10 WAYS TO LOWER LLM INFERENCE COST

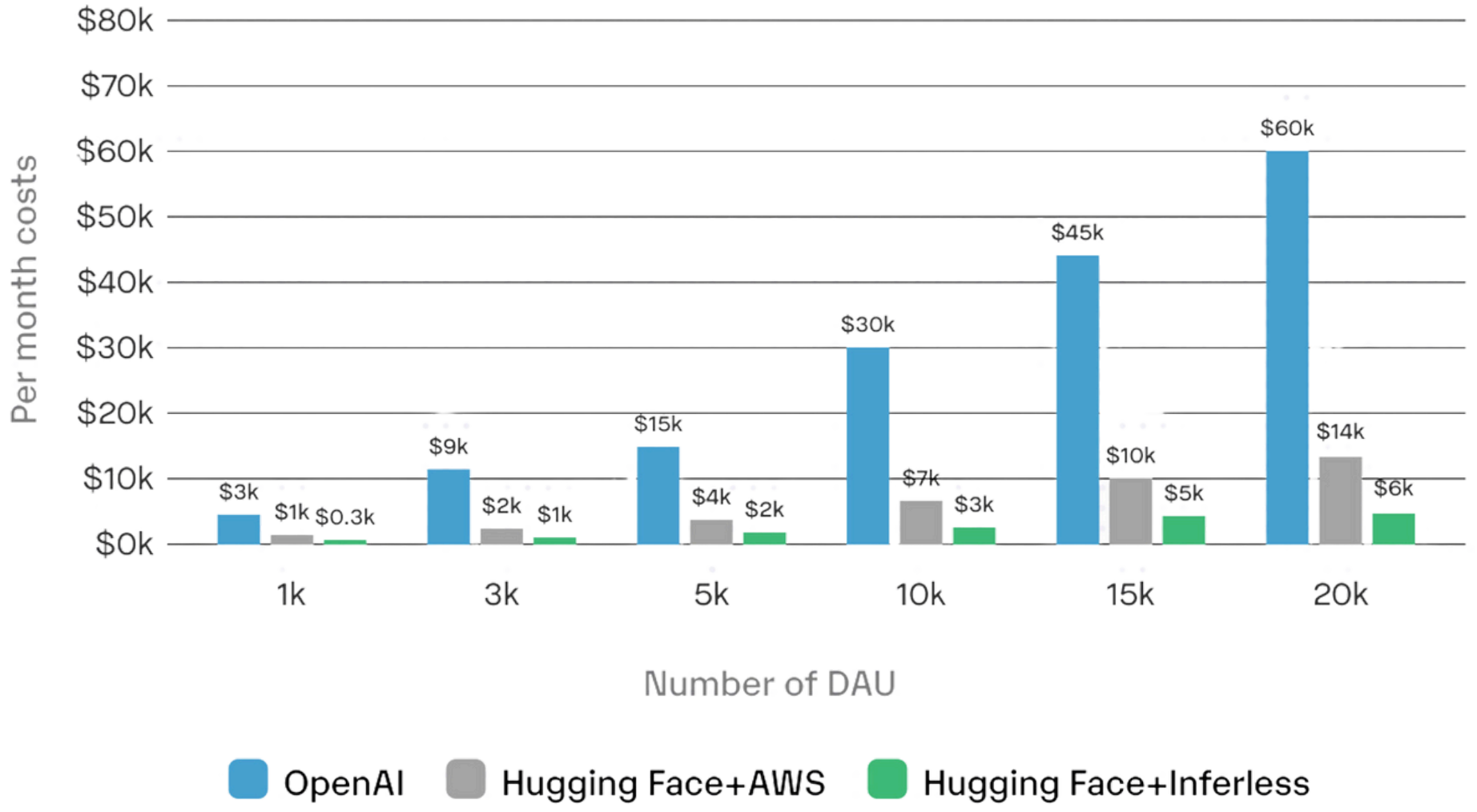
Bhavishya Pandit

OPTIMIZING LLM COSTS IS KEY TO SCALING AI

Large Language Models (LLMs) like GPT-4 have revolutionized AI, but their operational costs can be high. Reducing these costs is crucial for businesses, developers, and researchers who want to scale AI effectively and sustainably.

Why Should We Care?

- **Scalability**: Lowering costs helps businesses scale AI applications without breaking the bank.
- **Profitability**: More affordable inference = reduced operational expenses and better margins.
- **Sustainability**: Optimizing inference reduces energy consumption, contributing to greener AI practices.



