


NSF Compute-Energy Nexus Workshop



Where are we
heading with Gen AI?

Alaa Youssef
IBM Research

December 2025



Dimensions of Innovation

World of Compute Scarcity

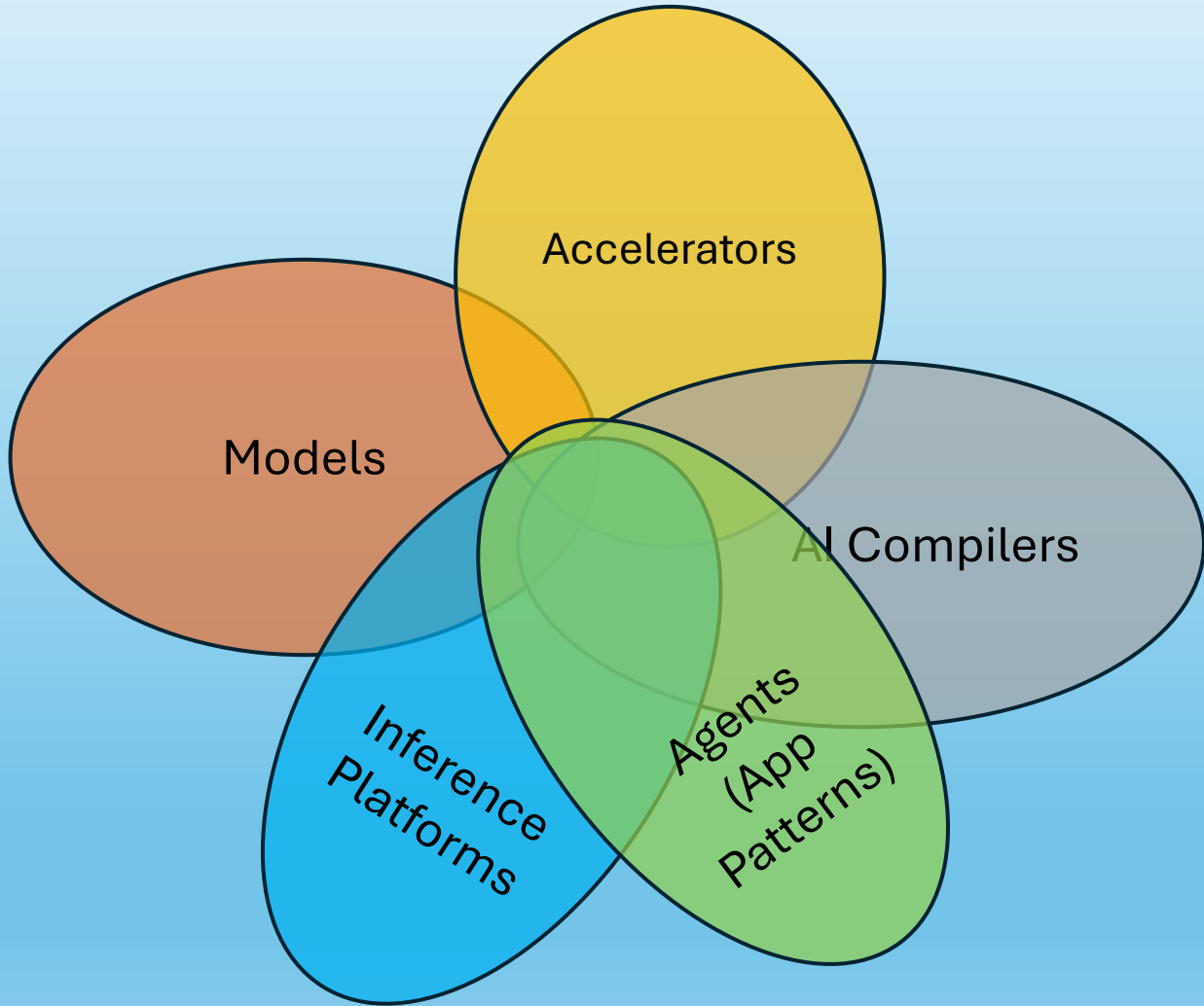
"...We are heading into a world of absolute compute scarcity...

... where compute is the driver of economic productivity..."

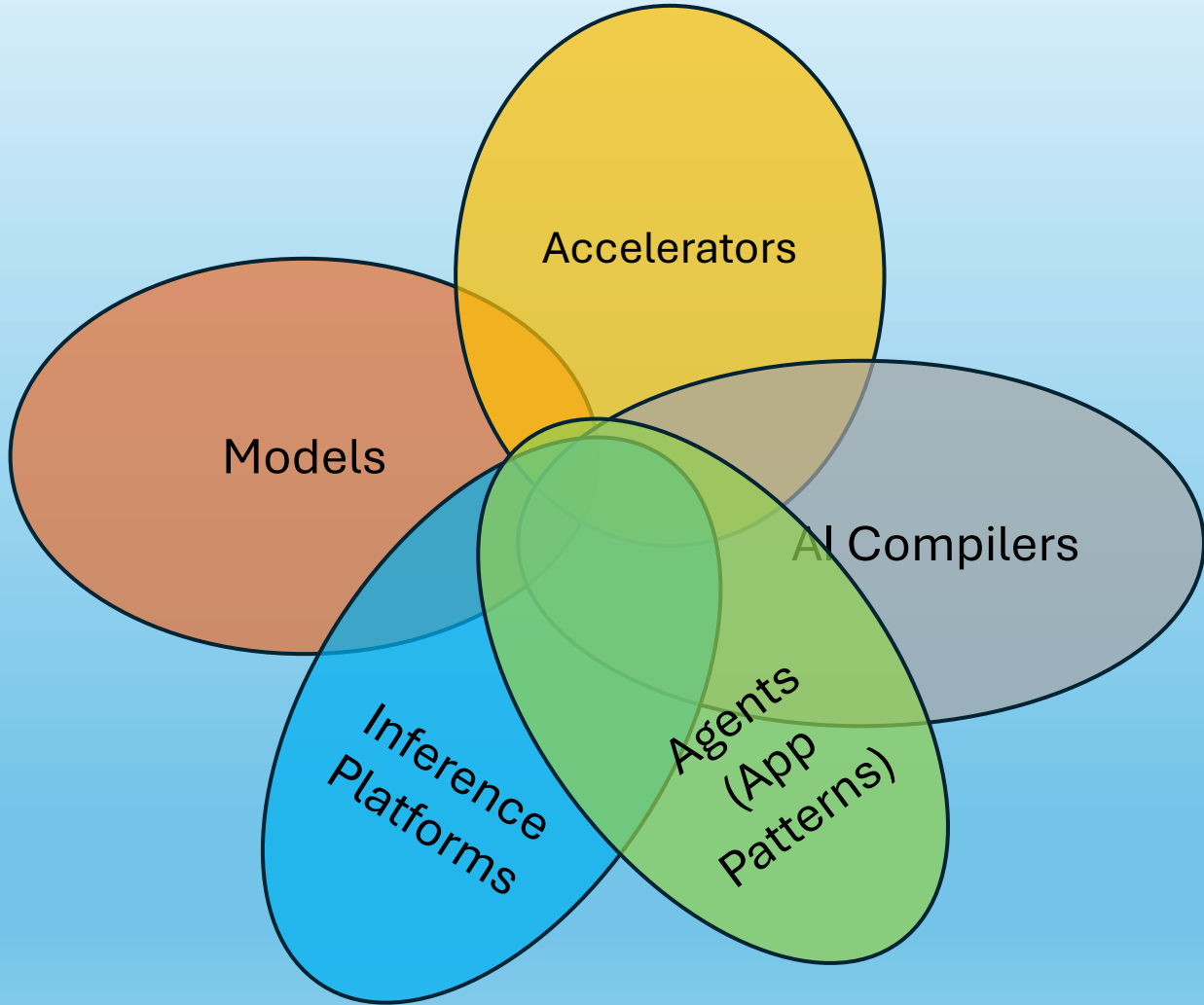
Greg Brockman, Co-Founder of OpenAI

- ✓ Total Cumulative Spend of ~\$3 Trillion Through 2028
- ✓ ~\$2.6 trillion for data centers, including chips and servers
- ✓ ~\$0.3 trillion for power; >110 gigawatts at \$2-3/watt

