WIKIPEDIA
The Free Encyclopedia

DeepSeek

Hangzhou DeepSeek Artificial Intelligence Basic Technology Research Co., Ltd.,[2][3][a] doing business as **DeepSeek,**[b] is a Chinese artificial intelligence company that develops open-source large language models (LLMs). Based in Hangzhou, Zhejiang, it is owned and funded by the Chinese hedge fund High-Flyer. DeepSeek was founded in July 2023 by High-Flyer co-founder Liang Wenfeng, who also serves as the CEO for both companies.

Hangzhou DeepSeek Artificial Intelligence Basic Technology Research Co., Ltd.

Native name	杭州深度求索人工智能基础技术研究有限公司
Company type	Privately held company
Industry	Information technology Artificial intelligence
Founded	17 July 2023[1]
Founder	Liang Wenfeng
Headquarters	Hangzhou, Zhejiang, China
Key people	Liang Wenfeng (CEO)
Owner	High-Flyer
Number of employees	Under 200
Website	deepseek.com (https://deepseek.com)

The DeepSeek-R1 model provides responses comparable to other contemporary large language models, such as OpenAI's GPT-4o and o1.[5] Its training cost is reported to be significantly lower than other LLMs. The company claims that it trained R1 for US$6 million compared to $100 million for OpenAI's GPT-4 in 2023,[6] and approximately one tenth of the computing power used for Meta's comparable model, LLaMA 3.1.[6][7][8][9] DeepSeek's success against larger and more established rivals has been described as "upending AI".[10][11]

DeepSeek's models are "open weight", which provides less freedom for modification than true open source software.[12][13] The company reportedly recruits AI researchers from top Chinese universities[10] and hires from outside the computer science field to diversify its models' knowledge and abilities.[7]

The low cost of training and running the language model was attributed to Chinese firms' lack of access to Nvidia chipsets, which were restricted by the US as part of the ongoing trade war between the two countries. This breakthrough in reducing expenses while increasing efficiency and maintaining the model's performance in the AI industry sent "shockwaves" through the market. It threatened the dominance of AI leaders like Nvidia and contributed to the largest drop in US stock market history, with Nvidia alone losing $600 billion in market value.[14][15]

DeepSeek's AI models were developed amid United States sanctions on China and other countries restricting access to chips used to train LLMs. These were intended to restrict the ability of these countries to develop advanced AI systems.[16][17] Lesser restrictions were later announced that would affect all but a few countries.[18]

Reports indicate that it applies content moderation in accordance with local regulations, limiting responses on topics such as the Tiananmen Square massacre and Taiwan's political status.[19][20] DeepSeek models that have been uncensored also display bias towards Chinese government viewpoints on controversial topics such as Xi Jinping's human rights record and Taiwan's political status.[21][22] However, users who have downloaded the models and hosted them on their own devices and servers have reported successfully removing this censorship.[23][24]

History

Founding and early years (2016–2023)

In February 2016, High-Flyer was co-founded by AI enthusiast Liang Wenfeng, who had been trading since the 2007–2008 financial crisis while attending Zhejiang University.[25]

The company began stock-trading using a GPU-dependent deep learning model on October 21, 2016. Prior to this, they used CPU-based models, mainly linear models. Most trading was driven by AI by the end of 2017.[26]

In 2019, Liang established High-Flyer as a hedge fund focused on developing and using AI trading algorithms. By 2021, High-Flyer exclusively used AI in trading,[27] often using Nvidia chips.[28]

Initial computing cluster Fire-Flyer began construction in 2019 and finished in 2020, at a cost of 200 million yuan. It contained 1,100 GPUs interconnected at a rate of 200 Gbps. It was 'retired' after 1.5 years in operation.

In 2021, Liang began stockpiling Nvidia GPUs for an AI project.[28] According to 36Kr, Liang acquired 10,000 Nvidia A100 GPUs[29] before the United States restricted chip sales to China.[27] Computing cluster Fire-Flyer 2 began construction in 2021 with a budget of 1 billion yuan.[26]

It was reported that in 2022, Fire-Flyer 2's capacity had been used at over 96%, totaling 56.74 million GPU hours. 27% was used to support scientific computing outside the company.[26]

During 2022, Fire-Flyer 2 had 5000 PCIe A100 GPUs in 625 nodes, each containing 8 GPUs. At the time, they exclusively used PCIe instead of the DGX version of A100, since at the time the models they trained could fit within a single 40 GB GPU VRAM, so there was no need for the higher bandwidth of DGX (i.e. they required only data parallelism but not model parallelism).[30] Later, they incorporated NVLinks and NCCL, to train larger models that required model parallelism.[31][32]

On 14 April 2023,[33] High-Flyer announced the start of an artificial general intelligence lab dedicated to research developing AI tools separate from High-Flyer's financial business.[34][35] Incorporated on 17 July 2023,[1] with High-Flyer as the investor and backer, the lab became its own company, DeepSeek.[27][36][35] Venture capital firms were reluctant to provide funding, as they considered it unlikely that the venture would be able to quickly generate an "exit".[27]

On 16 May 2023, the company Beijing DeepSeek Artificial Intelligence Basic Technology Research Company, Limited. was incorporated. It was later taken under 100% control of Hangzhou DeepSeek Artificial Intelligence Basic Technology Research Co., Ltd, which was incorporated 2 months after.

Model releases (2023–present)

On 2 November 2023, DeepSeek released its first model, DeepSeek Coder.

On 29 November 2023, DeepSeek released the DeepSeek-LLM series of models.[37]:section 5

On 9 January 2024, they released 2 DeepSeek-MoE models (Base and Chat).[38]

In April 2024, they released 3 DeepSeek-Math models: Base, Instruct, and RL.[39]

DeepSeek-V2 was released in May 2024.

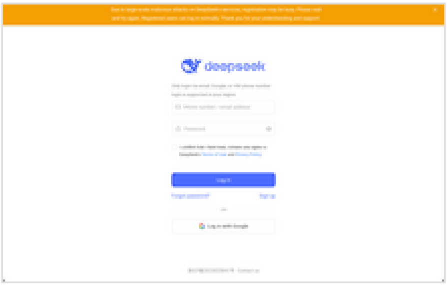
In June 2024, the DeepSeek-Coder V2 series was released.[40]

DeepSeek V2.5 was released in September and updated in December 2024.[41]

On 20 November 2024, DeepSeek-R1-Lite-Preview became accessible via API and chat.[42][43]

In December 2024, they released a base model DeepSeek-V3-Base and a chat model DeepSeek-V3.[31]

On 20 January 2025, DeepSeek released its first free DeepSeek chatbot, based on the DeepSeek-R1 model, for iOS and Android; by 27 January, DeepSeek had surpassed ChatGPT as the most downloaded free app on the iOS App Store in the United States,[10] causing Nvidia's share price to drop by 18%.[44][45]



The DeepSeek login page shortly after a cyberattack that occurred following its January 20 launch

Company operation

Based in Hangzhou, Zhejiang, DeepSeek is owned and funded by the Chinese hedge fund High-Flyer.

DeepSeek was founded in July 2023 by High-Flyer co-founder Liang Wenfeng, who also serves as its CEO.

