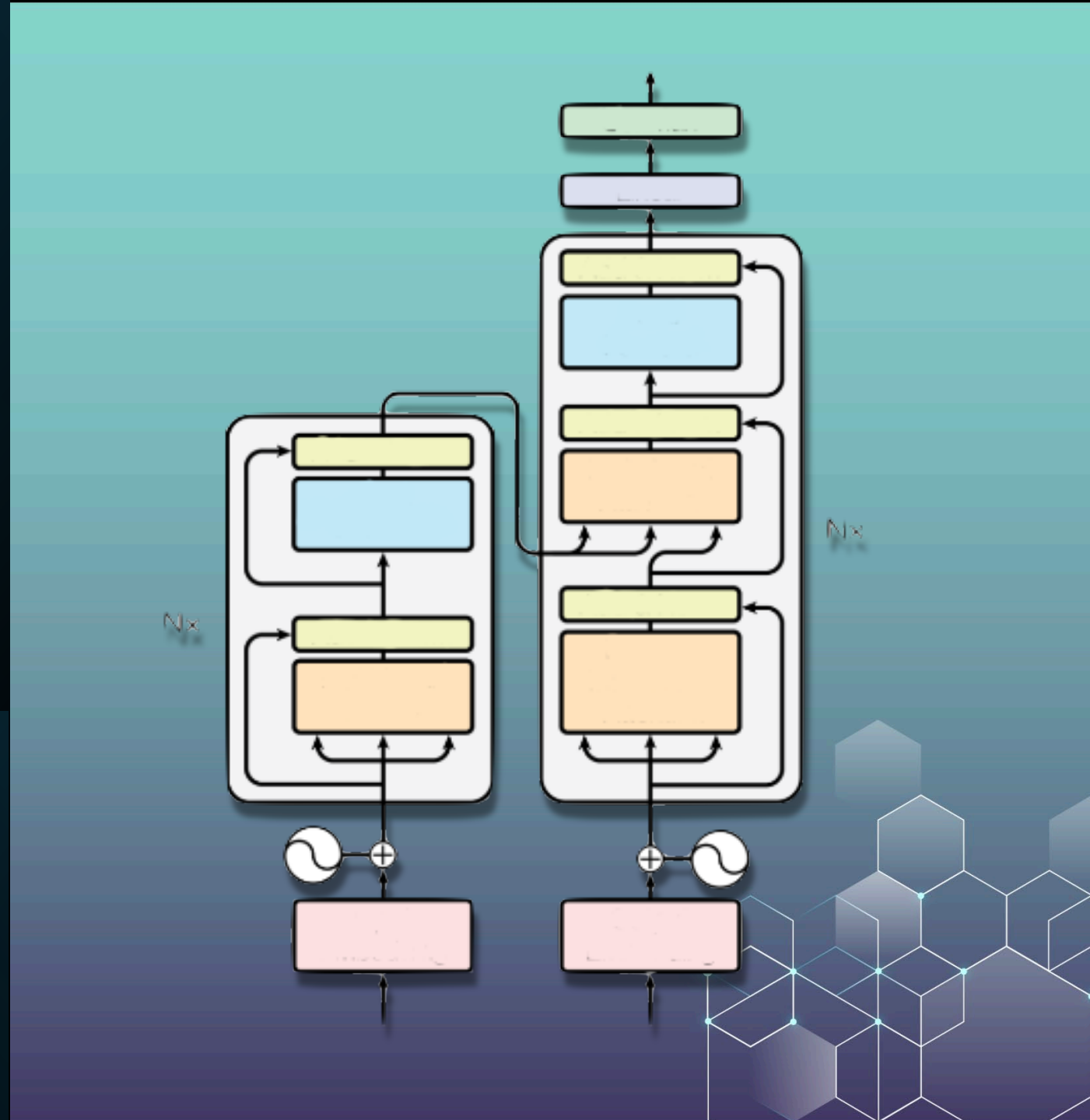
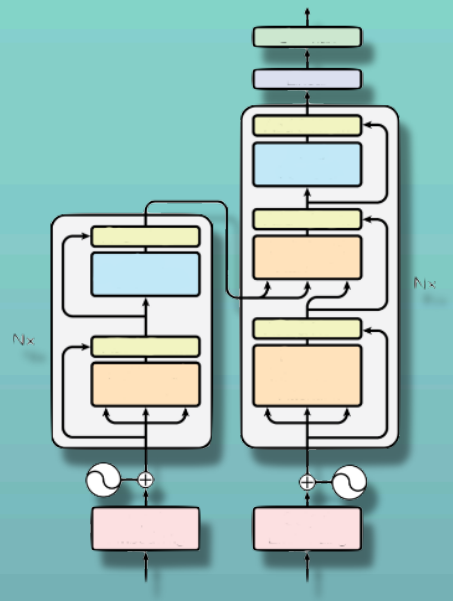
Scaling NVM Storage For Generative AI

George Williams, FMS 2025

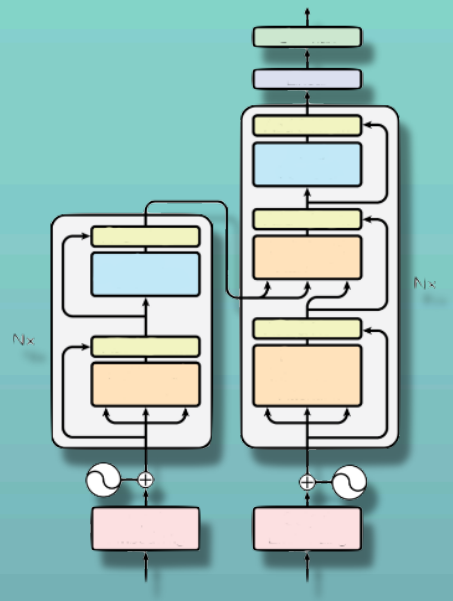
In The Beginning There Was Transformer

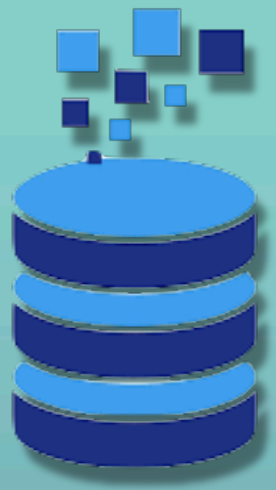




2017 2020 2022 2025

"Attention Is All You Need"

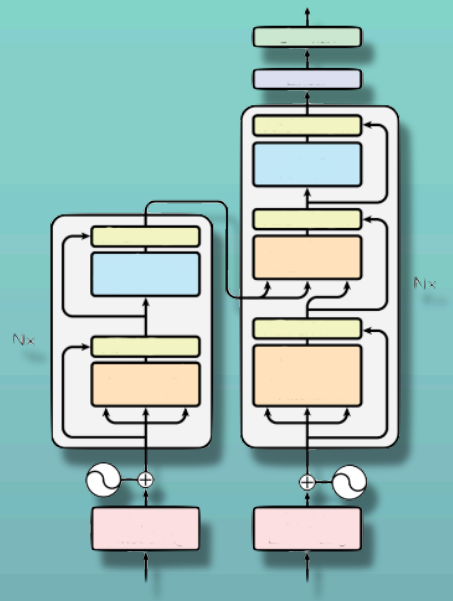


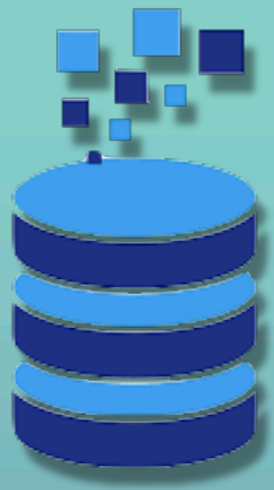


2017 2020 2022 2025

"Attention Is All You Need"

Retrieval Augmented Generation





2017

2020

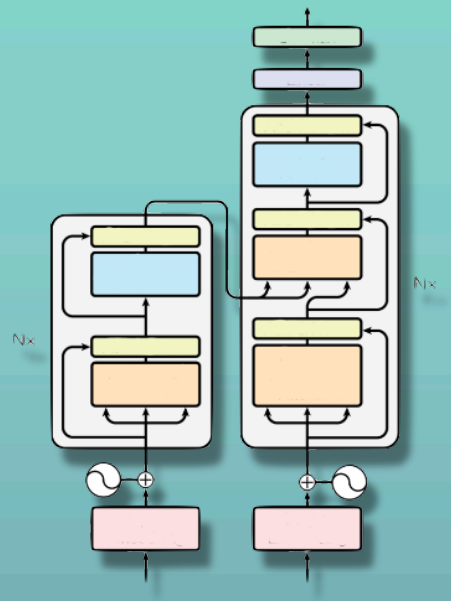
2022

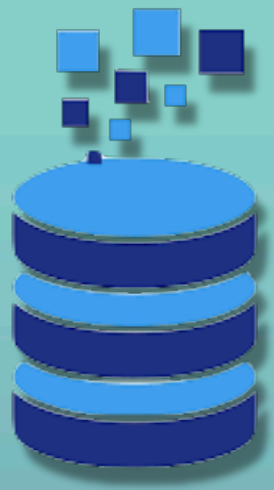
2025

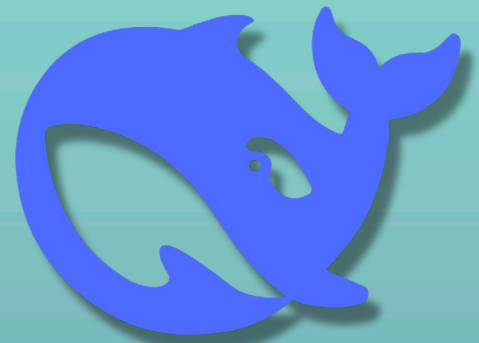
"Attention Is All You Need"

Retrieval Augmented Generation

OpenAI's ChatGPT







2017

2020

2022

2025

"Attention Is All You Need"

