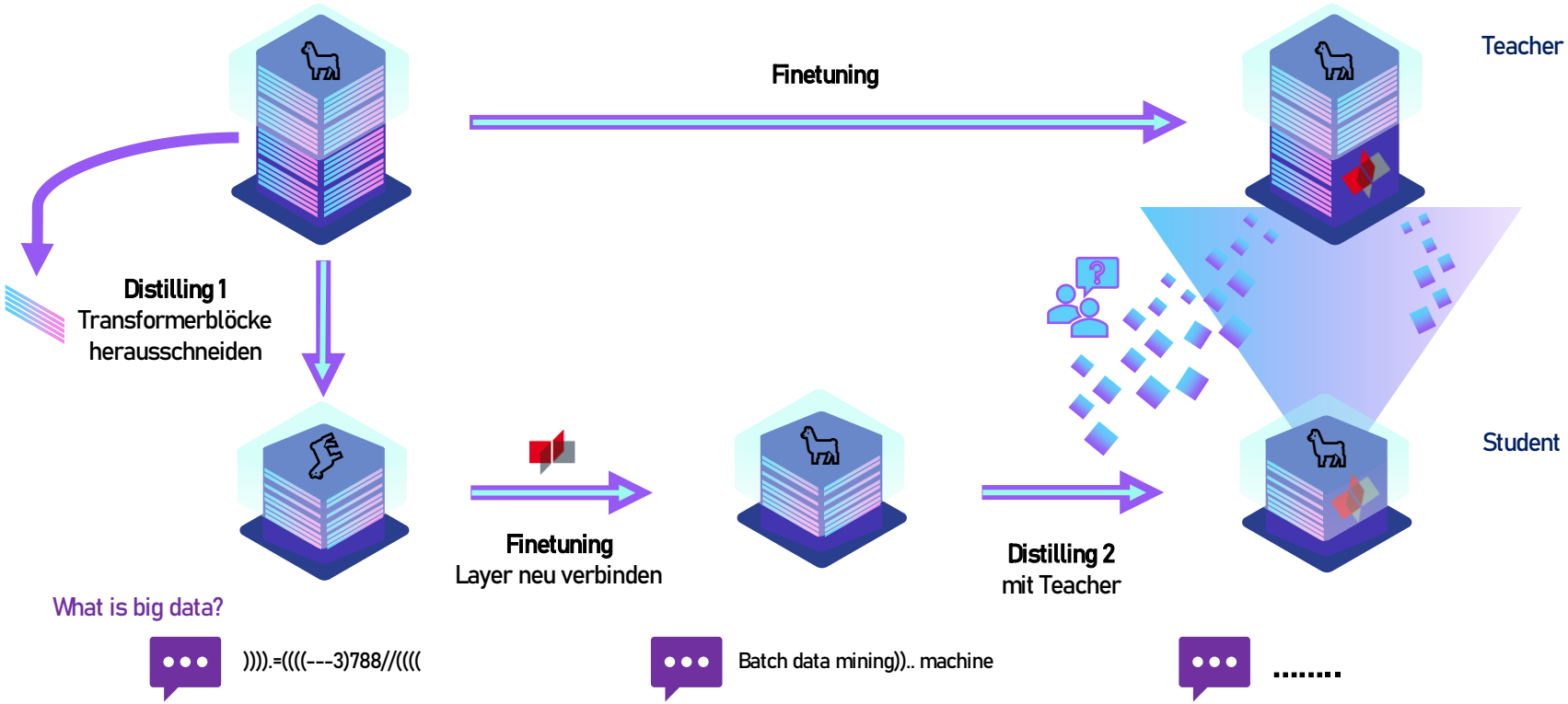


QLoRA

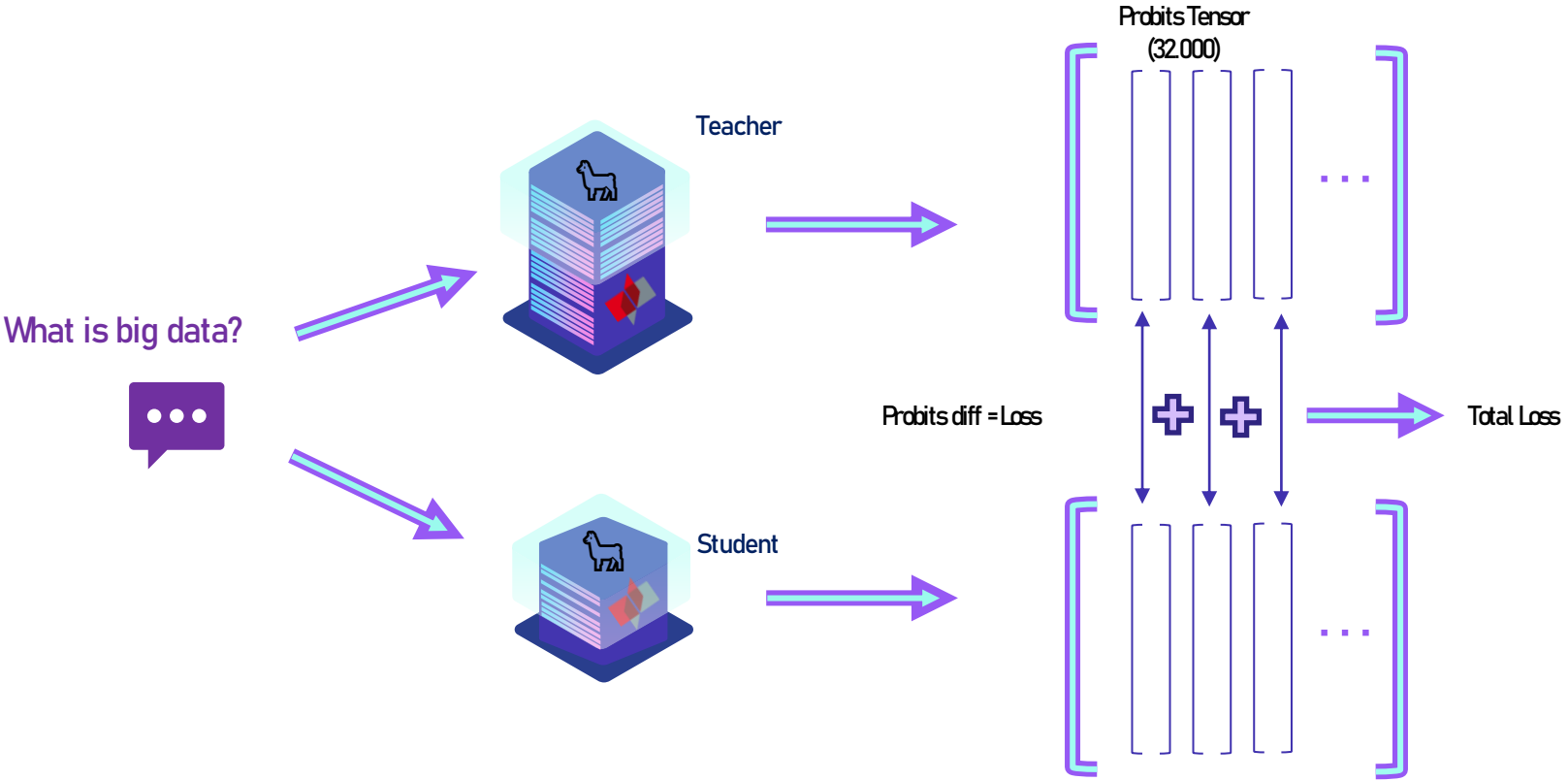


Finetuning Llama 2

Distilling Llama 2



Distilling Llama 2



Evaluierungsprozess

What is big data?