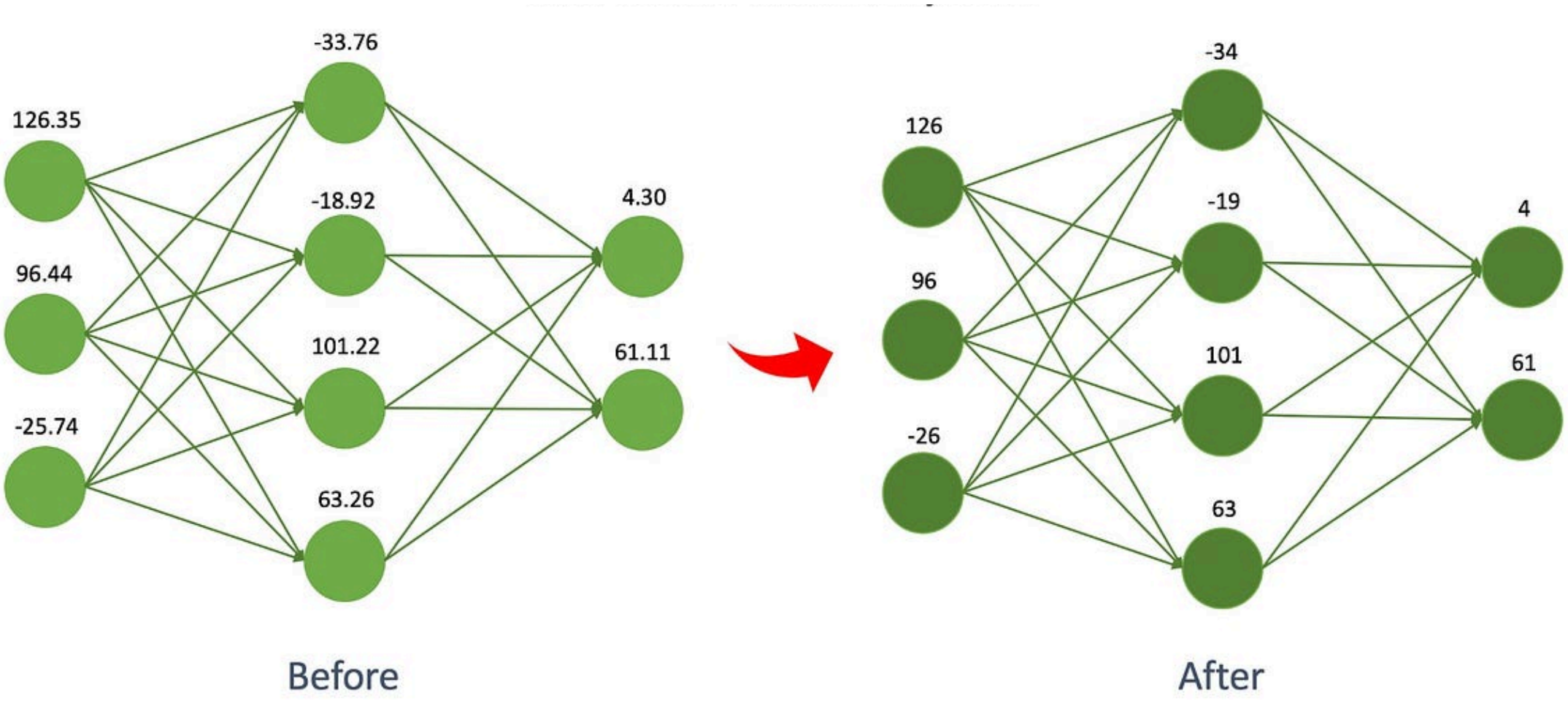
Credits: Inferless

Bhavishya Pandit

QUANTIZATION

What is it?
Quantization reduces the precision of the model weights, typically from 32-bit floating point to 8-bit integers, resulting in reduced memory usage and faster computations.

Why it matters:
Lower precision means fewer resources needed for storage and computation, reducing costs while maintaining a good level of model performance.