Retrieval Augmented Generation

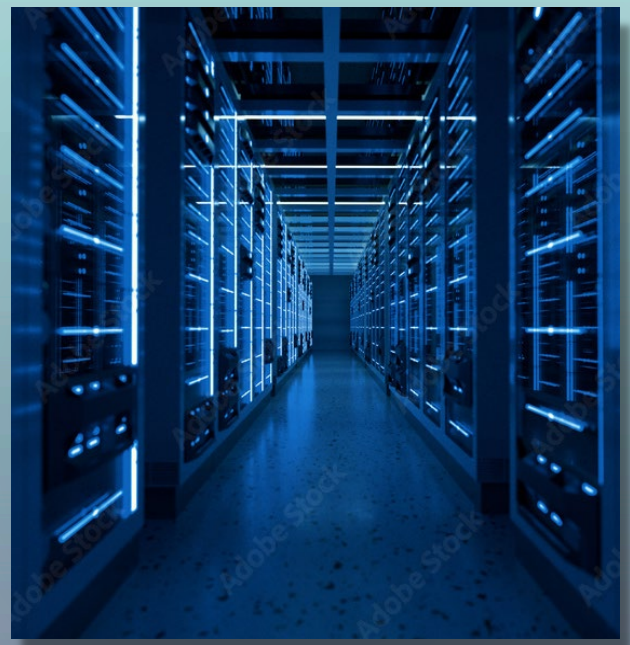
OpenAI's ChatGPT

The DeepSeek Moment

2025

Challenges and Opportunities Scaling NVM

Data Center



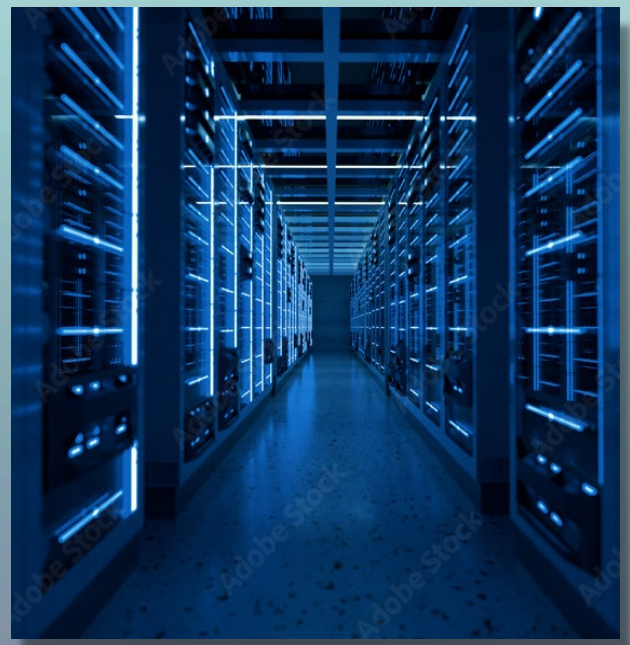
Enterprise

Edge

2025

Challenges and Opportunities Scaling NVM

Data Center



Orchestration, Performance

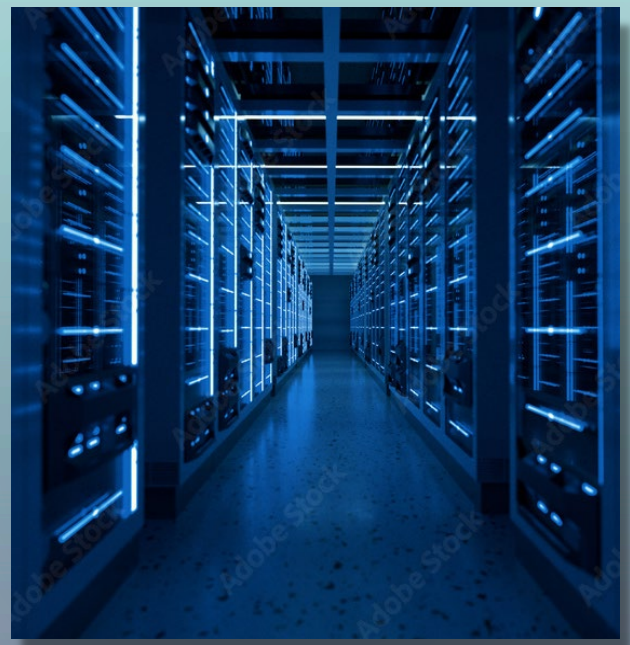
Enterprise

Edge

2025

Challenges and Opportunities Scaling NVM

Data Center



Orchestration, Performance

Enterprise

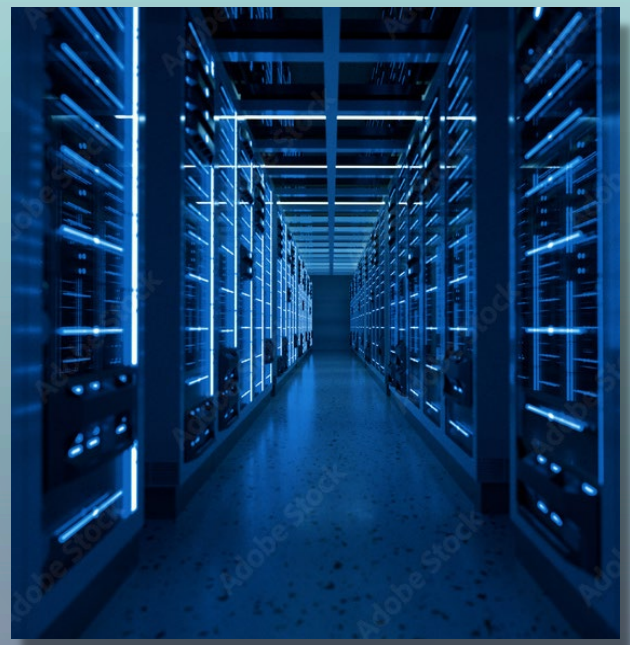
Content-Awareness, Trust

Edge

2025

Challenges and Opportunities Scaling NVM

Data Center



Orchestration, Performance

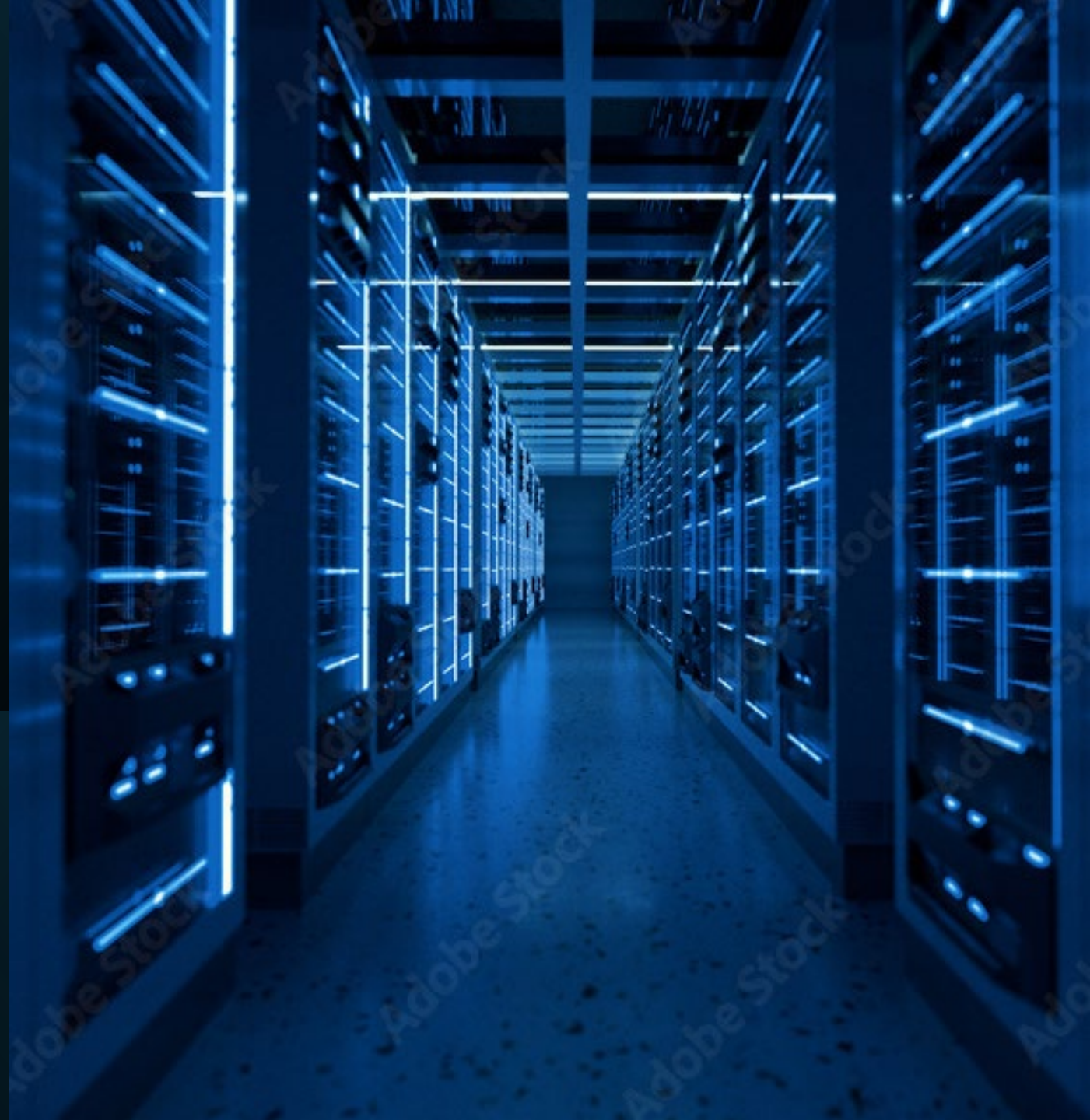
Enterprise

Content-Awareness, Trust

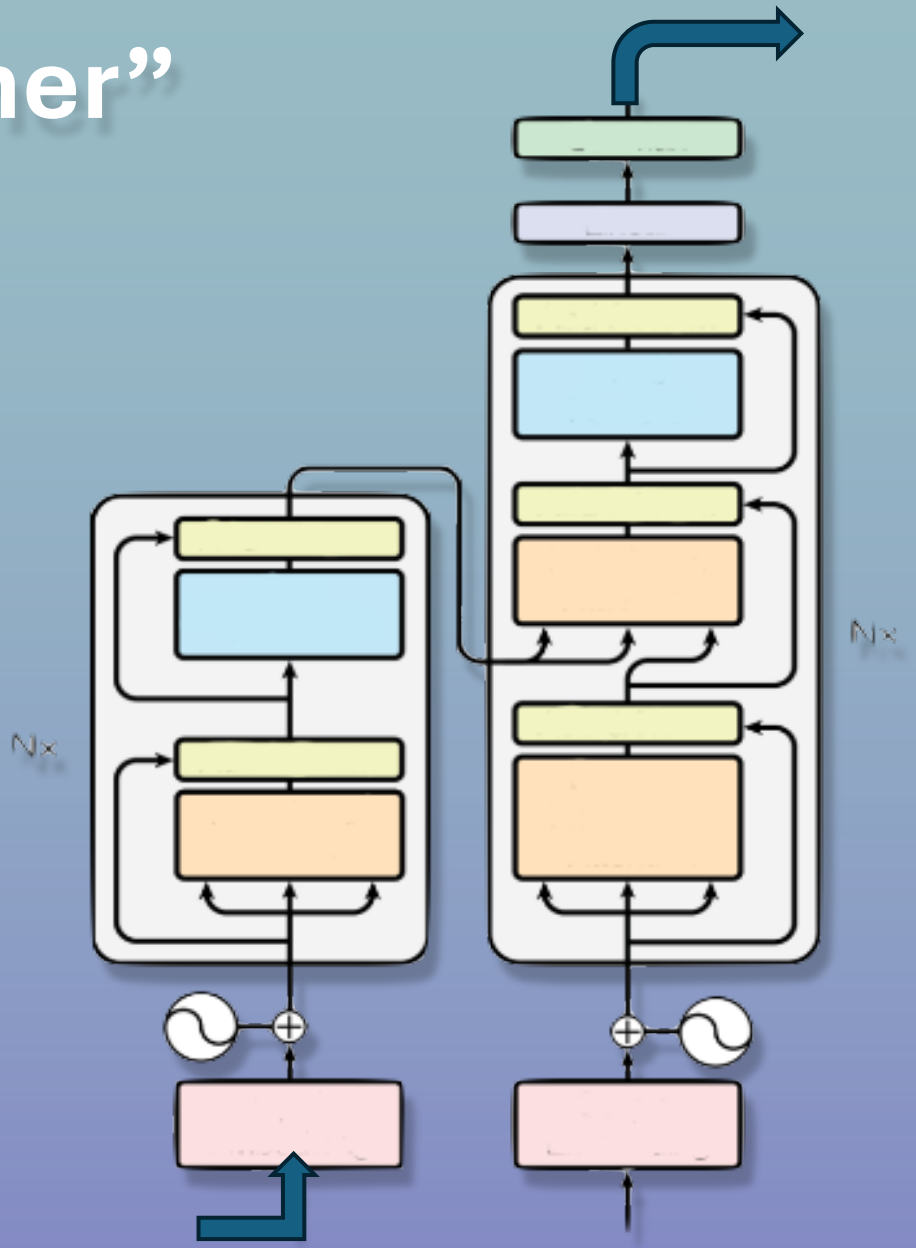
Edge

Innovation, Power

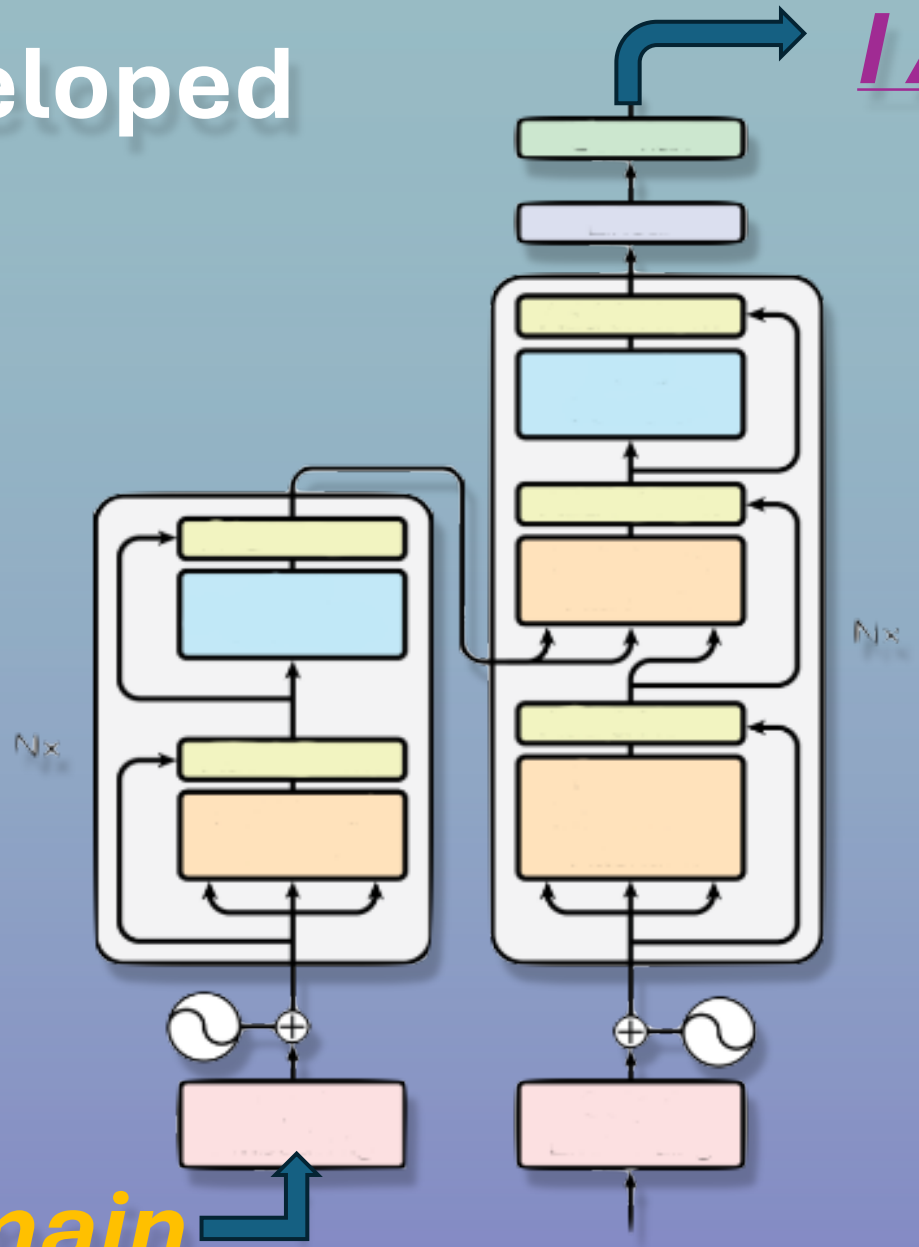
Some GenAI Inference Basics



The “Transformer”



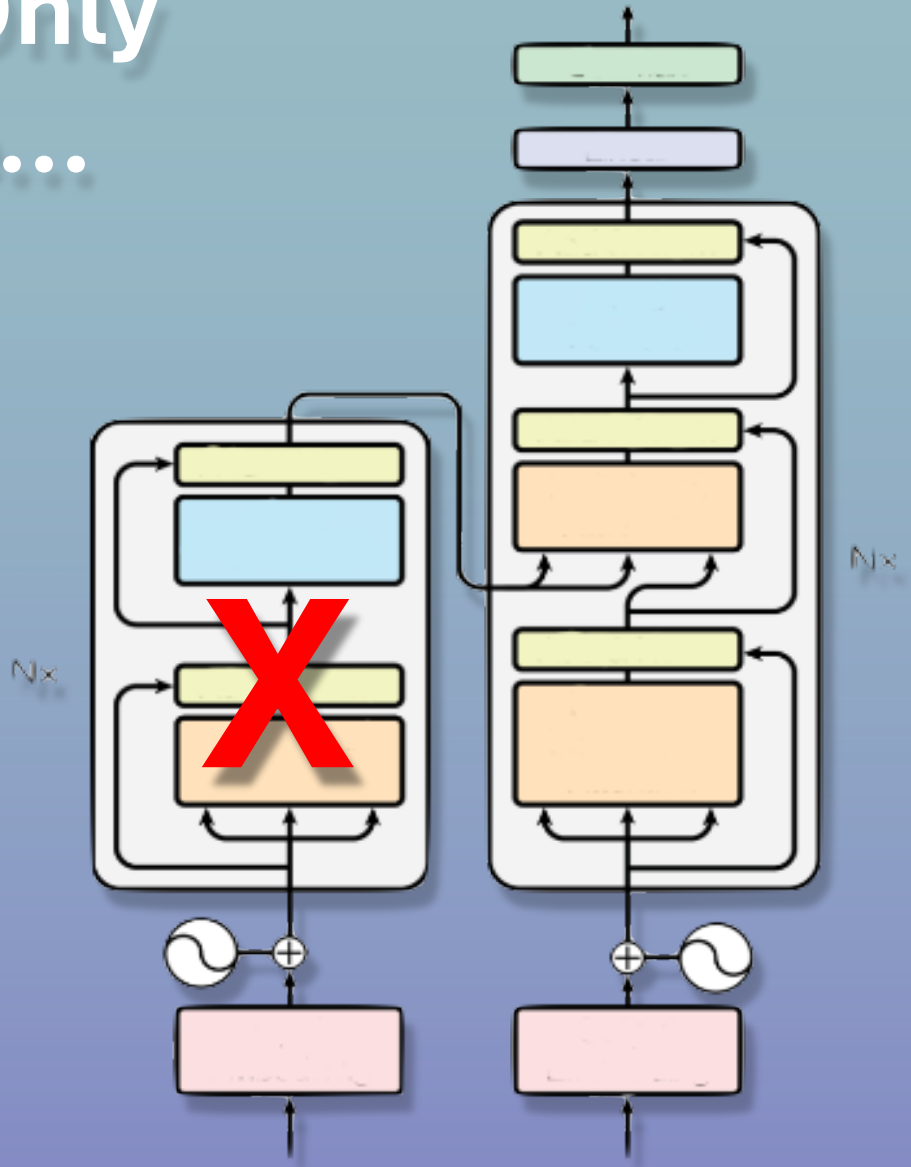
Originally Developed For Language Translation...



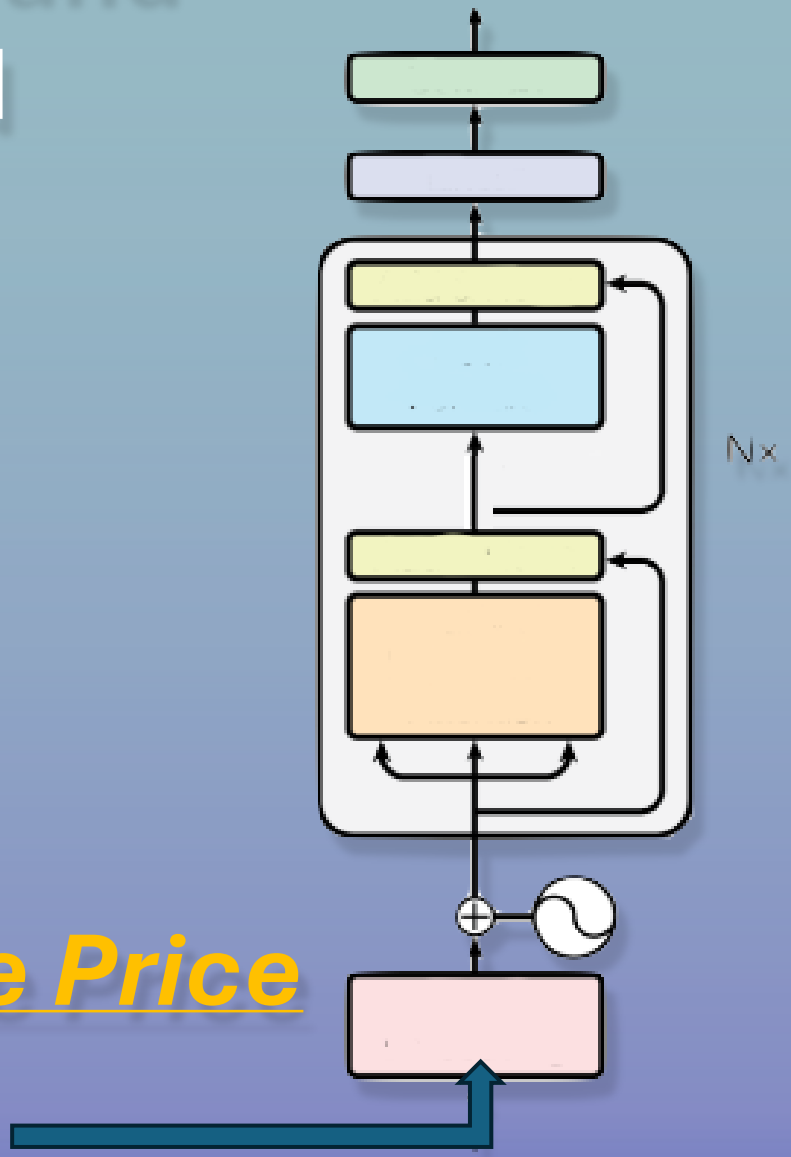
I Am A Human

Je Suis Un Humain

Today's Apps Only Need One Side...

