NEWS — DEEPSEEK — ARTIFICIAL INTELLIGENCE — HUANG

DeepSeek floats out update of LLM amidst AI datacentre bubble fears

Chinese giants reining back on GPUs ahead of next reasoning model release

JOE FAY

March 25, 2025 . 3:51 PM — 2 min read

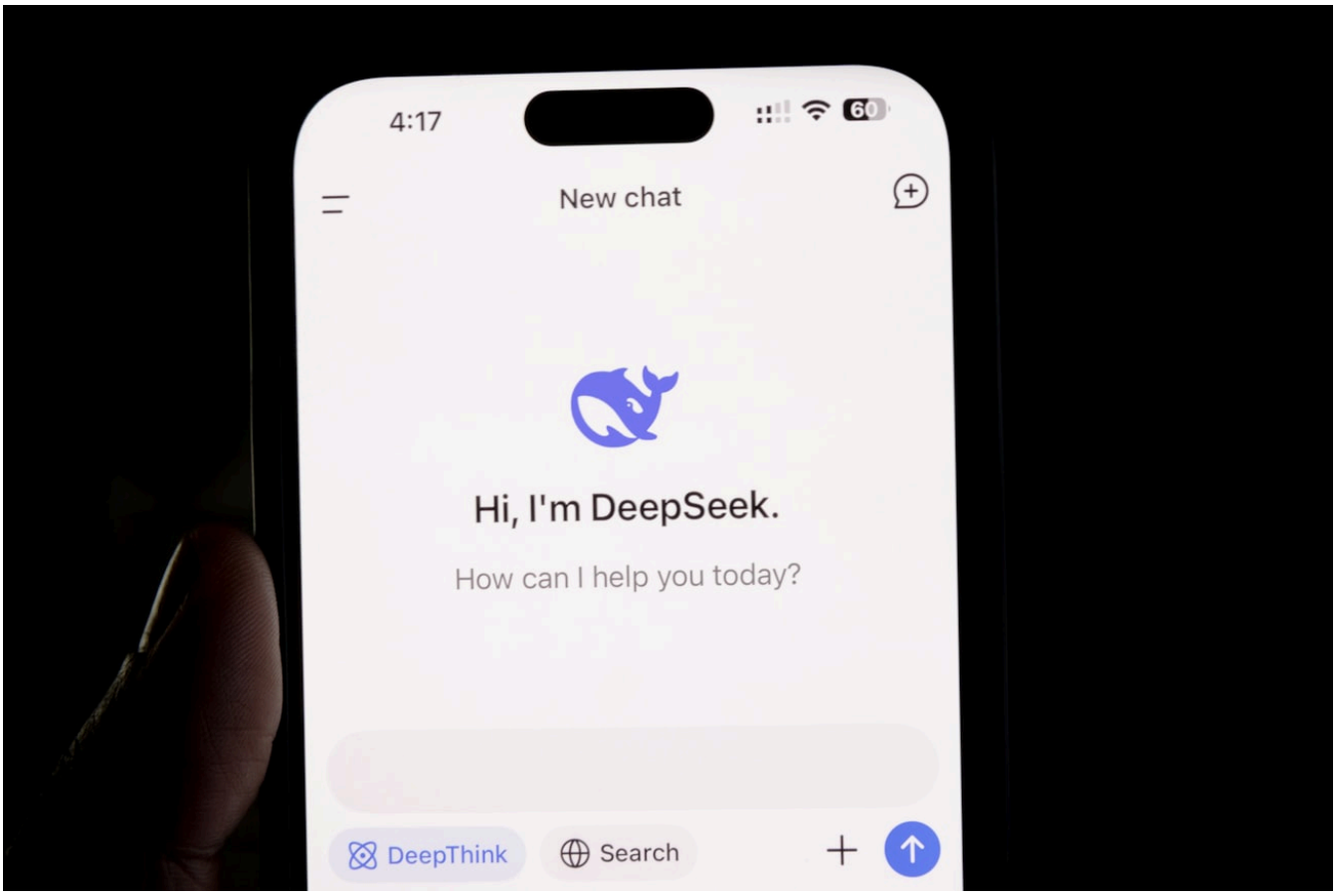

Photo by Solen Feyissa / Unsplash

DeepSeek has pushed out an upgrade to its V3 LLM which c
substantial performance improvements over its predecessor and coincides

with rumbles from China about the prospect of an AI datacentre bubble.

The 685bn parameter model – suffixed 0324 - pitched up on Hugging Face overnight and claimed some "notable improvements" over its predecessor, which appeared in December.

These include a quartet of benchmark claims, including 81.2, up from 75.9, on the MMLU-Pro benchmark used to gauge performance on knowledge-based tasks. GPT 4.5 remains the leader in this category, at 86.1.

The reasoning benchmark GPQA come in at 68.4, up from 59.1 – ChatGPT remains in the front.

Its math reasoning benchmark, AIME, jumps from 39.6 to 59.4, ahead of nearest rival GPT 4.5 on 36.7. But perhaps developers will be more interested in the LiveCodeBench score of 49.2, up from 39.2. By comparison, nearest competitors for the code weaving benchmark, GPT 4.5 and Claude-Sonnet 3.7 come in at 44.4 and 42.2, respectively.

As well as more aesthetically pleasing front ends, the makers also claim better "Chinese writing proficiency" as well as enhanced report analysis requests with more detailed outputs for Chinese Seach capabilities.

V3's stablemate, Deepseek R1, caused what seemed to be a temporary existential crisis for the AI sector, when its parent company claimed it had been trained on a fraction of that of rival, mainly US models.

Details haven't emerged on the infrastructure behind 0324, but its predecessor was also reputedly trained on a far more economic rig.

OpenUK CEO Amanda Brock said the lightweight nature of the model and its architecture, meant "The potential to move an LLMs' functioning away

from the data centre and allows its running on a local device. This shift is critical. Much of the early legislation and policy approach focused on parameter size and this is increasingly the wrong focus."

She added, "If the compute capacity reduces and the function is localised to a device not the data centre the apparent dearth of data centres may be a panic that passes quickly." She also noted the release of R2 is now rumoured to be set for April. "Logically its advances are likely to follow the V3-0324 trend. This will likely be a significant challenge to closed AI like OpenAI's ChatGPT-5."

Coincidentally, Alibaba chairman Joe Tsai told investors in Hong Kong he was "astounded" by the amount of investment being poured into AI infrastructure in the US, Reuters reported. He said he was seeing "the beginning of some kind of bubble" in datacentre capacity.

His comments came just days after Tencent CSO James Mitchell said it was slowing down its deployment of GPUs following its implementation of... DeepSeek.

That will all be news to Jensen Huang who opened GTC last week with claims that we are in an era of reasoning, agentic AI, that will require ever more, ever more powerful GPUs.

See also: DeepSeek R1 is now on AWS as hyperscalers move fast to offer Chinese lab's models

Related