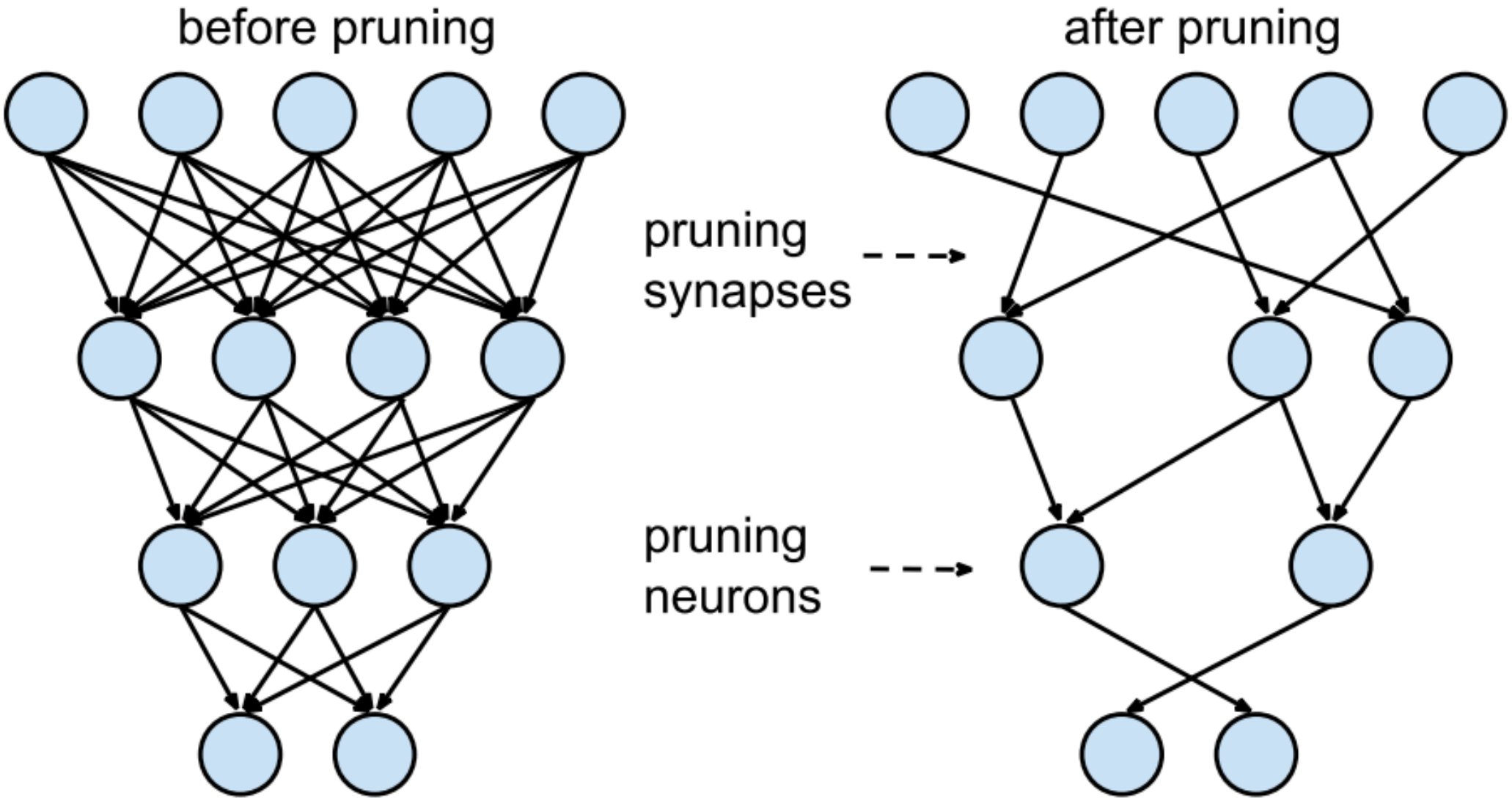


PC : Maarten Grootendorst

PRUNING

What is it?
Pruning involves removing unnecessary weights (parameters) from the model that don't contribute significantly to its predictions.

Why it matters:
A smaller model requires less computation, reducing the infrastructure cost while maintaining performance in many cases.

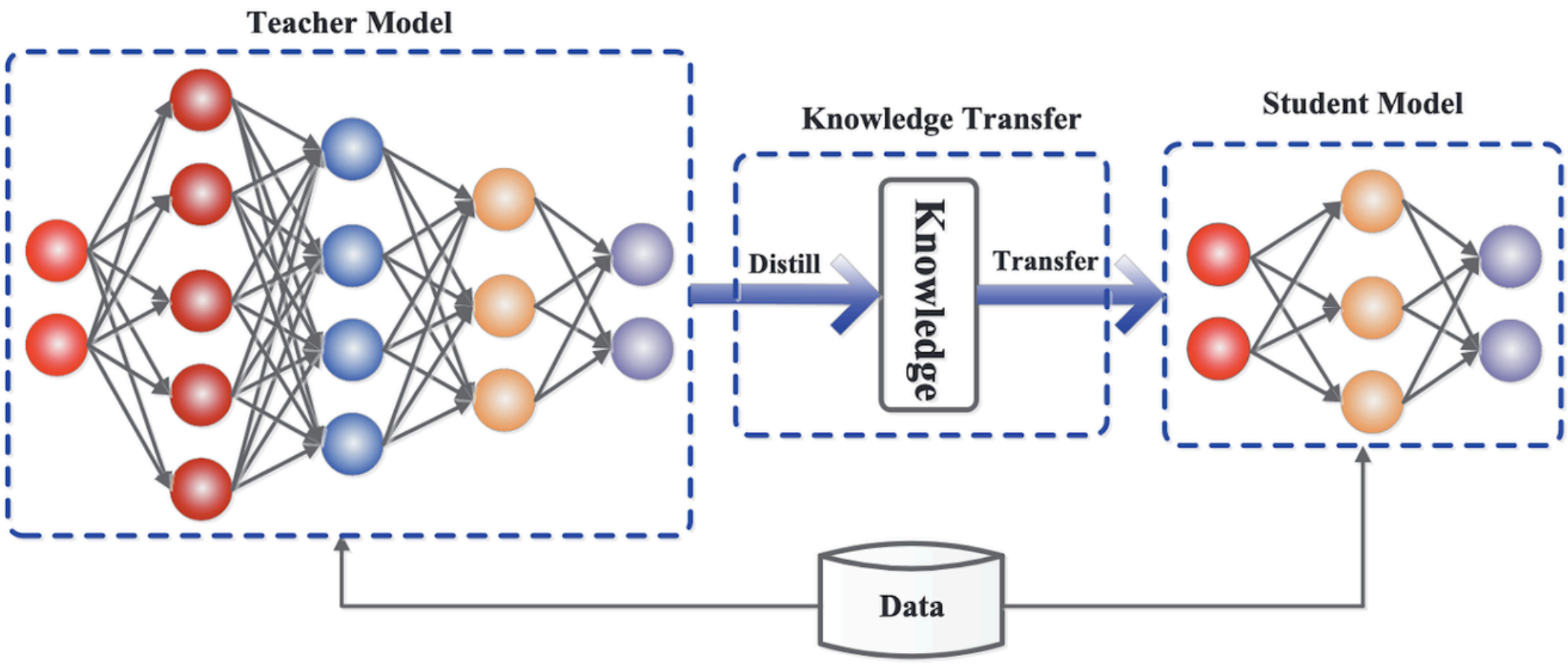


PC : novita.ai

KNOWLEDGE DISTILLATION

What is it?
This technique transfers knowledge from a large, complex model (the teacher) to a smaller, simpler model (the student).

Why it matters:
Distillation allows you to deploy smaller, faster models with lower operational costs while retaining much of the original model's accuracy.

