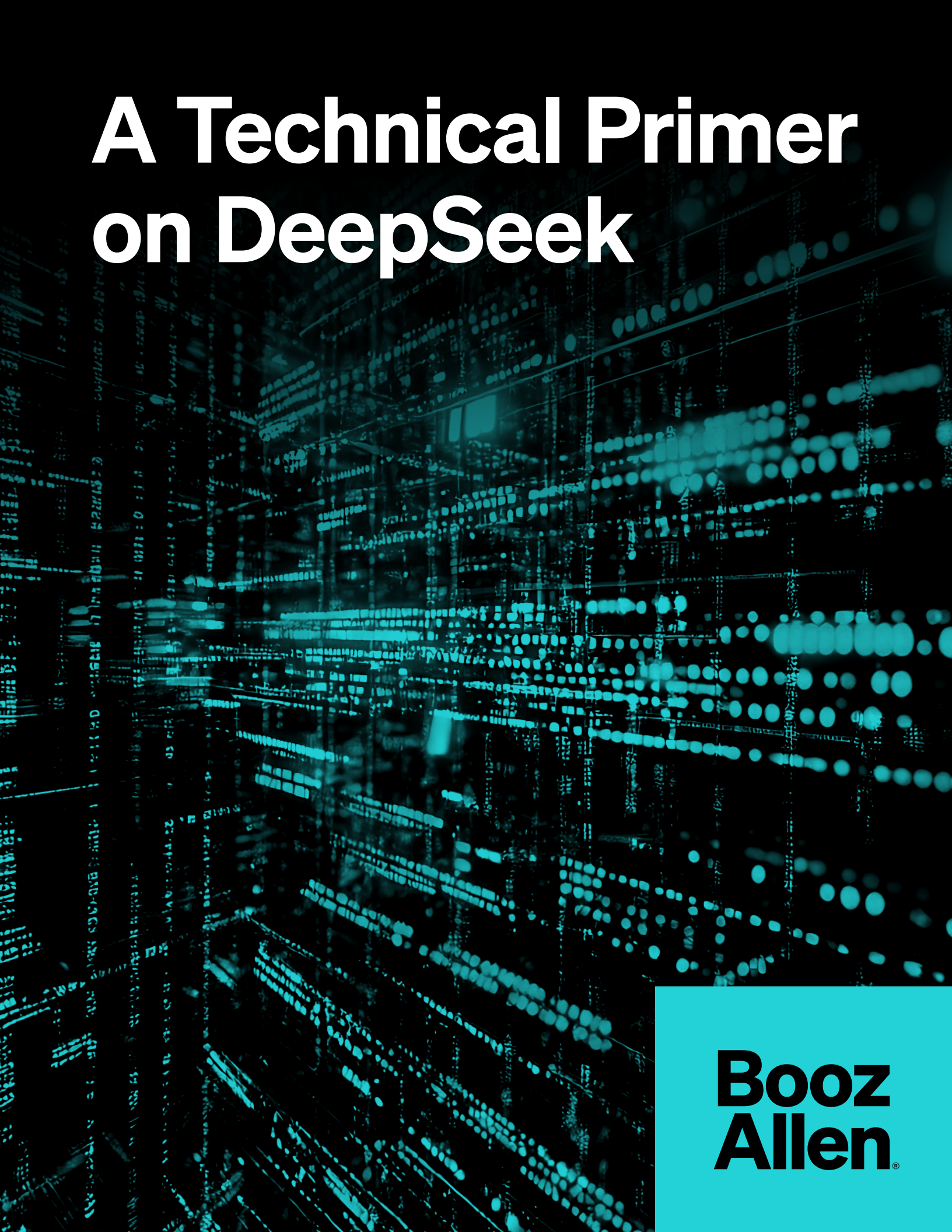
A Technical Primer on DeepSeek



Booz Allen®

Table of Contents

Overview

- **DeepSeek is a China-based AI startup that has led a well-funded effort to develop advanced large language models (LLM) using a large team (100+) of experienced developers.** Public interest stems from their newest models being released for free with what the company claims is performance comparable to OpenAI, Anthropic, and Meta LLMs at a fraction of the price and training time.

