Following the money in AI

The capex curve

Big tech capex in 2025 alone is projected to exceed 13 years of the Apollo Program (inflation adjusted)

Massive graphics processing unit (GPU) purchases

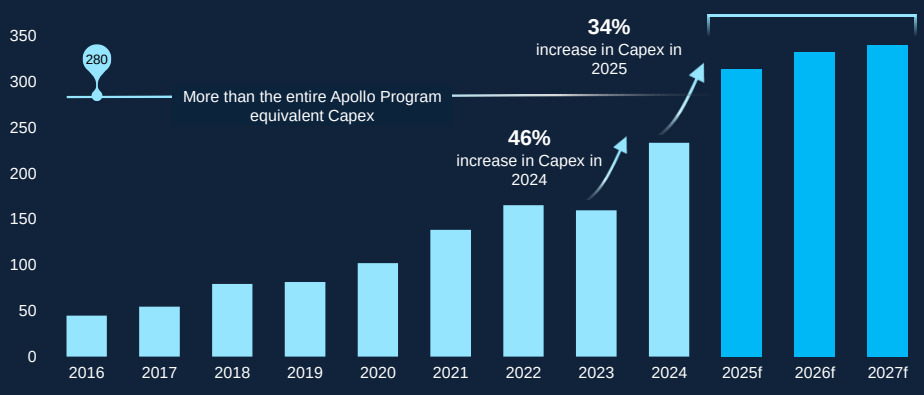
In 2024, Most of the Hyperscaler's ramped up GPU capex to catchup with OpenAI. The logic being the more training data, and larger the model and more compute you apply, the more effective the model will become.

Prisoner's dilemma

However, many big players also felt forced into an "arms race": if competitor A expands clusters, competitor B can't risk lagging behind

2025 "Reckoning?"

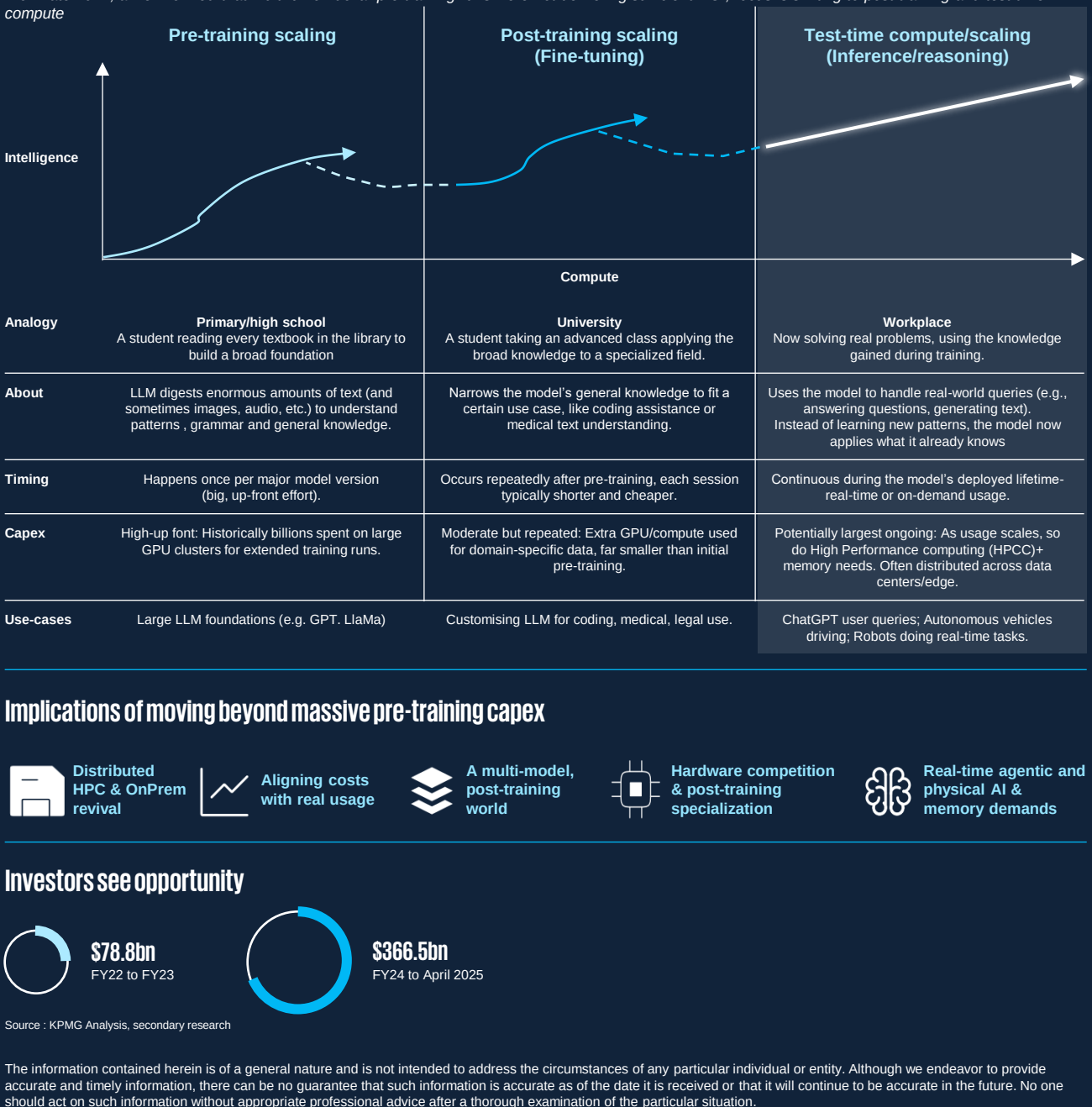
Some see 2025 as a year where labs must show dramatic new breakthroughs in Large Language Models (LLMs) to justify further giant capex.