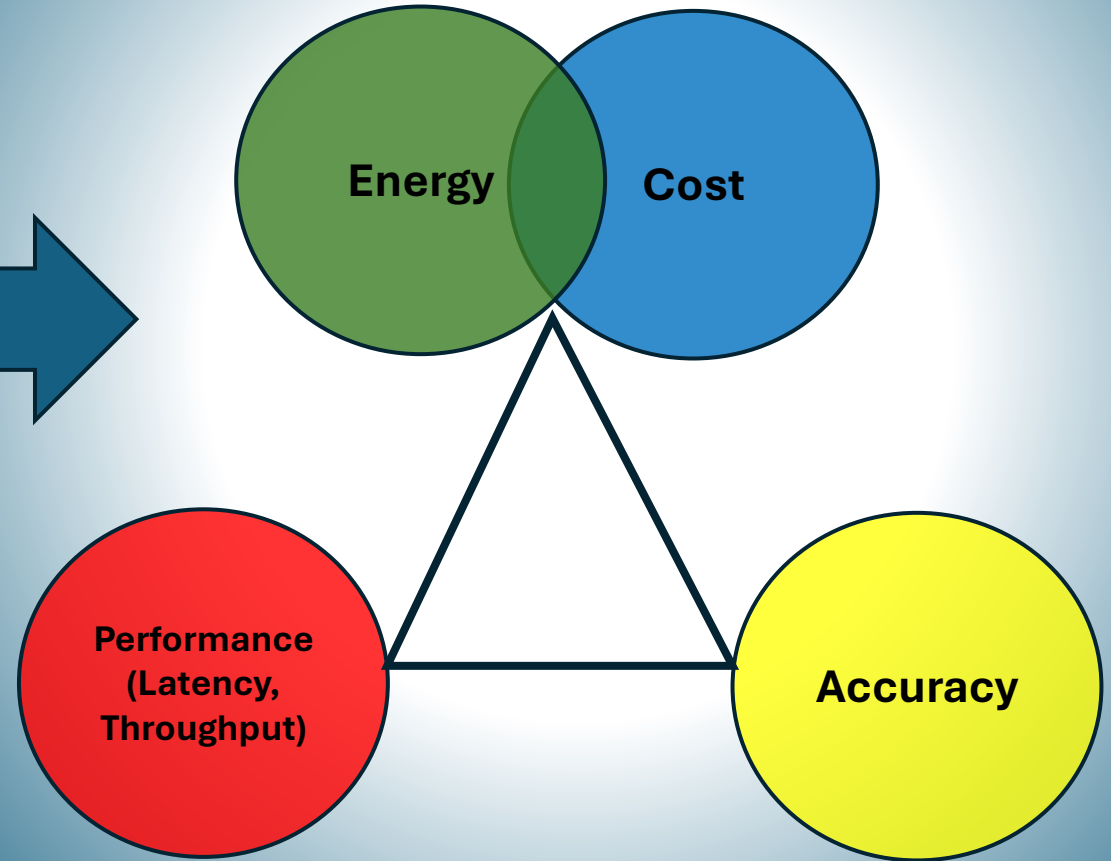
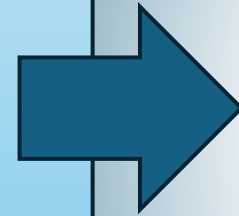
Dimensions of Innovation



