Congressional Research Service
Informing the legislative debate since 1914

IN FOCUS

July 1, 2025

DeepSeek and the Race to Develop Artificial Intelligence

DeepSeek is an artificial intelligence (AI) start-up founded in 2023 and based in Hangzhou, China. DeepSeek's CEO, Liang Wenfeng, also created High-Flyer, a hedge fund that is reportedly the sole funder of DeepSeek. Some have asserted that this funding "has allowed DeepSeek to pursue ambitious AI projects without the pressure of external investors"; others have claimed that the company is "state-subsidized" and "state-controlled." The company has been releasing a series of open-source large language models (LLMs)—AI systems that aim to model language, sometimes with millions or billions of parameters (i.e., numbers in the model that determine how inputs are converted to outputs). DeepSeek's latest AI models, DeepSeek-V3 (V3) and DeepSeek-R1 (R1), and the company's free chatbot app released in January 2025, have sparked a wave of interest and concern from some U.S. and international stakeholders, including the U.S. technology industry and some Members of Congress. This In Focus describes DeepSeek's purported advances in AI model development and sets out related analysis and issues for Congress on broader implications of the race to develop AI.

DeepSeek's AI Models and Reported Advances

The V3 and R1 models from DeepSeek are broadly similar in structure and function to other LLMs from U.S.-based companies. In the AI context, a *benchmark* has been described as "a particular combination of dataset[s] ... and a metric ... representing one or more specific tasks or sets of abilities" used by researchers as "a shared framework for the comparison of methods." V3 is a 671-billion-parameter general-purpose LLM that, according to the company's reporting, achieved scores on multiple benchmarks that were comparable to those of OpenAI's GPT-4o and Anthropic's Claude 3.5 Sonnet. (The number of parameters for GPT-4o has not been disclosed, but for comparison, GPT-4 reportedly has over one trillion parameters, 10 times the number in the GPT-3 model.) Benchmarks reported in DeepSeek's R1 evaluations included those assessing performance on general reasoning tasks, graduate-level questions and answers, and mathematics problem solving.

Using V3 as a base model, DeepSeek next developed R1, an LLM with *reasoning* capabilities (i.e., models that use a *chain-of-thought* technique "to refine their thinking process, try different strategies, and recognize their mistakes") similar to those of OpenAI's o1 model. DeepSeek further released six smaller models (i.e., fewer parameters). The smaller models were *distilled* from R1 based on other models. As one AI expert describes, model distillation "typically involves generating responses from the stronger model to train a weaker model so that the weaker model improves." DeepSeek researchers used responses generated from R1 to train versions of Alibaba's Qwen and Meta's Llama models, ranging in size from 1.5 billion to 70 billion parameters. Subsequently, other companies have released smaller, cheaper reasoning models (e.g., OpenAI's o1-mini and Baidu's Ernie X1). Among the distinguishing aspects of the DeepSeek models, compared to those from U.S. AI companies and LLMs, are DeepSeek's reported advances in the design and training of their AI models, their purported lower costs to train and run the models, and their decision to release the models as open source.

Model Architecture and Learning Methods

DeepSeek reports using a number of innovative and optimized approaches in model architectures and learning methods. For example, the *mixture-of-experts (MoE)* architecture uses a subset (i.e., "expert") of its parameters, rather than all parameters (as with "dense" models), for each input. Each expert is smaller and more specialized, so less memory is needed to train the model, and it is less expensive to run once deployed. DeepSeek also reportedly used a parallelism algorithm, called *DualPipe*, allowing for concurrent rather than sequential computation and communication steps, increasing the efficiency of the model training. A mixed-precision training technique called FP8 (using 8 rather than 16 or 32 bits to represent numbers in model training) reportedly reduced memory usage and computational load, allowing for faster data processing without loss of accuracy. DeepSeek reports using an automated reinforcement learning (RL) loop (*group-relative policy optimization*) to achieve efficiency, stability, and scalability benefits over other RL algorithms used in LLM training (e.g., many companies use RL along with human testers providing feedback to improve model performance, which is more time and labor intensive).

Through these and other improvements, DeepSeek focused on efficient and cost-effective model design and training, given their reported hardware constraints. Some of these innovations also have been demonstrated by U.S. companies. For example, in 2022, Microsoft, Google, and Meta reported cost reductions and performance improvements using MoE, compared with dense models; GPT-4 has been rumored to be an MoE model.

Hardware and Training Efficiency

A V3 technical paper reported training costs of less than $5.6 million using a cluster of 2,048 H800 chips from Nvidia. (In comparison, OpenAI's GPT-4 training costs have been estimated at around $79 million.) The technical paper cites "engineering optimizations"—leaving out precise explanations—for how the model achieves the reported performance with such low training costs. One AI expert noted "it's likely that DeepSeek's steady refinement of MoE is a key factor." Some analysts have questioned the accuracy of comparing this cost with that of other AI

models, asserting that $5.6 million is attributable to just the GPU (i.e., chip) rental cost, excluding such costs as acquiring chips, which would be tens of millions of dollars.