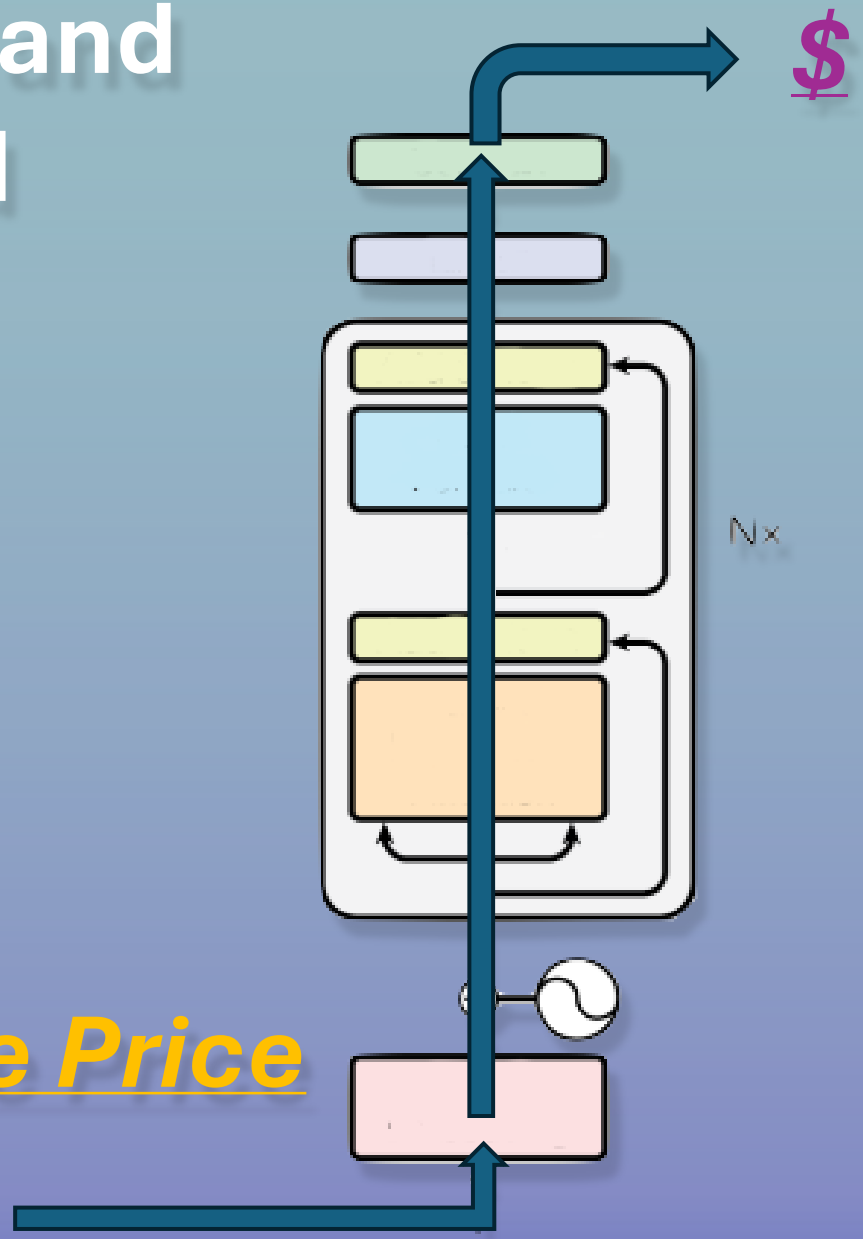


For Chatbots and AI Coding and Agents...



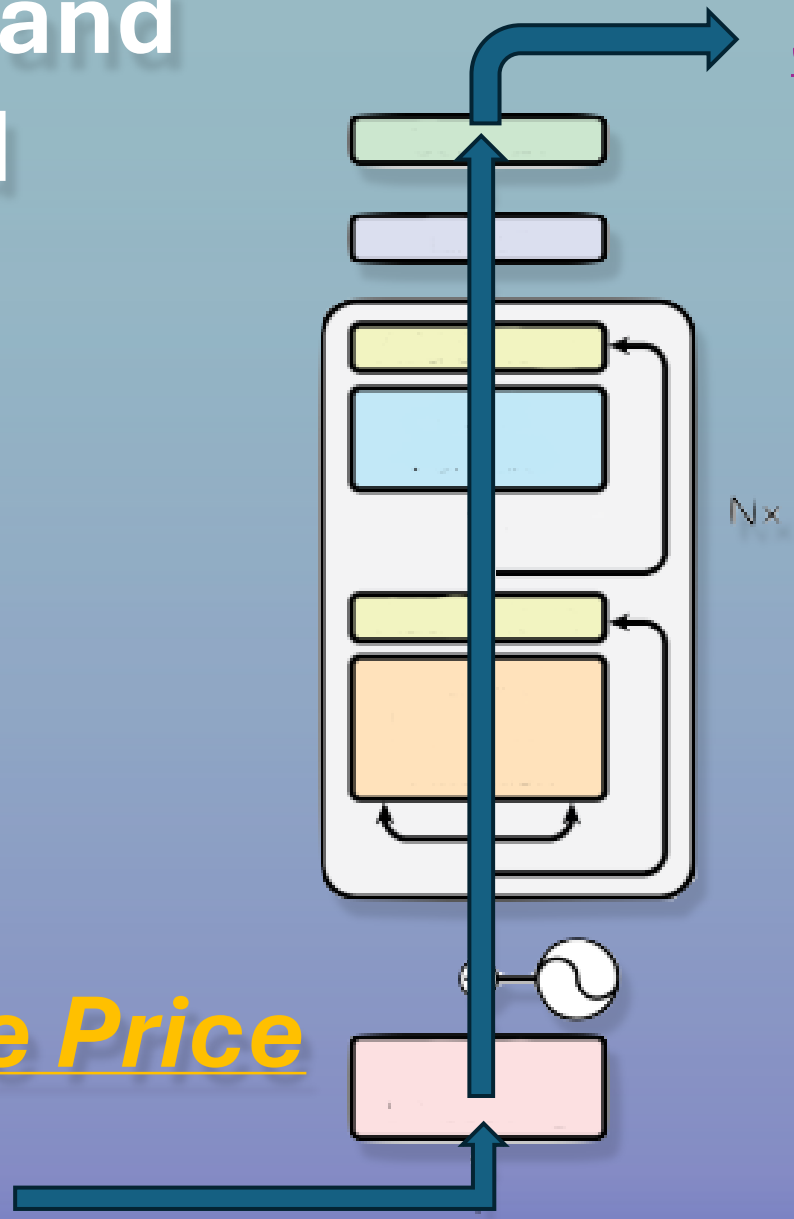
What's The Price Of Eggs?

For Chatbots and AI Coding and Agents...



What's The Price Of Eggs?

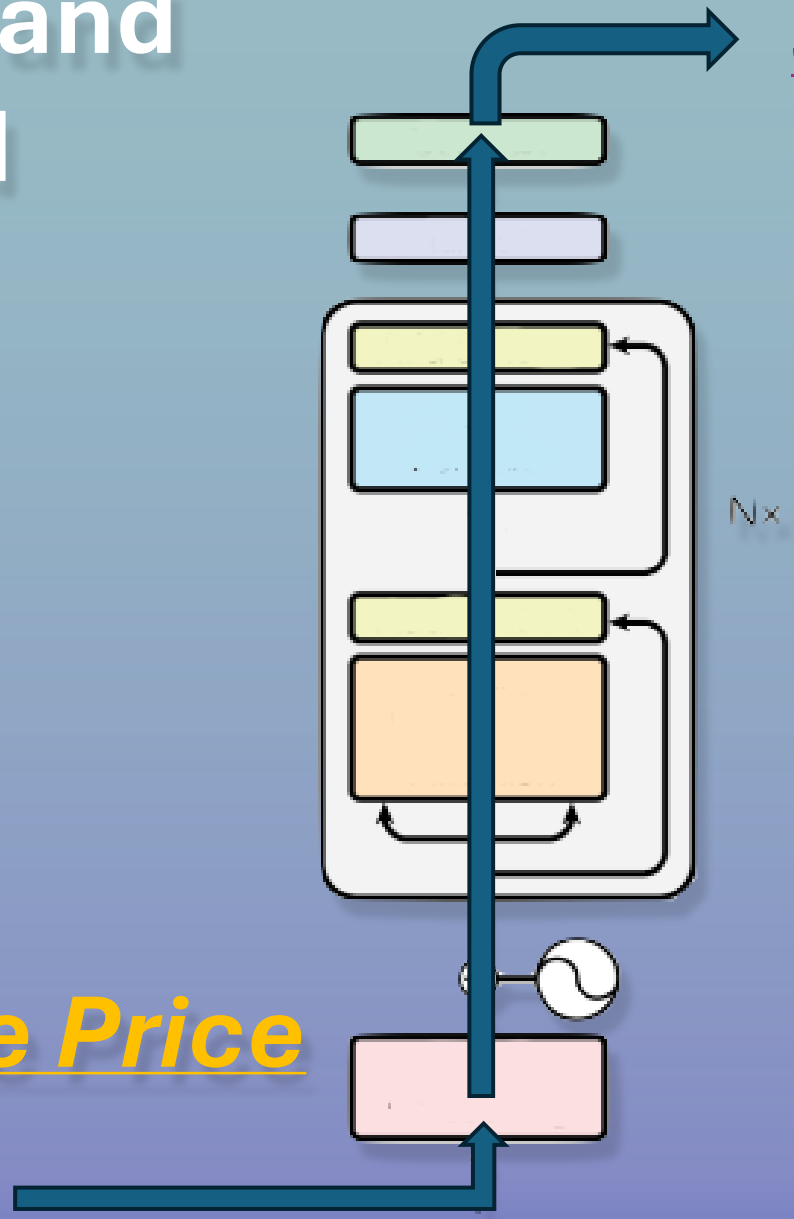
For Chatbots and AI Coding and Agents...



Response Is Generated 1 Token At A Time

What's The Price Of Eggs?

For Chatbots and AI Coding and Agents...

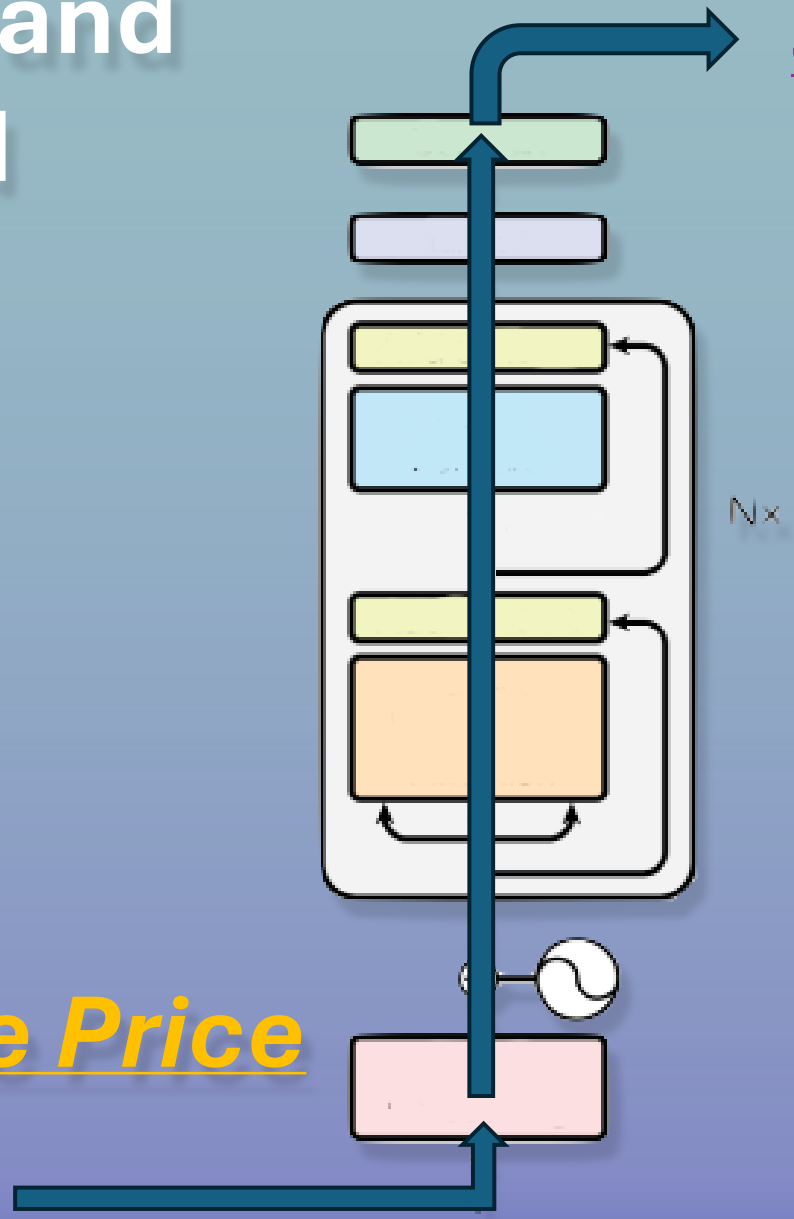


$ 29

Response Is Generated 1 Token At A Time

What's The Price Of Eggs?

For Chatbots and AI Coding and Agents...

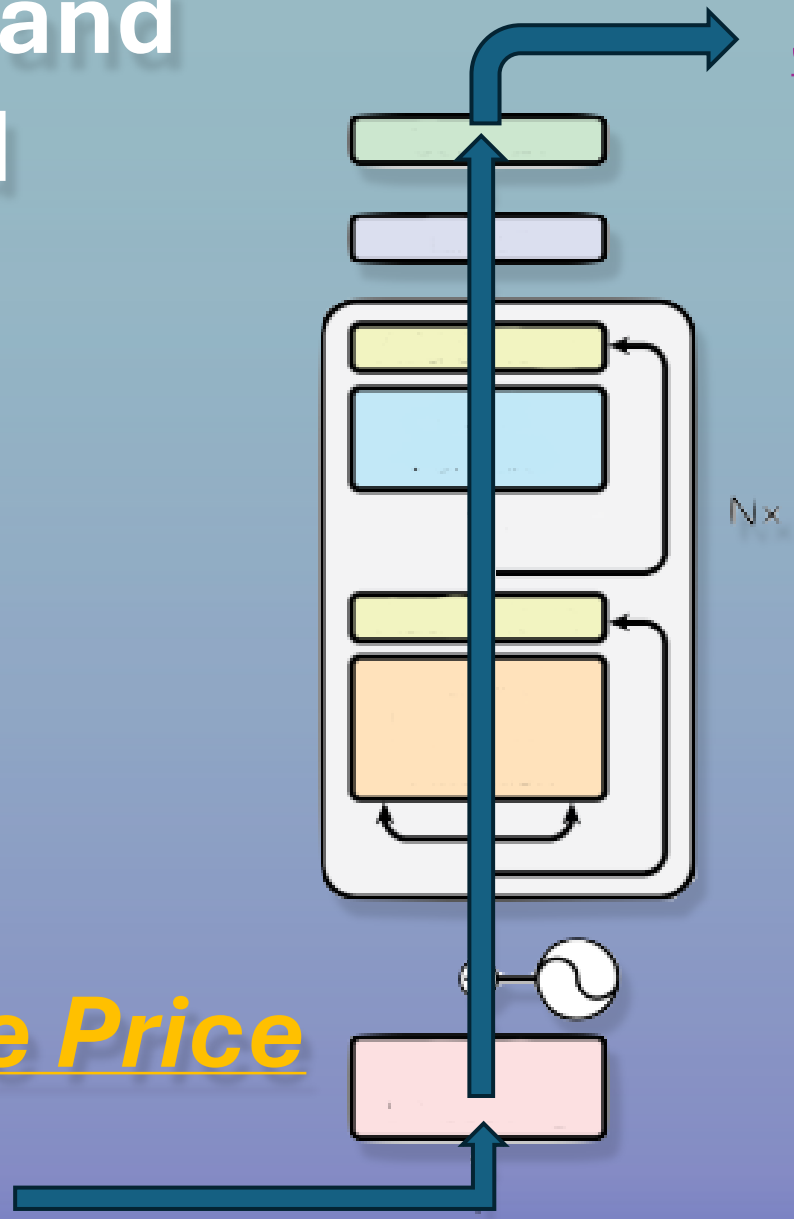


$ 29.

Response Is Generated 1 Token At A Time

What's The Price Of Eggs?

For Chatbots and AI Coding and Agents...

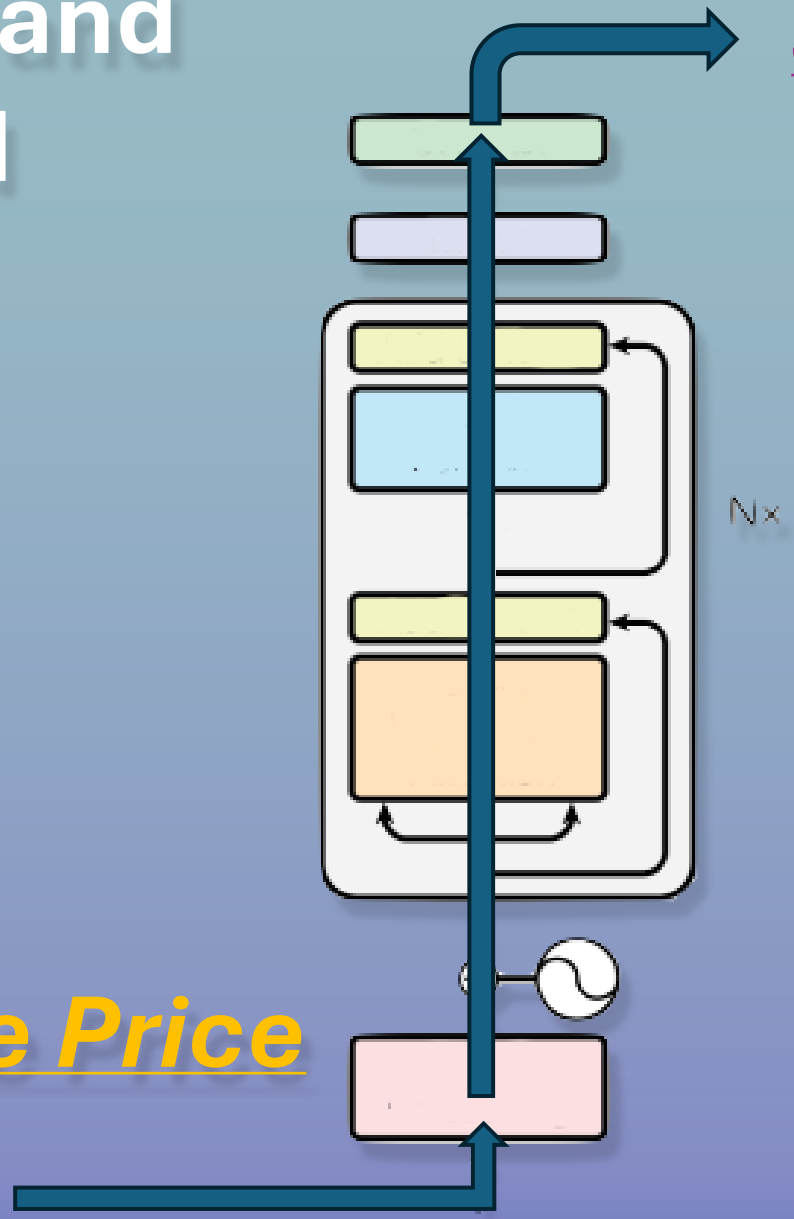


$ 29.9

Response Is Generated 1 Token At A Time

What's The Price Of Eggs?