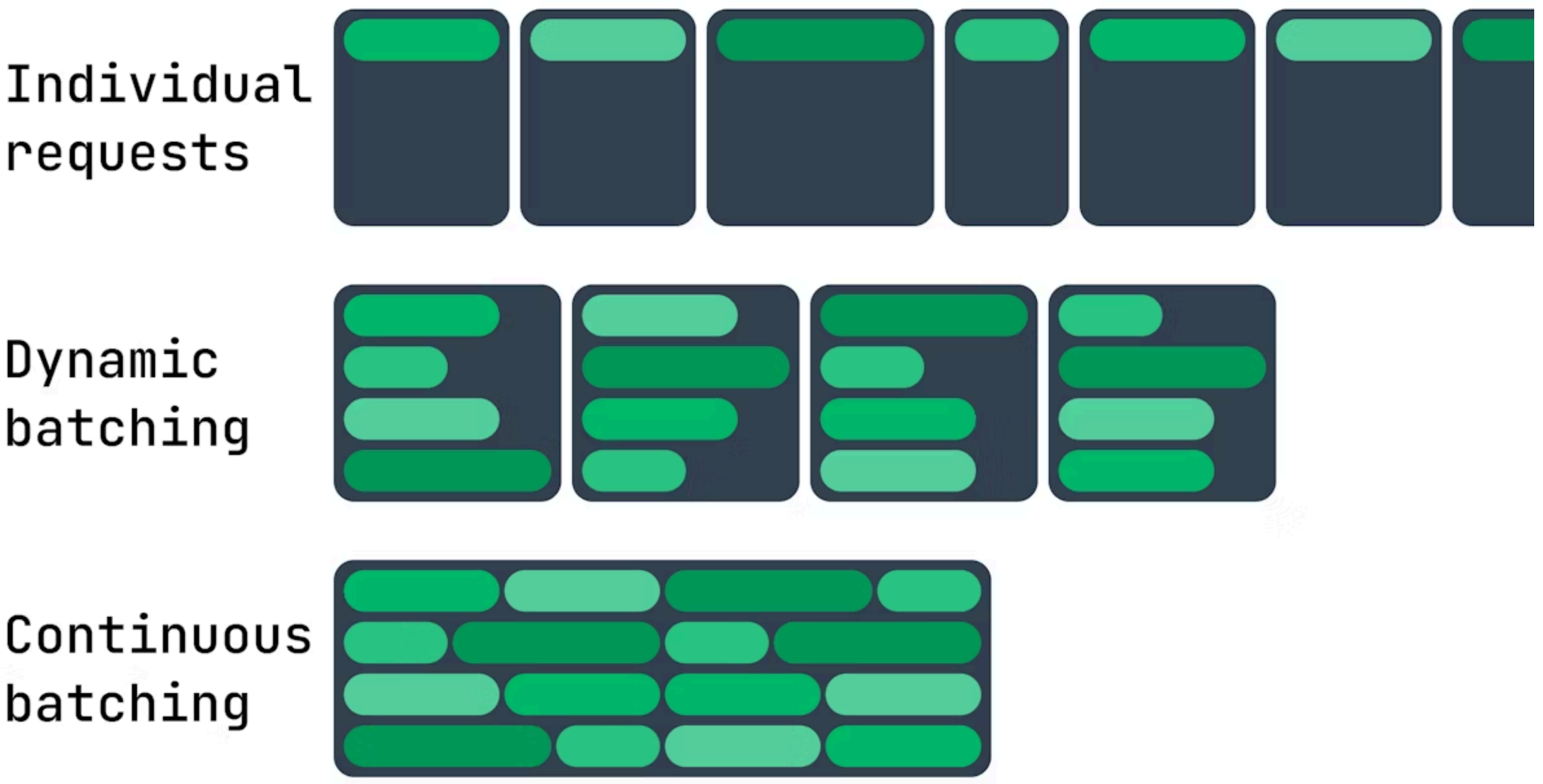


PC : DataCamp

BATCHING

What is it?
Batching involves processing multiple inputs together in one go rather than handling each input separately.

Why it matters:
It optimizes resource usage (such as GPU or CPU cycles) by reducing the overhead per request, which leads to a significant cost reduction in real-time inference.