- **"DeepSeek" is conflated with multiple algorithms of the same namesake, but it is the DeepSeek-R1 LLM—a 671B model—that is the focus of media attention.** It has been trained with a multi-stage pipeline of Reinforcement Learning (RL), Supervised Fine-Tuning (SFT), and possibly distillation methods to learn from a larger teacher model.

- **The cost to train DeepSeek is publicized as $6 million, which is derived from the older, DeepSeek-V3 base model.** It is not easy to verify the cost, and, at face value, it likely is a snapshot of a single, pristine training run. Their paper makes this explicit, but it has been overlooked in reactions that fail to account for significant experimentation, prior development, and infrastructure costs.

- Their training process applies a variety of artificial intelligence (AI), optimization, and hardware innovations derived from non-DeepSeek published research to train an LLM with less computational infrastructure. **DeepSeek's modifications, improvements, and assembly of these methods are meaningful, but no single extraordinary development appears to have occurred.**

- **Key details are missing, particularly around the training pipeline, datasets used to fine-tune the models, and technical implementation that drove efficiency.** For example, OpenAI has claimed DeepSeek's may have inappropriately acquired their intellectual property via distillation in violation of the company's terms of service. At the same time, DeepSeek's transparency goes far beyond the overwhelming majority of Western labs, with only a few (primarily non-profits such as EleutherAI and the Allen Institute for AI) disclosing more.

[1]https://www.nbcnews.com/tech/tech-news/openai-says-deepseek-may-inapproriately-used-data-rcna189872

First Take Summary

DeepSeek represents a significant advancement in **AI efficiency** by optimizing training and inference into a **scalable AI development pipeline**. By combining **Mixture-of-Experts (MoE), RL-based fine-tuning, advanced distillation techniques, and graphics processing unit (GPU)-level engineering**, DeepSeek has demonstrated a viable alternative to the **resource-heavy training approaches** used by other LLM providers.

- **MoE:** DeepSeek's MoE selectively activates specialized "experts" per token, reducing computational overhead while maintaining performance. It optimizes **GShard sparse-gating** and **load-balancing techniques** to prevent inefficiencies, ensuring efficient expert utilization and full token processing during training.
- **RL and Group Relative Policy Optimization (GRPO)**: DeepSeek's training pipeline replaces traditional **SFT** with **GRPO**, an RL variant that **removes the need for a separate value model**, reducing memory overhead and computational complexity. This allows DeepSeek to **improve reasoning without requiring extensive human-annotated ranking datasets**, a cost-intensive step in other LLMs.
- **DualPipe System:** DeepSeek introduces a **parallelized GPU scheduling and workload management framework**, enabling **simultaneous forward and backward passes** during training. This innovation **reduces idle compute time, optimizes GPU utilization**, and speeds up both training and inference, making DeepSeek's model development pipeline significantly more efficient.
- **Distillation Techniques**: DeepSeek has successfully **distilled the reasoning and computational abilities of its larger models into smaller, high-performance variants**, such as Qwen models (1.5B to 70B parameters). These distilled models **outperform OpenAI-o1-mini and Claude-3.5 in math, coding, and reasoning tasks**, proving that **high efficiency does not necessarily require massive-scale architectures**. This approach allows for **smaller, more cost-effective AI models** that retain strong reasoning capabilities.

When assessing the overall performance of DeepSeek, the widely cited **$6 million training cost** applies only to **DeepSeek-V3**, rather than the more advanced **DeepSeek-R1**. Nevertheless, its efficiency innovations **still challenge the assumption** that **massive capital investment is required** to develop state-of-the-art AI models. DeepSeek's **inference efficiency claims are supported by its MoE-based selective activation**, which **drastically reduces power consumption and memory requirements** compared to dense models like GPT-4. The performance benchmarks highlight DeepSeek's **strengths in reasoning, math, and coding tasks**, with results **exceeding OpenAI-o1-mini and Claude-3.5** in multiple structured problem-solving tests, although its general conversational abilities remain unverified. However, DeepSeek's **lack of transparency** regarding training data sources, fine-tuning methodology, and full infrastructure details raises **questions about the reproducibility of its efficiency claims**.