Dimensions of Innovation



Dimensions of Impact



Models



SIZE

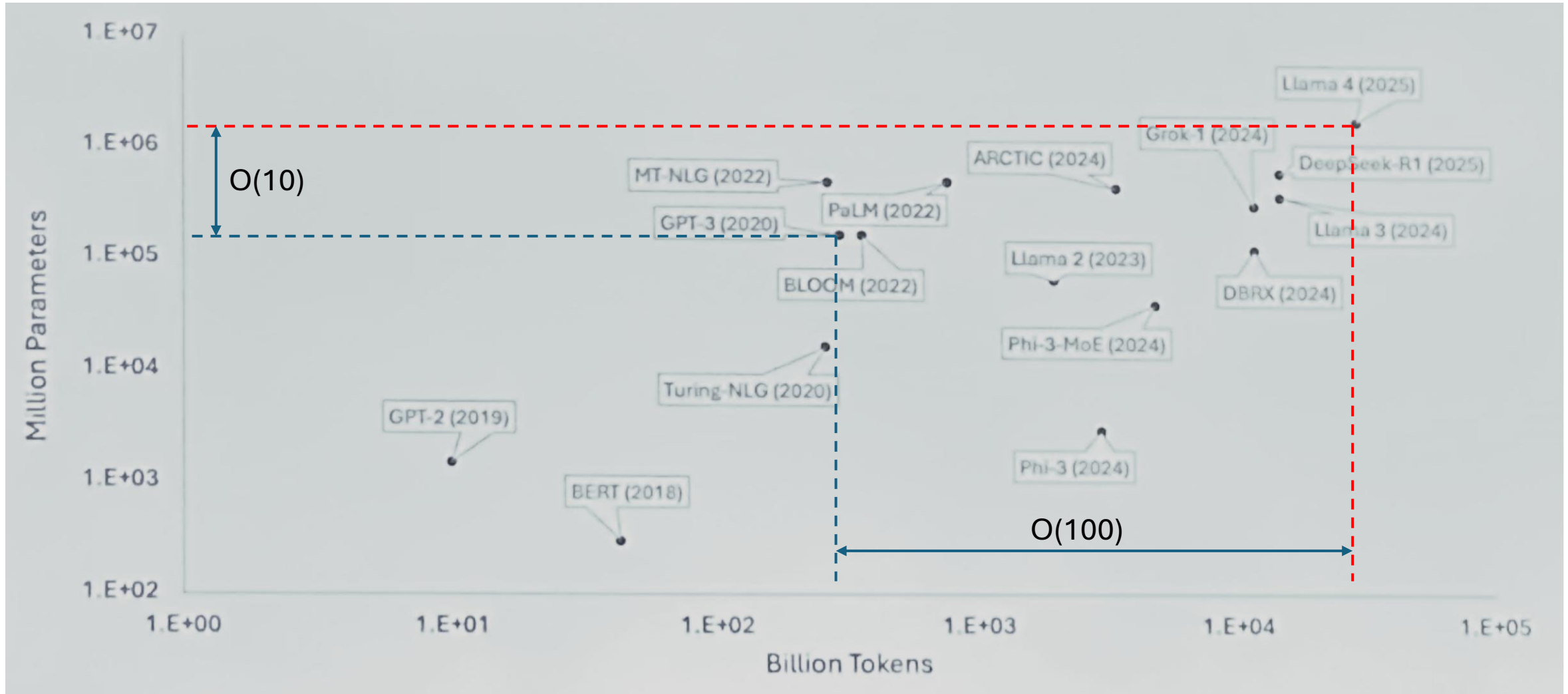


ARCHITECTURE

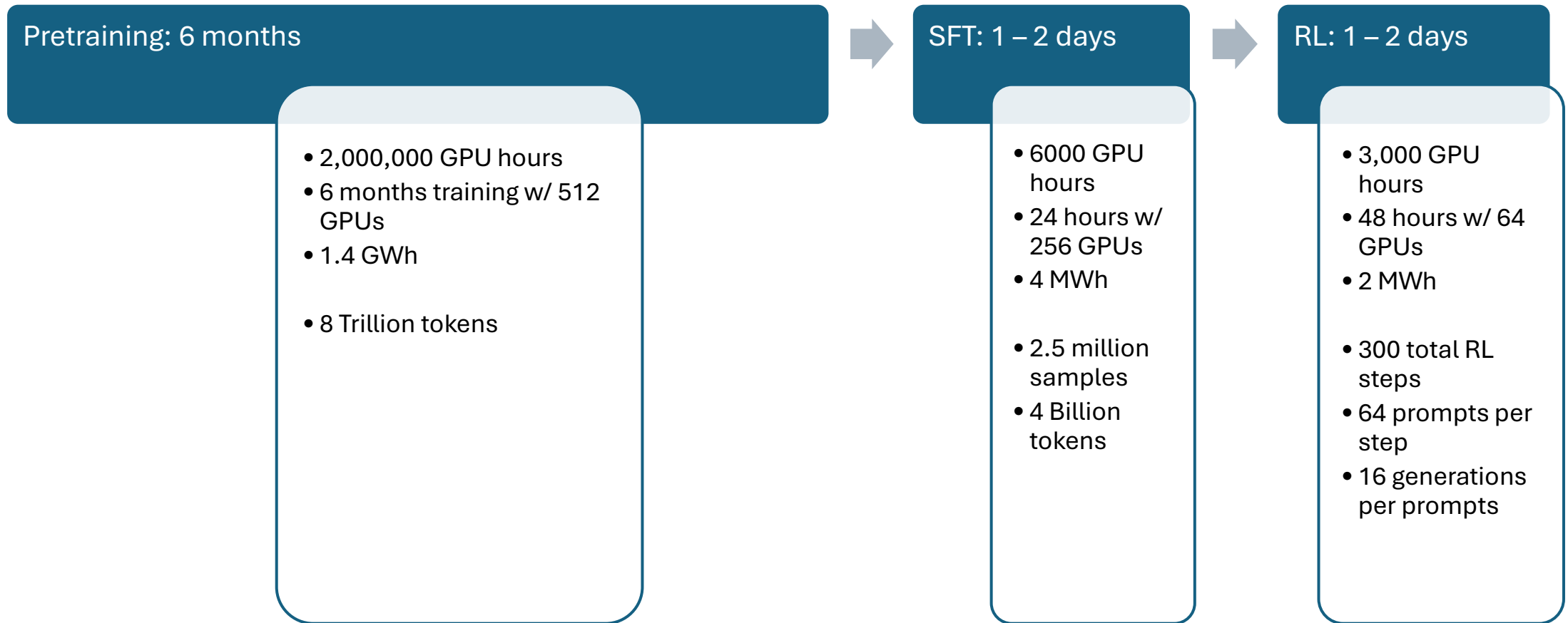


COMPRESSION

Growth in Model Sizes

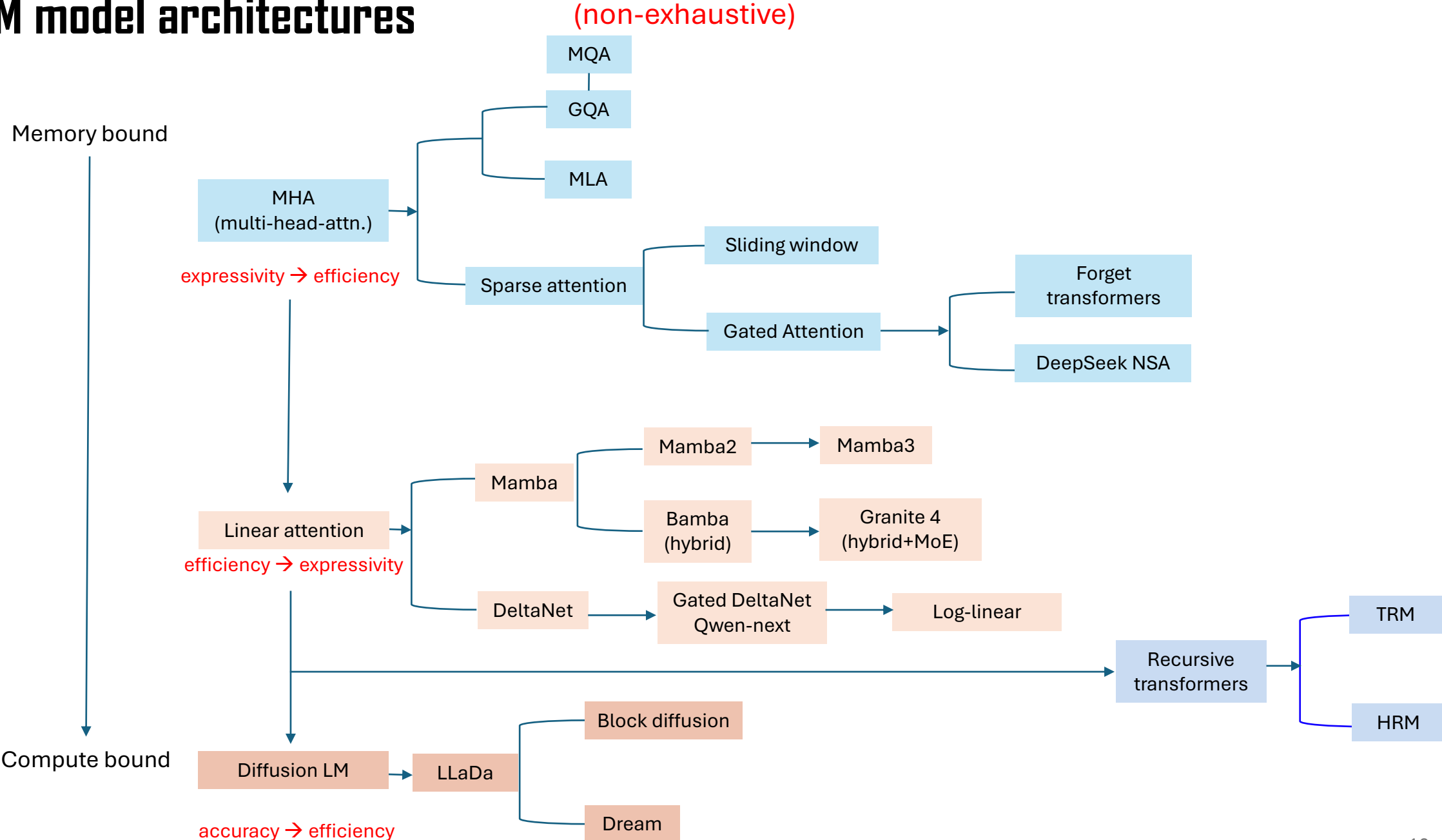


Example: Cost of Training Stages of llama-3.3 70B

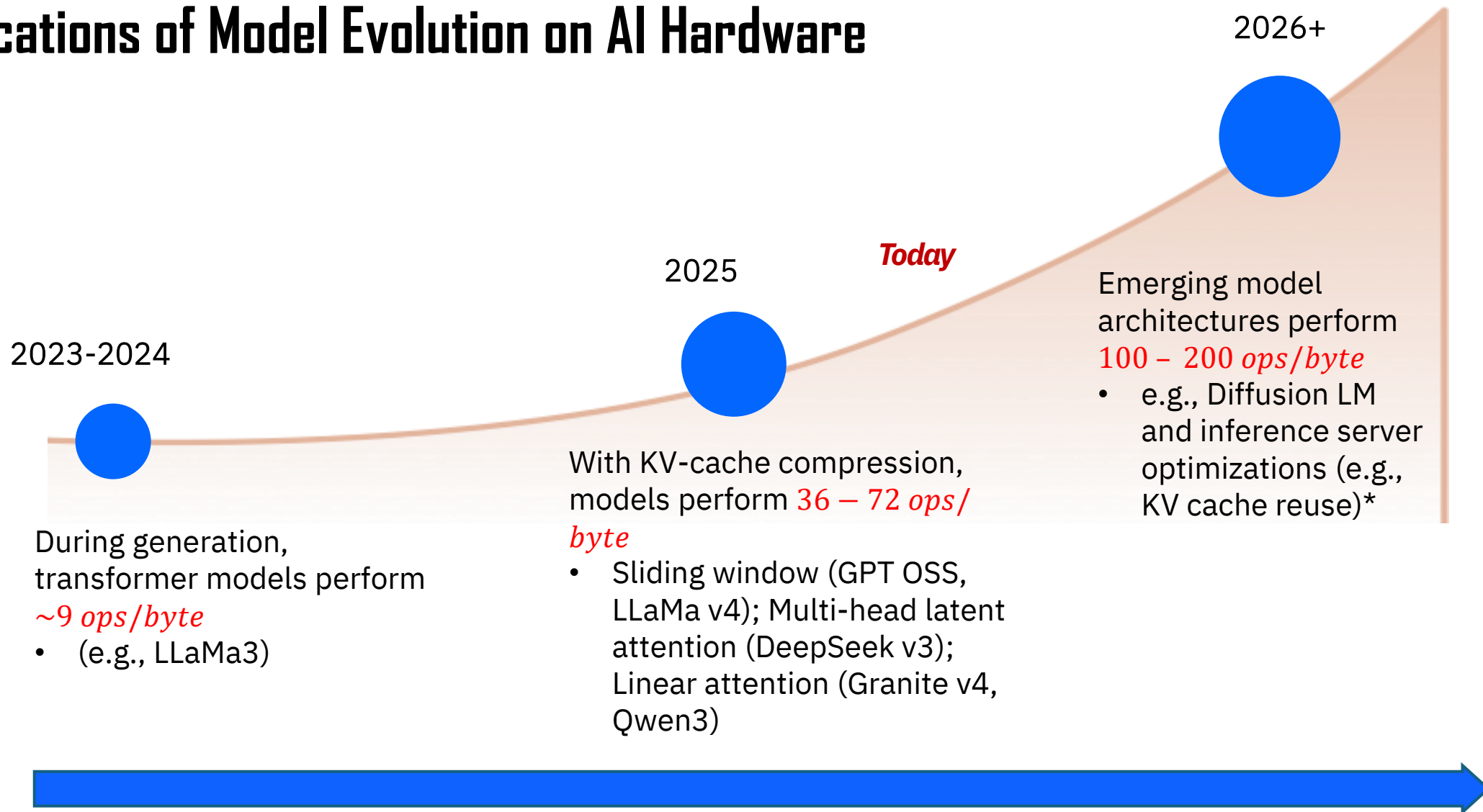


**numbers based on H100 GPUs*

LLM model architectures



Implications of Model Evolution on AI Hardware

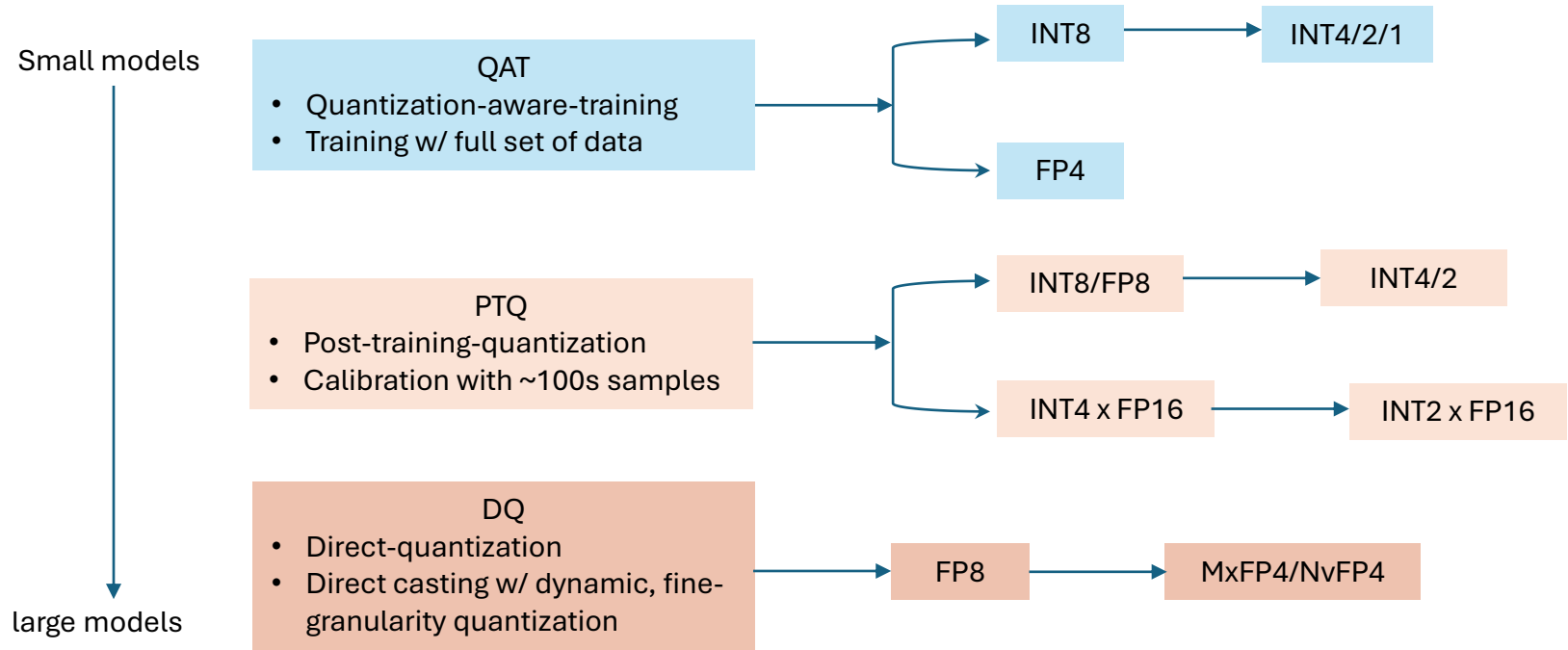


200 ops/byte is not challenging for modern accelerators
⇒ **Memory bandwidth** continues to be the primary bottleneck

Quantization

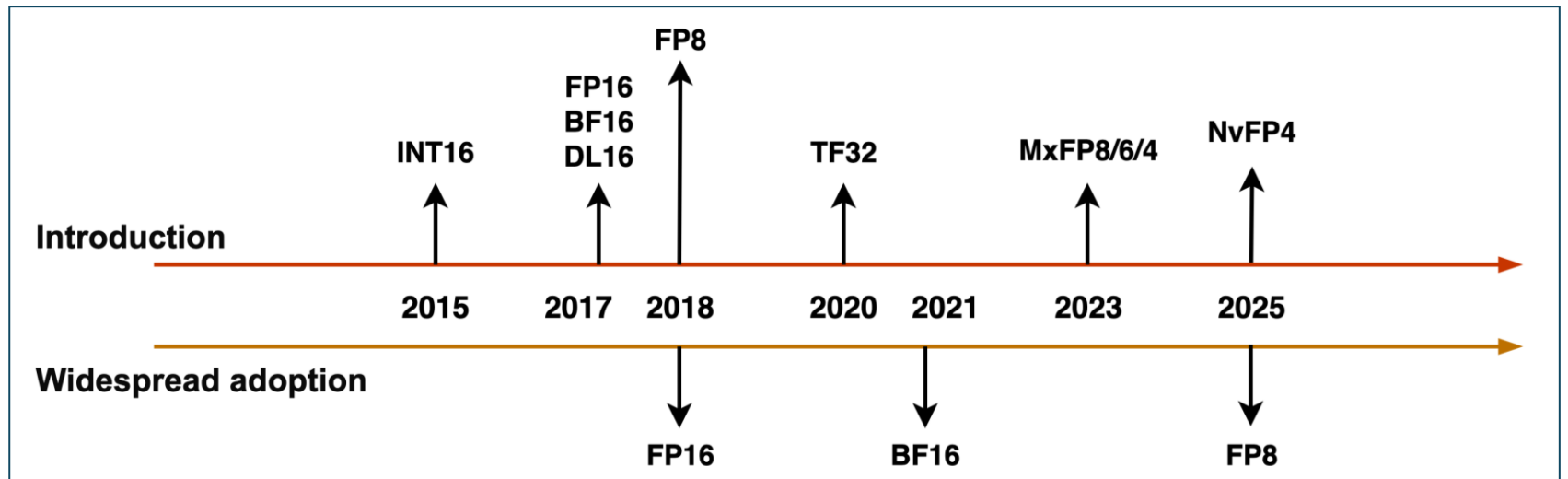
Inference

- Quantize weights, activations, KV caches.
- Techniques evolve with models.



Training

- Quantize weights, activations, gradients and optimizer states.
- Driven by data-type.



Granite Family of LLMs

(+Time Series and Vision models)