As of May 2024, Liang owned 84% of DeepSeek through two shell corporations.[note 1][46]

Strategy

DeepSeek is focused on research and has not detailed plans for commercialization.[47]

This allows its technology to avoid the most stringent provisions of China's AI regulations, such as requiring consumer-facing technology to comply with government controls on information.[7]

DeepSeek's hiring preferences target technical abilities rather than work experience; most new hires are either recent university graduates or developers whose AI careers are less established.[35][7]

Likewise, the company recruits individuals without any computer science background to help its technology understand more knowledge areas,[10] such as poetry and China's notoriously difficult college admissions exams (Gaokao).[7]

Training framework

High-Flyer/DeepSeek operates at least two computing clusters, Fire-Flyer (萤火一号) and Fire-Flyer 2 (萤火二号). Fire-Flyer 2 consists of co-designed software and hardware architecture. On the hardware side, Nvidia GPUs use 200 Gbps interconnects. The cluster is divided into two "zones", and the platform supports cross-zone tasks. The network topology was two fat trees, chosen for high bisection bandwidth. On the software side are:[32][26]

- `3FS` (Fire-Flyer File System): A distributed parallel file system, specifically designed for asynchronous random reads. It uses Direct I/O and RDMA Read. In contrast to standard Buffered I/O, Direct I/O does not cache data. Caching is useless for this case, since each data read is random, and isn't reused.[48]
- `hfreduce`: Library for asynchronous communication, originally designed to replace Nvidia Collective Communication Library (NCCL).[30] It is mainly used for allreduce, especially of gradients during backpropagation. It is asynchronously run on the CPU to avoid blocking kernels on the GPU.[32] It uses two-tree broadcast like NCCL.[30]
- `hfai.nn`: Software library of commonly used operators for neural network training, similar to `torch.nn` in PyTorch.
- `HaiScale Distributed Data Parallel` (DDP): Parallel training library that implements various forms of parallelism such as Data Parallelism (DP), Pipeline Parallelism (PP), Tensor Parallelism (TP), Experts Parallelism (EP), Fully Sharded Data Parallel (FSDP) and Zero Redundancy Optimizer (ZeRO). It is similar to PyTorch DDP, which uses NCCL on the backend.
- `HAI Platform`: Various applications such as task scheduling, fault handling, and disaster recovery.[49]

As of 2022, Fire-Flyer 2 had 5000 PCIe A100 GPUs in 625 nodes, each containing 8 GPUs.[30] They later incorporated NVLinks and NCCL, to train larger models that required model parallelism.[31][32]

Development and release history

Major versions of DeepSeek models. SFT stands for supervised finetuning.

Major versions	Release date	Major variants	Remarks
DeepSeek Coder	2 Nov 2023	Base (pretrained); Instruct (with instruction-finetuned)	The architecture is essentially the same as Llama.
DeepSeek-LLM	29 Nov 2023	Base; Chat (with SFT)	
DeepSeek-MoE	9 Jan 2024	Base; Chat	Developed a variant of mixture of experts (MoE).
DeepSeek-Math	Apr 2024	Base	Initialized with DS-Coder-Base-v1.5
		Instruct (with SFT)	
		RL (using a process reward model)	Developed Group Relative Policy Optimization (GRPO), a variant of Proximal Policy Optimization (PPO).
DeepSeek V2	May 2024	DeepSeek-V2, DeepSeek-V2-Chat DeepSeek-V2-Lite, DeepSeek-V2-Lite-Chat DeepSeek-Coder-V2 DeepSeek-V2.5	Developed multi-head latent attention (MLA). Also used mixture of experts (MoE).
DeepSeek V3	Dec 2024	DeepSeek-V3-Base DeepSeek-V3 (a chat model)	The architecture is essentially the same as V2.
DeepSeek R1	20 Nov 2024	DeepSeek-R1-Lite-Preview	Only accessed through API and a chat interface.
	20 Jan 2025	DeepSeek-R1 DeepSeek-R1-Zero	Initialized from DeepSeek-V3-Base and sharing the V3 architecture.
		Distilled models	Initialized from other models, such as Llama, Qwen, etc. Distilled from data synthesized by R1 and R1-Zero.[50]

The first DeepSeek models were essentially the same as Llama,[37] which were dense decoder-only Transformers. Later models incorporated Mixture of Experts, and then multi-head latent attention.[38][40]

A decoder-only Transformer consists of multiple identical decoder layers. Each of these layers features two main components: an attention layer and a FeedForward network (FFN) layer.[40] In the attention layer, the traditional multi-head attention mechanism has been enhanced with multi-head latent attention.

This update introduces compressed latent vectors to boost performance and reduce memory usage during inference.

Meanwhile, the FFN layer adopts a variant of the mixture of experts (MoE) approach, effectively doubling the number of experts compared to standard implementations. It distinguishes between two types of experts: shared experts, which are always active to encapsulate general knowledge, and routed experts, where only a select few are activated to capture specialized information.

Overview of models

DeepSeek's models are "open weight", which provides less freedom for modification than true open source software.[51][52]

DeepSeek Coder

DeepSeek Coder is a series of 8 models, 4 pretrained (`Base`) and 4 instruction-finetuned (`Instruct`). They all have 16K context lengths. The code for the model was made open-source under the MIT License, with an additional license agreement ("DeepSeek license") regarding "open and responsible downstream usage" for the model.[53]

The training program was:[54][55][56]

1. Pretraining: 1.8T tokens (87% source code, 10% code-related English (GitHub markdown and Stack Exchange), and 3% code-unrelated Chinese).
2. Long-context pretraining: 200B tokens. This extends the context length from 4K to 16K. This produced the `Base` models.
3. Supervised finetuning (SFT): 2B tokens of instruction data. This produced the `Instruct` models.

They were trained on clusters of A100 and H800 Nvidia GPUs, connected by InfiniBand, NVLink, NVSwitch.[54]

DeepSeek Coder properties[54]:Table 2[57]

Params.	n_{layers}	d_{model}	$d_{\text{intermediate}}$	n_{heads}	$n_{\text{kv-heads}}$
1.3B	24	2048	5504	16	16
5.7B	32	4096	11008	32	1[note 2]
6.7B	32	4096	11008	32	32
33B	62	7168	19200	56	7[note 2]

DeepSeek-LLM

The DeepSeek-LLM series was released in November 2023. It has 7B and 67B parameters in both Base and Chat forms. The accompanying paper claimed benchmark results higher than most open source LLMs at the time, especially Llama 2.[37]:section 5 The model code was under MIT license, with DeepSeek license for the model itself.[58]

The architecture was essentially the same as the Llama series. They used the pre-norm decoder-only Transformer with RMSNorm as the normalization, SwiGLU in the feedforward layers, rotary positional embedding (RoPE), and grouped-query attention (GQA). Both had vocabulary size 102,400 (byte-level BPE) and context length of 4096. They trained on 2 trillion tokens of English and Chinese text obtained by deduplicating the Common Crawl.[37]