Open Source

DeepSeek considers its V3 and R1 models to be open source and has made the models' weights, inference code, and technical documentation publicly available. However, the company did not release its training codes or complete training datasets. DeepSeek's approach to classifying their models as open source is similar to that of other companies, such as Meta (Llama models) and Alibaba (Qwen models). However, there is debate in the academic community over a definition of *open source.* R1 also has sparked efforts by some researchers to create a fully open-source reproduction.

Selected Issues for Congress

Transparency and Ethical Development and Use

Concerns about a lack of transparency on how DeepSeek's models were trained and fine-tuned have led to questions around lawful and ethical development and use. For example, OpenAI claimed that DeepSeek used outputs of ChatGPT to train its AI models, which is forbidden by OpenAI's terms of use. Other experts assert that such a claim is difficult to prove. Critiques of Chinese-developed models have reported that they are subject to testing for "core socialist values" by China's internet regulator and restrict some output, such as about the 1989 Tiananmen Square crackdown or the political status of Taiwan, raising concerns about a lack of ethical development and a lack of information supporting democratic values.

Development Costs and Investments

DeepSeek claims to have developed a high-performing AI model with less powerful chips and less money than comparable models from U.S. companies. Subsequently, some stakeholders have raised questions about the necessity of large investments in AI infrastructure (e.g., data centers with advanced chips). On the other hand, the continued growth in AI research and development (R&D) has prompted multibillion dollar investment commitments for AI infrastructure, such as up to $500 billion for the U.S. Stargate joint venture and up to $207 billion (€200 billion) in the European Union. For more on AI infrastructure, see CRS In Focus IF12899, *Data Centers and Cloud Computing: Information Technology Infrastructure for Artificial Intelligence*.

Export Controls

DeepSeek claims to have used Nvidia's H800 chips, which are slower than advanced chips such as Nvidia's H100. In November 2022, the Department of Commerce's Bureau of Industry and Security imposed controls on the export of certain advanced computing components to China, including Nvidia's H100 chips. Nvidia then released the slower H800 chips, which DeepSeek claims to have used in the V3 and R1 models. Later, in October 2023, the H800 chips were added to the U.S. export ban. This has sparked a new round of debates about the efficacy of export controls for slowing AI development in other countries and whether such restrictions have instead spurred more cost-effective approaches for AI training. The efficacy of export control enforcement also has been questioned, with reports of White House and Federal Bureau of Investigation probes into whether DeepSeek bought advanced, restricted Nvidia chips through third parties in Singapore. Some reporting has speculated that DeepSeek may be considering using data centers in Southeast Asia to gain remote access to Nvidia AI chips. For more on export controls, see CRS Report R47684, *Export Controls—International Coordination: Issues for Congress*.

Security

At the end of January 2025, DeepSeek's R1-powered chatbot took the top spot as the most downloaded free app in the United States on Apple's App Store, a position previously held by OpenAI. Soon after, DeepSeek announced that it was halting new user registrations because of "large-scale malicious attacks" on its servers, which are mostly located in China and store user data, according to its privacy policy. Researchers also have reported that the R1 model failed to block harmful behavior prompts in security tests, asserting that "R1 lacks robust guardrails, making it highly susceptible to algorithmic jailbreaking and potential misuse" (e.g., overriding model constraints meant to prevent the production of harmful content, such as information on making bioweapons). Misuse of AI models is not limited to R1, however; for example, hacker groups have been reportedly using Google's Gemini in malicious attacks in the United States.

The federal government has taken varied actions on the DeepSeek models and app. Actions include a National Security Council evaluation of potential security implications, an investigative report and hearing by a House select committee, prohibitions on use by the House Chief Administrative Officer, and bans by agencies on use on government devices. Some Members of the 119th Congress have introduced legislation to ban the use and download of DeepSeek's models on government devices (H.R. 1121/S. 765). Regulators in other countries have been evaluating potential privacy risks (e.g., the United Kingdom, France, and Ireland) or banning the app from Apple and Google app stores (e.g., Italy) or government devices (e.g., Australia).

The "AI Race"

Advances in the R&D and deployment of AI technologies have led to growing interest in which country is "winning" a purported "AI race." While the United States has historically dominated AI research and model development, recent analysis has suggested that China-based models are catching up. The release of DeepSeek's models has highlighted concerns regarding the safety and security of AI technologies. Beyond efforts to control access to resources for AI development in countries of concern, the United States might consider ways to collaborate with allies to support the development of AI systems that support U.S. national security priorities. As Congress considers AI legislation and oversight activities, central to the debate may be ethical and transparent AI development; optimal levels of public and private sector investments; and mechanisms to support U.S. AI innovation, secure deployment, and technological competitiveness.

Laurie Harris, Analyst in Science and Technology Policy

Disclaimer