DeepSeek rushes to launch new AI model as China goes all in

Reuters — DeepSeek is looking to press home its advantage.

The Chinese startup triggered a $1 trillion-plus sell-off in global equities markets last month with a cut-price AI reasoning model that outperformed many Western competitors.

Now, the Hangzhou-based firm is accelerating the launch of the successor to January's R1 model, according to three people familiar with the company.

Deepseek had planned to release R2 in early May but now wants it out as early as possible, two of them said, without providing specifics.

The company says it hopes the new model will produce better coding and be able to reason in languages beyond English. Details of the accelerated timeline for R2's release have not been previously reported.

DeepSeek did not respond to a request for comment for this story.

Rivals are still digesting the implications of R1, which was built with less-powerful Nvidia chips but is competitive with those developed at the costs of hundreds of billions of dollars by U.S. tech giants.

"The launch of DeepSeek's R2 model could be a pivotal moment in the AI industry," said Vijayasimha Alilughatta, chief operating officer of Indian tech services provider Zensar. DeepSeek's success at creating cost-effective AI models "would likely spur companies worldwide to accelerate their own efforts … breaking the stranglehold of the few dominant players in the field," he said.

R2 is likely to worry the U.S. government, which has identified leadership of AI as a national priority. Its release may further galvanize Chinese authorities and companies, dozens of which say they have started integrating DeepSeek models into their products.

Little is known about DeepSeek, whose founder Liang Wenfeng became a billionaire through his quantitative hedge fund High-Flyer. Liang, who was described by a former employer as "low-key and introverted," has not spoken to any media since July 2024.

Reuters interviewed a dozen former employees, as well as quant fund professionals knowledgeable about the operations of DeepSeek and its parent company High-Flyer. It also reviewed state media articles, social-media posts from the companies and research papers dating back to 2019.

They told a story of a company that functioned more like a research lab than a for-profit enterprise and was unencumbered by the hierarchical traditions of China's high-pressure tech industry, even as it became responsible for what many investors see as the latest breakthrough in AI.

DIFFERENT PATH

Liang was born in 1985 in a rural village in the southern province of Guangdong. He later obtained communication engineering degrees at the elite Zhejiang University.

One of his first jobs was running a research department at a smart imaging firm in Shanghai. His then-boss, Zhou Chaoen, told state media on Feb. 9 that Liang had hired prize-winning algorithm engineers and operated with a "flat management style."

At DeepSeek and High-Flyer, Liang has similarly shunned the practices of Chinese tech giants known for rigid top-down management, low pay for young employees and “996” — working from 9 a.m. to 9 p.m. six days a week.

Liang opened his Beijing office within walking distance of Tsinghua University and Peking University, China’s two most prestigious education institutions. He regularly delved into technical details and was happy to work alongside Gen-Z interns and recent graduates that comprised the bulk of its workforce, according to two former employees. They also described usually working eight-hour days in a collaborative atmosphere.

“Liang gave us control and treated us as experts. He constantly asked questions and learned alongside us,” said 26-year-old researcher Benjamin Liu, who left the company in September. “DeepSeek allowed me to take ownership of critical parts of the pipeline, which was very exciting.”

Liang did not respond to questions sent via DeepSeek.

While Baidu and other Chinese tech giants were racing to build their consumer-facing versions of ChatGPT in 2023 and profit off of the global AI boom, Liang told Chinese media outlet Waves last year that he deliberately avoided spending heavily on app development, focusing instead on refining the AI model’s quality.

Both DeepSeek and High-Flyer are known for paying generously, according to three people familiar with its compensation practices. At High-Flyer, it is not uncommon for a senior data scientist to make 1.5 million yuan annually, while competitors rarely pay more than 800,000, said one of the people, a rival quant fund manager who knows Liang.

The largesse was funded by High-Flyer, which became one of China’s most successful quant funds and, even after a

government crackdown on the sector, still manages tens of billions of yuan, according to two people in the industry.

COMPUTING POWER

DeepSeek's success with a low-cost AI model is based on High-Flyer's decade-long and substantial investment in research and computing power, three people said.

The quant fund was an earlier pioneer in AI trading and a top executive said in 2020 that High-Flyer was going "all in" on AI by re-investing 70% of its revenue, mostly into AI research.