For Chatbots and AI Coding and Agents...

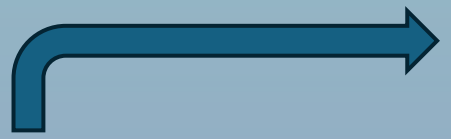


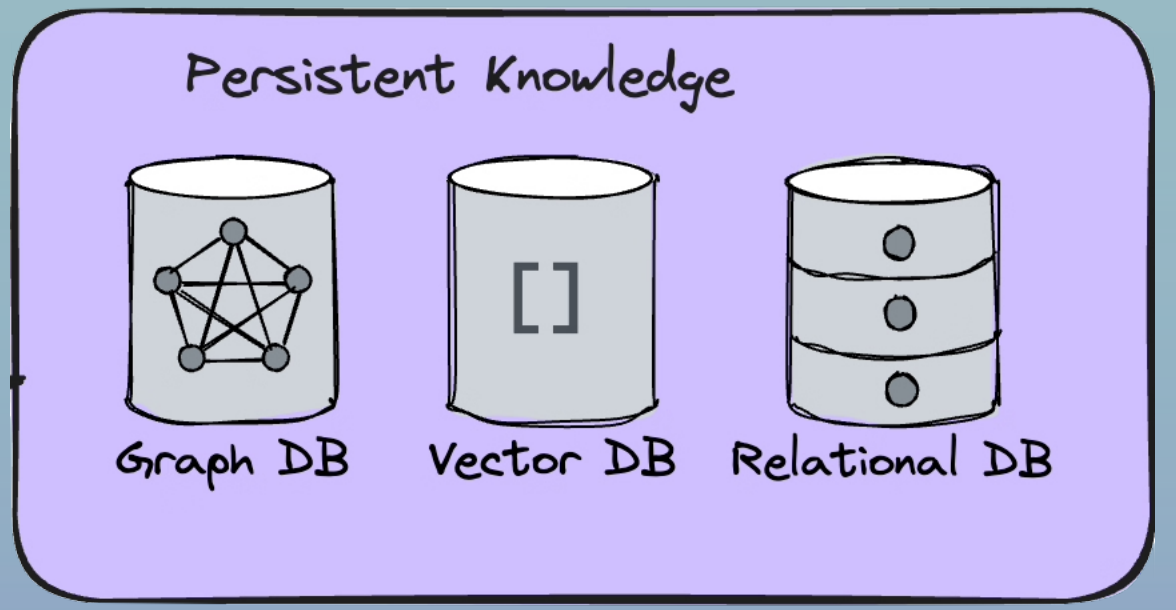
$ 29.99

Response Is Generated 1 Token At A Time

What's The Price Of Eggs?

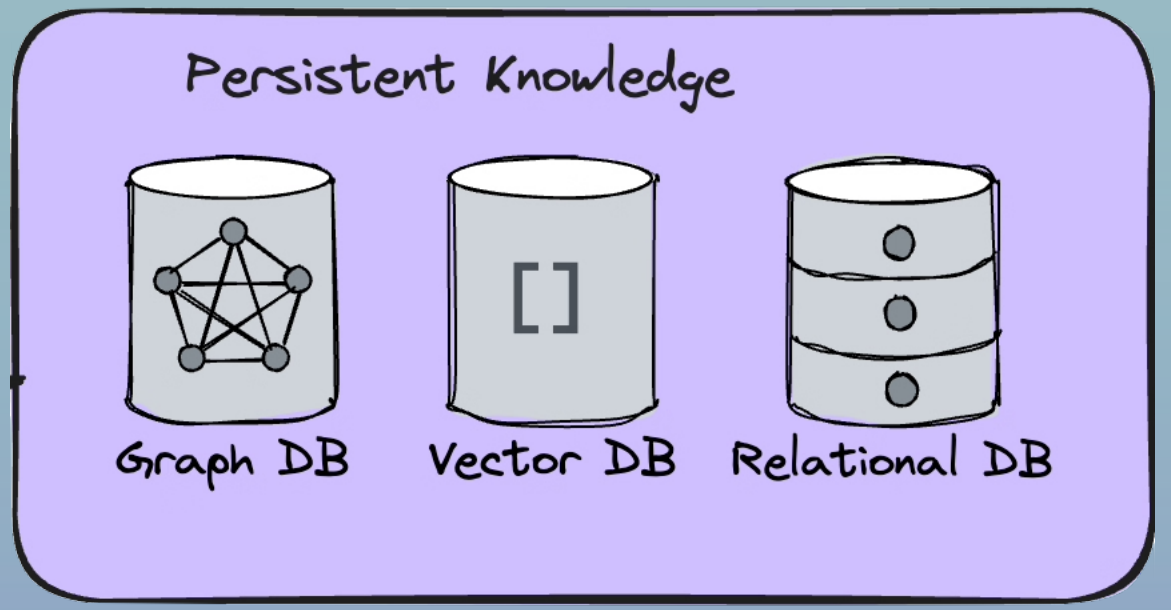
Retrieval Augmented Generation





What's The Price Of Eggs?

Retrieval Augmented Generation



What's The Price Of Eggs?

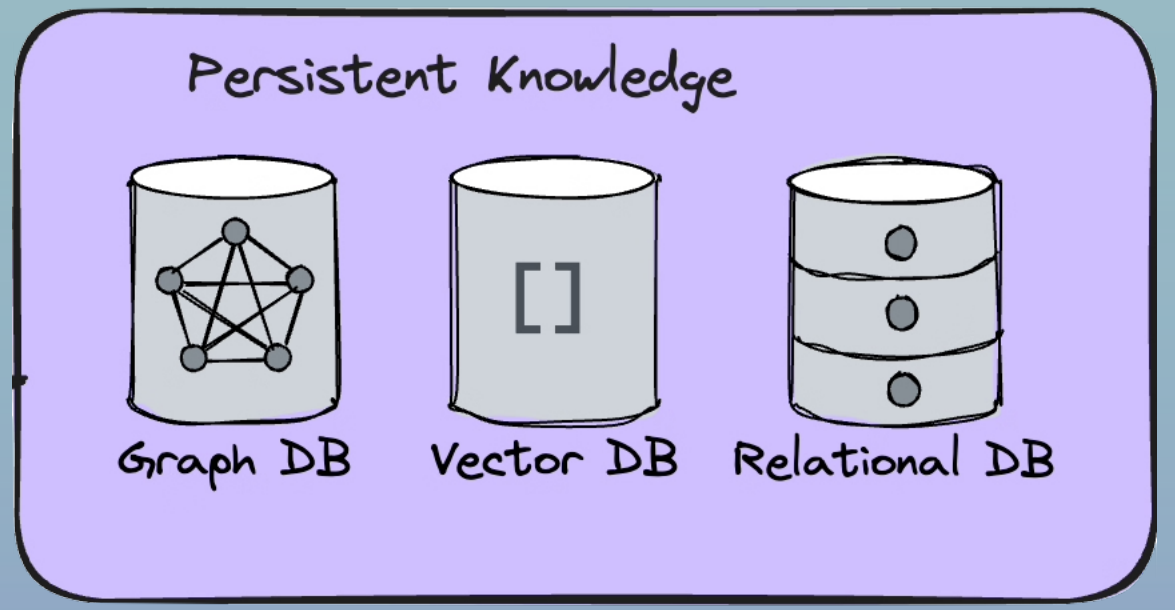
Eggs Come in A Dozen.

Eggs Are $3.99. Scrambled Eggs Are Good.

Additional ”context"

- *from an external data source*
- *not typically visible to user*
- *can be noisy*

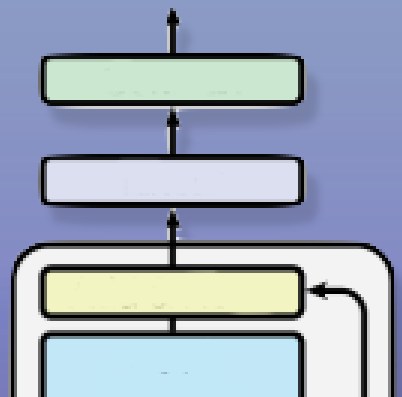
Retrieval Augmented Generation



What's The Price Of Eggs?

Eggs Come in A Dozen.

Eggs Are $3.99. Scrambled Eggs Are Good.

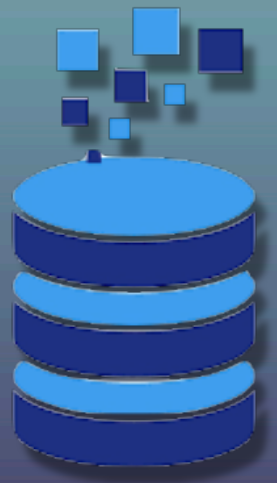


Important Inference Metrics

- Context Window Size > Input Tokens + Output Tokens
- Time To First Token (TTFT)
- Token Rate – (Tokens / Second)

Important Inference Metrics

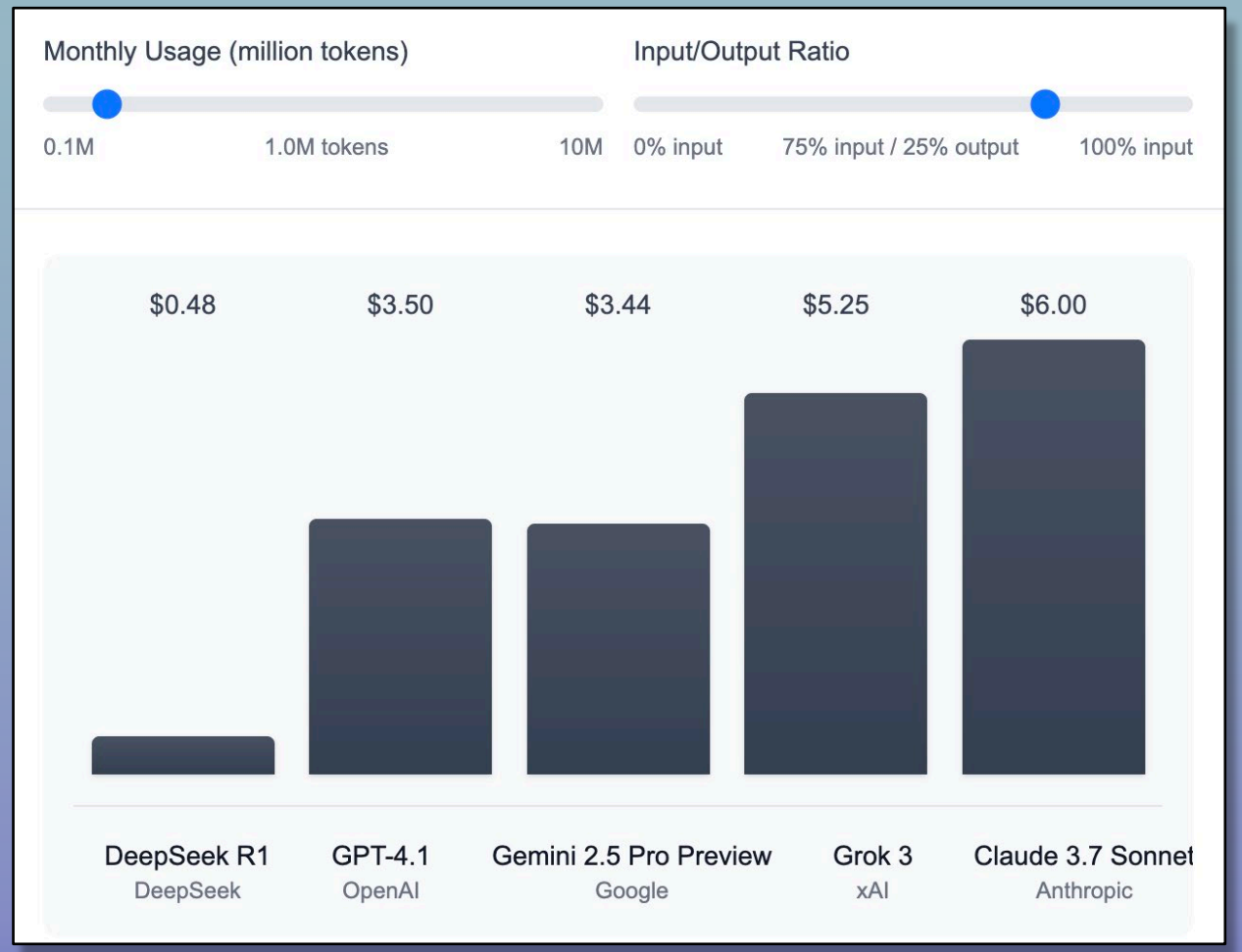
- Context Window Size > Input Tokens + Output Tokens
- Time To First Token (TTFT)
- Token Rate – (Tokens / Second)