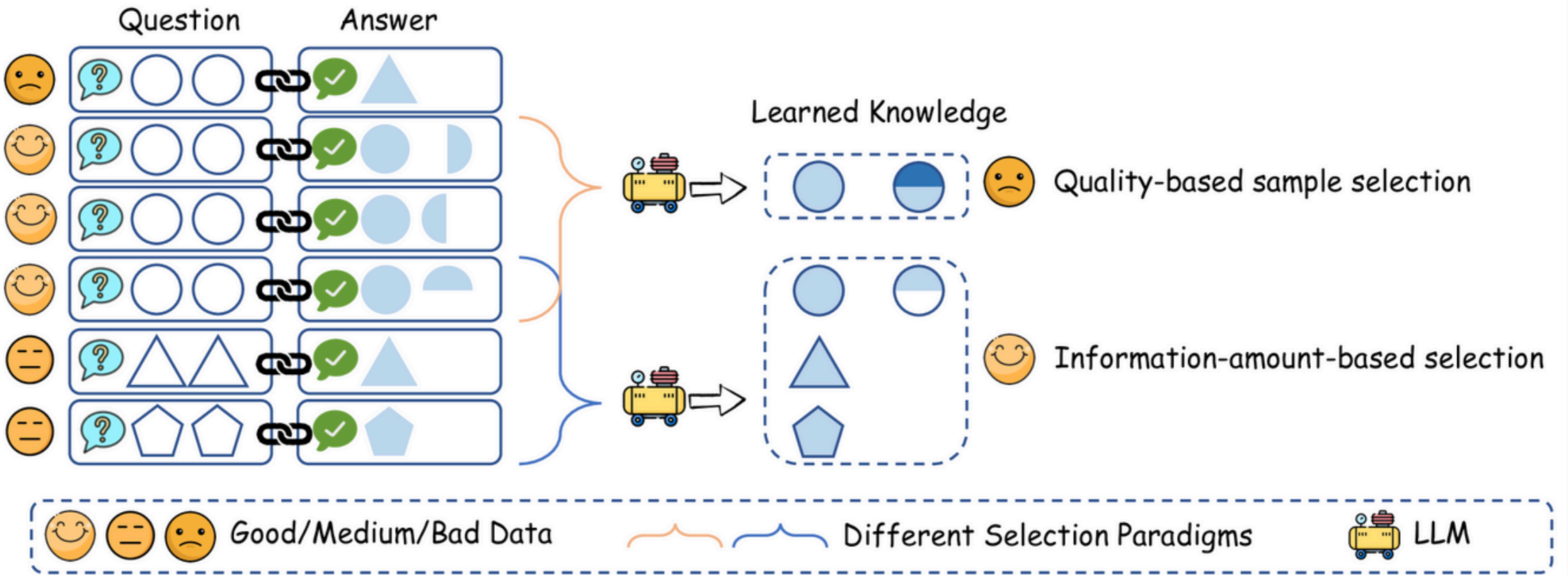


PC : Baseten

MODEL COMPRESSION

What is it?
Model compression reduces the size of the model through techniques like weight sharing, low-rank factorization, and knowledge distillation.

Why it matters:
Smaller models mean lower storage, reduced memory bandwidth, and faster inference, ultimately lowering both cloud and hardware costs.

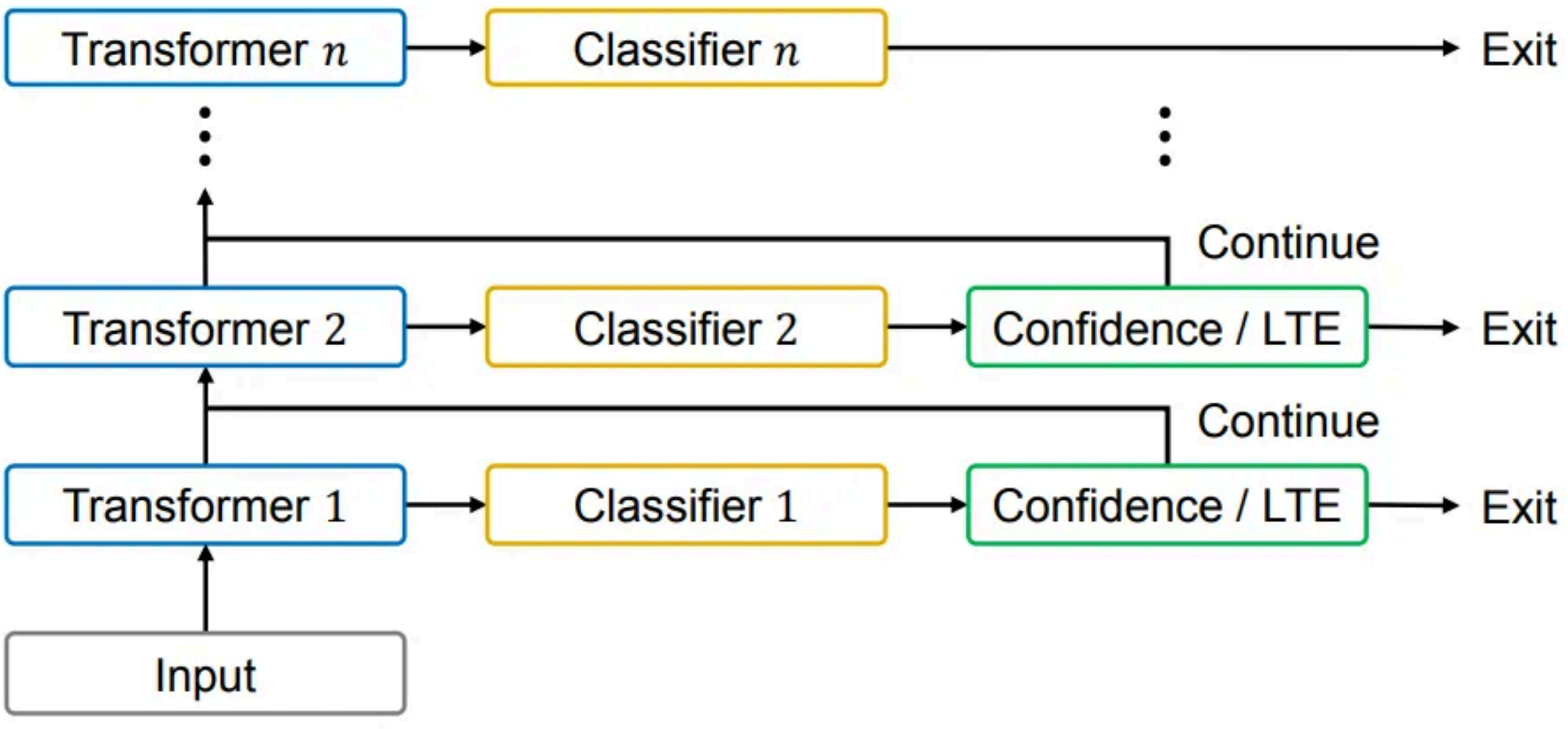


PC : AIModels.fyi

EARLY EXITING

What is it?
Early exiting allows the model to stop processing once a certain level of confidence or decision threshold is met, instead of running the entire model.