DeepSeek LLM properties[37]:Table 2

Params.	n_{layers}	d_{model}	$d_{\text{intermediate}}$	n_{heads}	$n_{\text{kv-heads}}$
7B	30	4096	11008	32	32
67B	95	8192	22016	64	8[note 2]

The Chat versions of the two Base models was released concurrently, obtained by training Base by supervised finetuning (SFT) followed by direct policy optimization (DPO).[37]

MoE

DeepSeek-MoE models (Base and Chat), each have 16B parameters (2.7B activated per token, 4K context length). The training was essentially the same as DeepSeek-LLM 7B, and was trained on a part of its training dataset. They claimed performance comparable to a 16B MoE as a 7B non-MoE. It is a variant of the standard sparsely-gated MoE, with "shared experts" that are always queried, and "routed experts" that might not be. They found this to help with expert balancing. In standard MoE, some experts can become overused, while others are rarely used, wasting space. Attempting to balance expert usage causes experts to replicate the same capacity. They proposed the shared experts to learn core capacities that are often used, and let the routed experts learn peripheral capacities that are rarely used.[38]

Math

DeepSeek-Math includes 3 models: Base, Instruct, and RL. Math was trained as follows:[39]

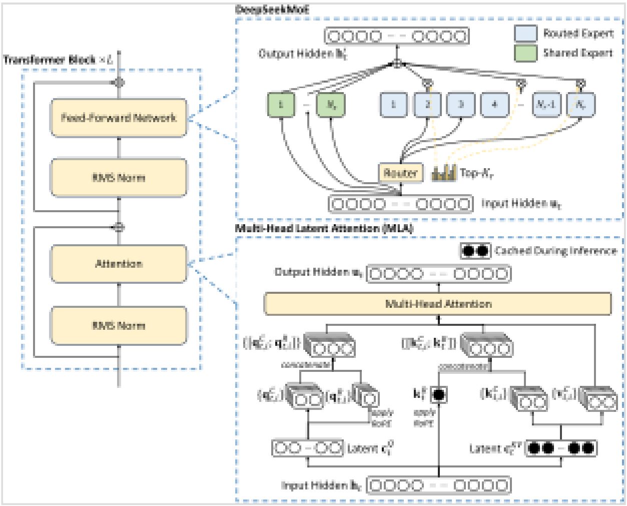
1. Initialize with a previously pretrained DeepSeek-Coder Base v1.5 7B.
2. Further pretrain with 500B tokens (6% DeepSeekMath Corpus, 4% AlgebraicStack, 10% arXiv, 20% GitHub code, 10% Common Crawl). This produced Base.
3. Train an instruction-following model by SFT Base with 776K math problems and tool-use-integrated step-by-step solutions. This produced Instruct.
4. Reinforcement learning (RL): The reward model was a process reward model (PRM) trained from Base according to the Math-Shepherd method.[59] This reward model was then used to train Instruct using Group Relative Policy Optimization (GRPO) on a dataset of 144K math questions "related to GSM8K and MATH". The reward model was continuously updated during training to avoid reward hacking. This resulted in RL.

V2

In May 2024, DeepSeek released the DeepSeek-V2 series. The series includes 4 models, 2 base models (DeepSeek-V2, DeepSeek-V2 Lite) and 2 chatbots (Chat). The two larger models were trained as follows:[60]

1. Pretrain on a dataset of 8.1T tokens, using 12% more Chinese tokens than English ones.
2. Extend context length from 4K to 128K using YaRN.[61] This resulted in DeepSeek-V2.

3. SFT with 1.2M instances for helpfulness and 0.3M for safety. This resulted in Chat SFT, which was not released.
4. RL using GRPO in two stages. The first stage was trained to solve math and coding problems. This stage used 1 reward model, trained on compiler feedback (for coding) and ground-truth labels (for math). The second stage was trained to be helpful, safe, and follow rules. This stage used 3 reward models. The helpfulness and safety reward models were trained on human preference data. The rule-based reward model was manually programmed. All trained reward models were initialized from Chat (SFT). This resulted in the released version of Chat.



The architecture of V2, showing both shared-routed MoE and MLA[60]:Figure 2

They opted for 2-staged RL, because they found that RL on reasoning data had "unique characteristics" different from RL on general data. For example, RL on reasoning could improve over more training steps.[60]

The two V2-Lite models were smaller, and trained similarly. DeepSeek-V2 Lite-Chat underwent only SFT, not RL. They trained the Lite version to help "further research and development on MLA and DeepSeekMoE".[60]

Architecturally, the V2 models were significantly different from the DeepSeek LLM series. They changed the standard attention mechanism by a low-rank approximation called multi-head latent attention (MLA), and used the previously published mixture of experts (MoE) variant.[38]

DeepSeek V2 properties[60]:Section 3.1.2, Appendix B[62][63]

Name	Params.	Active params	n_{layers}	Context length	$n_{\text{shared experts}}$	$n_{\text{routed experts}}$
V2-Lite	15.7B	2.4B	27	32K	2	64
V2	236B	21B	60	128K	2	160

The *Financial Times* reported that it was cheaper than its peers with a price of 2 RMB for every million output tokens. The University of Waterloo Tiger Lab's leaderboard ranked DeepSeek-V2 seventh on its LLM ranking.[36]

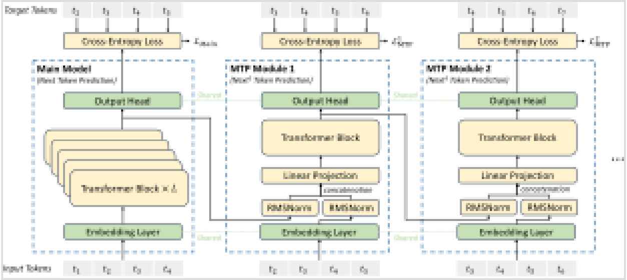
The DeepSeek-Coder V2 series included V2-Base, V2-Lite-Base, V2-Instruct, and V20-Lite-Instruct.. Training:[40][note 3]

1. Base models were initialized from corresponding intermediate checkpoints after pretraining on 4.2T tokens (not the version at the end of pretraining), then pretrained further for 6T tokens, then context-extended to 128K context length.
2. DeepSeek-Coder and DeepSeek-Math were used to generate 20K code-related and 30K math-related instruction data, then combined with an instruction dataset of 300M tokens. This was used for SFT.
3. RL with GRPO. The reward for math problems was computed by comparing with the ground-truth label. The reward for code problems was generated by a reward model trained to predict whether a program would pass the unit tests.