Source: Capital IQ
Note: 2025-2027 based on analyst projections of six largest hyperscalers; where companies have different financial period ends, we have sought to align with calendar year.

Concern growing on compute return

In 2023 and 2024 Hyperscalers (five out of six major hyperscalers) spent over 50% of operating cash flows (OCF) on Capex.



Source: Capital IQ
Note: 2025-2027 based on analyst projections of six largest hyperscalers; where companies have different financial period ends, we have sought to align with calendar year.

The pivot to test time compute

From late 2024, a view formed that multi-billion dollar pre-training runs were not delivering sufficient ROI; focus is shifting to post-training and test-time compute

	Pre-training scaling	Post-training scaling (Fine-tuning)	Test-time compute/scaling (Inference/reasoning)
Analogy	**Primary/high school** A student reading every textbook in the library to build a broad foundation	**University** A student taking an advanced class applying the broad knowledge to a specialized field.	**Workplace** Now solving real problems, using the knowledge gained during training.
About	LLM digests enormous amounts of text (and sometimes images, audio, etc.) to understand patterns , grammar and general knowledge.	Narrows the model's general knowledge to fit a certain use case, like coding assistance or medical text understanding.	Uses the model to handle real-world queries (e.g., answering questions, generating text). Instead of learning new patterns, the model now applies what it already knows
Timing	Happens once per major model version (big, up-front effort).	Occurs repeatedly after pre-training, each session typically shorter and cheaper.	Continuous during the model's deployed lifetime-real-time or on-demand usage.
Capex	High-up font: Historically billions spent on large GPU clusters for extended training runs.	Moderate but repeated: Extra GPU/compute used for domain-specific data, far smaller than initial pre-training.	Potentially largest ongoing: As usage scales, so do High Performance computing (HPCC)+ memory needs. Often distributed across data centers/edge.
Use-cases	Large LLM foundations (e.g. GPT. LlaMa)	Customising LLM for coding, medical, legal use.	ChatGPT user queries; Autonomous vehicles driving; Robots doing real-time tasks.

Implications of moving beyond massive pre-training capex

- **Distributed HPC & OnPrem revival**
- **Aligning costs with real usage**
- **A multi-model, post-training world**
- **Hardware competition & post-training specialization**
- **Real-time agentic and physical AI & memory demands**

Investors see opportunity

Source : KPMG Analysis, secondary research