Why it matters:
By avoiding unnecessary computation, this technique significantly reduces latency and cost, especially in real-time applications.

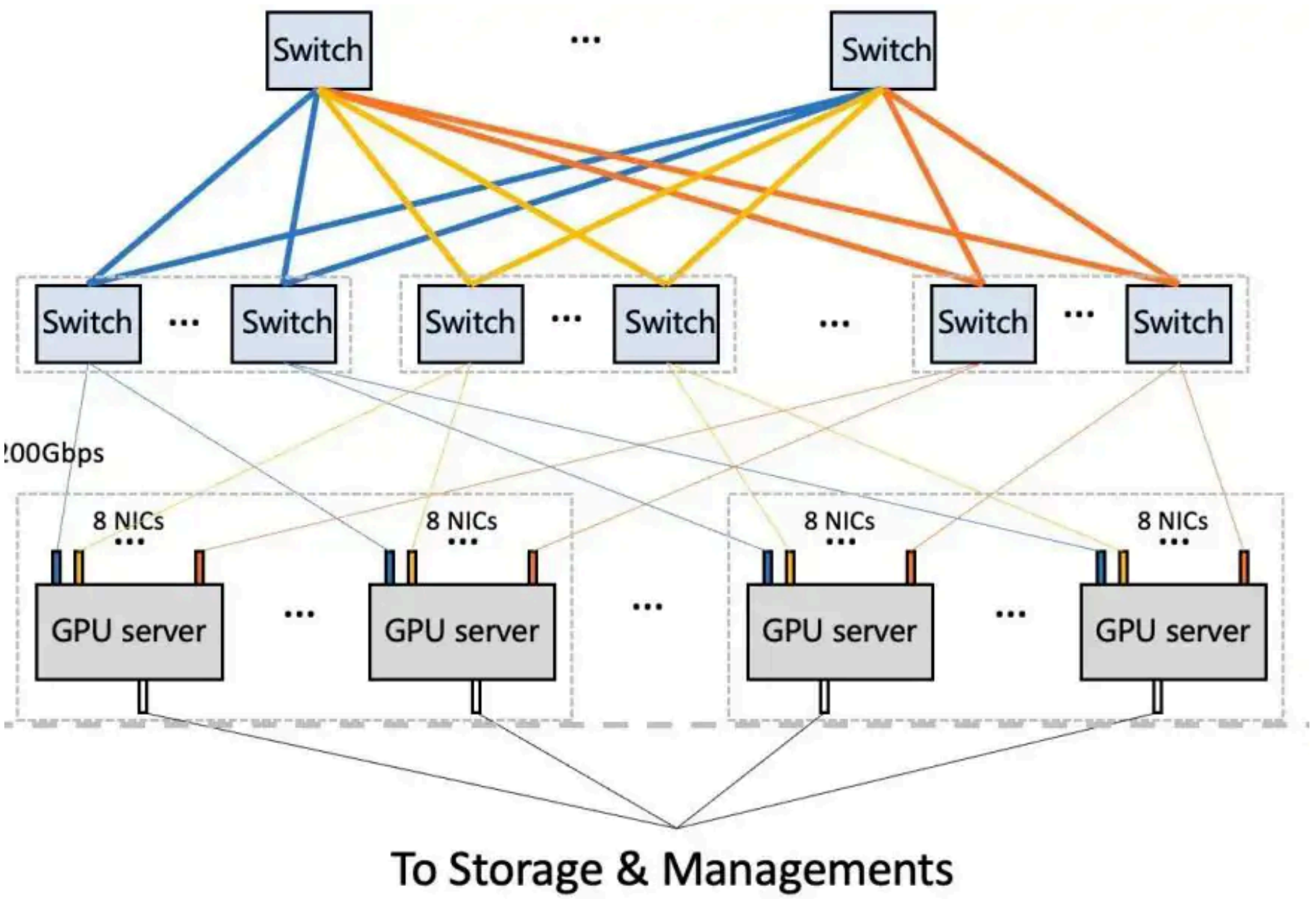


PC : Towards Data Science

OPTIMIZED HARDWARE

What is it?
Using specialized hardware like TPUs or GPUs that are optimized for AI workloads can dramatically reduce the cost of running LLMs.

Why it matters:
These processors are designed to handle the large-scale matrix operations of LLMs more efficiently, resulting in faster processing and reduced infrastructure costs.