DeepSeek-V2.5 was made by combining DeepSeek-V2-Chat and DeepSeek-Coder-V2-Instruct.[41]

V3

DeepSeek-V3-Base and DeepSeek-V3 (a chat model) use essentially the same architecture as V2 with the addition of multi-token prediction, which (optionally) decodes extra tokens faster but less accurately. Training:[31]



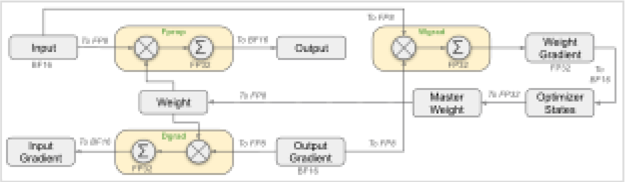
Multi-Token Prediction

1. Pretraining on 14.8T tokens of a multilingual corpus, mostly English and Chinese. It contained a higher ratio of math and programming than the pretraining dataset of V2.
2. Extend context length twice, from 4K to 32K and then to 128K, using YaRN.[61] This produced DeepSeek-V3-Base.
3. SFT for 2 epochs on 1.5M samples of reasoning (math, programming, logic) and non-reasoning (creative writing, roleplay, simple question answering) data. Reasoning data was generated by "expert models". Non-reasoning data was generated by DeepSeek-V2.5 and checked by humans.
   - The "expert models" were trained by starting with an unspecified base model, then SFT on both <problem, original response> data, and synthetic <system prompt, prompt, problem, R1 response> data generated by an internal DeepSeek-R1-Lite model. The system prompt asked R1 to reflect and verify during thinking. Then the expert models were RL using an undisclosed reward function.
   - Each expert model was trained to generate just synthetic reasoning data in one specific domain (math, programming, logic).
   - Expert models were used instead of R1 itself, since the output from R1 itself suffered "overthinking, poor formatting, and excessive length".
4. Model-based reward models were made by starting with a SFT checkpoint of V3, then finetuning on human preference data containing both final reward and chain-of-thought leading to the final reward. The reward model produced reward signals for both questions with objective but free-form answers, and questions without objective answers (such as creative writing).
5. An SFT checkpoint of V3 was trained by GRPO using both reward models and rule-based reward. The rule-based reward was computed for math problems with a final answer (put in a box), and for programming problems by unit tests. This produced DeepSeek-V3.

DeepSeek V3 properties[31]:Section 4.2[64]

Name	Params.	Active params	n_{layers}	Context length	$n_{\text{shared experts}}$	$n_{\text{routed experts}}$
V3	671B	37B	61	128K	1	256

The DeepSeek team performed extensive low-level engineering to improve efficiency. They used mixed-precision arithmetic. Much of the forward pass was performed in 8-bit floating point numbers (5E2M: 5-bit exponent and 2-bit mantissa) rather than the standard 32-bit, requiring special GEMM routines to accumulate accurately. They used a custom 12-bit float (E5M6) only for the inputs to the linear layers after the attention modules. Optimizer states were in 16-bit (BF16). They minimized communication latency by extensively overlapping computation and communication, such as dedicating 20 streaming multiprocessors out of 132 per H800 for only inter-GPU communication. They lowered communication by rearranging (every 10 minutes) the exact machine each expert was on so as to avoid querying certain machines more often than others, adding auxiliary load-balancing losses to the training loss function, and other load-balancing techniques.[31]



Mixed-precision framework for V3[31]:Figure 6

After training, it was deployed on clusters of H800 GPUs. The 8 H800 GPUs within a cluster were connected by NVLink, and the clusters were connected by InfiniBand.[31]

Total cost of training the DeepSeek-V3 model[31]:Table 1

Stage	Cost (in one thousand GPU hours)	Cost (in one million USD$)
Pre-training	2,664	5.328
Context extension	119	0.24
Fine-tuning	5	0.01
Total	2,788	5.576

The cost has been discussed[65][66][67] and called misleading, because it covers only parts of the true cost.[68]

Benchmark tests show that V3 outperformed Llama 3.1 and Qwen 2.5 while matching GPT-4o and Claude 3.5 Sonnet.[35][69][70][71]

R1

DeepSeek-R1-Lite-Preview[42][43][note 4] was trained for logical inference, mathematical reasoning, and real-time problem-solving. DeepSeek claimed that it exceeded performance of OpenAI o1 on benchmarks such as American Invitational Mathematics Examination (AIME) and MATH.[72] However, *The Wall Street Journal* reported that on 15 problems from the 2024 edition of AIME, the o1 model reached a solution faster.[73]

DeepSeek-R1 and DeepSeek-R1-Zero[74] were initialized from DeepSeek-V3-Base and share its architecture. DeepSeek-R1-Distill models were instead initialized from other pretrained open-weight models, including LLaMA and Qwen, then fine-tuned on synthetic data generated by R1.[50]

DeepSeek-R1-Zero was trained exclusively using GRPO RL without SFT. Unlike previous versions, it used no model-based reward. All reward functions were rule-based, "mainly" of two types (other types were not specified): accuracy rewards and format rewards. Accuracy reward was checking whether a boxed answer is correct (for math) or whether a code passes tests (for programming). Format reward was checking whether the model puts its thinking trace within a <think>...</think> tag.[50]

Template for `DeepSeek-R1-Zero`

A conversation between User and Assistant. The user asks a question, and the Assistant solves it. The assistant first thinks about the reasoning process in the mind and then provides the user with the answer. The reasoning process and answer are enclosed within <think> </think> and <answer> </answer> tags, respectively, i.e.,

<think> reasoning process here </think> <answer> answer here </answer>. User: <prompt>. Assistant:

– <prompt> is replaced with the specific reasoning question during training.

R1-Zero has issues with readability and mixing languages. R1 was trained to address these issues and further improve reasoning:[50]

1. SFT DeepSeek-V3-Base on "thousands" of "cold-start" data all with the standard format of `|special_token|<reasoning_process>|special_token|<summary>`, designed to improve model output readability.
2. Apply the same GRPO RL process as R1-Zero, adding a "language consistency reward" to encourage it to respond monolingually. This produced an un released internal model.
3. Synthesize 600K reasoning data from the internal model, with rejection sampling (i.e. if the generated reasoning had a wrong final answer, then it is removed). Synthesize 200K non-reasoning data (writing, factual QA, self-cognition, translation) using DeepSeek-V3.
4. SFT DeepSeek-V3-Base on the 800K synthetic data for 2 epochs.
5. Apply the same GRPO RL process as R1-Zero with rule-based reward (for reasoning tasks), but also model-based reward (for non-reasoning tasks, helpfulness, and harmlessness). This produced DeepSeek-R1.

Distilled models were trained by SFT on 800K data synthesized from DeepSeek-R1, in a similar way as step 3. They were not trained with RL.[50]

Significance

DeepSeek's success against larger and more established rivals has been described as "upending AI".[10][75]

The DeepSeek-R1 model provides responses comparable to other contemporary large language models, such as OpenAI's GPT-4o and o1.[76] Its training cost is reported to be significantly lower than other LLMs.

The company claims that it trained R1 for US$6 million compared to $100 million for OpenAI's GPT-4 in 2023,[6] and approximately one tenth of the computing power used for Meta's comparable model, LLaMA 3.1.[6][7][77][78]

Domestically

DeepSeek models offer performance for a low price, and became the catalyst for China's AI model price war.