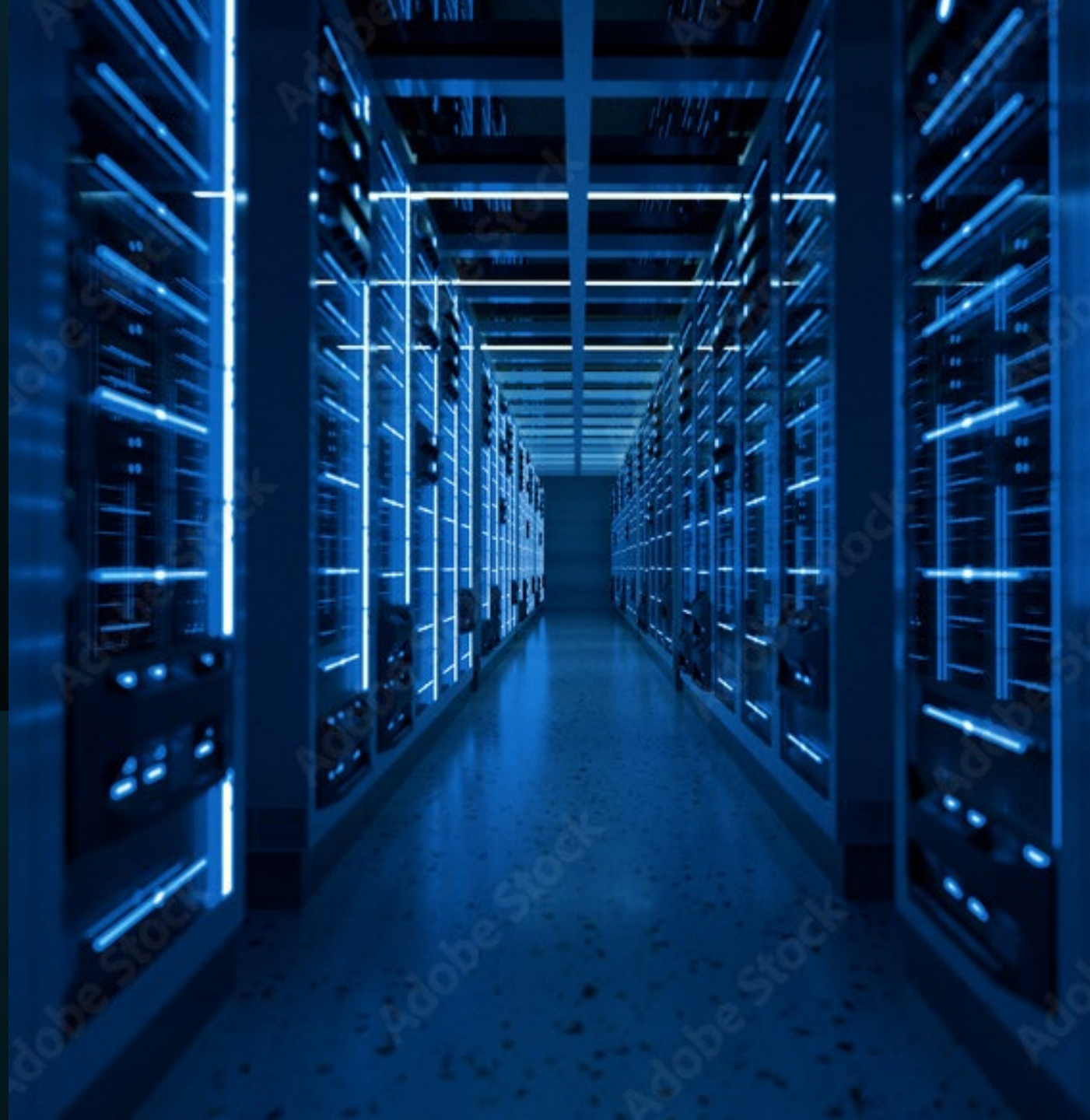
Storage Bottlenecks Are Possible

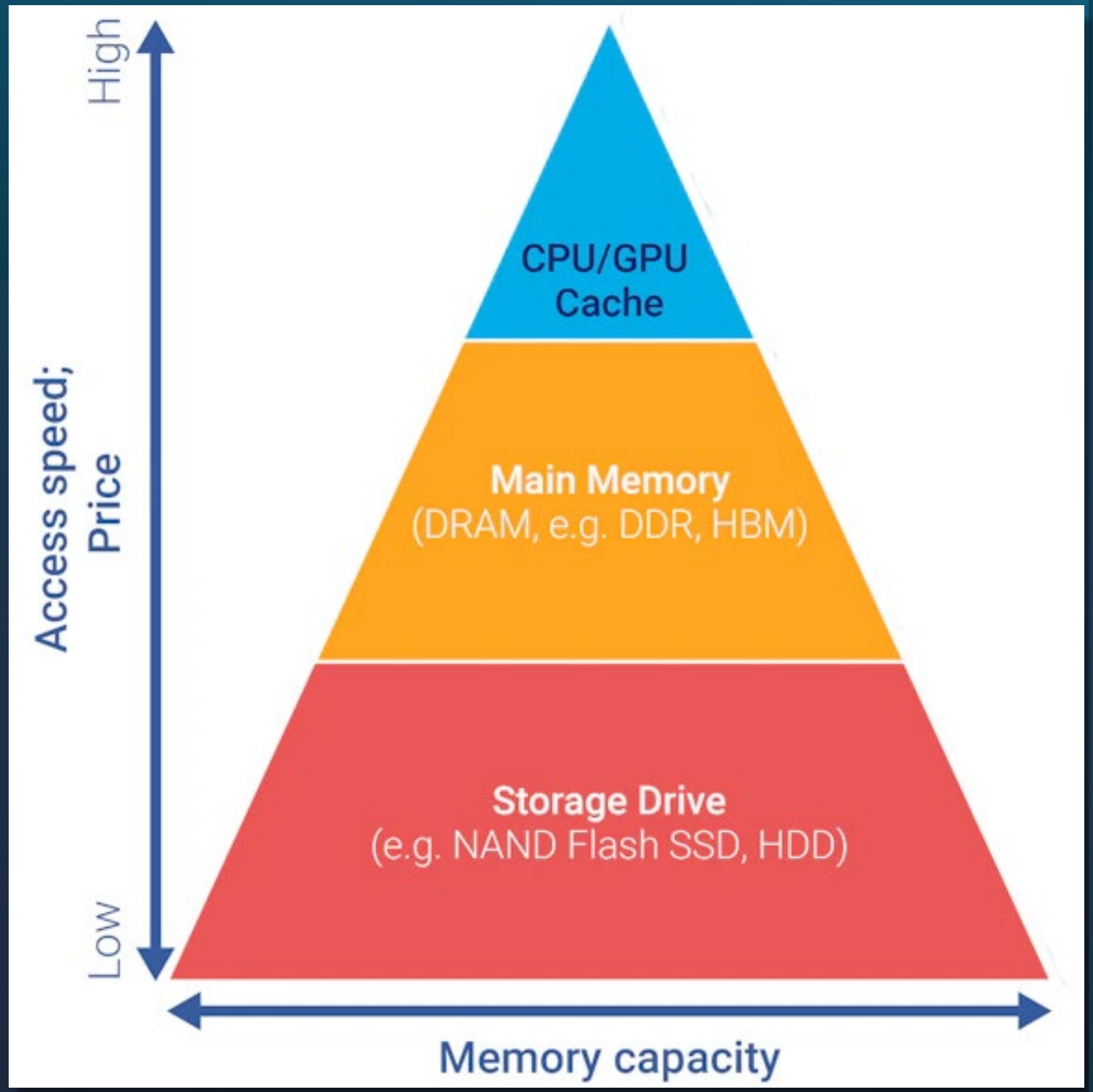
- Model Weights
- RAG Data Sources
- KV-Cache (Pre-Cache)

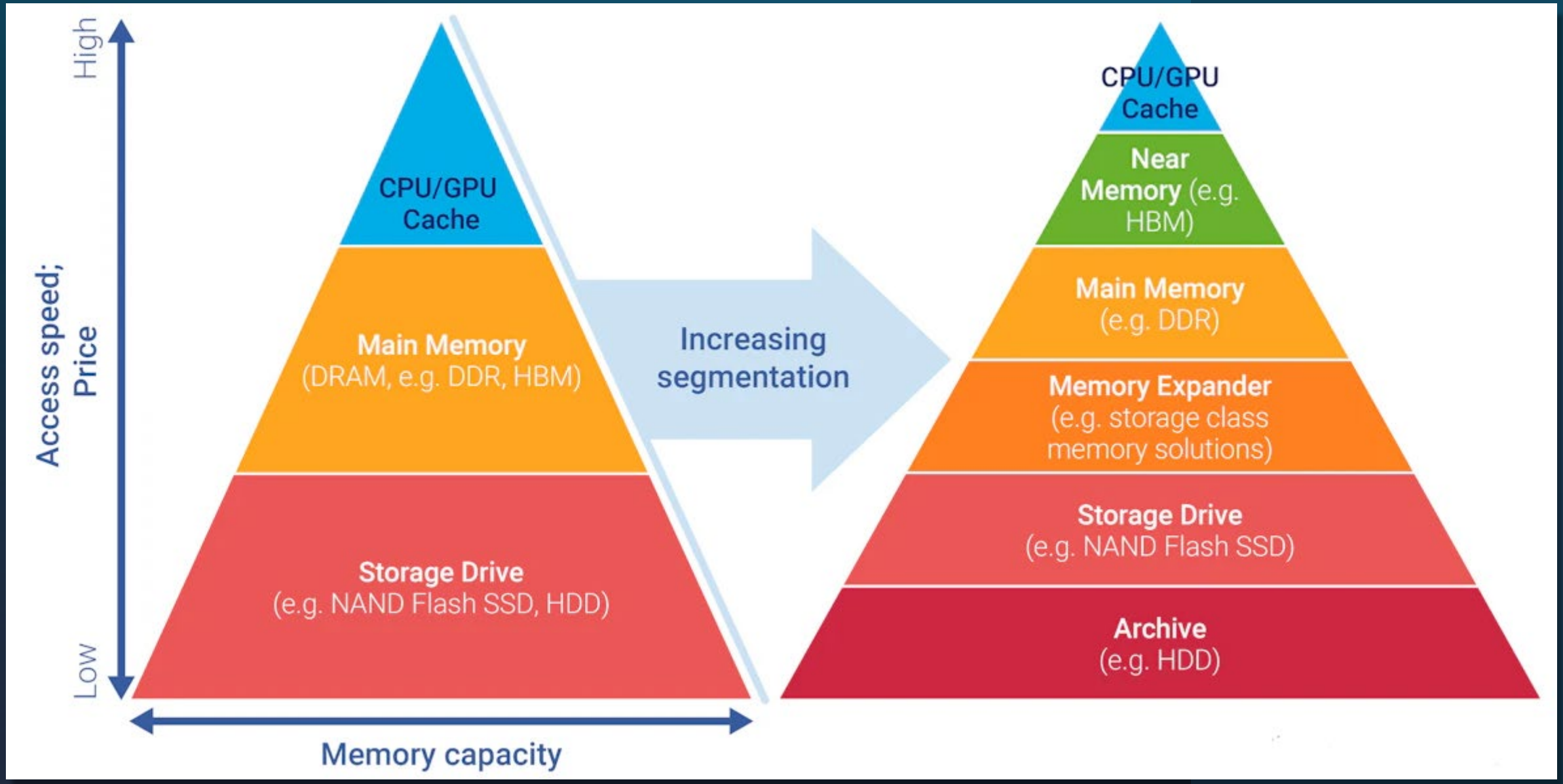
Tokens As Commodity

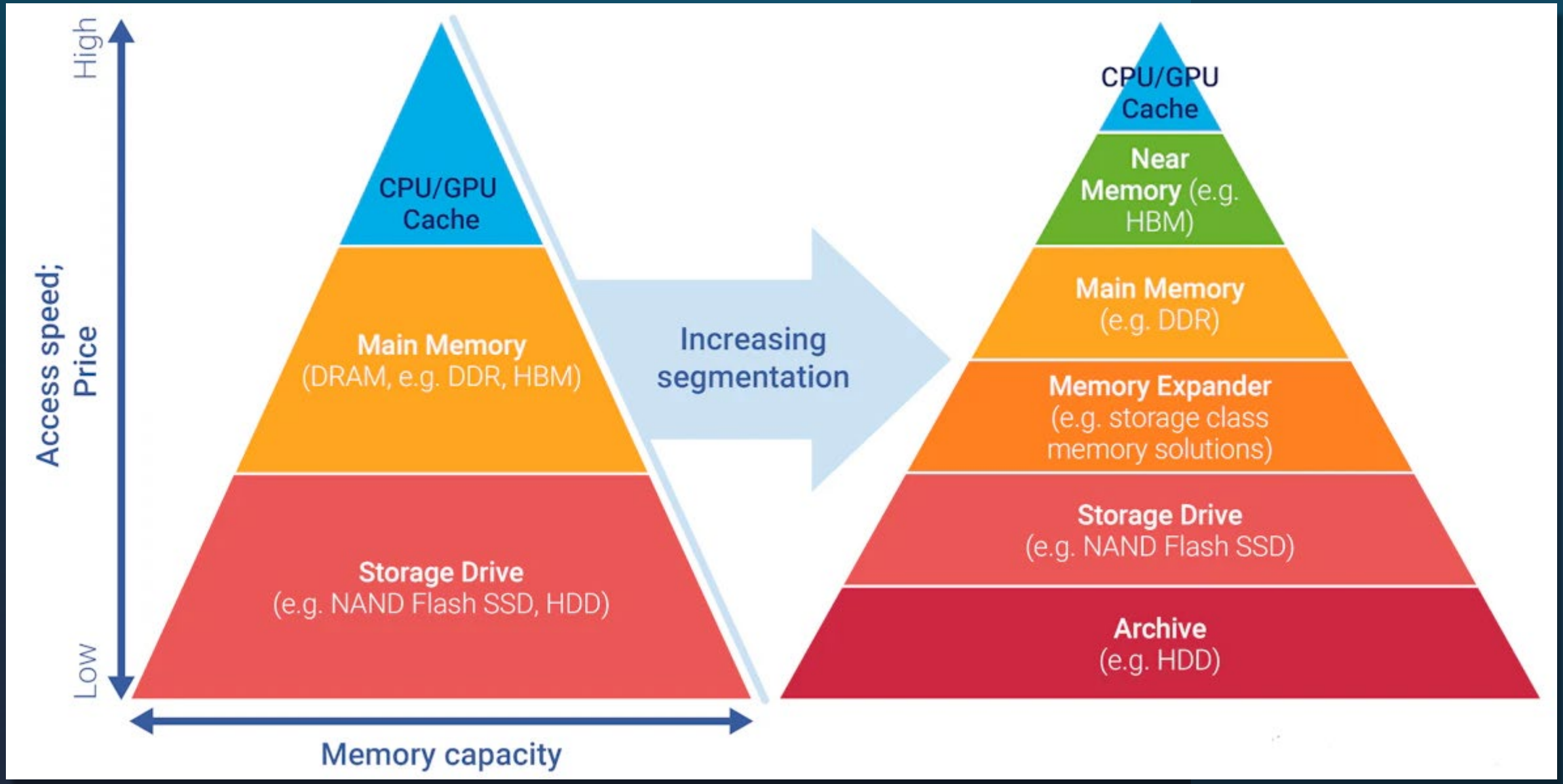


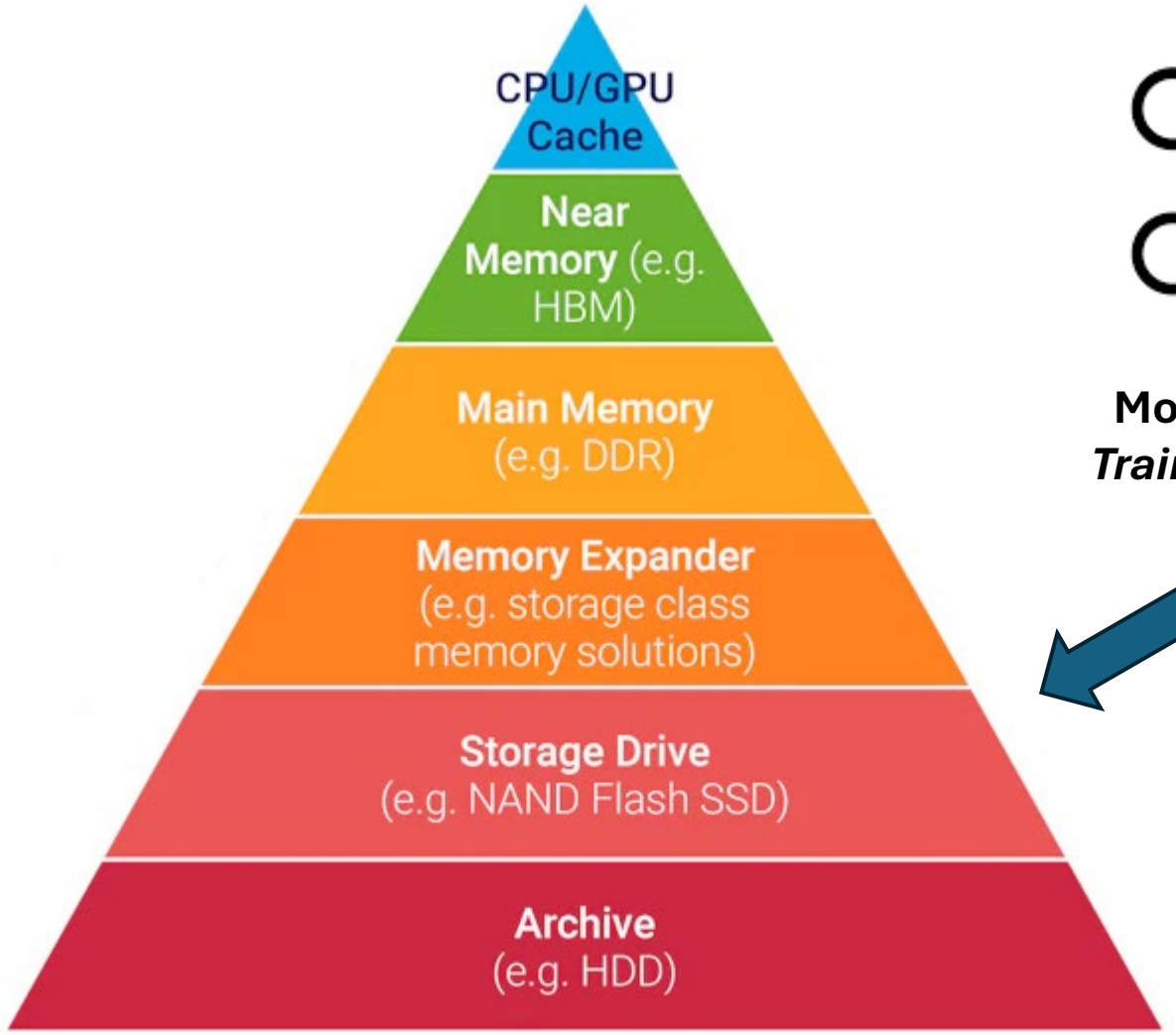
Scaling GenAI NVM In The Datacenter

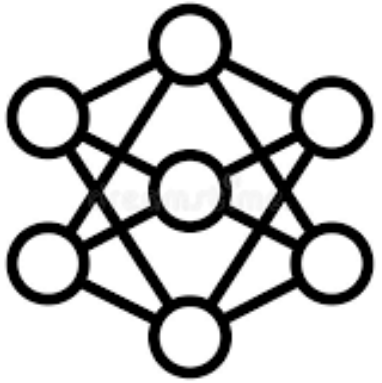



High
Access speed;
Price
Low
CPU/GPU
Cache
Main Memory
(DRAM, e.g. DDR, HBM)
Storage Drive
(e.g. NAND Flash SSD, HDD)
Memory capacity

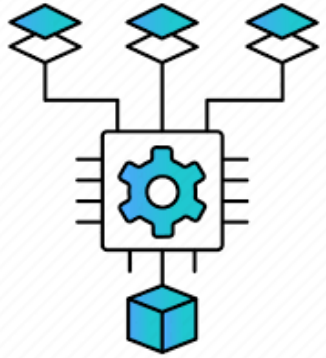

High
Low
Access speed;
Price
CPU/GPU
Cache
Main Memory
(DRAM, e.g. DDR, HBM)
Storage Drive
(e.g. NAND Flash SSD, HDD)
Memory capacity
Increasing
segmentation
CPU/GPU
Cache
Near
Memory (e.g.
HBM)
Main Memory
(e.g. DDR)
Memory Expander
(e.g. storage class
memory solutions)
Storage Drive
(e.g. NAND Flash SSD)
Archive
(e.g. HDD)


High
Low
Access speed;
Price
CPU/GPU
Cache
Main Memory
(DRAM, e.g. DDR, HBM)
Storage Drive
(e.g. NAND Flash SSD, HDD)
Memory capacity
Increasing
segmentation
CPU/GPU
Cache
Near
Memory (e.g.
HBM)
Main Memory
(e.g. DDR)
Memory Expander
(e.g. storage class
memory solutions)
Storage Drive
(e.g. NAND Flash SSD)
Archive
(e.g. HDD)