100 Fragen

Musterantwort

Antwort LLM 1

Antwort LLM 2

Word2Vec

Cosinusähnlichkeit

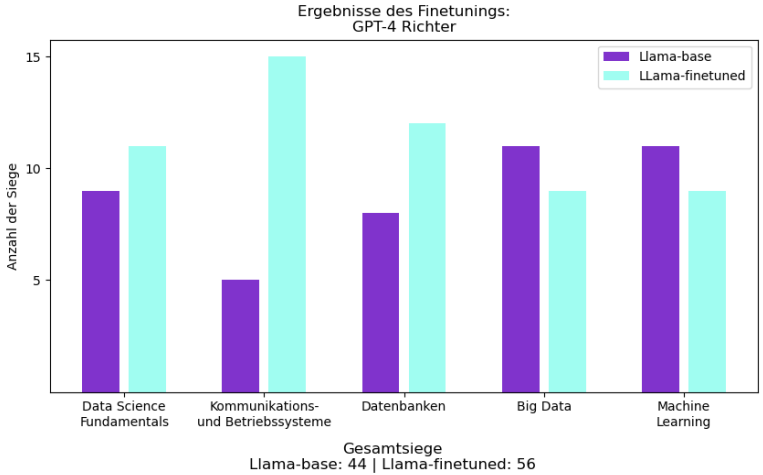
GPT - 4

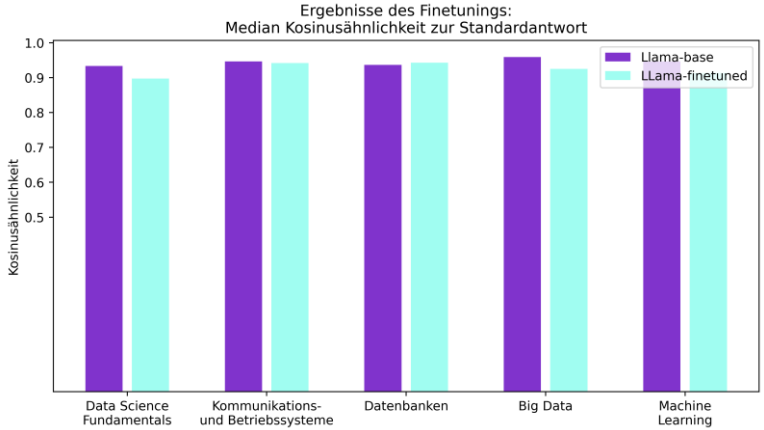
Cos Winner

LLM Winner

Erklärung LLM

Ergebnisse Finetuning





Total GPT Wins (Base): **44**
Total GPT Wins (Fine-tuned): **56**

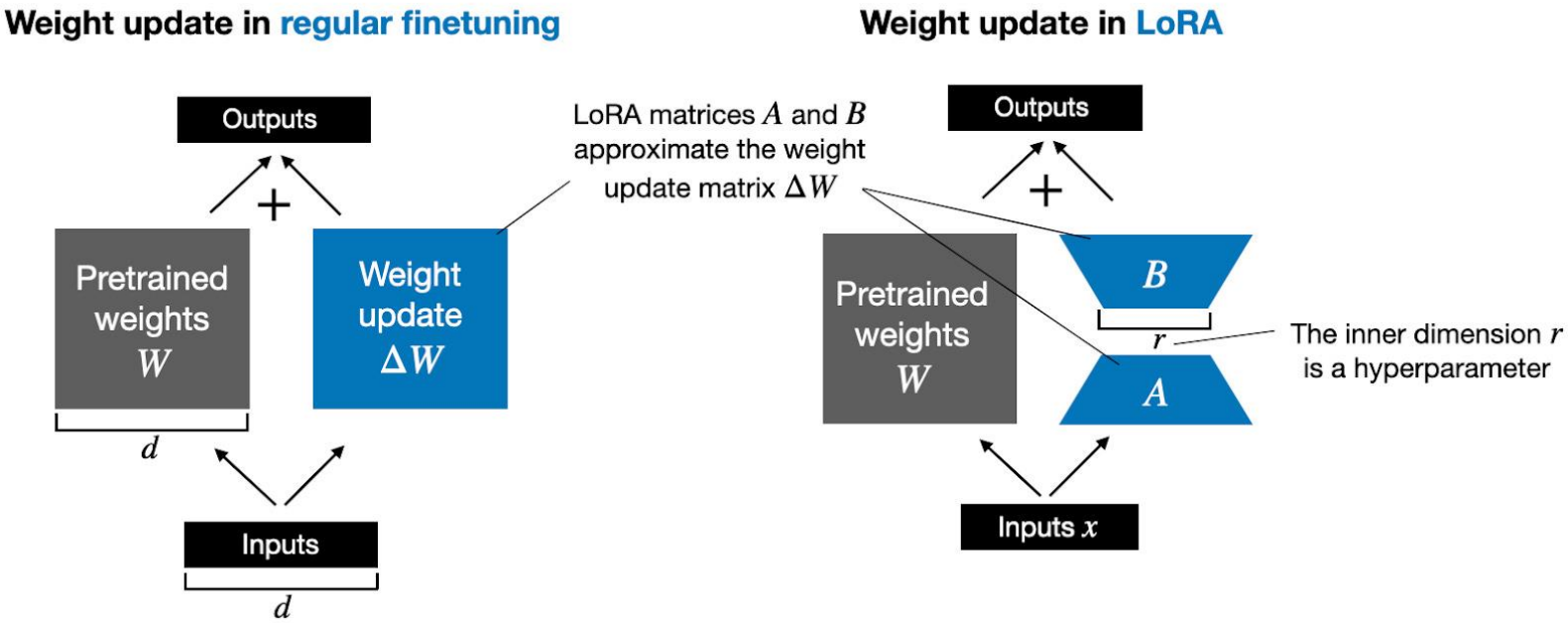
Total Similarity Wins (Base): **66**
Total Similarity Wins (Fine-tuned): **34**