It was dubbed the "Pinduoduo of AI", and other Chinese tech giants such as ByteDance, Tencent, Baidu, and Alibaba cut the price of their AI models. Despite its low price, it was profitable compared to its money-losing rivals.[47]

Overseas

DeepSeek's AI models were developed amid United States sanctions on China and other countries restricting access to chips used to train LLMs intended to restrict the ability of these countries to develop advanced AI systems.[16][17][18]

Controversies

Reports indicate that DeepSeek models applies content restrictions in accordance with local regulations, limiting responses on topics such as the Tiananmen Square massacre and Taiwan's political status.[79][20]

DeepSeek models that have been uncensored also display heavy bias towards Chinese government viewpoints on controversial topics such as Xi Jinping's human rights record and Taiwan's political status.[80][81]

See also

Free and open-source software portal

Business portal

China portal

- Artificial intelligence industry in China
- OpenAI

Notes

a. Chinese: 杭州深度求索人工智能基础技术研究有限公司[4]

b. Chinese: 深度求索; pinyin: *Shēndù Qiúsuǒ*

1. 宁波程信柔兆企业管理咨询合伙企业（有限合伙） and 宁波程恩企业管理咨询合伙企业（有限合伙）
2. The number of heads does not equal the number of KV heads, due to GQA.
3. Inexplicably, the model named `DeepSeek-Coder-V2 Chat` in the paper was released as `DeepSeek-Coder-V2-Instruct` in HuggingFace.
4. At that time, the `R1-Lite-Preview` required selecting "Deep Think enabled", and every user could use it only 50 times a day.

References

1. "DeepSeek突传消息！" (https://finance.sina.com.cn/jjxw/2025-02-01/doc-inehyqcx9694053.shtml). *Sina Corp*. 1 February 2025. Retrieved 1 February 2025.

2. "Hangzhou DeepSeek Artificial Intelligence Basic Technology Research Co., Ltd" (https://www.bloomberg.com/profile/company/2544189D:CH). *Bloomberg L.P.*
3. "DeepSeek Coder Model Service Agreement" (https://chat.deepseek.com/downloads/DeepSeek%20Coder%20Model%20Service%20Agreement_1019.pdf) (PDF), *DeepSeek*, 19 October 2023
4. "全国互联网安全管理平台" (https://beian.mps.gov.cn/#/query/webSearch?code=33010502011812). *beian.mps.gov.cn*. Retrieved 9 February 2025.
5. Gibney, Elizabeth (23 January 2025). "China's cheap, open AI model DeepSeek thrills scientists" (https://www.nature.com/articles/d41586-025-00229-6). *Nature*. doi:10.1038/d41586-025-00229-6 (https://doi.org/10.1038%2Fd41586-025-00229-6). ISSN 1476-4687 (https://search.worldcat.org/issn/1476-4687). PMID 39849139 (https://pubmed.ncbi.nlm.nih.gov/39849139).
6. Vincent, James (28 January 2025). "The DeepSeek panic reveals an AI world ready to blow" (https://www.theguardian.com/commentisfree/2025/jan/28/deepseek-r1-ai-world-chinese-chatbot-tech-world-western). *The Guardian*.
7. Metz, Cade; Tobin, Meaghan (23 January 2025). "How Chinese A.I. Start-Up DeepSeek Is Competing With Silicon Valley Giants" (https://www.nytimes.com/2025/01/23/technology/deepseek-bd-ai-chips.html?smid=fb-nytimes&smtyp=cur&fbclid=IwY2xjawIEynFleHRuA2FlbQIxMQABHZYKXN7GJpUyNRsaGEDQVadxRBarp-aBp1GhiuRe3B57Ehe6HYv7oiK78Q_aem_KTeDgqjV_-R80owNNWOBCQ). *The New York Times*. ISSN 0362-4331 (https://search.worldcat.org/issn/0362-4331). Retrieved 27 January 2025.
8. Cosgrove, Emma (27 January 2025). "DeepSeek's cheaper models and weaker chips call into question trillions in AI infrastructure spending" (https://www.businessinsider.com/explaining-deepseek-chinese-models-efficiency-scaring-markets-2025-1). *Business Insider*.
9. Erdil, Ege (17 January 2025). "How has DeepSeek improved the Transformer architecture?" (https://epoch.ai/gradient-updates/how-has-deepseek-improved-the-transformer-architecture). *Epoch AI*. Retrieved 3 February 2025.
10. Metz, Cade (27 January 2025). "What is DeepSeek? And How Is It Upending A.I.?" (https://www.nytimes.com/2025/01/27/technology/what-is-deepseek-china-ai.html). *The New York Times*. ISSN 0362-4331 (https://search.worldcat.org/issn/0362-4331). Retrieved 27 January 2025.
11. Roose, Kevin (28 January 2025). "Why DeepSeek Could Change What Silicon Valley Believe About A.I." (https://www.nytimes.com/2025/01/28/technology/why-deepseek-could-change-what-silicon-valley-believes-about-ai.html) *The New York Times*. ISSN 0362-4331 (https://search.worldcat.org/issn/0362-4331). Retrieved 28 January 2025.
12. Delbert, Caroline (31 January 2025). "DeepSeek Is Cracking the 'Black Box' of Corporate AI Wide Open" (https://www.popularmechanics.com/science/a63633889/deepseek-open-weight/). *Popular Mechanics*. Retrieved 12 February 2025.
13. Gibney, Elizabeth (23 January 2025). "China's cheap, open AI model DeepSeek thrills scientists" (https://www.nature.com/articles/d41586-025-00229-6). *Nature*. Retrieved 12 February 2025.
14. DeepSeek sends shock waves across Silicon Valley (https://liberationnews.org/deepseek-sends-shock-waves-across-silicon-valley/) Liberation News – The Newspaper of the Party for Socialism and Liberatio
15. DeepSeek: Tech firm suffers biggest drop in US stock market history as low-cost Chinese AI company bites Silicon Valley (https://news.sky.com/story/deepseek-us-tech-stocks-tumble-on-fears-of-cheaper-chinese-ai-13297788) Sky News
16. Sevastopulo, Demetri; Hille, Kathrin (7 October 2022). "US hits China with sweeping tech export controls" (https://www.ft.com/content/6825bee4-52a7-4c86-b1aa-31c100708c3e). *Financial Times*. Retrieved 5 February 2025.