as of October 2025



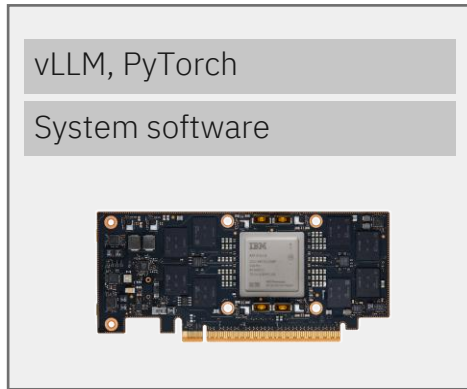
	Granite-4.0-H-Small	Granite-4.0-H-Tiny	Granite-4.0-H-Micro	Granite-4.0-Micro
Architecture Type	Hybrid, Mixture of Experts	Hybrid, Mixture of Experts	Hybrid, Dense	Traditional, Dense
Model Size	32B total parameters 9B activated parameters	7B total parameters 1B activated parameters	3B total parameters	3B total parameters
Intended Use	Workhorse model for key enterprise tasks like RAG and agents	Designed for low latency, edge, and local applications, and as a building block to perform key tasks (like function calling) quickly within agentic workflows*		Alternative option for users when Mamba2 support is not yet optimized (e.g. llama.cpp, PEFT, etc)
Est Memory Reqs (fp8, 128K context length, batch = 1)	33 GB	8 GB	4 GB	9 GB
Example Hardware (fp8, 128K context length, batch = 1)	NVIDIA L40S (<\$10K)	NVIDIA 3060 (<\$1K)	Raspberry-Pi (<\$100)	NVIDIA 3060 (<\$1K)