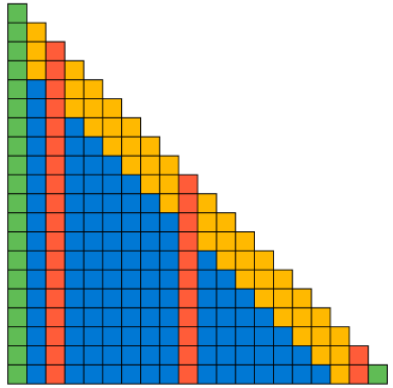




Model Weights *and* *Training Checkpoints*

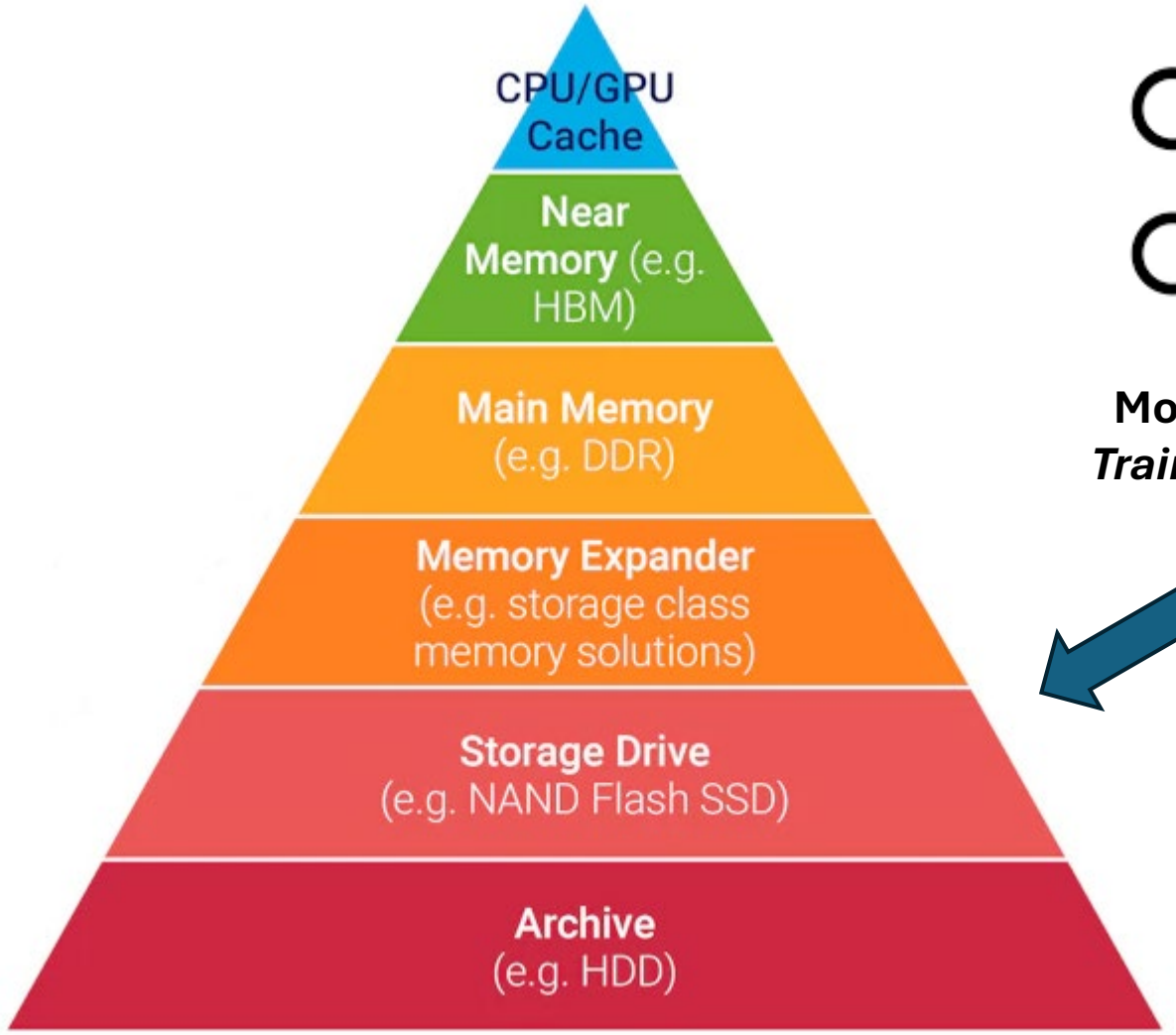


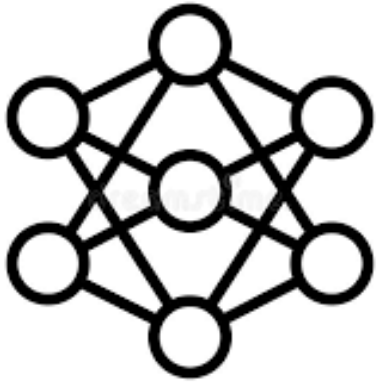
Training Data



KV Pre-Cache

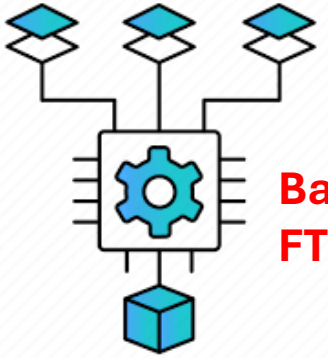
RAG Data Sources





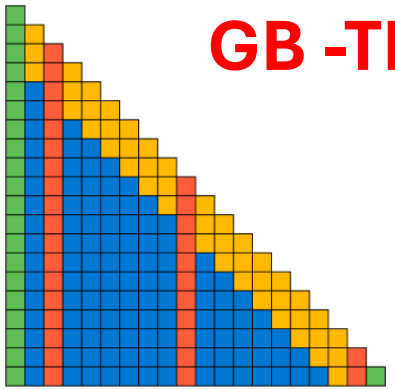
GB - TB

Model Weights *and* *Training Checkpoints*



Base: GB - TB
FT: MB - GB

Training Data

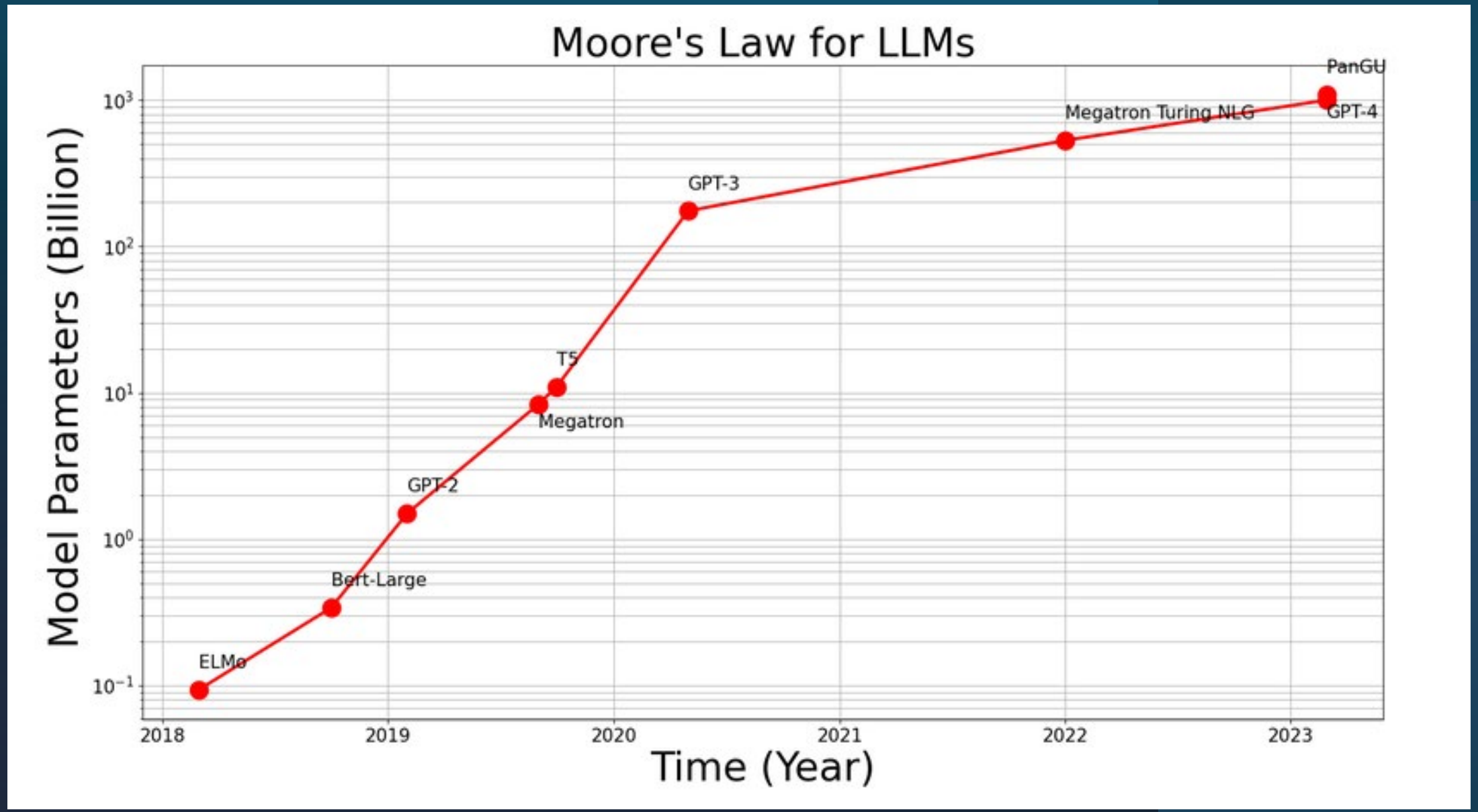


GB -TB

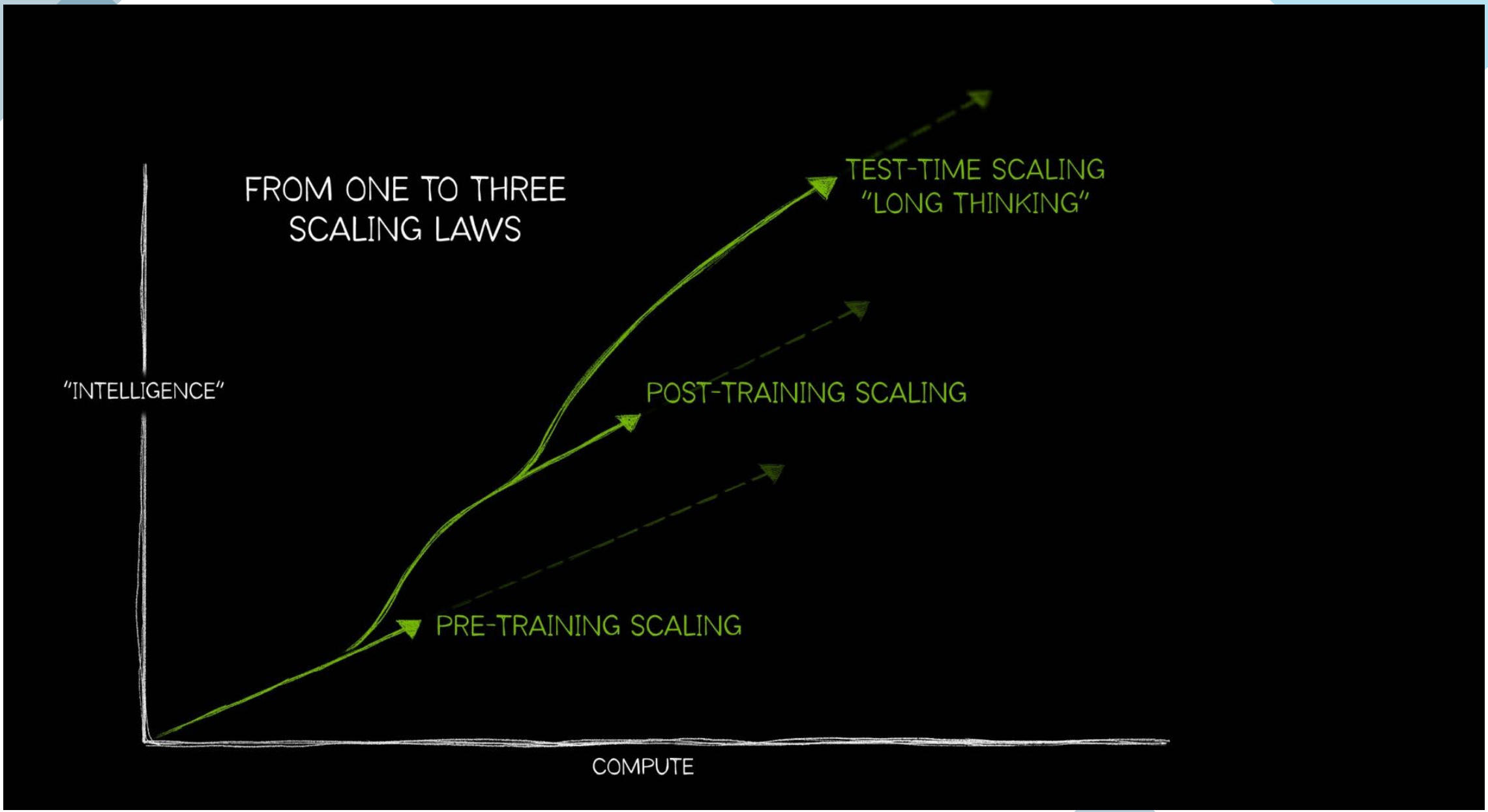
KV Pre-Cache

TB - PB

RAG Data Sources



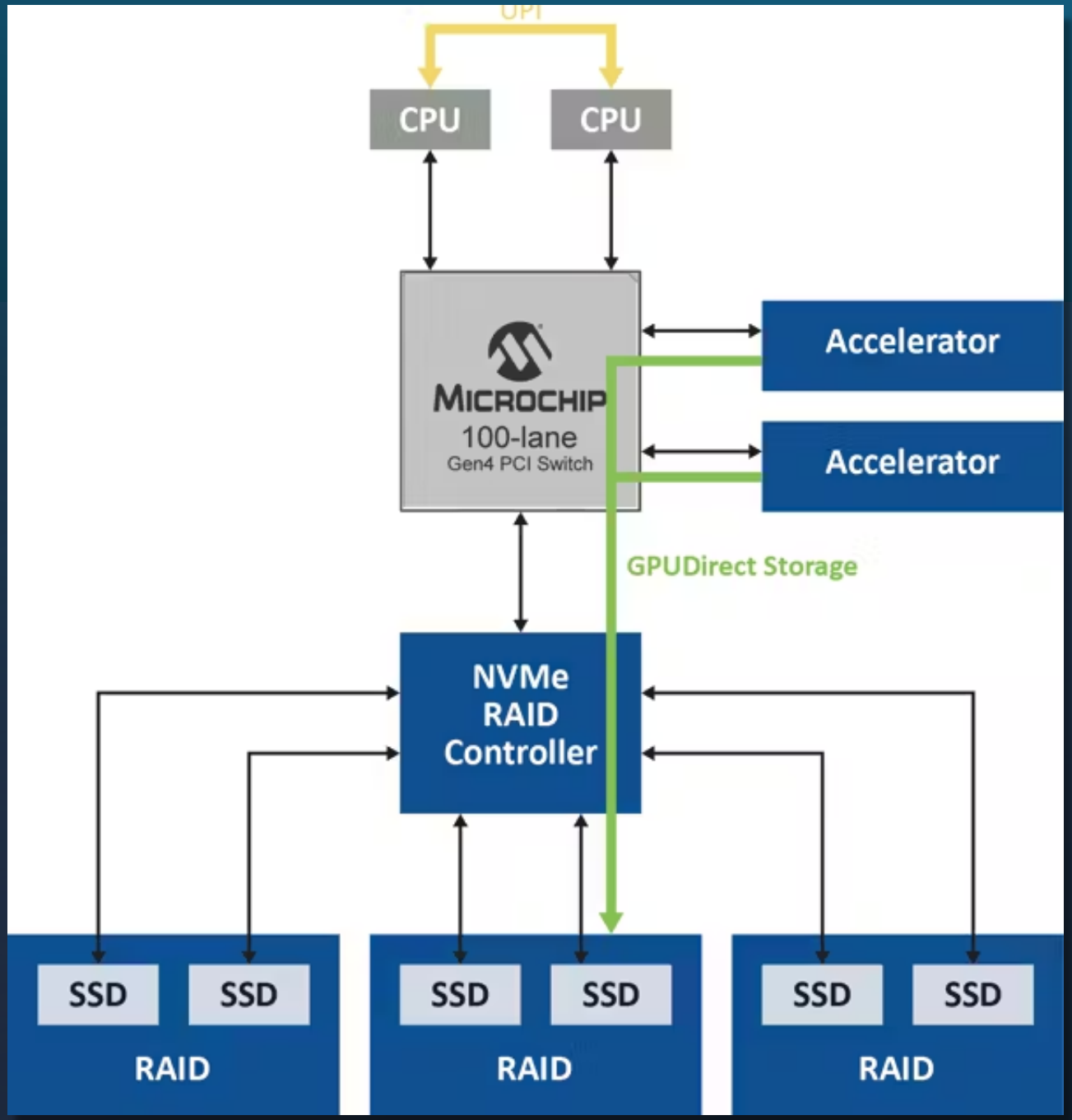
ChatGPT in the Age of Generative AI: A Concise Survey, 2023



Nvidia GDC 2025

Nvidia GDC 2025

NVMe + RAID + GPUDirect



Example:
RAID5
9 NVMe
45W/110R GB/s
(Graid Tech)