GPT - 4

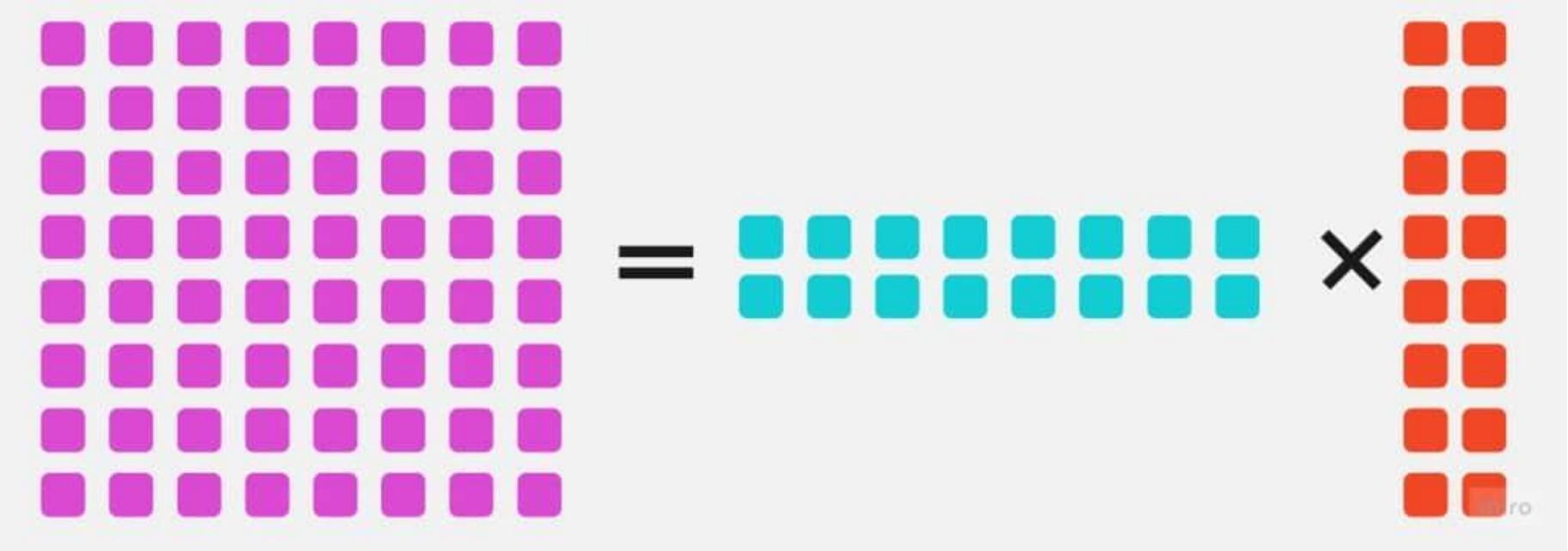
-

Backup Folien:

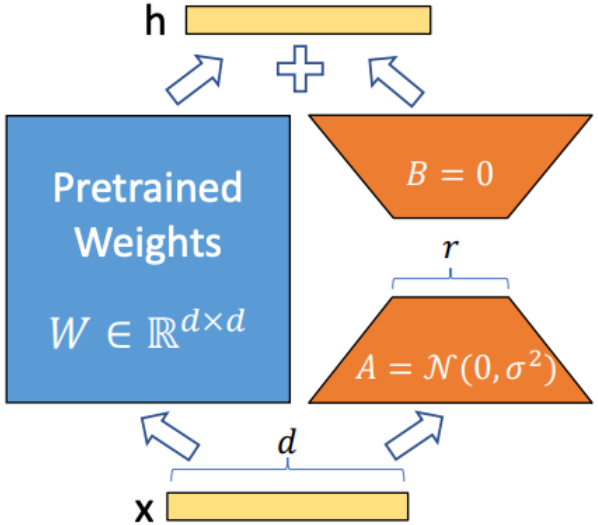
LoRA


Weight update in regular finetuning
Outputs
Pretrained weights W
Weight update ΔW
d
Inputs
d
LoRA matrices A and B approximate the weight update matrix ΔW
Weight update in LoRA
Outputs
Pretrained weights W
B
r
A
The inner dimension r is a hyperparameter
Inputs x