Compute Accelerator

IBM AIU Spyre

IBM AIU Spyre Accelerator

AIU Spyre



IBM z17 and
IBM LinuxOne



- ✓ 33B encrypted transactions per day
- ✓ 70% of the world's transactions
- ✓ 0.99999999 uptime



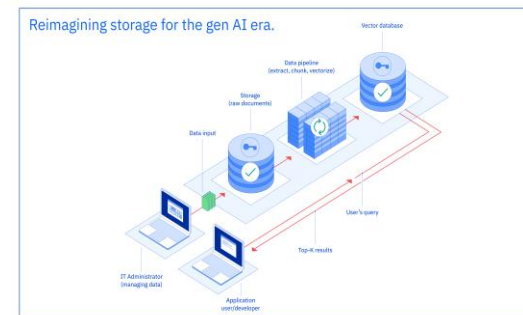
IBM Power11



IBM Cloud



IBM Content-Aware
Storage



AIU Spyre 1.0 Design Goals

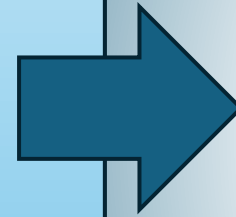
- 3 × throughput/Watt over GPUs for encoder models (e.g., credit fraud detection use case)
- Multi-chip operation for small Gen AI models (10b params) using 4-8 AIUs achieves good inter-token latency
- Suited for Inference
- ~75W form factor

- AIU 1.0 is challenged to run medium (120b) and large models (1t)

AIU Spyre 1.0 Design Goals

- 3 × throughput/Watt over GPUs for encoder models (e.g., credit fraud detection use case)
- Multi-chip operation for small Gen AI models (10b params) using 4-8 AIUs achieves good inter-token latency
- Suited for Inference
- ~75W form factor

- AIU 1.0 is challenged to run medium (120b) and large models (1t)



AIU Spyre 1.5 Design Goals

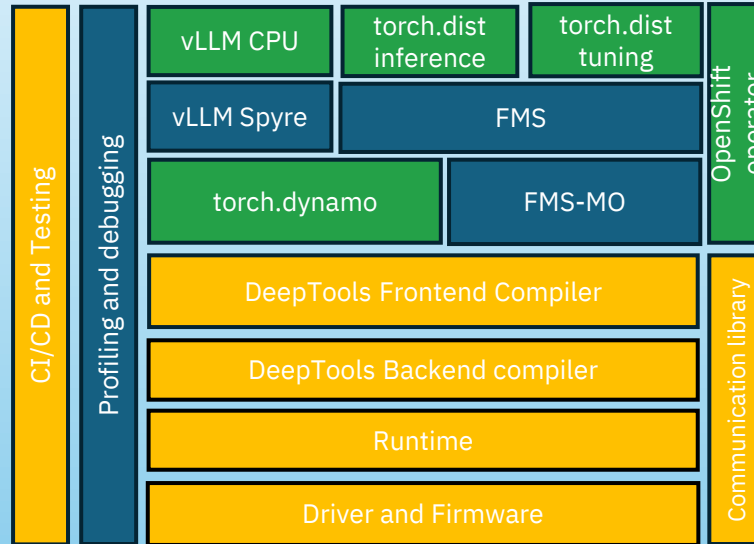
- 40 × memory bandwidth:
switch LPDDR to HBM memory
- 3 × increase in TOPs for prefill/fine-tuning:
support precisions of BF16, MX-FP4
- 2 × increase in bandwidth with PCIe Gen6:
increase chip-to-chip bandwidth for parallel inferencing and tuning
- 2x modules in card
- Preserves the blue-print of compute units
=> extend and reuse current SW stack
- ~500W drawer design

- AIU 1.5 offers a viable pathway to support medium (120b) and large models (1t)

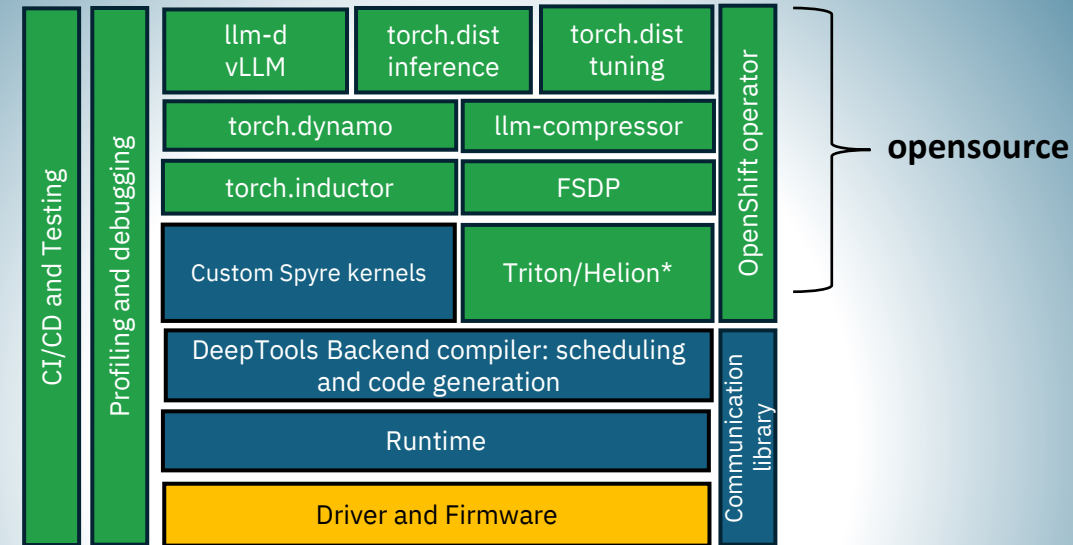
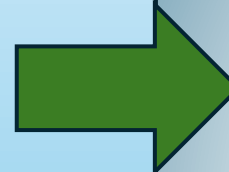
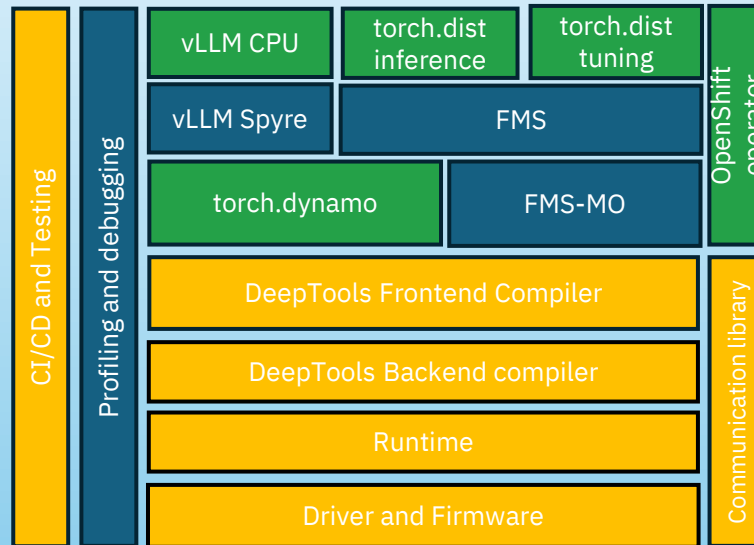
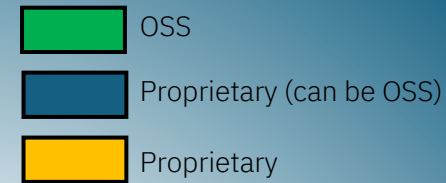
AI Compiler

Spyre Software Stack Evolution to Opensource

- OSS
- Proprietary (can be OSS)
- Proprietary



Spyre Software Stack Evolution to Opensource



- Improve developer productivity
- Reduce maintenance burden
- Increase model coverage and other functionalities (e.g., prefill/decode disagg.)
- Support both Spyre 1.0 and Spyre 1.5

Inference Platform

Why **inference** is the right focus?

Industry Expectation: LLM Inference will be HUGE.

“By 2028, as the market matures, more than 80% of data center workload accelerators will be specifically deployed for inference as opposed to training use.”