In conclusion, DeepSeek's emergence is **not just about one model—it's about reshaping the playbook for AI development**. Collectively, DeepSeek's combination of efficiencies in algorithms, framework, and hardware is significant. If its approach proves sustainable, DeepSeek's model could **shift AI development away from hyperscale cloud dependency**, making high-performance AI **more affordable, decentralized, and accessible across various industries**. Whether its methods are truly sustainable or not, it has already **forced the AI industry to reconsider the economics of model training, optimization, and deployment**.

[2] Liu, Aixin, et al. "DeepSeek-V3 Technical Report." arXiv:2412.19437v (2024).

[3] Guo, Daya, et al. "DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning." arXiv:2501.12948 (2025).

Claim

DeepSeek claims that, with a budget of $6 million, it achieves performance comparable to big proprietary LLMs like OpenAI at a fraction of the cost and compute. Scientifically, it claims that the DeepSeek-R1-Zero LLM is the first open research to validate that LLMs can be trained solely through RL after pretraining. This is cost-critical because the DeepSeek-R1-Zero algorithm can match the performance of some LLMs without needing SFT, which is a bottleneck. Bypassing SFT affords DeepSeek to be trained without explicitly teaching the model through expensive, manual examples. This, among other algorithmic, framework, and hardware innovations, has enabled DeepSeek to be trained faster and with less compute.

However, there are nuances to this claim. There are two models that are conflated. Both DeepSeek-R1-Zero and DeepSeek-R1 are trained with RL, but the more powerful LLM is DeepSeek-R1, which is not trained exclusively with pure RL. The DeepSeek-R1-Zero LLM mixes different languages or lacks markdown formatting to highlight answers, making the outputs difficult to read. As a result, the DeepSeek-R1 LLM trains on a tiny number of supervised samples that have been carefully engineered to kick off a four-stage process called "cold start." DeepSeek-R1 has been evaluated on 21 benchmarks covering English understanding, coding, mathematics, and Chinese. They compare their results to Claude-3.5-Sonnet-1022, GPT-4o-0513, OpenAI-o1-mini, and OpenAI-o1-1217 (see Table 1). However, DeepSeek-R1 was focused on model reasoning for tasks such as coding, mathematics, and logical reasoning where the problems are well-defined and the solutions are verifiable by another computer program.

It is difficult to validate the claims DeepSeek makes around its RL, SFT, and safety/alignment claims because their data and exact techniques have not been published. They discuss that RL can be used to improve reasoning with problems with well-defined solutions and safety/alignment with safety feedback after answering, but they do not say where that data was sourced from or how they produced it themselves. Similarly, they claim to perform SFT through rejection sampling and "supervised data from DeepSeek-V3" but do not publish samples from this data.

	MATH: AIME v 2024 (pass@1*)	ENGLISH: GPQA Diamond (pass@1)	CODE: LiveCode Bench (pass@1)
OpenAI-o1-mini	63.6	60.0	53.8
Claude-3.5-Sonnet-1022	16.0	65.0	38.9
GPT-4o-0513	9.3	49.9	32.9
DeepSeek V3	39.2	59.1	36.2
DeepSeek-R1-Zero (trained with RL)	71.0	73.3	50.0
DeepSeek-R1 (trained with SFT & RL)	79.8	71.5	65.9

*Table 1: An Abridged Set of Three Evaluation Benchmarks (Data Source: DeepSeek-R: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning, 22 Jan. 2025.) * pass@1 is a metric that measures the algorithm's ability to get the correct answer on the first attempt.*

DeepSeek Models

It is important to realize that when people refer to "DeepSeek," they are likely referring to DeepSeek-R1. However, DeepSeek models encompass at least 13 different LLMs and related methodologies that have been published on arXiv and released open source since the beginning of January 2024. The timeline below highlights the evolution of DeepSeek's capabilities, which are directly released as downloadable models and code on Github.