17. Nellis, Stephen; Cherney, Max A. (31 August 2023). "US curbs AI chip exports from Nvidia and AMD to some Middle East countries" (https://www.reuters.com/technology/us-restricts-exports-some-nvidia-chips-middle-east-countries-filing-2023-08-30/). *Reuters*. Retrieved 4 February 2025.
18. Hawkins, Mackenzie; Leonard, Jenny (8 January 2025). "Biden to Further Limit Nvidia AI Chip Exports in Final Push" (https://www.bloomberg.com/news/articles/2025-01-08/biden-to-further-limit-nvidia-amd-ai-chip-exports-in-final-push?srnd=phx-technology). *Bloomberg*. Retrieved 4 February 2025.
19. Field, Matthew; Titcomb, James (27 January 2025). "Chinese AI has sparked a $1 trillion panic – and it doesn't care about free speech" (https://www.telegraph.co.uk/business/2025/01/27/Bd-deepseek-ai-has-sparked-a-1-trillion-panic/). *The Daily Telegraph*. ISSN 0307-1235 (https://search.worldcat.org/issn/0307-1235). Retrieved 27 January 2025.
20. Lu, Donna (28 January 2025). "We tried out DeepSeek. It worked well, until we asked it about Tiananmen Square and Taiwan" (https://www.theguardian.com/technology/2025/jan/28/we-tried-out-deepseek-it-works-well-until-we-asked-it-about-tiananmen-square-and-taiwan). *The Guardian*. ISSN 0261-3077 (https://search.worldcat.org/issn/0261-3077). Retrieved 30 January 2025.
21. Colville, Alex (10 February 2025). "DeepSeeking Truth" (https://chinamediaproject.org/2025/02/10/deepseeking-truth/). *China Media Project*. Retrieved 12 February 2025.
22. Yang, Ziyi (31 January 2025). "Here's How DeepSeek Censorship Actually Works - And How to Get Around It" (https://www.wired.com/story/deepseek-censorship/). *Wired*. Retrieved 12 February 2025.
23. Booth, Robert; Milmo, Dan (28 January 2025). "Chinese AI chatbot DeepSeek censors itself in realtime, users report" (https://www.theguardian.com/technology/2025/jan/28/chinese-ai-chatbot-deepseek-censors-itself-in-realtime-users-report). *The Guardian*. ISSN 0261-3077 (htps://search.worldcat.org/issn/0261-3077). Retrieved 13 February 2025.
24. published, David Nield (29 January 2025). "Want to try DeepSeek without the privacy worries? Perplexity AI just launched it on its iOS and web apps" (https://www.techradar.com/computing/artificial-intelligence/want-to-try-deepseek-without-the-privacy-worries-perplexity-ai-just-launched-it-on-its-ios-and-web-apps). *TechRadar*. Retrieved 13 February 2025.
25. Chen, Caiwei (24 January 2025). "How a top Chinese AI model overcame US sanctions" (https://www.technologyreview.com/2025/01/24/1110526/china-deepseek-top-ai-despite-sanctions/). *MIT Technology Review*. Archived (https://web.archive.org/web/20250125180427/https://www.technologyreview.com/2025/01/24/1110526/china-deepseek-top-ai-despite-sanctions/) from the original on 25 January 2025. Retrieved 25 January 2025.
26. "幻方 | 幻方历程" (https://www.high-flyer.cn/history/). *High-Flyer* (in Chinese (China)). Retrieved 2 February 2025.
27. Ottinger, Lily (9 December 2024). "Deepseek: From Hedge Fund to Frontier Model Maker" (https://www.chinatalk.media/p/deepseek-from-hedge-fund-to-frontier). *ChinaTalk*. Archived (https://web.archive.org/web/20241228030725/https://www.chinatalk.media/p/deepseek-from-hedge-fund-to-frontier) from the original on 28 December 2024. Retrieved 28 December 2024.
28. Olcott, Eleanor; Wu, Zijing (24 January 2025). "How small Chinese AI start-up DeepSeek shocked Silicon Valley" (https://www.removepaywall.com/search?url=https://www.ft.com/content/747a7b11-dcba-4aa5-8d25-403f56216d7e). *Financial Times*. Retrieved 31 January 2025.
29. Leswing, Kif (23 February 2023). "Meet the $10,000 Nvidia chip powering the race for A.I." (https://www.cnbc.com/2023/02/23/nvidias-a100-is-the-10000-chip-powering-the-race-for-ai-.html) *CNBC*. Retrieved 30 January 2025.
30. "hfreduce | 高性能的多卡并行通信工具" (https://www.high-flyer.cn/blog/hf-reduce/). *High-Flyer*. 4 March 2020. Retrieved 3 February 2025.

31. DeepSeek-AI; Liu, Aixin; Feng, Bei; Xue, Bing; Wang, Bingxuan; Wu, Bochao; Lu, Chengda; Zhao, Chenggang; Deng, Chengqi (27 December 2024), *DeepSeek-V3 Technical Report*, arXiv:2412.19437 (https://arxiv.org/abs/2412.19437)
32. An, Wei; Bi, Xiao; Chen, Guanting; Chen, Shanhuang; Deng, Chengqi; Ding, Honghui; Dong, Kai; Du, Qiushi; Gao, Wenjun; Guan, Kang; Guo, Jianzhong; Guo, Yongqiang; Fu, Zhe; He, Ying; Huang, Panpan (17 November 2024). "Fire-Flyer AI-HPC: A Cost-Effective Software-Hardware Co-Design for Deep Learning" (https://ieeexplore.ieee.org/document/10793193/). *IEEE Xplore*: 1–23. doi:10.1109/SC41406.2024.00089 (https://doi.org/10.1109%2FSC41406.2024.00089). ISBN 979-8-3503-5291-7.
33. "独家|幻方量化回应市场关注：AGI不是用来炒股的，"和金融没关系" " (https://www.yicai.com/news/101732215.html). *Yicai*. Retrieved 3 February 2025.
34. Yu, Xu (17 April 2023). "[Exclusive] Chinese Quant Hedge Fund High-Flyer Won't Use AGI to Trade Stocks, MD Says" (https://www.yicaiglobal.com/news/exclusive-chinese-quant-fund-high-flyer-will-not-use-agi-to-trade-stocks-managing-director-says). *Yicai Global*. Archived (https://web.archive.org/web/20231231030712/https://www.yicaiglobal.com/news/exclusive-chinese-quant-fund-high-flyer-will-not-use-agi-to-trade-stocks-managing-director-says) from the original on 31 December 2023. Retrieved 28 December 2024.
35. Jiang, Ben; Perezi, Bien (1 January 2025). "Meet DeepSeek: the Chinese start-up that is changing how AI models are trained" (https://www.scmp.com/tech/tech-trends/article/3293050/meet-deepseek-chinese-start-changing-how-ai-models-are-trained). *South China Morning Post*. Archived (https://web.archive.org/web/20250122160046/https://www.scmp.com/tech/tech-trends/article/3293050/meet-deepseek-chinese-start-changing-how-ai-models-are-trained) from the original on 22 January 2025. Retrieved 1 January 2025.
36. McMorrow, Ryan; Olcott, Eleanor (9 June 2024). "The Chinese quant fund-turned-AI pioneer" (https://www.ft.com/content/357f3c68-b866-4c2e-b678-0d075051a260). *Financial Times*. Archived (https://web.archive.org/web/20240717030903/https://www.ft.com/content/357f3c68-b866-4c2e-b678-0d075051a260) from the original on 17 July 2024. Retrieved 28 December 2024.
37. DeepSeek-AI; Bi, Xiao; Chen, Deli; Chen, Guanting; Chen, Shanhuang; Dai, Damai; Deng, Chengqi; Ding, Honghui; Dong, Kai (5 January 2024), *DeepSeek LLM: Scaling Open-Source Language Models with Longtermism*, arXiv:2401.02954 (https://arxiv.org/abs/2401.02954)
38. Dai, Damai; Deng, Chengqi; Zhao, Chenggang; Xu, R. X.; Gao, Huazuo; Chen, Deli; Li, Jiashi; Zeng, Wangding; Yu, Xingkai (11 January 2024), *DeepSeekMoE: Towards Ultimate Expert Specialization in Mixture-of-Experts Language Models*, arXiv:2401.06066 (https://arxiv.org/abs/2401.06066)
39. Shao, Zhihong; Wang, Peiyi; Zhu, Qihao; Xu, Runxin; Song, Junxiao; Bi, Xiao; Zhang, Haowei; Zhang, Mingchuan; Li, Y. K. (27 April 2024), *DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models*, arXiv:2402.03300 (https://arxiv.org/abs/2402.03300).
40. DeepSeek-AI; Zhu, Qihao; Guo, Daya; Shao, Zhihong; Yang, Dejian; Wang, Peiyi; Xu, Runxin; Wu, Y.; Li, Yukun (17 June 2024), *DeepSeek-Coder-V2: Breaking the Barrier of Closed-Source Models in Code Intelligence*, arXiv:2406.11931 (https://arxiv.org/abs/2406.11931)
41. "deepseek-ai/DeepSeek-V2.5 · Hugging Face" (https://huggingface.co/deepseek-ai/DeepSeek-V2.5). *Hugging Face*. 3 January 2025. Retrieved 28 January 2025.
42. "Deepseek Log in page" (https://chat.deepseek.com/sign_in). *DeepSeek*. Retrieved 30 January 2025.