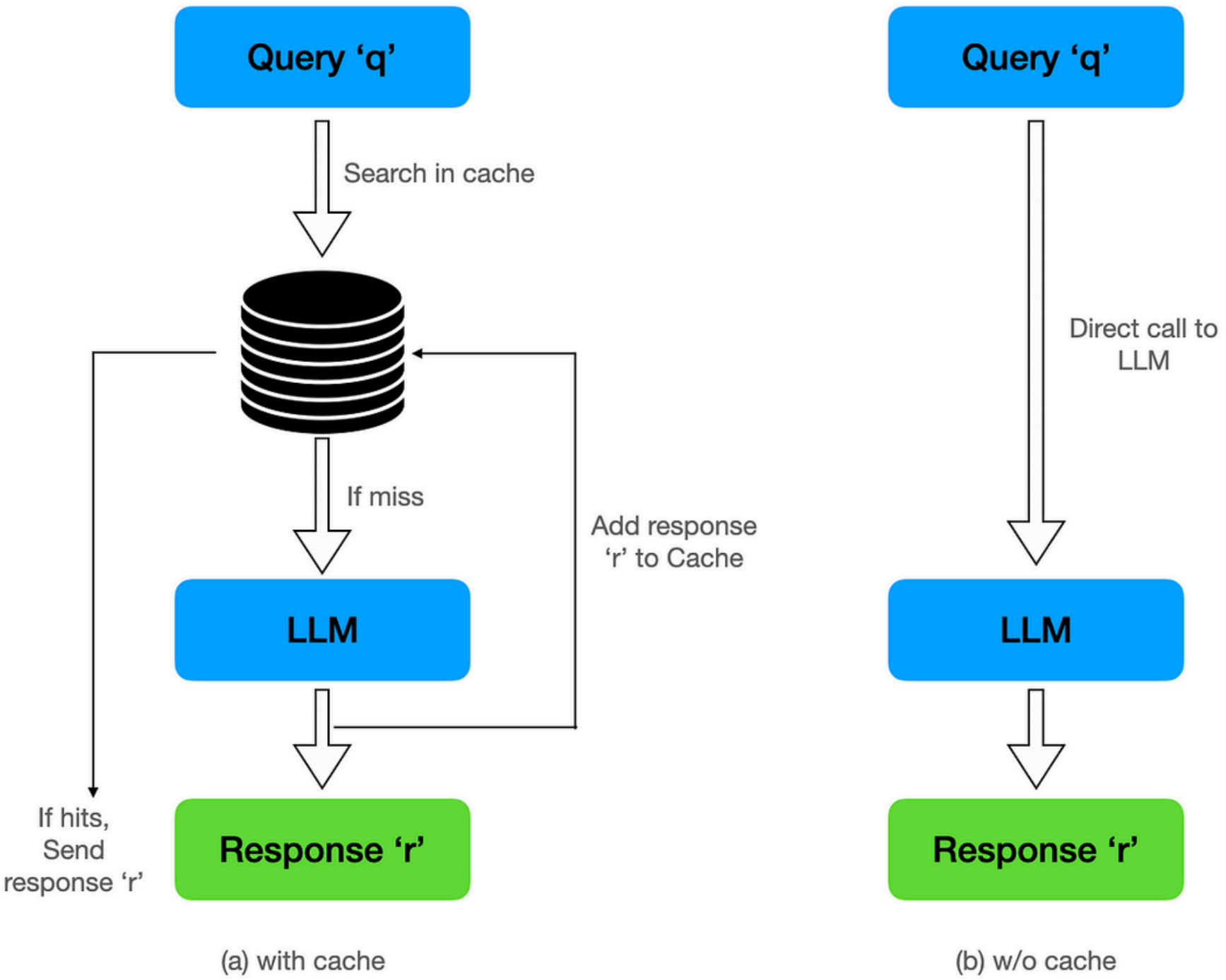


PC : NADDOD

CACHING

What is it?
Caching stores the results of common queries or computations, so you don't have to recompute them each time.

Why it matters:
By reusing pre-computed responses, caching reduces redundant inference calls, lowering both computational costs and response times.