1. 5 January 2024, DeepSeek-MoE (Towards Ultimate Expert Specialization in Mixture-of-Experts Language Models)

2. 5 January 2024, DeepSeek-LLM (Scaling Open-Source Language Models with Longtermism)

3. 27 April 2024, DeepSeekMath (Pushing the Limits of Mathematical Reasoning in Open Language Models)

4. 11 March 2024, DeepSeek-VL (Towards Real-World Vision-Language Understanding)

5. 26 January 2024, DeepSeek-Coder (When the Large Language Model Meets Programming—The Rise of Code Intelligence)

6. 15 August 2024, DeepSeek-Prover-V1.5 (Harnessing Proof Assistant Feedback for Reinforcement Learning and Monte-Carlo Tree Search)

7. 5 July 2024, Methodology for DeepSeek MoE Implementation (Let the Expert Stick to His Last: Expert-Specialized Fine-Tuning for Sparse Architectural Large Language Models)

8. 17 June 2024, DeepSeek-Coder-V2 (Breaking the Barrier of Closed-Source Models in Code Intelligence)

9. 19 June 2024, DeepSeek-V2 (A Strong, Economical, and Efficient Mixture-of-Experts Language Model)

10. 12 November 2024, JanusFlow (Harmonizing Autoregression and Rectified Flow for Unified Multimodal Understanding and Generation)

11. 13 December 2024, DeepSeek-VL2 (Mixture-of-Experts Vision-Language Models for Advanced Multimodal Understanding)

12.27 December 2024, DeepSeek-V3 (Technical Report)

13. 22 January 2025, DeepSeek-R1 (Incentivizing Reasoning Capability in LLMs via Reinforcement Learning)

The timeline reveals an industrial research lab that has made aggressive, iterative steps in research in a short amount of time. The key steppingstones that accelerated DeepSeek-R1 are: DeepSeek-MoE (January 2024), where they save computation by using an MoE framework; DeepSeekMath (April 2024), which introduced their novel RL algorithm GRPO; and DeepSeek-V3 (December 2024), which is the backbone of the DeepSeek-R1 LLM. To provide a sense of pace, OpenAI's timeline of GPT-related work began in September 2019, culminating with the release of ChatGPT in November 2022, a three-year delta of AI development. For Llama (Meta), the time between the initial model (Llama-65B) and the latest (Llama 3.3 70B) is nearly two years. For Claude (Anthropic), the time is 11 months. However, a careful look at DeepSeek's papers indicates a focus on computational savings and domain-specific tasks performant on math and coding. Additionally, they benefited from being a later player in this space, since they already knew about many strategies that are essential to effective LLM development, along with strategies that have empirically not worked well. It is a mistake to think of the DeepSeek models as a grassroots effort from a few, little-known AI researchers. Rather, this is a large, concerted effort from a development team with significant experience in the optimization, performance, and AI space.

[4] OpenAI published research on GPT-2 on 19 September 2019. Almost three years later, it published InstructGPT (27 January 2022), which introduced the framework of using SFT; it released ChatGPT almost a year later on 30 November 2022. A few months later, OpenAI released GT-4 on 14 March 2023.

[5] Llama-65B (Meta) was introduced in February 2023, followed by Llama-2 70B in July, Llama 4 70B in April 2024, Llama 3.1 405B in July 2024, and Llama 3.3 70B in December 2024.

Architecture

The simplest element of DeepSeek-R1 is the architecture itself. The architecture mimics DeepSeek-V3 (27 December 2024), which is exactly the same as DeepSeek-V2 (19 June 2024). This means that no improvement of the transformer architecture, itself, was the focus for algorithmic improvement in DeepSeek-R1. The DeepSeek-V3 paper indicates that it uses Multi-Head Latent Attention (MHLA), but a comparison against the V3 paper, published seven months earlier, also shows the same feature. Regardless, optimizing the attention mechanism is one of the key ways that DeepSeek made transformers more efficient, since attention requires each token (in a sentence) to be compared to every other token to determine its relevance. As the length of the passage increases, the computational complexity of the attention computation scales quadratically. The goal of MHLA is to compress the keys and values. This, in addition to low-rank compression, which attenuates memory, provides some computational speed over the traditional Multi-Head Attention conventionally seen in transformers.

Initially, a machine learning technique, Mixture of Experts has enjoyed a resurgence as a means for training large language models using significantly less compute.

