23 February 2026

Artificial Intelligence: Unpacking the investment risk and opportunity

Truffle Asset Management

AUTHOR: Danesh Ranchhod, Equity Analyst

Since OpenAI's release of ChatGPT, Artificial Intelligence (AI) has become one of the biggest themes in global markets. Enthusiasm has been high but more recently, fears regarding disruption risks have escalated.

Some of the new AI applications demonstrate the potential to disrupt traditional software businesses, while uncertainty remains regarding the complexity and pace of development. Some investors also fear that the rapid spending on AI infrastructure may be creating a bubble.

Given the industry's uncertainty and ongoing evolution, we continue to assess how AI fits into our portfolios. This article explains the foundation concepts of AI and its ecosystem. We focus particularly on the hardware sector as we aim to provide context on our investment positioning in this space.

How AI actually works (in simple terms)

What is a Large Language Model?

At the core of AI applications are Large Language Models (LLMs). It is important to understand that LLMs such as ChatGPT, Claude and Perplexity are not "intelligent" in the human sense. Instead, they are highly advanced prediction engines designed to identify patterns within vast datasets, which in turn enables them to respond "intelligently".

LLMs are designed to convert words into numeric codes known as "tokens". Through a training process of analysing massive datasets, the LLM will learn statistical relationships between these tokens. When a user asks their LLM a question, it does not retrieve a stored answer from a database. Instead, it predicts the most likely next token – one token at a time – based on the patterns it learned during training. It then translates the token into formative words. A trained LLM is effectively a probability machine that has processed significant portions of the internet to mimic human linguistic patterns. The more data it has processed and the more complex the statistical relationships learned, the more sophisticated its responses become.

The LLM lifecycle

Every LLM or AI model goes through two main stages - each driving specific hardware requirements.

1. **Training – building the "brain"**

This phase is expensive and compute-intensive. During training, trillions of statistical relationships are calculated to define the model's parameters and how language works.

These calculations require specialised chips called Graphics Processing Units (GPUs), which can perform many calculations simultaneously. As developers race to build more powerful and sophisticated models, the demand for compute power has grown rapidly (approximately 4.5x annually), requiring frequent updates to GPU chip design (roughly every 18 months).

2. Inference – using the "brain"

This phase is where the model generates real-time responses to users.

While inference (predicting a word) is less compute-intensive than training, it requires extremely large short-term memory. Because the GPU must retrieve the entire model each time from short-term memory for every word it generates, the bandwidth of memory chips becomes critical. If we think of memory as the data pipeline and the GPU as the engine, the requirement becomes clear. Today, the pipeline is too narrow relative to the engine's capacity. During inference, GPUs are therefore waiting for data to arrive, sitting partially idle not because of insufficient compute power, but because memory bandwidth cannot feed them quickly enough.

As LLMs grow in complexity, through greater parameters and reasoning capabilities, and generate longer outputs (including video and image generation), memory infrastructure becomes even more important.

Where are the bottlenecks?

There are three primary constraints in evolving the AI ecosystem:

1. Compute power (GPUs).
2. Energy supply.
3. Memory bandwidth (the volume of data that moves from short-term memory to the GPU in a unit of time).

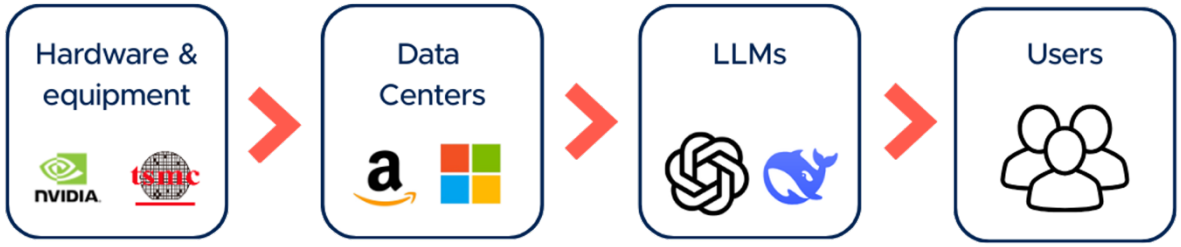
Energy supply and memory bandwidth are emerging as the most pressing medium-term challenges. While GPU computation power has improved rapidly, memory bandwidth has not kept pace. This creates inefficiencies in data centres with idle, expensive GPUs. In practical terms, memory bandwidth has become the bottleneck. As a result, expensive GPU infrastructure is underutilised, which in turn depresses data centre efficiency and weakens overall unit economics.

