ARCHITECTURE NOTES · RESOURCE

DeepSeek V4 and the new economics of inference

DeepSeek V4 did not change the model layer. It changed what a margin on the model layer costs. The interesting part of the AI stack is now everywhere else.

Companion to: DeepSeek V4 and the new economics of inference.

What this deck covers

A year ago, frontier-tier inference cost was treated as a soft floor. Vendors were not literally aligned on price, but the assumption inside most AI product roadmaps was that the cost per million tokens for a top-bracket model would stay in the same order of magnitude for a while. Long enough, anyway, to underwrite the unit economics of agentic SaaS at the prices the market was paying.

- What DeepSeek V4 actually launched
- Why the pricing is the story
- Mixture-of-experts, briefly and honestly
- The 1M context window, and what it actually changes

What DeepSeek V4 actually launched

01 **DeepSeek V4 is the fourth-generation release**

DeepSeek V4 is the fourth-generation release from a Chinese AI lab that has been quietly competitive on inference economics for the last eighteen months.

02 **It ships in two configurations**

It ships in two configurations.

03 **V4-Pro is the flagship**

V4-Pro is the flagship.

Why the pricing is the story

01 **The reason to watch DeepSeek V4**

The reason to watch DeepSeek V4 is not the benchmark headline.

02 **It is the gap between what**

It is the gap between what frontier teams have been charging and what serving the same capability appears to actually cost in 2026.

03 **When a closed lab quotes a**

When a closed lab quotes a price per million input tokens, that price reflects model cost, infrastructure cost, R&D amortisation, sales and platform overhead, and a margin.

Mixture-of-experts, briefly and honestly

01 **The architecture choice is the part**

The architecture choice is the part that gets simplified into nonsense in most coverage, so it is worth stating cleanly.

02 **A dense transformer activates every parameter**

A dense transformer activates every parameter on every forward pass.

03 **A mixture-of-experts (MoE) transformer routes each**

A mixture-of-experts (MoE) transformer routes each token through a small subset of "experts" - independent feed-forward networks - chosen by a gating function.

The 1M context window, and what it actually changes

01 **V4-Pro ships with a million-token context**
V4-Pro ships with a million-token context window.

02 **The frontier labs have caught up**
The frontier labs have caught up here too, with Claude Opus 4.

03 **7 and Gemini in the same**
7 and Gemini in the same band, and GPT-5.

Why China is competing on efficiency, not on GPU scale

01 **The strategic posture is the part**

The strategic posture is the part that most US-centric coverage gets wrong.

02 **The Chinese AI labs do not**

The Chinese AI labs do not have access to the latest US GPU stack at scale.

03 **Export controls have tightened across H100,**

Export controls have tightened across H100, B200, and the Blackwell generation.

Why inference economics matter more than benchmarks

01 **A benchmark answers one question: how**

A benchmark answers one question: how does the model perform on this specific harness, with this specific prompt template, on this specific date.

02 **A unit-economics answer addresses a different**

A unit-economics answer addresses a different question: at the price you can serve this model, what set of products become defensible.

03 **The first question is interesting for**

The first question is interesting for researchers and for vendors writing pitch decks.

NEXT STEP

Read the full architecture note

DeepSeek V4 did not change the model layer. It changed what a margin on the model layer costs. The interesting part of the AI stack is now everywhere else.

arthea.ai/book ->