Most important to the architecture is implementing MoE. MoE is an old idea—originally introduced in the early 1990s by colleagues of Geoff Hinton—that has been increasingly applied for LLMs. The idea is that different parts of a model (i.e., neural network) can be thought of as "experts" and that these experts specialize in different tasks when it comes to the data. This is ideal for computation because only certain experts are used and called upon during training. MoE models have had a substantial resurgence over the past three years in LLMs, led by work at Google, Mistral, and Databricks. Usually, MoE layers are used that combine and task multiple feed-forward neural networks (FFNs) as experts. A special gating function is used to activate which expert should be turned on. DeepSeek-V3 (and, in turn, DeepSeek-R1) compare their "DeepSeekMoE" to a sparse-gating function that is specifically known as "GShard," a known method for token choice gating. The general idea of sparse-gating for MoE is to activate a selected subset of experts (i.e., FFNs) to process each individual input token. DeepSeekMoE asserts that it uses finer-grained experts. Furthermore, the use of "auxiliary loss-free load-balancing" is not new and was developed by Shazeer et al in 2017 . The reason is to prevent routing collapse, where the data and computation may degrade on the way to the experts. They introduce a series of training parameters to monitor the expert load on the whole batch while training, thus balancing the experts in the transformer. In addition, in another way to save computational costs, DeepSeek-V3 restricts the routing of data to at most M nodes. These methods together enable DeepSeek-V3 to be fully trained with all tokens where no tokens are dropped during MoE training.

[6] Claude 2(Anthropic) in July 2023, Claude 2 Opus in March 2024, and Claude 3.5 Sonnet in June 2024.

[7] Cai, Weilin, et al. "A survey on mixture of experts." arXiv preprint arXiv:2407.06204 (2024).

[8] Shazeer, Noam, et al. "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer." arXiv preprint arXiv:1701.06538 (2017).

DeepSeek-R1 Reinforcement Learning with Human Feedback

Pipeline

Most of the training procedures for DeepSeek-R1 are identical to the DeepSeek-V3 paper (27 December 2024). There are four stages: 1) Cold Start Data, 2) RL with GRPO, 3) Rejection Sampling and SFT, and 4) RL for All Scenarios.

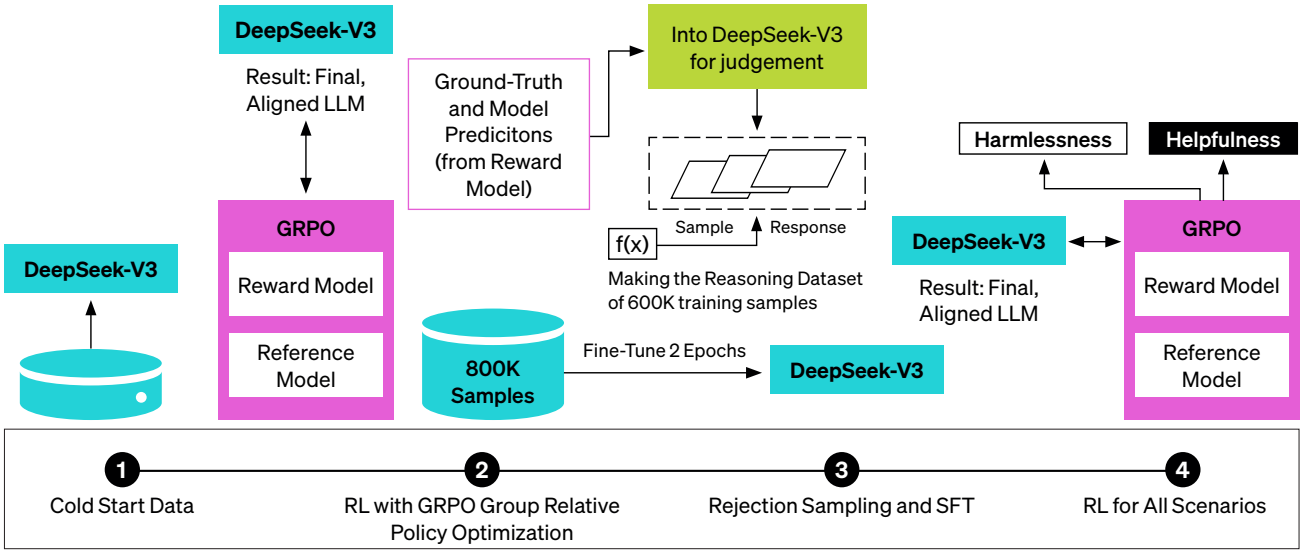


Figure 1: An Overview of the Multi-Stage Training Pipeline for DeepSeek-R1. This pipeline uses a combination of supervised fine training and reinforcement learning.