To date, technological improvements, spanning both model optimisation and hardware architecture, have meaningfully reduced the cost per unit of AI compute. However, these efficiency gains have been more than offset by accelerating demand. As models grow larger and use cases proliferate, total consumption continues to scale faster than the underlying efficiency improvements.

In other words, while cost per token or per inference may decline, aggregate compute and memory requirements continue to rise in absolute terms. The net effect is that overall infrastructure demand continues to expand despite ongoing technological progress.

Where are the profit pools in the AI ecosystem today?

To understand where we invest, it is worthwhile to segment the value chain into three broad categories.



1. Model developers (LLM Companies):

Companies include: OpenAI, Anthropic, xAI, Google, and Meta.

These companies build AI models which are offered via subscription and licensing as enterprise solutions.

The challenge: Monetisation paths are uncertain. Despite growing adoption, operating costs are extremely high, and revenue is constrained as many users remain on free tiers. OpenAI (ChatGPT), for instance, has projected a breakeven by 2029, with revenues of $145bn from c$20bn in 2025.

The industry hopes that "Agentic AI" - autonomous agents capable of complex workflows - will eventually drive enterprise revenue to bridge this gap.

2. Infrastructure providers (Hyperscalers)

Companies include: Microsoft, Amazon, Google and Oracle.

These entities build and operate the massive data centres required for AI training and inference.

Planned AI capital expenditure in this space is significant, with hyperscalers estimated to spend approximately USD 650 billion globally in 2026. However, the return on investment (ROI) is uncertain. AI data centres are expensive to build, with 50-60% of costs allocated to GPUs that have a short depreciation cycle of just 4-6 years.

We are concerned that capital in these businesses is being deployed into assets with compressed returns. For example, Microsoft's return on incremental capital is likely to

decline significantly as it shifts cash flow from high-margin software to short-cycle, capital-heavy AI infrastructure.

3. Hardware and Equipment Providers

Companies include: Nvidia, TSMC, Samsung, SK Hynix, Micron, and ASML

These entities represent the "pick and shovel" enablers, supplying semiconductors (directly or indirectly) and power equipment /generation to the hyperscalers.

Historically, in disruptive technology cycles, suppliers of critical components have achieved profitability first. Currently, the bulk of the profit pool therefore resides with these entities.

Hardware providers also capture immediate value. Approximately 60% of a data centre's build cost goes toward server racks (GPU, CPUs and memory chips), with semiconductor companies ranking among the most profitable players in this space.

Our positioning in the AI space

Our exposure to the AI theme has been concentrated in the third category – the enablers of AI. We have specifically focused on the semiconductor space. Below, we share insight into the investment case for some of these companies.

Taiwan Semiconductor Manufacturing Company (TSMC)

TSMC manufactures advanced logic or processor chips, effectively the brains of data centres, and are also used in smartphones, automotive systems and consumer electronics. TSMC is purely a manufacturer and will work with microchip designers.

We own TSMC for several reasons:

- **Dominant market position:** High capital intensity and decades of Research & Development create immense barriers to entry. Since the 1990s, the industry has consolidated to four players from over 20, with TSMC having a 90% global dominance in the most advanced chip manufacturing.
- **Improving earnings quality:** TSMC faces limited competition as the dominant supplier for the most advanced chips and earnings volatility has reduced compared to previous cycles.
- **Margin improvement:** Growing scale, productivity improvements and manufacturing efficiency continue to improve profitability.
- **Capital discipline:** While capital expenditures are high, assets are long-lived, and TSMC's prudent management of capacity additions means returns on incremental capital are attractive. We estimate above 20% in the medium-term.