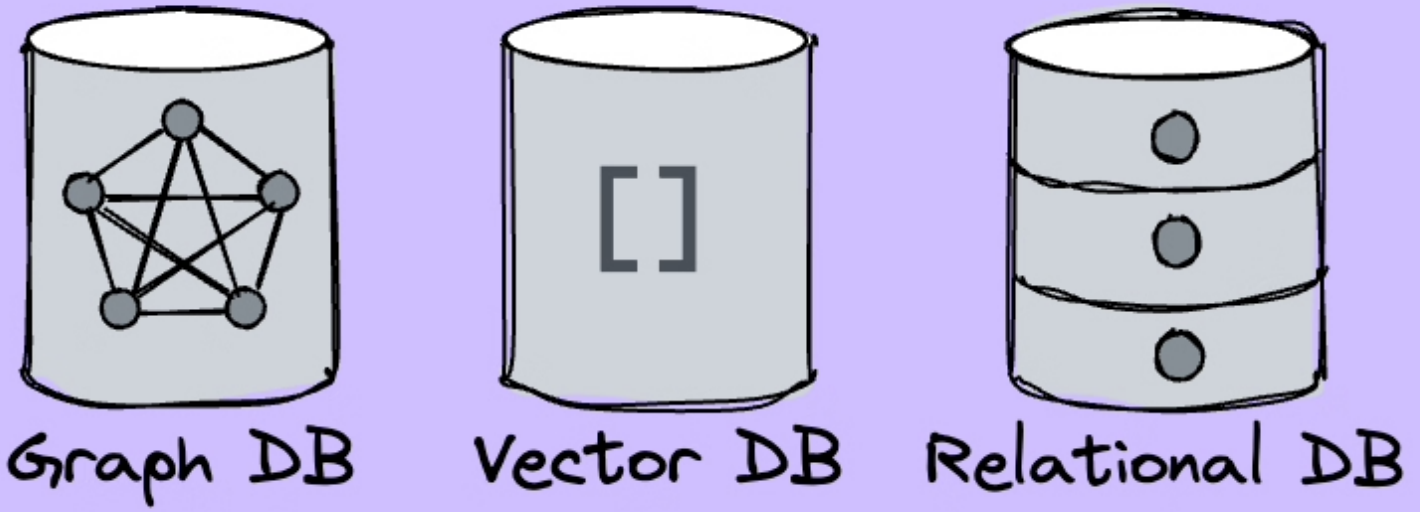
Electronic Design, 2022

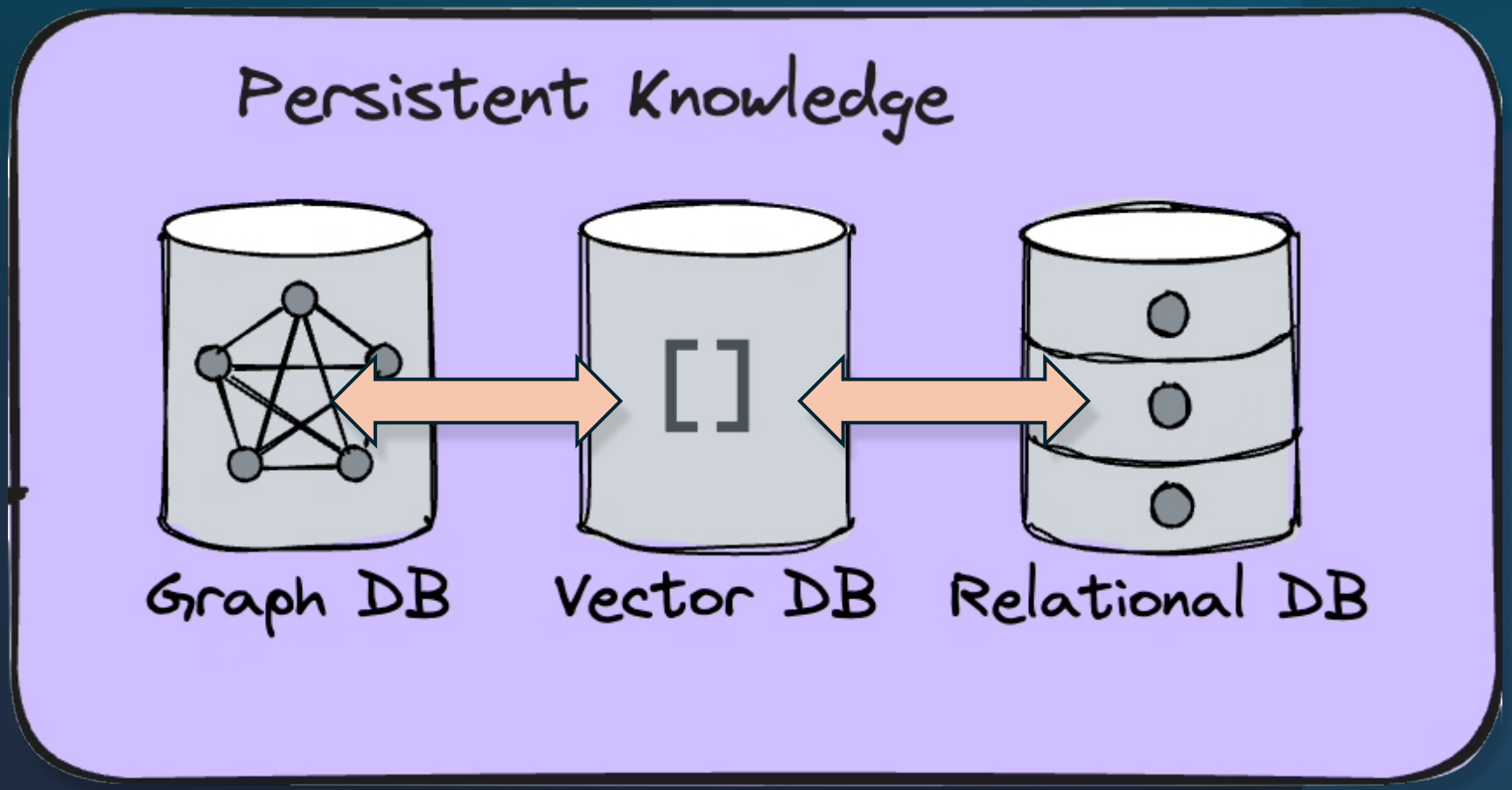
Scaling NVM Storage In The Enterprise

of respondents report that half or more of their organization's data is dark.

Splunk, The State of Enterprise Dark Data, 2025

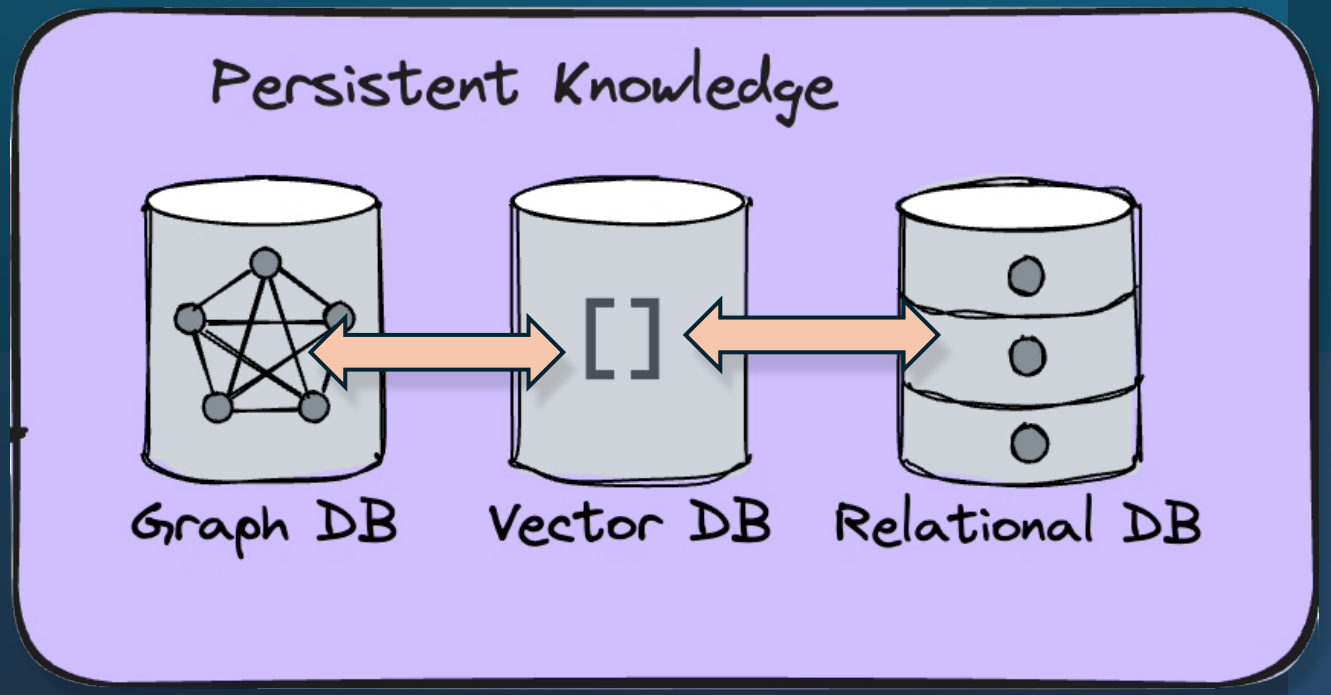
Persistent Knowledge





Compound Systems Are Blurring the Boundaries...

OSS Leading The Way



neo4j

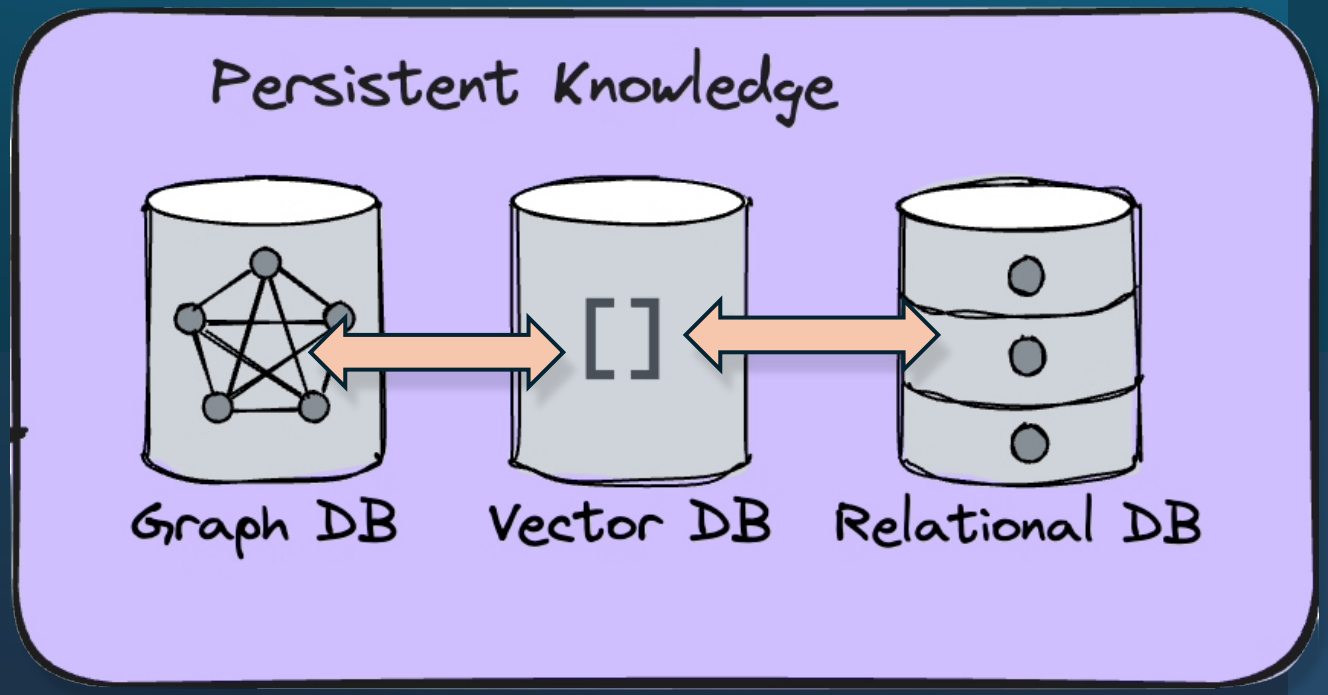
PostgreSQL

mongoDB

elasticsearch

Weaviate

OSS Leading The Way



Scaling On Public Cloud HW

neo4j

PostgreSQL

mongoDB

elasticsearch

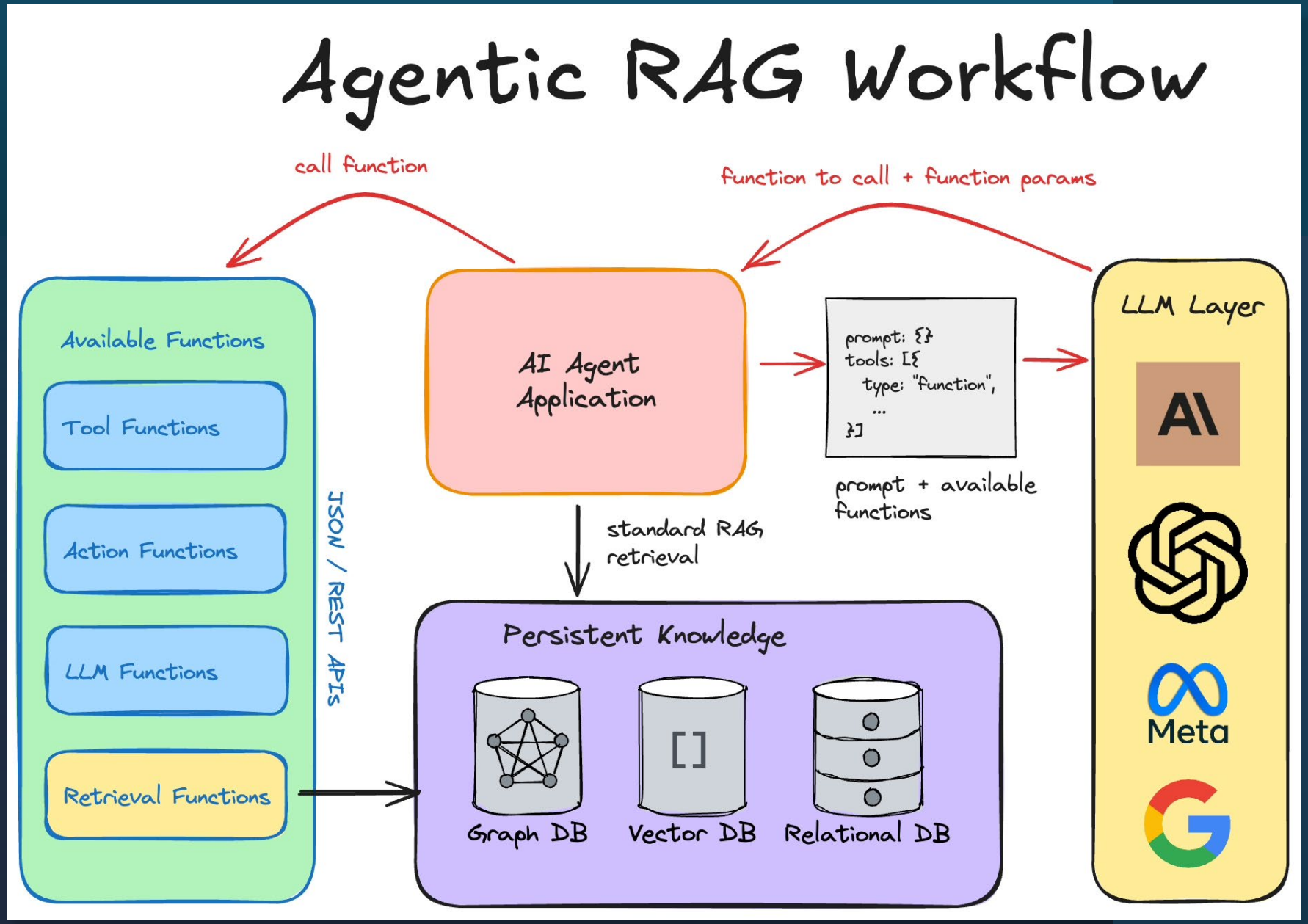
Weaviate

Azure

Google Cloud

aws

Agentic RAG Workflow



npr

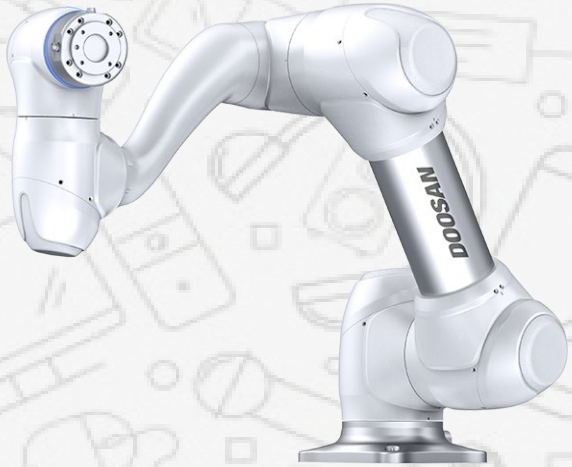
International regulators probe how DeepSeek is using data. Is the app safe to use?