In phase 1, thousands of long Chain-of-Thought (CoT) data points are used to fine-tune the model as the initial RL actor. Thousands may sound like a lot of data, but LLMs typically use millions of data points to begin SFT. This is the carefully designed, readable pattern DeepSeek uses for the data formatting: | special token | <reasoning_process> | special_token | <summary>

Reminder: RL is used here because the rules are discrete (e.g., 1 or 2, but not 1.5), and normal deep learning (via backpropagation) cannot work with discrete values.

Here the reasoning process is CoT. In phase 2, GRPO foregoes the "Value Model," or the “critic,” in a typical three-model Proximal Policy Optimization (PPO) framework. This reduces training resources. It adopts a rule-based reward system. For DeepSeek-R1, the rewards are accuracy, format, and language consistency, where all three are summed to define the objective of the RL model. Instead of training a Value Model, it calculates the average reward of multiple grouped outputs generated from the same question as the baseline. In phase 3, once the RL model has converged, they save the checkpoint (i.e., the model file) and create a curated set of new reasoning data.

The training is exactly the same as what is done with DeepSeek-V3. The reasoning dataset of 600,000 samples is developed by a statistical technique called “rejection sampling,” which is akin to Monte Carlo estimation. The idea is to randomly sample from a proxy data distribution if it is not possible to sample from a complex, target distribution. For the non-reasoning data, which includes writing, factual question-answering, self-cognition, and translation, they reuse parts of the SFT dataset from DeepSeek-V3 (14.8T tokens) leading to 200,000 training samples. Phase 4 is the second phase of RL training for "helpfulness" and "harmlessness" and is the most

ambiguous of the four. It uses the same reward model for reasoning data (accuracy, format) as was done in DeepSeek-R1-Zero and unspecified reward models for non-reasoning data (general data) to capture human preference in complex, nuanced scenarios.

Helpfulness: Utility and relevance of response.

Harmlessness: Potential risks, biases, harmful content. No information about how they evaluated.

Reinforcement Learning

To understand the RL strategy DeepSeek-R1 uses, we need a short tutorial on Reinforcement Learning with Human Feedback (RLHF)—a common tool in training LLMs, such as ChatGPT, to train an LLM based on human preference.

Typically, a set of prompts (i.e., questions) are provided to an LLM that generates several outputs. A human annotator is then hired to rank/score each output according to alignment or preference. This is a supervised process, SFT, which DeepSeek-R1-Zero bypasses. The goal of the RL algorithm is to fine-tune the LLM in a way that does not require human input.

For RL, the LLM is the policy model to be optimized. For ChatGPT, and many LLMs, the RL algorithm of choice is PPO. Several algorithms are available, where the design choices are based on the rules of update. Given how the LLM is the actual policy being optimized, the LLM will produce the "action" in conventional RL terminology by sampling the "environment" by generating text. The "action" is predicting the next token. The act of doing so changes the current "state" by extending the current sequence and adding on to it. When finished, the output is then rewarded by rating the quality of the complete sentence, and the policy is updated by the rules of PPO.

The DeepSeek-R1-Zero algorithm was trained purely by applying RL in a novel method developed in April 2024 by DeepSeek called "GRPO." GRPO allows for computational savings by removing a Value Model, which is a neural network from the PPO framework. In PPO, the Value Model is a function that calculates expected long-term reward of a given state and is kept as a baseline. It basically estimates the value, or the improvement the given action provided given the current state. It also prevents the reward from being over-optimized. Hence, it is sometimes called a "critic." It is one of three models, or neural networks in PPO, and can add to the training time. Part of DeepSeek's innovation is a clever way of using RL to make sure the output follows desired rules of syntactic correctness that end up encouraging more complex behavior/reasoning to satisfy these simple-to-validate rules.