- **Relative safety vs Nvidia:** We view TSMC as a safer investment than Nvidia (a chip designer), given the company faces competitive risks from hyperscalers designing their own chips. TSMC, as a manufacturer, is more insulated from this risk.
- **AI bubble risk:** Elements of today's AI investment cycle exhibit characteristics commonly associated with speculative bubbles. Valuations across parts of the value chain are elevated, in some cases despite limited visibility into current earnings or sustained operating losses. For some AI companies, the path to durable profitability remains uncertain, with future margins potentially far lower than implied by current market expectations. That said, even if portions of the ecosystem prove to be overcapitalised or mispriced, the underlying demand for high-performance computing is not new. It reflects a structural, multi-decade trend driven by increasing data intensity, more complex workloads, and rising computational requirements across industries. While equity market outcomes may diverge sharply across participants, the long-term trajectory of compute demand itself remains intact.

Memory companies: Samsung Electronics, SK Hynix and Micron

Samsung Electronics, SK Hynix and Micron Technology produce the bulk of short- and long-term memory chips used in data centres, PCs, smartphones, automotives and other consumer electronics. Like the logic chip industry, the memory chip market has consolidated from 19 players in the 1990's.

Short-term memory (DRAM) has historically represented the larger revenue base and profit for these companies. While Samsung Electronics also operates smartphone and consumer electronics businesses, as well as a foundry business, we expect its memory division to be the key medium-term earnings driver for the tech group.

AI servers increasingly rely on advanced Dynamic Random Access Memory (DRAM), specifically HBM (High Bandwidth Memory). As GPUs become more powerful, performance is increasingly constrained by memory throughput. To operate efficiently, these processors require substantially higher bandwidth memory capable of delivering data at speeds that match computational intensity.

We currently own Samsung with our investment case premised on several factors, including:

- **Structural demand shift:** Historically, the demand for memory has been cyclical, AI's distinct cycle is driven by the skyrocketing DRAM content per server (from single-digit gigabytes to tens of gigabytes) and continued scaling of GPUs. The demand for memory intensity is increasing.

- **Inference growth:** Similarly, as AI usage grows, inference (real-time responses) becomes a larger part of demand relative to training and is particularly memory-intensive.
- **Supply constraints:** HBM is harder to manufacture, requiring significantly more silicon than legacy DRAM. Hence, the legacy DRAM supply has also tightened, supporting pricing and further compounding profits for manufacturers.
- **Extended profit cycle:** Historically, memory industry profit cycles have been sharp and short in duration, given the pace of technological development, which led to quick price revisions as memory supply increased and caught up with demand. However, in the current AI cycle, demand dynamics appear structurally different. Inference is likely to grow exponentially; hence, the medium-term runway for memory demand will continue. This supports tighter supply-demand conditions and suggests the potential for a more prolonged and durable profit cycle than in prior periods.

Valuation opportunity: Samsung trades at a meaningful discount to peers Micron and SK Hynix and has the flexibility to shift production from less profitable foundry capacity to higher-margin memory to meet near-term demand.

Conclusion: staying selective

AI is transformative, but it is also capital-intensive and volatile as an investment.

Rather than chasing the most visible AI application, we remain selective, focusing on the "enablers" of the technology ecosystem over the "disrupters" or downstream application developers. We specifically like hardware and semiconductor providers with strong market positions and attractive returns on capital.

Our investment process is anchored on a fundamental valuation-based approach. When assessing companies in the AI industry, we focus on returns on incremental capital deployed and the sustainability of those returns through and post-cycle. We remain mindful of the risk that capital expenditure growth could moderate, and the implications this may have for valuation compression across the sector. In recognition of this risk, we actively manage position sizing within the portfolio to ensure exposure remains proportionate to the underlying uncertainty.

We are currently reviewing the software sector, which is experiencing significant volatility due to AI disruption concerns, and will provide further insight as a second part to this article.

AUTHOR

Danesh Ranchhod | Equity Analyst

Danesh joined Truffle in March 2025 as an equity analyst covering global stocks. Danesh brings a wealth of experience in research, having worked as an equity analyst at Franklin Templeton Emerging Markets for 14 years, covering companies in South Africa and Frontier Africa across various sectors. He began his career as a performance and attribution analyst, first at JP Morgan Administration Services and then Investec Asset Management (now Ninety One). He has also worked as a buy-side trader for RECM and in roles across operations and investments at Optis Investment Management.

Disclaimer

info@truffle.co.za | +11 035 7337 | truffle.co.za