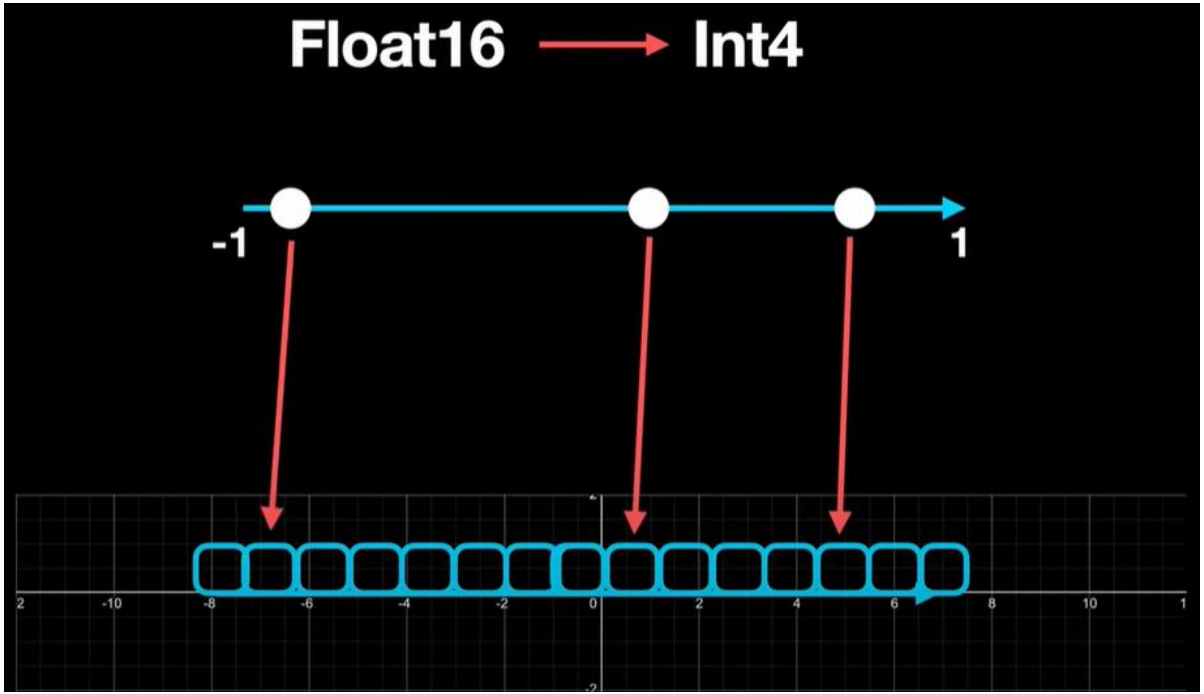
LoRA



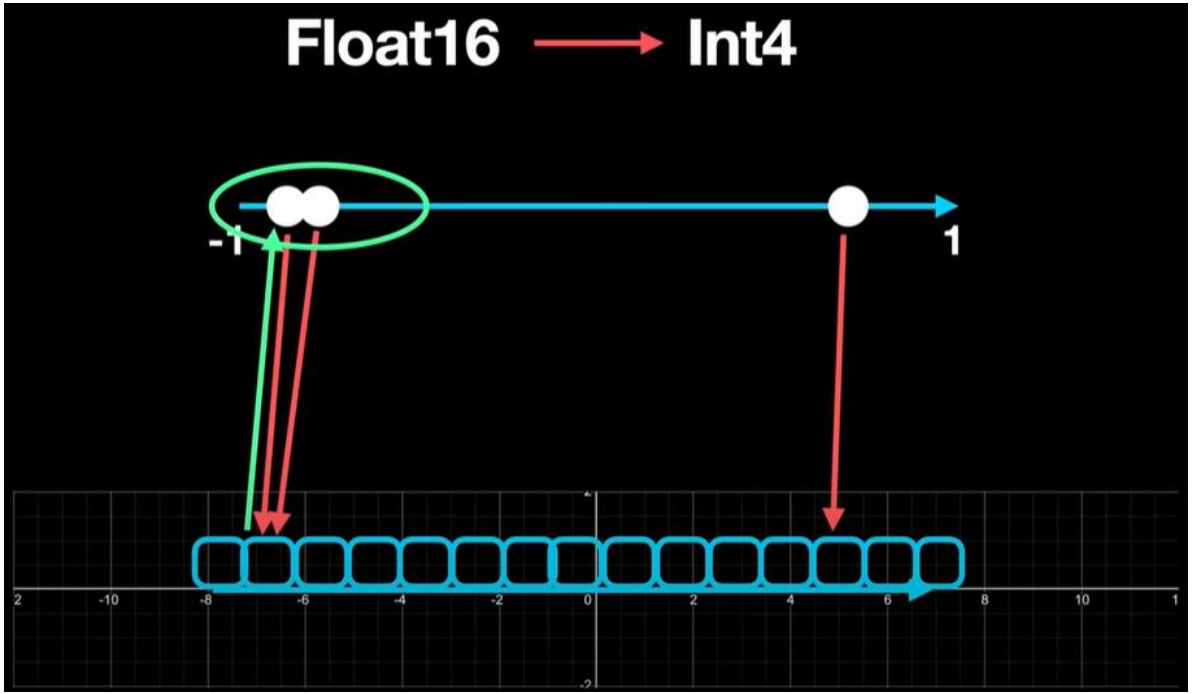
LoRA


h
Pretrained Weights
$W \in \mathbb{R}^{d \times d}$
$B = 0$
r
$A = \mathcal{N}(0, \sigma^2)$
d
x