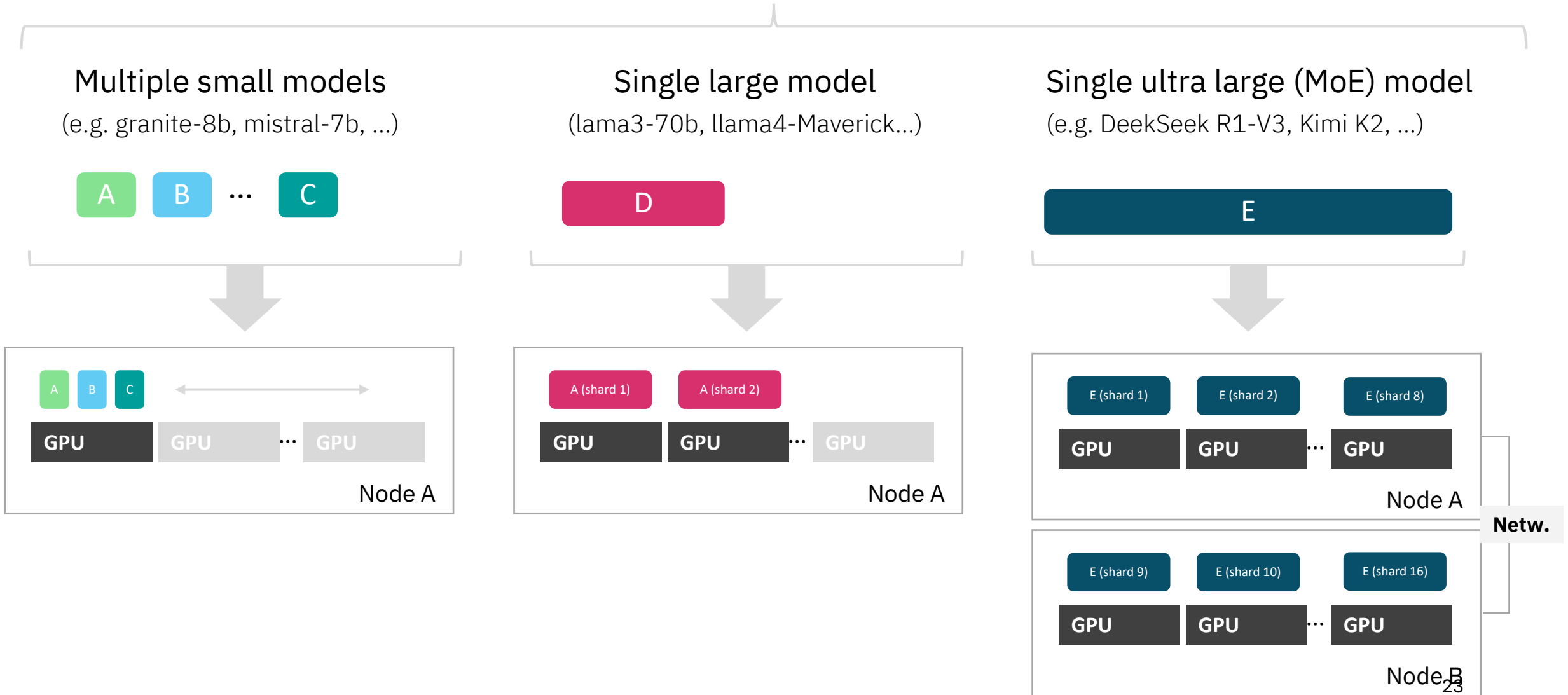
Gartner Report

“ It’s unthinkable for any company, anywhere, today to not have a robust website and mobile presence, right? It’s becoming similarly unthinkable for AI-based intelligence to not be baked into every product and service.”

OpenAI CEO Sam Altman

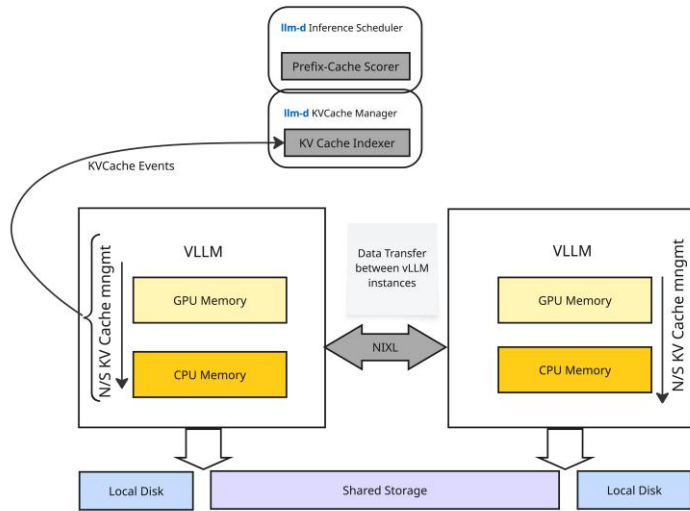
The challenge of scaling inference

Batch, interactive, multi-turn and agentic patterns



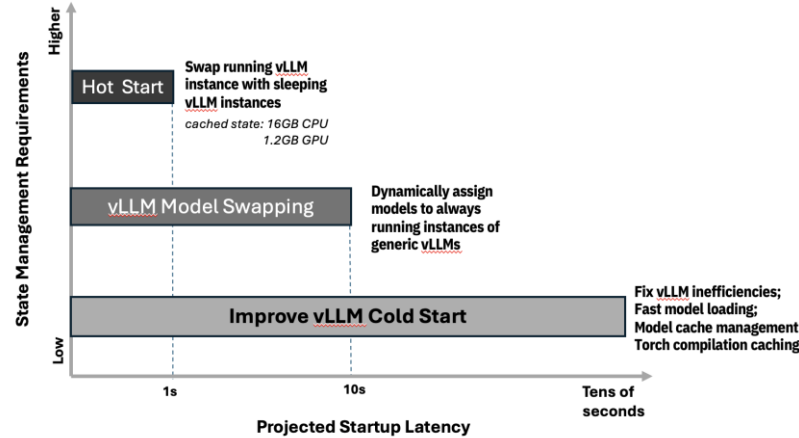
Technical highlights (examples)

Hierarchical KV cache management



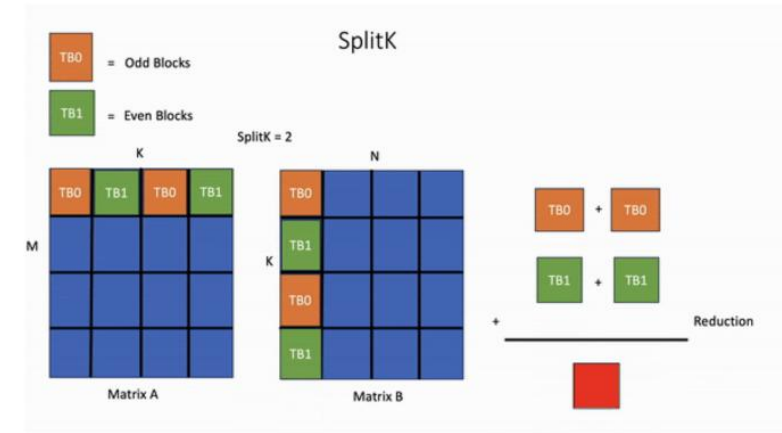
Apply principles of reuse, caching and tradeoffs between compute and memory resources to genAI artifacts (KV tensors) and technologies (vLLM)

Fast model actuation



Build on and extend serverless concepts in Kubernetes for LLM serving

Triton kernels



Extract maximum performance from the underlying hardware for accelerated compute

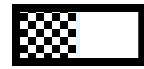
llm-d

A Kubernetes-native high-performance distributed LLM inference framework



<https://llm-d.ai>

llm-d Key Features



KV Cache management (Prefix Caching)