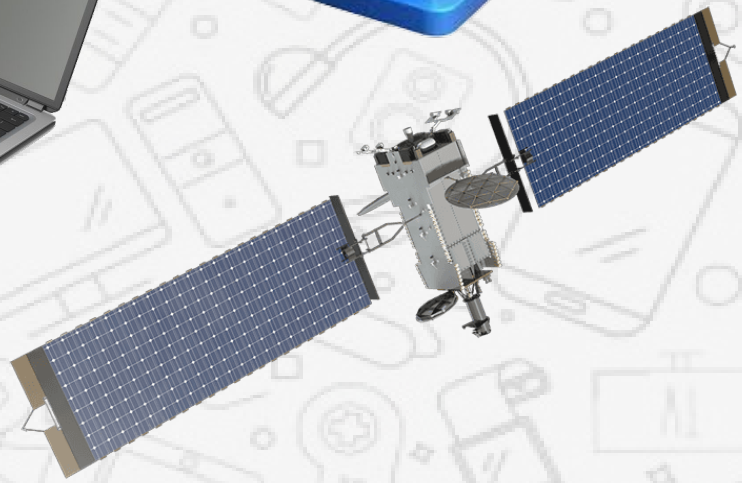
AP

The Czech Republic bans DeepSeek in state administration over cybersecurity concerns

Scaling GenAI NVM At The Edge

Edge Devices





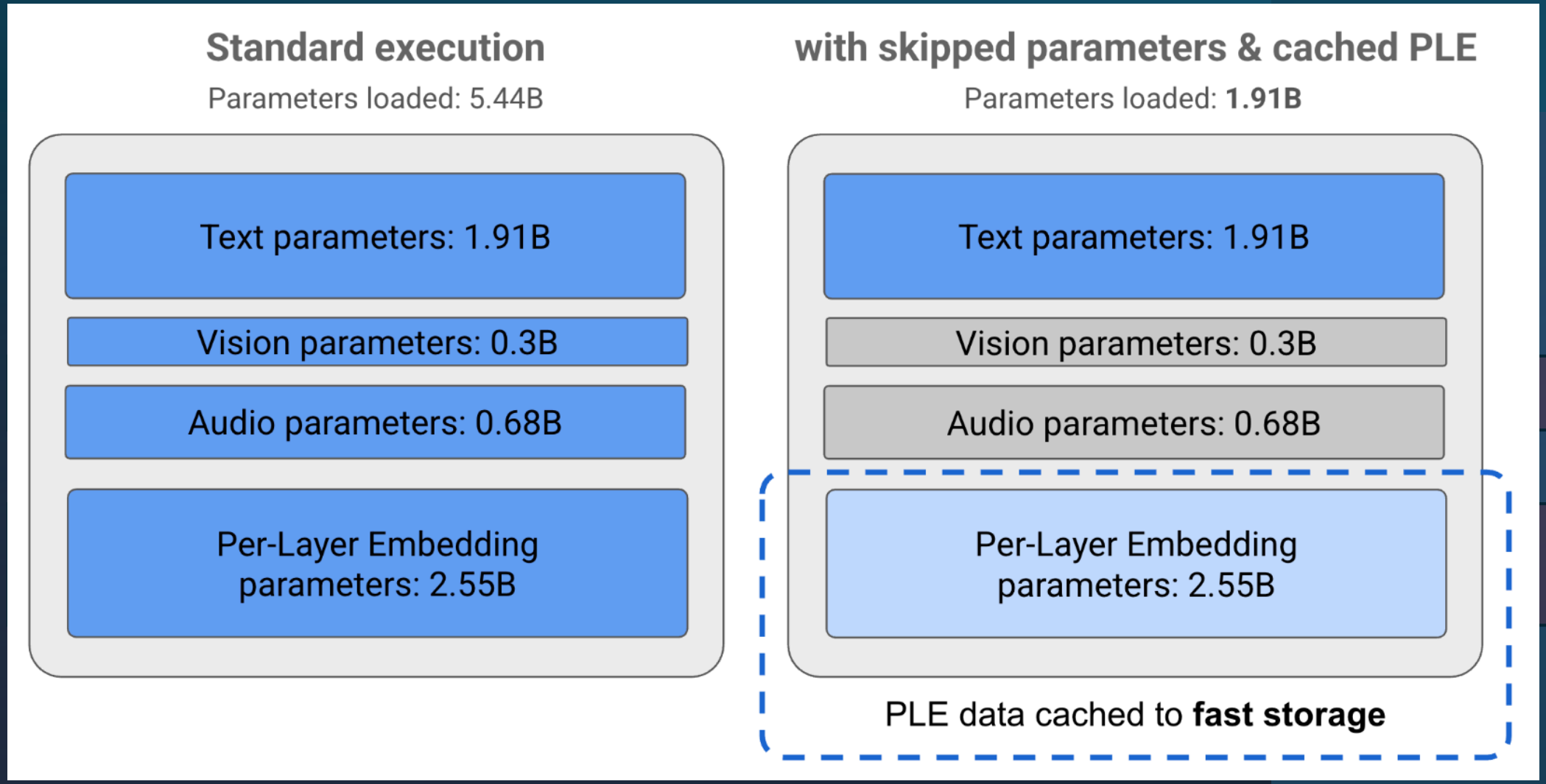
30,000,000,000,000

QualComm, 2025

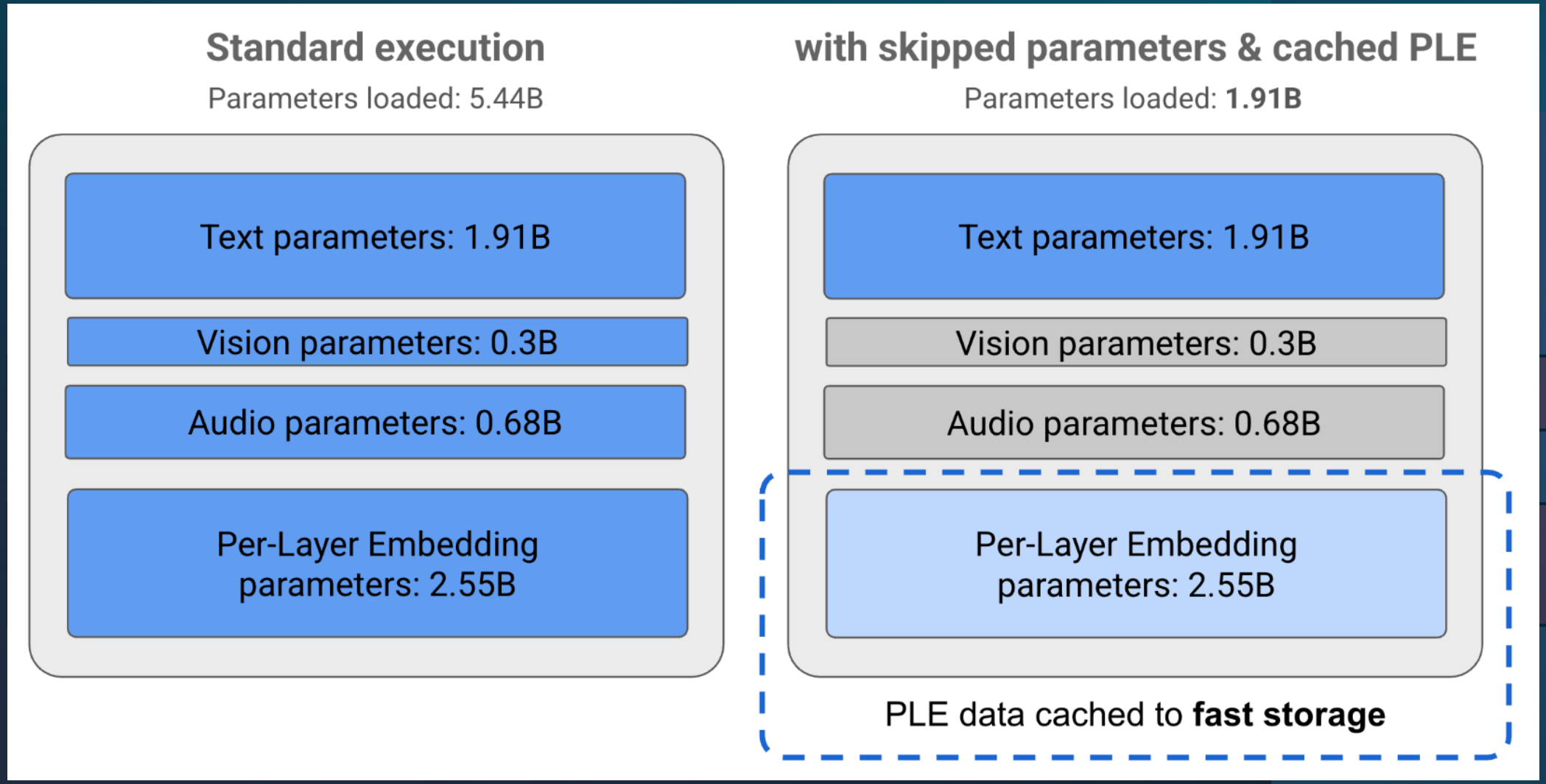
Feature	SLMs	LLMs
Size	<1B parameters	10B–70B+ parameters
Speed	Fast, <50ms latency (edge deployment)	Slower, 200–500ms (cloud-dependent)
Cost to Run	Low (can run locally) ~ 2Mvs.20M+	High (cloud or multi-GPU needed) 50M–100M+
Accuracy	Great for basic tasks	Best for complex tasks
Privacy	Better (can run offline)	Depends on platform/API
Context Length	Short (2K–4K tokens)	Long (up to 1M tokens in 2025)
Energy Efficiency	60–70% lower carbon footprint	High energy demand (160% rise in data center power by 2030)
Accuracy	92%+ in domain-specific tasks (e.g., NoBroker's multilingual customer service)	85% in general tasks; prone to "hallucinations" (~15% error rate)

Feature	SLMs	LLMs
Size	<1B parameters	10B–70B+ parameters
Speed	Fast, <50ms latency (edge deployment)	Slower, 200–500ms (cloud-dependent)
Cost to Run	Low (can run locally) ~ 2Mvs.20M+	High (cloud or multi-GPU needed) 50M–100M+
Accuracy	Great for basic tasks	Best for complex tasks
Privacy	Better (can run offline)	Depends on platform/API
Context Length	Short (2K–4K tokens)	Long (up to 1M tokens in 2025)
Energy Efficiency	60–70% lower carbon footprint	High energy demand (160% rise in data center power by 2030)
Accuracy	92%+ in domain-specific tasks (e.g., NoBroker's multilingual customer service)	85% in general tasks; prone to "hallucinations" (~15% error rate)

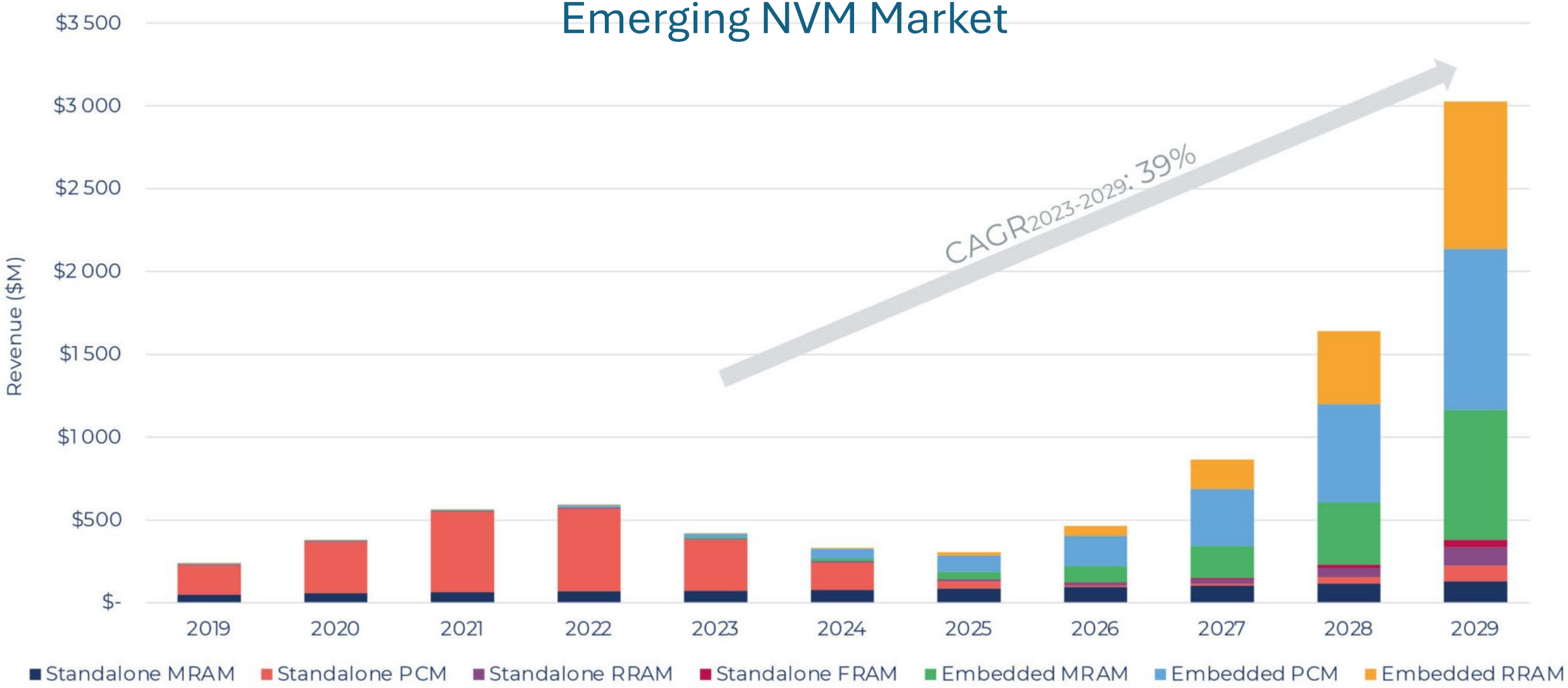
Google’s Gemma3N Mobile GenAI



Google’s Gemma3N Mobile GenAI

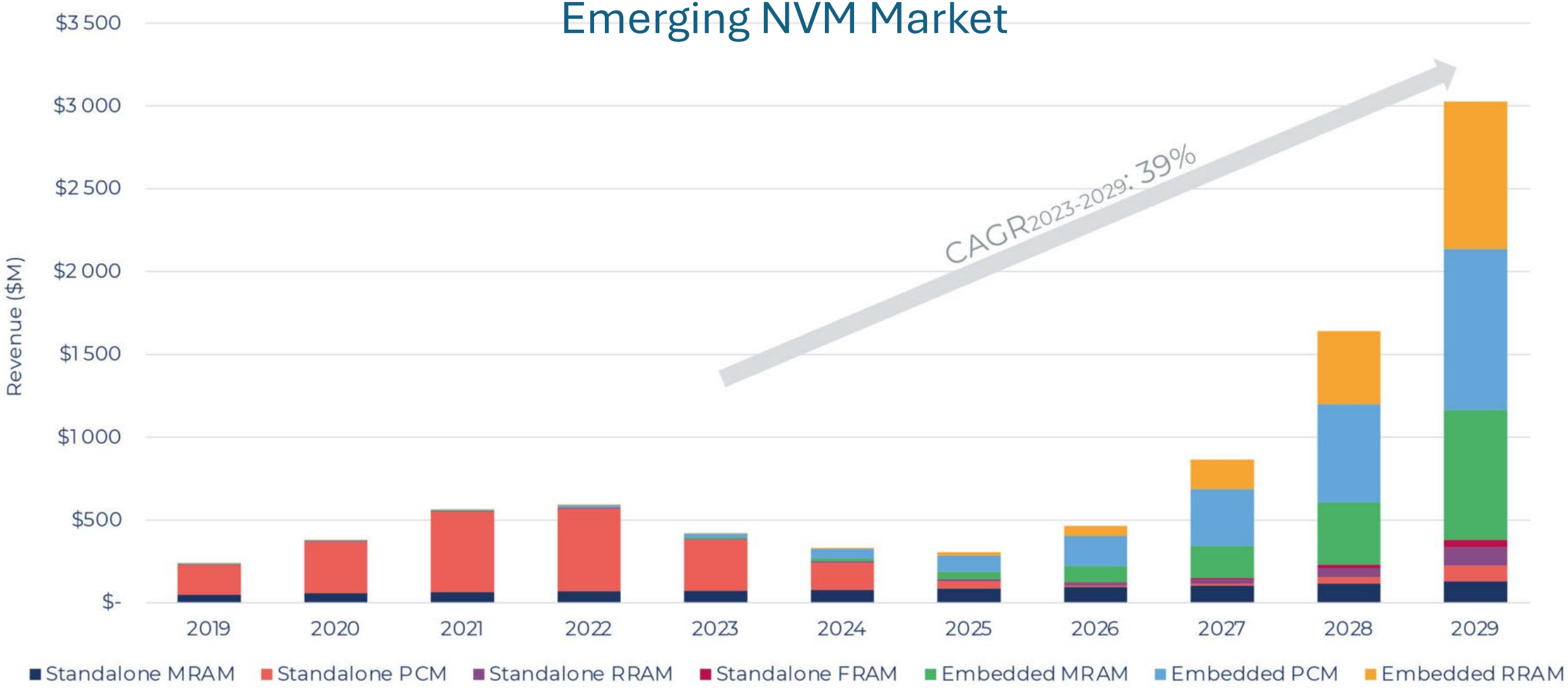


Emerging NVM Market



Yole Group, 2025

Emerging NVM Market



Yole Group, 2025