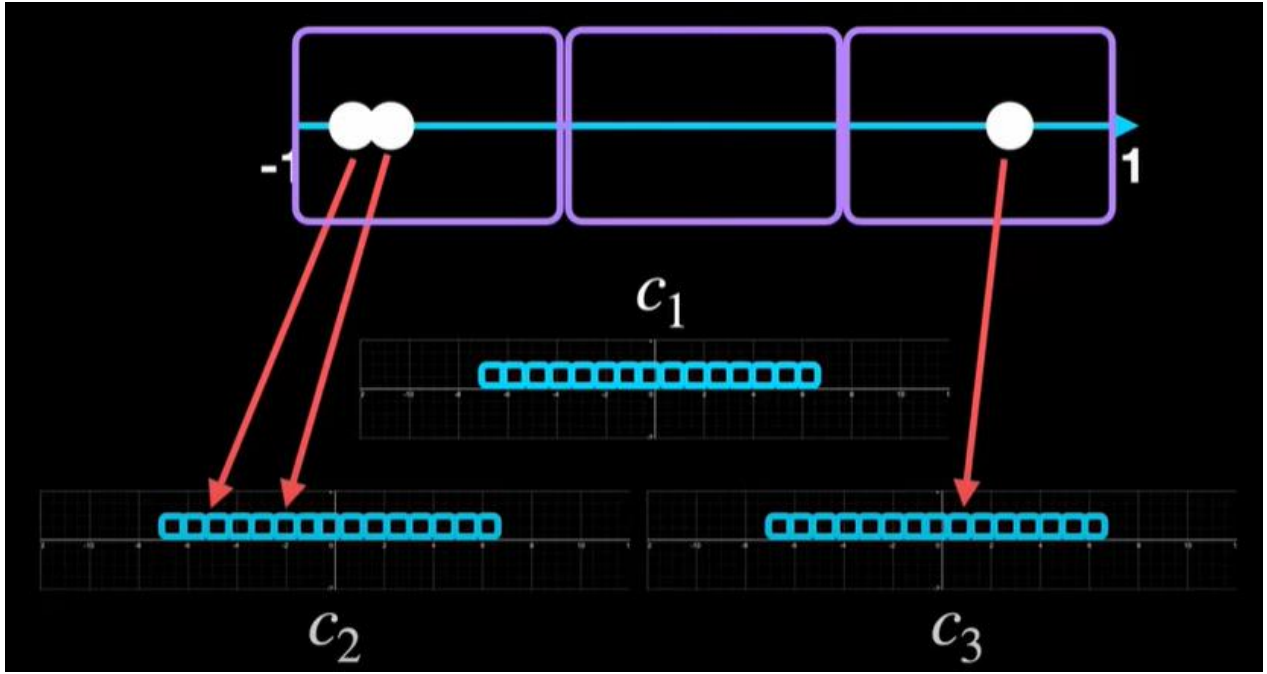
QLoRA: 4-Bit-Normal Float



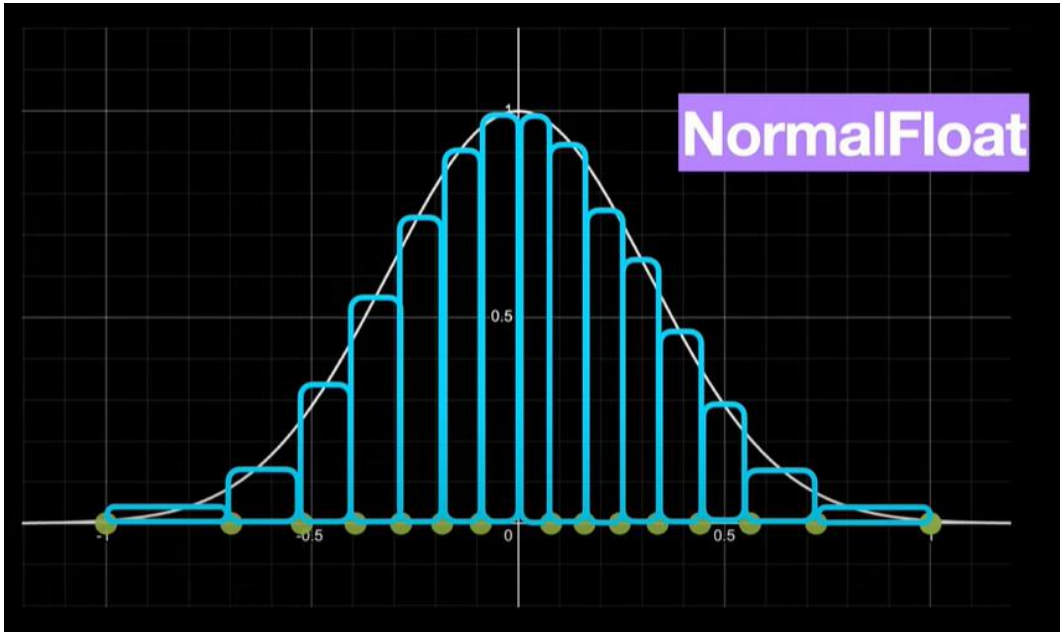
QLoRA: 4-Bit-Normal Float



QLoRA: 4-Bit-Normal Float



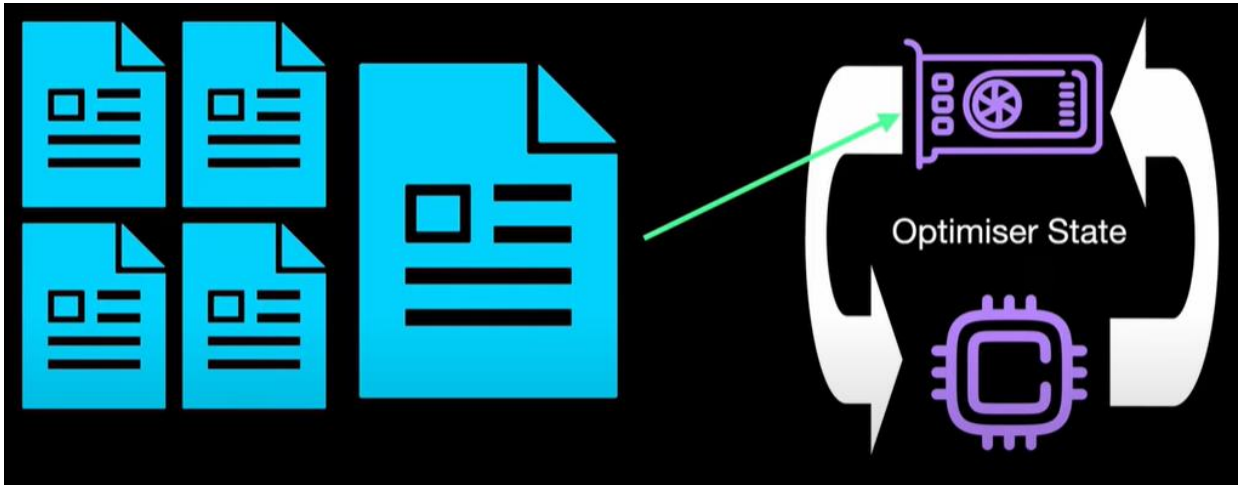
QLoRA: 4-Bit-Normal Float



QLoRA: Double Quantization

Double Quantization We introduce *Double Quantization* (DQ), the process of quantizing the quantization constants for additional memory savings. While a small blocksize is required for precise 4-bit quantization [13], it also has a considerable memory overhead. For example, using 32-bit constants and a blocksize of 64 for $\mathbf{W}$, quantization constants add $32/64 = 0.5$ bits per parameter on average. Double Quantization helps reduce the memory footprint of quantization constants.

QLoRA: Paged Optimizers



Ergebnisse Destillation

What is the purpose of memory abstraction through base and limit registers, and what are its advantages and disadvantages?	Memory abstraction through base and limit registers provides a simple solution for dynamic relocation. The base register stores the start address, and the limit register stores the program's length, allowing the program to run in memory without static relocation. The advantage is that it effectively resolves memory abstraction issues. However, the disadvantage is that every memory access requires an addition and a comparison, increasing overhead. Notably, systems like the Intel 8088 only had a base register, presenting challenges in effectively managing memory limits.	