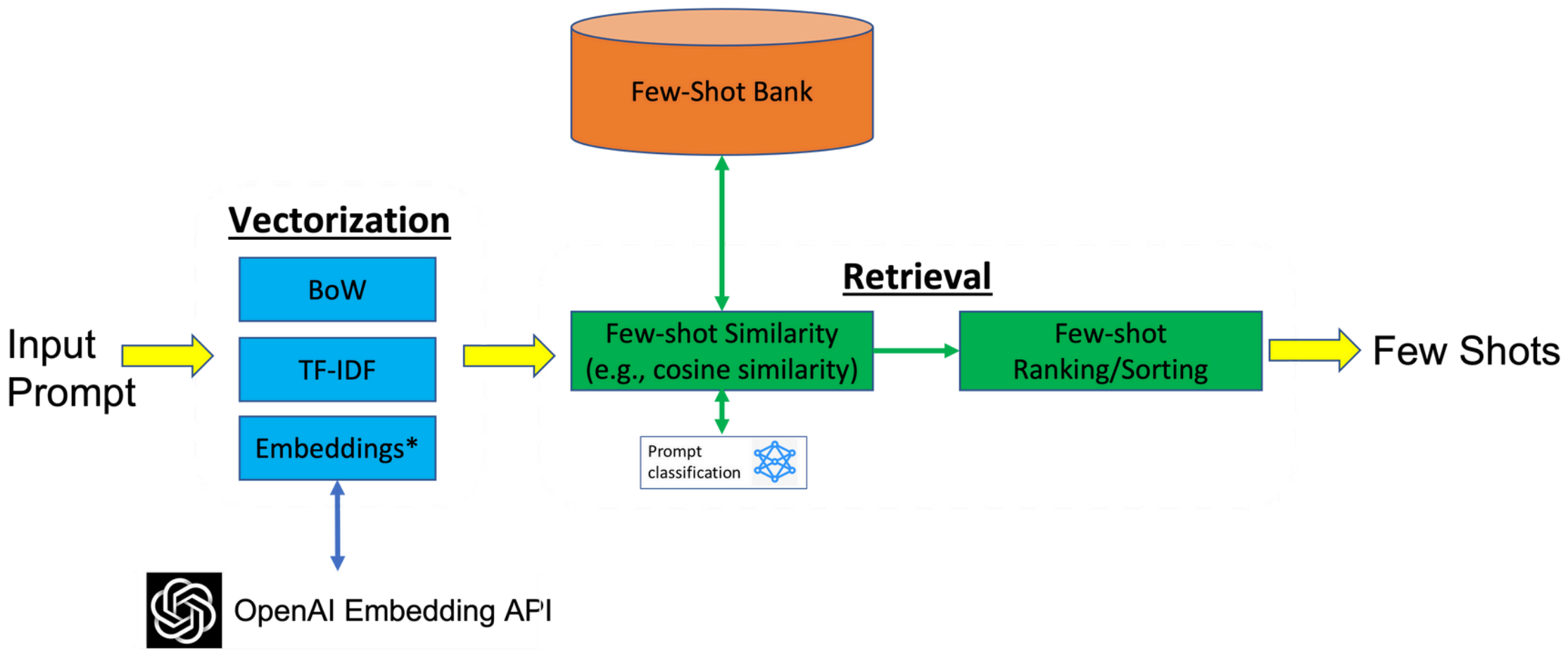


PC : Medium

PROMPT ENGINEERING

What is it?
Fine-tuning how you design and structure your input prompts can improve model performance and reduce the need for multiple inference calls.

Why it matters:
Optimizing prompts ensures that you get the best result from fewer computations, reducing the overall cost of each request.

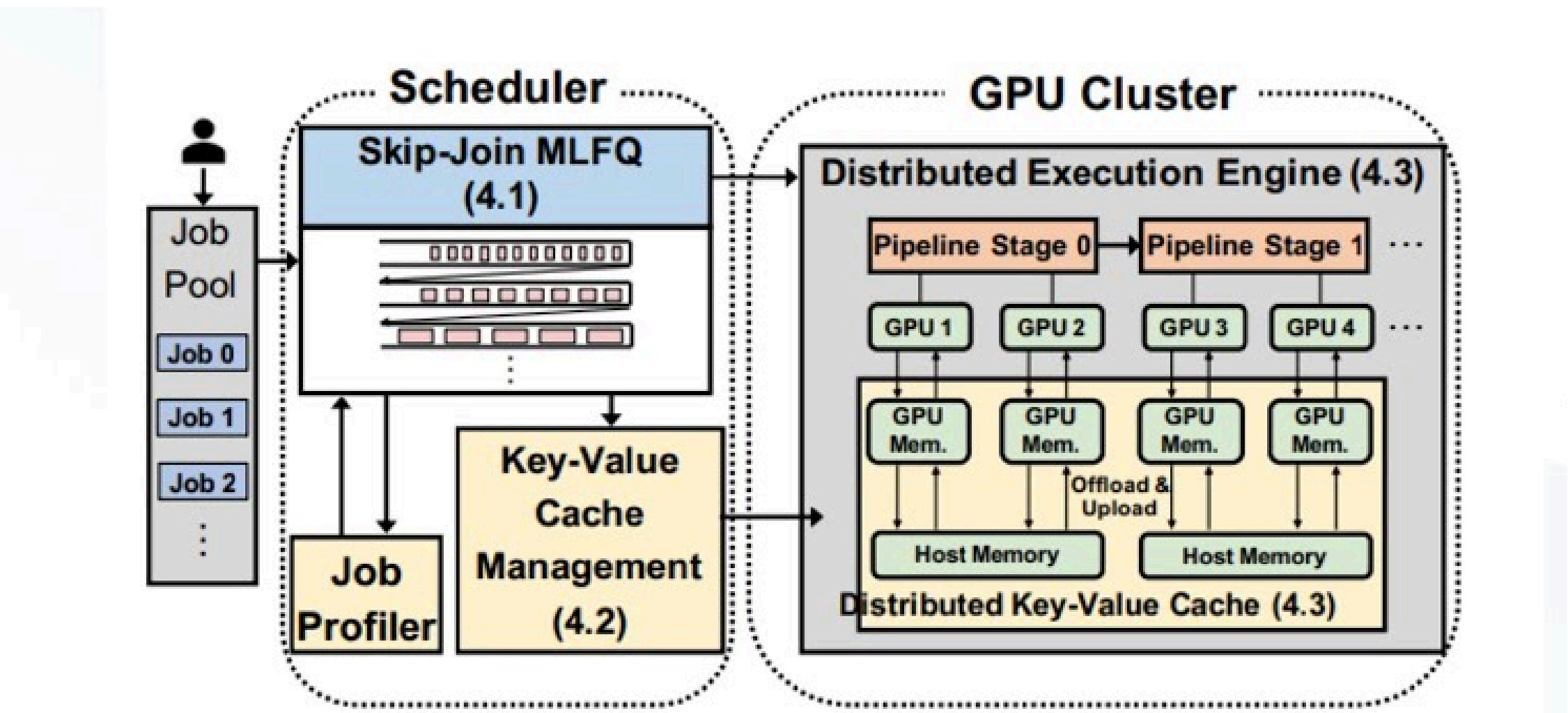


PC : Microsoft Learn

DISTRIBUTED INFERENCE

What is it?
Distributed inference breaks down a large task into smaller sub-tasks that are processed across multiple machines or GPUs.

Why it matters:
This helps balance the load, reduce latency, and scale operations efficiently, all while lowering infrastructure costs.



PC : Reddit

Follow to stay updated on Generative AI

SAVE **LIKE** **SHARE**

Bhavishya Pandit