43. "News | DeepSeek-R1-Lite Release 2024/11/20: 🚀 DeepSeek-R1-Lite-Preview is now live: unleashing supercharged reasoning power!" (https://web.archive.org/web/20241120141324/https://api-docs.deepseek.com/news/news1120). *DeepSeek API Docs*. Archived from the original (https://api-docs.deepseek.com/news/news1120) on 20 November 2024. Retrieved 28 January 2025.
44. Field, Hayden (27 January 2025). "China's DeepSeek AI dethrones ChatGPT on App Store: Here's what you should know" (https://www.cnbc.com/2025/01/27/chinas-deepseek-ai-tops-chatgpt-app-store-what-you-should-know.html). *CNBC*.
45. Picchi, Aimee (27 January 2025). "What is DeepSeek, and why is it causing Nvidia and other stocks to slump?" (https://www.cbsnews.com/news/what-is-deepseek-ai-china-stock-nvidia-nvda-asml/). *CBS News*.
46. "大模型价格又砍一刀 这次"屠夫"竟是量化私募？" (https://www.cls.cn/detail/1672635). *www.cls.cn*. 10 May 2024. Retrieved 3 February 2025.
47. Schneider, Jordan (27 November 2024). "Deepseek: The Quiet Giant Leading China's AI Race" (https://www.chinatalk.media/p/deepseek-ceo-interview-with-chinas). *ChinaTalk*. Retrieved 28 December 2024.
48. "幻方力量 | 高速文件系统 3FS" (https://www.high-flyer.cn/blog/3fs/). *High-Flyer*. 13 June 2019. Retrieved 3 February 2025.
49. "HFAiLab/hai-platform" (https://github.com/HFAiLab/hai-platform), *High-Flyer*, 2 February 2025, retrieved 3 February 2025
50. DeepSeek-AI; Guo, Daya; Yang, Dejian; Zhang, Haowei; Song, Junxiao; Zhang, Ruoyu; Xu, Runxin; Zhu, Qihao; Ma, Shirong (22 January 2025), *DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning*, arXiv:2501.12948 (https://arxiv.org/abs/2501.12948)
51. Delbert, Caroline (31 January 2025). "DeepSeek Is Cracking the 'Black Box' of Corporate AI Wide Open" (https://www.popularmechanics.com/science/a63633889/deepseek-open-weight/). *Popular Mechanics*. Retrieved 12 February 2025.
52. Gibney, Elizabeth (23 January 2025). "China's cheap, open AI model DeepSeek thrills scientists" (https://www.nature.com/articles/d41586-025-00229-6). *Nature*. Retrieved 12 February 2025.
53. "DeepSeek-Coder/LICENSE-MODEL at main · deepseek-ai/DeepSeek-Coder" (https://github.com/deepseek-ai/DeepSeek-Coder/blob/main/LICENSE-MODEL). *GitHub*. Archived (https://web.archive.org/web/20250122195853/https://github.com/deepseek-ai/deepseek-coder/blob/main/LICENSE-MODEL) from the original on 22 January 2025. Retrieved 24 January 2025.
54. Guo, Daya; Zhu, Qihao; Yang, Dejian; Xie, Zhenda; Dong, Kai; Zhang, Wentao; Chen, Guanting; Bi, Xiao; Wu, Y. (26 January 2024), *DeepSeek-Coder: When the Large Language Model Meets Programming – The Rise of Code Intelligence*, arXiv:2401.14196 (https://arxiv.org/abs/2401.14196)
55. "DeepSeek Coder" (https://deepseekcoder.github.io/). *deepseekcoder.github.io*. Retrieved 27 January 2025.
56. *deepseek-ai/DeepSeek-Coder* (https://github.com/deepseek-ai/deepseek-coder/), DeepSeek, 27 January 2025, retrieved 27 January 2025
57. "deepseek-ai/deepseek-coder-5.7bmqa-base · Hugging Face" (https://huggingface.co/deepseek-ai/deepseek-coder-5.7bmqa-base). *Hugging Face*. Retrieved 27 January 2025.
58. *deepseek-ai/DeepSeek-LLM* (https://github.com/deepseek-ai/DeepSeek-LLM), DeepSeek, 27 January 2025, retrieved 27 January 2025
59. Wang, Peiyi; Li, Lei; Shao, Zhihong; Xu, R. X.; Dai, Damai; Li, Yifei; Chen, Deli; Wu, Y.; Sui, Zhifang (19 February 2024), *Math-Shepherd: Verify and Reinforce LLMs Step-by-step without Human Annotations*, arXiv:2312.08935 (https://arxiv.org/abs/2312.08935).