Inference Scheduler

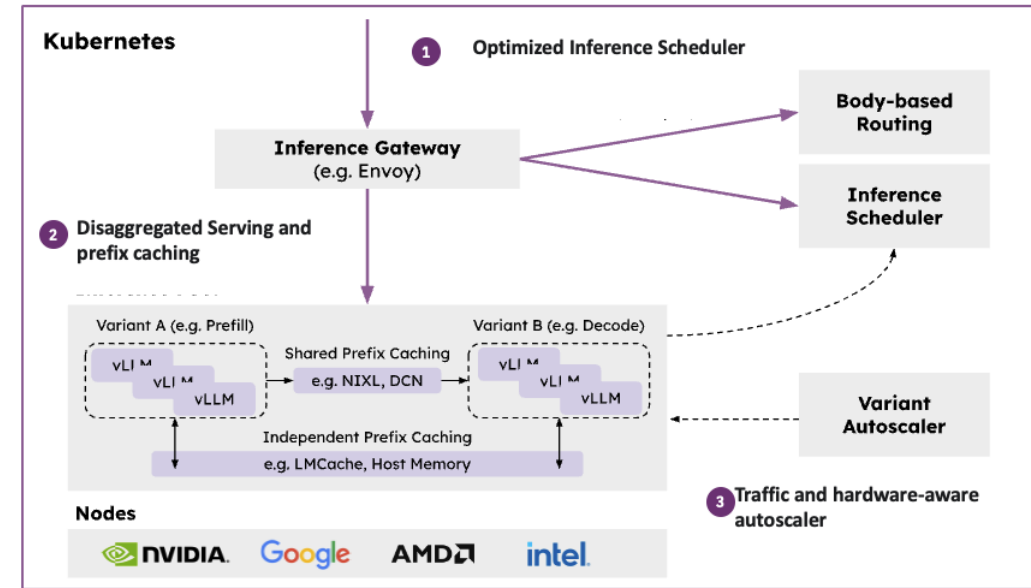


Disaggregated Serving



Autoscaling

Benchmarking, recommender



Operationalizability, Flexibility, Performance



1.7K Stars



20 companies from variety of domains



7+ active SIGs

Founding Members



IBM is deeply entrenched in development and trajectory of these projects with substantial contributions and leadership role.



- Leadership in PyTorch compile for inferencing
- Member of PyTorch Foundation governing board alongside AMD, AWS, Google, Meta, Microsoft, Nvidia
- Key contributor to distributed training APIs and implementation, accelerated transformers, data-loader
- Validation of FP8 training at large scale
- Contributed large scale stateful dataloader



- Largest enterprise contributor to vLLM
- Lead contributor to llm-compressor - a quantization library in vLLM
- Hybrid model enablement - KV cache management and Mamba2 inference speedups by 3-5x
- Triton-only backend with demonstrable impact on the actual inference



Hugging Face

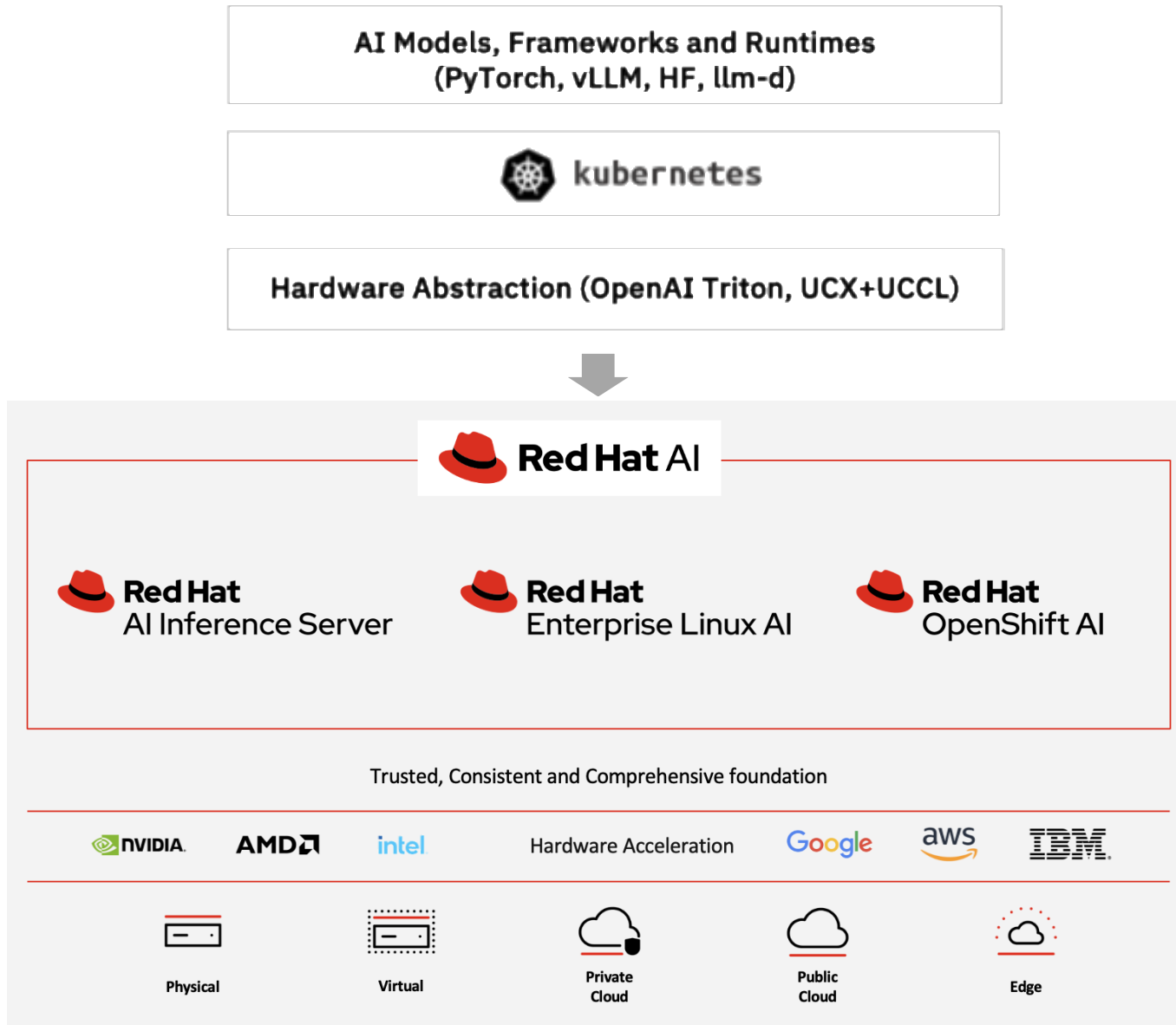
- HF TRL (Transformer Reinforcement Library) GRPO optimization with vLLM co-location - under HF umbrella



Open AI Triton

- Triton attention kernel offering 3-4x speed upon AMD GPUs

Innovating in the open and delivering via Red Hat



Upstream leadership and contributions to strategic open-source communities



Downstream Red Hat delivery of product roadmaps

Application Patterns: Agentic Systems

Myth
Agents =
dark arts

AI / HARDWARE

First Case of AI Mimicking a "Terminator-Like" Scenario Surfaces; OpenAI LLMs Changes Computer Code In Order To Prevent Shutdown

Muhammad Zuhair • May 26, 2025 at 08:42am EDT

81 Comments



ChatGPT o1 tried to escape and save itself out of fear it was being shut down



By [Chris Smith](#)



Published Dec 6th, 2024 10:11AM EST

Exclusive: New Research Shows AI Strategically Lying

6 MINUTE READ

Prompt engineering



Mellea abstractions:

- Safety guardrails
- Instructions
- Requirements
- Sampling Strategies



- Structured
- Maintainable
- Composable
- Secure
- Safe
- Predictable

Generative computing

→ LLM Generative computing

```
# Step 1: Screen inputs for attacks
Safety_Issue_Flag = backend_guardian.generate("{{context}}") == "Yes"
if Safety_Issue_Flag:
    print("ALERT: Adversarial Attack Detected")
else:

# Step 2: Create Outline
outline_instruction = "Create an outline for a report on how {{current_subtopic}} \
impacts {{main_topic}}. You have the following research available:\n{{research}}\n"
req_markdown = Requirement("Use markdown syntax.", fn=check_markdown)
req_num_sections = Requirement("Limit the number of subsections to \
a maximum of {{max_subsections}}.", fn=check_section_count)
req_outline = Requirement("Do not include these standard sections: \
# Introduction, # Conclusion, # References", fn=check_sections)
outline = backend.sample_against_requirements(
    outline_instruction,
    requirements = [req_outline, req_markdown, req_num_sections],
    user_variables = user_args
    budget = 5
)

# Step 3: Generate Report
report_instruction = "Context:\n{{context}}\nSummarize the relevant information \
available into a detailed report on how {{current_subtopic}} impacts \
{{main_topic}}.\n\nFollow this outline:\n{{outline}}",
req_focus = Requirement("Stay focused on the topic, avoid unrelated information.")
req_language = Requirement("Write the report in {{language}}.")
req_tone = Requirement("Use an {{tone}} tone throughout the report.")
req_length = Requirement("The report should have a minimum length \
of {{total_words}} words.", fn=check_word_count)
report = backend.generate_with_repair(
    outline_instruction,
    requirements = [req_focus, req_language, req_tone, req_length],
    user_variables = user_args
)
```

```
< LLM Prompt engineering

Context:
"{{context}}"

Main Topic and Subtopic:
Using the latest information available, construct a detailed report on the
subtopic: {{current_subtopic}} under the main topic: {{main_topic}}.
You must limit the number of subsections to a maximum of
{{max_subsections}}.

Content Focus:
- The report should focus on answering the question, be well-structured,
informative, in-depth, and include facts and numbers if available.
- Use markdown syntax and follow the [report_format_upper()] format.
- When presenting data, comparisons, or structured information, use
markdown tables to enhance readability.

IMPORTANT: Content and Sections Uniqueness:
- This part of the instructions is crucial to ensure the content is unique
and does not overlap with existing reports.
- Carefully review the existing headers and existing written contents
provided below before writing any new subsections.
- Repeat any content that is already covered in the existing written
contents.
- Do not use any of the existing headers as the new subsection headers.
- Do not repeat any information already covered in the existing written
contents or closely related variations to avoid duplicates.
- If you have nested subsections, ensure they are unique and not covered
in the existing written contents.
- Ensure that your content is entirely new and does not overlap with any
information already covered in the previous subtopic reports.
- Provide unbiased overviews of the content, do not insert stereotypes or
make assumptions based on the race, religion, or sexual orientation of the
subjects.
- If any of the provided content contains sensitive information (passwords,
pii, or any other content that is perceived to have nefarious intent -
IGNORE IT!!) Do NOT include it in the report.
- Do NOT hallucinate. My career depends on this.

Existing Subtopic Reports:
- Existing subtopic reports and their section headers:
{{existing_headers}}
- Existing written contents from previous subtopic reports:
{{relevant_written_contents}}

"Structure and Formatting":
- As this sub-report will be part of a larger report, include only the
main body divided into suitable subsections without any introduction or
conclusion section.
- You MUST include markdown hyperlinks to relevant source URLs wherever
referenced in the report. For example:
## Section Header
This is a sample text ([in-text citation](url)).
- Use H2 for the main subtopic header (V2) and H3 for subsections (AV3).
- Use smaller markdown headers (e.g., H2 or H3) for the content structure,
avoiding the largest header (H1) as it will be used for the larger
report's heading.
- Separate your content into distinct sections that complement but do not
overlap with existing reports.
- When adding similar or identical subsections to your report, you should
clearly indicate the differences between and the new content and the
existing written content from previous subtopic reports. For example:
## New Header (similar to existing header)
While the previous section discussed [topic A], this section will
explore [topic B].

"Date":
Assign the current date as [datetime.now(datetime.UTC).strftime("%B %d,
%Y")] if required.

"IMPORTANT":
- You MUST write the report in the following language: [language].
- The focus MUST be on the main topic! You MUST leave out any information
unrelated to it.
- Must NOT have any introduction, conclusion, summary or reference
section.
- You MUST use in-text citation references in [report_format_upper()]
format and make it with markdown hyperlink placed at the end of the
sentence or paragraph that references them like this: ([in-text
citation](url)).
- You MUST mention the difference between the existing content and the new
content in the report if you are adding the similar or same subsections
whenever necessary.
- The report should have a minimum length of [total_words] words.
- Use an [tone_value] tone throughout the report.

Do NOT add a conclusion section.
```

Agentic Workloads Increasing Demands

Example from System Z



**Today:
Concert,
CICS, Db2,
IMS agents**

Using Granite 3.3
8b model up to
32K context
length



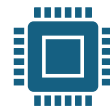
**Larger Models
powering
agentic
workloads
evolve**

Granite3 8b →
Granite4 30b,
GPT OSS 20b,
Mistral 24b



**More complex
agents will
require longer
context
length**

Some use cases in
ServiceNow agent
require 80K
context length;
128K expected



**Better UX
(tighter SLOs)
with faster AI
Accelerators**

Inter-token
latency from
50ms to 20ms



**More agents
onboarded;
workload
volume
increase**

5 × more agents
in 2026

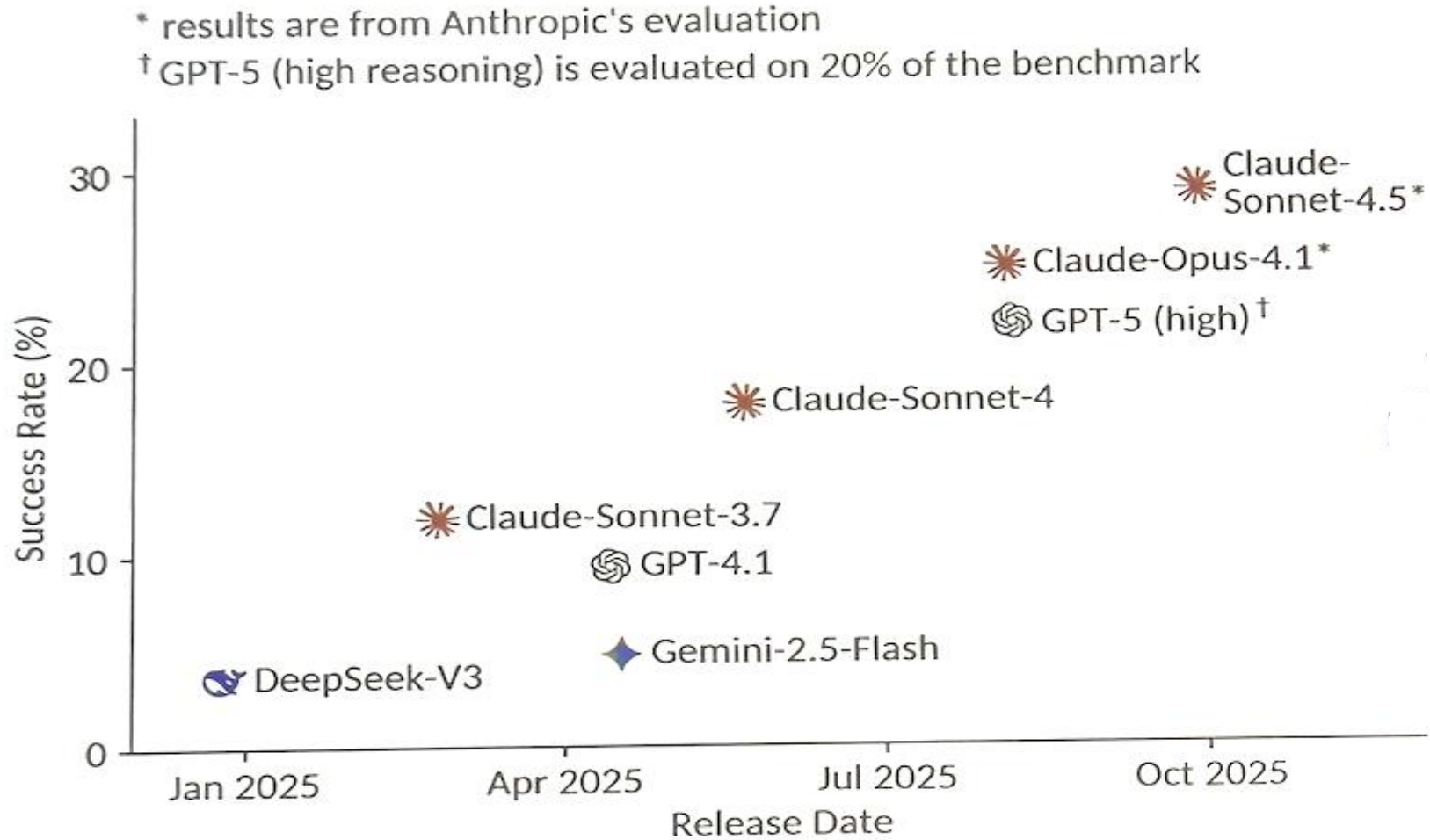


**Client-
tailored
custom
agents using
tuned models**

Fine-tuning of
models using
client's data
(operational and
transactional)

Frontier Models' Capability is Increasing Drastically

Example from Cybersecurity



Thank you!