High-Flyer spent 1.2 billion yuan on two supercomputing AI clusters in 2020 and 2021. The second cluster, Fire-Flyer II, was made up of around 10,000 Nvidia A100 chips, used for training AI models.

DeepSeek had not been established at that time, so the accumulation of computing power caught the attention of Chinese securities regulators, said a person with direct knowledge of officials' thinking.

"Regulators wanted to know why they need so many chips?" the person said. "How they were going to use it? What kind of impact would that have on the market?"

Authorities decided not to intervene, in a move that would prove crucial for DeepSeek's fortunes: the U.S. banned the export of A100 chips to China in 2022, at which point Fire-Flyer II was already in operation.

Beijing now celebrates DeepSeek, but has instructed it not to engage with the media without approval, according to a person familiar with Chinese official thinking.

Authorities had asked Liang to keep a low-profile because they were worried that too much hype in the media would draw

unnecessary attention, the person said.

China's cabinet and commerce ministry, as well as China's securities regulator, did not respond to requests for comment.

As one of the few companies with a large A100 cluster, High-Flyer and DeepSeek were able to attract some of China's best research talent, two former employees said.

"The key advantage of vast (computing) resources is that it allows for large-scale experimentation," said Liu, the former employee.

Some Western AI entrepreneurs, like Scale AI CEO Alexandr Wang, have claimed that DeepSeek had as many as 50,000 higher-end Nvidia chips that are banned for export to China. He has not produced evidence for the allegation or responded to Reuters' requests to provide proof.

DeepSeek has not responded to Wang's claims. Two former employees attributed the company's success to Liang's focus on more cost-effective AI architecture.

The startup used techniques like Mixture-of-Experts (MoE) and multihead latent attention (MLA), which incur far lower computing costs, its research papers show.

The MoE technique divides an AI model into different areas of expertise and activates only those related to a query, as opposed to more common architectures that use the entire model.

MLA architecture allows a model to process different aspects of one piece of information simultaneously, helping it detect key details more effectively.

While competitors like France's Mistral have developed models based on MoE, DeepSeek was the first firm to depend heavily on this architecture while achieving parity with more expensively built models.

DeepSeek’s pricing was 20 to 40 times cheaper than what OpenAI charged for equivalent models, analysts at Bernstein brokerage estimated in early February.

For now, Western and Chinese tech giants have signaled plans to continue heavy AI spending, but DeepSeek’s success with R1 and its earlier V3 model has prompted some to alter strategies.

OpenAI cut prices this month, while Google’s Gemini has introduced discounted tiers of access. Since R1’s launch, OpenAI has also released an O3-Mini model that relies on less computing power.

Adnan Masood of U.S. tech services provider UST told Reuters that his laboratory had run benchmarks that found R1 often used three times as many tokens, or units of data processed by the AI model, for reasoning as OpenAI’s scaled-down model.

STATE EMBRACE

Even before R1 gripped global attention, there were signs that DeepSeek had caught Beijing’s favor. In January, state media reported that Liang attended a meeting with Chinese Premier Li Qiang in Beijing as the designated representative of the AI sector, ahead of the leaders of better-known firms.

The subsequent fanfare over the cost competitiveness of its models has buoyed Beijing’s belief that it can out-innovate the U.S., with Chinese companies and government bodies embracing DeepSeek models at a pace that has not been offered to other firms.

At least 13 Chinese city governments and 10 state-owned energy companies say they have deployed DeepSeek into their systems, while tech giants Lenovo (0992.HK)

, opens new tab, Baidu (9888.HK), opens new tab and Tencent (0700.HK)

, opens new tab – owner of China's largest social media app WeChat – have integrated DeepSeek's models into their products.

Chinese leader Xi Jinping and Li "have signalled they endorse DeepSeek," said Alfred Wu, an expert on Chinese policymaking at Singapore's Lee Kuan Yew School of Public Policy. "Now everyone just endorses it."

The Chinese embrace comes as governments from South Korea to Italy remove DeepSeek from national app stores, citing privacy concerns.

"If DeepSeek becomes the go-to AI model across Chinese state entities, Western regulators might see this as another reason to escalate restrictions on AI chips or software collaborations," said Stephen Wu, an AI expert and founder of hedge fund Carthage Capital.

Further limits on advanced AI chips are a challenge that Liang has acknowledged.

"Our problem has never been funding," he told Waves in July. "It's the embargo on high-end chips."

Reports- Reuters