60. DeepSeek-AI; Liu, Aixin; Feng, Bei; Wang, Bin; Wang, Bingxuan; Liu, Bo; Zhao, Chenggang; Dengr, Chengqi; Ruan, Chong (19 June 2024), *DeepSeek-V2: A Strong, Economical, and Efficient Mixture-of-Experts Language Model*, arXiv:2405.04434 (https://arxiv.org/abs/2405.04434).
61. Peng, Bowen; Quesnelle, Jeffrey; Fan, Honglu; Shippole, Enrico (1 November 2023), *YaRN: Efficient Context Window Extension of Large Language Models*, arXiv:2309.00071 (https://arxiv.org/abs/2309.00071).
62. "config.json · deepseek-ai/DeepSeek-V2-Lite at main" (https://huggingface.co/deepseek-ai/DeepSeek-V2-Lite/blob/main/config.json). *Hugging Face*. 15 May 2024. Retrieved 28 January 2025.
63. "config.json · deepseek-ai/DeepSeek-V2 at main" (https://huggingface.co/deepseek-ai/DeepSeek-V2/blob/main/config.json). *Hugging Face*. 6 May 2024. Retrieved 28 January 2025.
64. "config.json · deepseek-ai/DeepSeek-V3 at main" (https://huggingface.co/deepseek-ai/DeepSeek-V3/blob/main/config.json). *Hugging Face*. 26 December 2024. Retrieved 28 January 2025.
65. "DeepSeek Debates: Chinese Leadership On Cost, True Training Cost, Closed Model Margin Impacts" (https://semianalysis.com/2025/01/31/deepseek-debates/). *SemiAnalysis*. 31 January 2025. Retrieved 13 February 2025.
66. "DeepSeek's AI costs far exceed $5.5 million claim, may have reached $1.6 billion with 50,000 Nvidia GPUs" (https://www.techspot.com/news/106612-deepseek-ai-costs-far-exceed-55-million-claim.html). *TechSpot*. 3 February 2025. Retrieved 13 February 2025.
67. "Research exposes DeepSeek's AI training cost is not $6M, it's a staggering $1.3B" (https://www.yahoo.com/news/research-exposes-deepseek-ai-training-165025904.html). *Yahoo News*. 31 January 2025. Retrieved 13 February 2025.
68. "Martin Vechev of INSAIT: "DeepSeek $6M Cost Of Training Is Misleading" " (https://therecursive.com/martin-vechev-of-insait-deepseek-6m-cost-of-training-is-misleading/). *TheRecursive.com*. 28 January 2025. Retrieved 13 February 2025.
69. Jiang, Ben (27 December 2024). "Chinese start-up DeepSeek's new AI model outperforms Meta, OpenAI products" (https://www.scmp.com/tech/tech-trends/article/3292507/chinese-start-deepseek-launches-ai-model-outperforms-meta-openai-products). *South China Morning Post*. Archived (https://web.archive.org/web/20241227191529/https://www.scmp.com/tech/tech-trends/article/3292507/chinese-start-deepseek-launches-ai-model-outperforms-meta-openai-products) from the original on 27 December 2024. Retrieved 28 December 2024.
70. Sharma, Shubham (26 December 2024). "DeepSeek-V3, ultra-large open-source AI, outperforms Llama and Qwen on launch" (https://venturebeat.com/ai/deepseek-v3-ultra-large-open-source-ai-outperforms-llama-and-qwen-on-launch/). *VentureBeat*. Archived (https://web.archive.org/web/20241227195503/https://venturebeat.com/ai/deepseek-v3-ultra-large-open-source-ai-outperforms-llama-and-qwen-on-launch/) from the original on 27 December 2024. Retrieved 28 December 2024.
71. Wiggers, Kyle (26 December 2024). "DeepSeek's new AI model appears to be one of the best 'open' challengers yet" (https://techcrunch.com/2024/12/26/deepseeks-new-ai-model-appears-to-be-one-of-the-best-open-challengers-yet/). *TechCrunch*. Archived (https://web.archive.org/web/20250102103526/https://techcrunch.com/2024/12/26/deepseeks-new-ai-model-appears-to-be-one-of-the-best-open-challengers-yet/) from the original on 2 January 2025. Retrieved 31 December 2024.
72. Franzen, Carl (20 November 2024). "DeepSeek's first reasoning model R1-Lite-Preview turns heads, beating OpenAI o1 performance" (https://venturebeat.com/ai/deepseeks-first-reasoning-model-r1-lite-preview-turns-heads-beating-openai-o1-performance/). *VentureBeat*. Archived (https://web.archive.org/web/20241122010413/https://venturebeat.com/ai/deepseeks-first-reasoning-model-r1-lite-preview-turns-heads-beating-openai-o1-performance/) from the original on 22 November 2024. Retrieved 28 December 2024.

73. Huang, Raffaele (24 December 2024). "Don't Look Now, but China's AI Is Catching Up Fast" (https://www.wsj.com/tech/ai/china-ai-advances-us-chips-7838fd20). *The Wall Street Journal*. Archived (https://web.archive.org/web/20241227183842/https://www.wsj.com/tech/ai/china-ai-advances-us-chips-7838fd20) from the original on 27 December 2024. Retrieved 28 December 2024.
74. "Release DeepSeek-R1 · deepseek-ai/DeepSeek-R1@23807ce" (https://github.com/deepseek-ai/DeepSeek-R1/commit/23807ced51627276434655dd9f27725354818974). *GitHub*. Archived (https://web.archive.org/web/20250121104009/https://github.com/deepseek-ai/DeepSeek-R1/commit/23807ced51627276434655dd9f27725354818974) from the original on 21 January 2025. Retrieved 21 January 2025.
75. Roose, Kevin (28 January 2025). "Why DeepSeek Could Change What Silicon Valley Believe About A.I." (https://www.nytimes.com/2025/01/28/technology/why-deepseek-could-change-what-silicon-valley-believes-about-ai.html) *The New York Times*. ISSN 0362-4331 (https://search.worldcat.org/issn/0362-4331). Retrieved 28 January 2025.
76. Gibney, Elizabeth (23 January 2025). "China's cheap, open AI model DeepSeek thrills scientists" (https://www.nature.com/articles/d41586-025-00229-6). *Nature*. doi:10.1038/d41586-025-00229-6 (https://doi.org/10.1038%2Fd41586-025-00229-6). ISSN 1476-4687 (https://search.worldcat.org/issn/1476-4687). PMID 39849139 (https://pubmed.ncbi.nlm.nih.gov/39849139).
77. Cosgrove, Emma (27 January 2025). "DeepSeek's cheaper models and weaker chips call into question trillions in AI infrastructure spending" (https://www.businessinsider.com/explaining-deepseek-chinese-models-efficiency-scaring-markets-2025-1). *Business Insider*.
78. Erdil, Ege (17 January 2025). "How has DeepSeek improved the Transformer architecture?" (https://epoch.ai/gradient-updates/how-has-deepseek-improved-the-transformer-architecture). *Epoch AI*. Retrieved 3 February 2025.
79. Field, Matthew; Titcomb, James (27 January 2025). "Chinese AI has sparked a $1 trillion panic – and it doesn't care about free speech" (https://www.telegraph.co.uk/business/2025/01/27/Bd-deepseek-ai-has-sparked-a-1-trillion-panic/). *The Daily Telegraph*. ISSN 0307-1235 (https://search.worldcat.org/issn/0307-1235). Retrieved 27 January 2025.
80. Colville, Alex (10 February 2025). "DeepSeeking Truth" (https://chinamediaproject.org/2025/02/10/deepseeking-truth/). *China Media Project*. Retrieved 12 February 2025.
81. Yang, Ziyi (31 January 2025). "Here's How DeepSeek Censorship Actually Works - And How to Get Around It" (https://www.wired.com/story/deepseek-censorship/). *Wired*. Retrieved 12 February 2025.

External links

- Official website (https://deepseek.com)
- DeepSeek (https://github.com/deepseek-ai) on GitHub
- DeepSeek (https://huggingface.co/deepseek-ai/DeepSeek-V2.5-1210) on Hugging Face
- Official API documentation (https://api-docs.deepseek.com/)
- Anthology of DeepSeek papers (https://huggingface.co/collections/Presidentlin/deepseek-papers-674c536aa6acddd9bc98c2ac)
- Research blog of High-Flyer (https://www.high-flyer.cn/blog/)

Retrieved from "https://en.wikipedia.org/w/index.php?title=DeepSeek&oldid=1275605176"