Memory abstraction is a technique used in computer architecture to separate the logic of managing memory from the rest of the system. sierpraws base and limit registers are two examples of memory abstraction techniques used to manage memory in computer systems. Here is the purpose, advantages, and disadvantages of these techniques:\n\nPurpose of Memory Abstraction:\n\nThe primary purpose of memory abstraction is to decouple the memory management logics from the rest of the system. By doing so, the system can focus on other tasks	Memory abstraction through base and limit registers is a technique used to manage memory in computer systems. Memory abstraction is needed because most microprocessors can address only a certain amount of memory (up to 4 GB in the 32-bit x86 architecture, up to 8 GB in the 64-bit x86 architecture and up to 16 TB in 64-bit architectures like SPARC64) and to provide a way to manage multiple addresses separately (different programs	2	0.941656	0.918323	1	The second answer is better as it expands more directly on the purpose of memory abstraction in the context of base and limit registers. It talks about how microprocessors can address certain amounts of memory and the necessity of managing multiple addresses, which aligns with the sample solution's mention of dynamic relocation and program running without static relocation. Although the answer could be more detailed, it is more relevant to the question about the purpose and also briefly touches on advantages and disadvantages, such as addressing limits of microprocessors. The first answer starts to move in the right direction by discussing the purpose of memory abstraction but cuts off abruptly without providing the same level of detail as the second answer, particularly failing to mention the role of base and limit registers specifically or their advantages and disadvantages.