GRPO drops the Value Model to leave only two neural networks to train, which are the Reward and Reference Models. To replace the Value Model, GRPO calculates the average reward of multiple (Grouped, "G") sampled outputs generated from the same question as the baseline. Figure 5 provides a comparison of the mathematics and the conceptual pipeline of the traditional PPO algorithm and DeepSeek's GRPO.

PPO Algorithm

$$\mathcal{J}_{PPO}(\theta) = \mathbb{E}[q \sim P(Q), o \sim \pi_{\theta_{old}}(O|q)] \frac{1}{|o|} \sum_{t=1}^{|o|} \min\left[\frac{\pi_\theta(o_t|q,o_{<t})}{\pi_{\theta_{old}}(o_t|q,o_{<t})} A_t, \text{clip}\left(\frac{\pi_\theta(o_t|q,o_{<t})}{\pi_{\theta_{old}}(o_t|q,o_{<t})}, 1-\varepsilon, 1+\varepsilon\right) A_t\right], \quad (1)$$

Same: Ratio of new and old policies

Same: The clipped ratio of new and old policies if the new gets too "far" from the old

GRPO Algorithm

New loss function

$$\mathcal{J}_{GRPO}(\theta) = \mathbb{E}[q \sim P(Q), \{o_i\}_{i=1}^{G} \sim \pi_{\theta_{old}}(O|q)]$$

$$\frac{1}{G}\sum_{i=1}^{G} \frac{1}{|o_i|} \sum_{t=1}^{|o_i|} \left\{\min\left[\frac{\pi_\theta(o_{i,t}|q,o_{i,<t})}{\pi_{\theta_{old}}(o_{i,t}|q,o_{i,<t})} \hat{A}_{i,t}, \text{clip}\left(\frac{\pi_\theta(o_{i,t}|q,o_{i,<t})}{\pi_{\theta_{old}}(o_{i,t}|q,o_{i,<t})}, 1-\varepsilon, 1+\varepsilon\right) \hat{A}_{i,t}\right] - \beta \mathbb{D}_{KL}\left[\pi_\theta || \pi_{ref}\right]\right\}, \quad (3)$$

Same **Same**

Trained RL Policy ReferenceRL Policy

For each group of outputs, G, do:

Figure 2: A Comparison of the Optimization Functions for PPO and DeepSeek's GRPO. Note how some components of PPO are the same as GRPO, except aggregated over several outputs (G). The use of grouping is to get away from using the third Value Model. In the PPO diagram, "old=reference" and "new – no theta = trained."

[9] Shao, Zhihong, et al. "Deepseekmath: Pushing the limits of mathematical reasoning in open language models." arXiv preprint arXiv:2402.03300 (2024).

Reward modeling trains the RL signal to drive the optimization. DeepSeek-R1-Zero uses accuracy rewards, which rewards the model if the answer is correct, and format rewards, which enforces the model to format its thinking process between the <think></think> tags.

Reward modeling for DeepSeek-R1 is different. Because the GRPO algorithm is being used to fine-tune cold-start examples (which is essentially SFT), there is an additional reward model that is a "language consistency reward." This is the proportion of target language words that are included in the CoT output. The accuracy reward, format reward, and language consistency reward are summed together for a total reward that trains GRPO until it reaches convergence.

DeepSeek authors claim that DeepSeek-R1-Zero demonstrates a natural reasoning ability over time. They lean on the average response length of DeepSeek-R1-Zero over the training time, showing how the length of the text response (hundreds to thousands of reasoning tokens) via CoT increases the longer it is trained. They note examples of reflection (when the model revisits and reevaluates its previous steps) and exploration of alternative approaches. They also note an "aha" moment where an intermediate (in the midst of training) DeepSeek-R1-Zero models attempts to calculate a math problem and stops to say, "Wait, wait. Wait. That's an aha moment I can flag here."

Note that the data used for all these experiments is not shared, as is the current behavior of other LLM providers. The DeepSeek paper makes explicit note of the use of cold start data that contains "Long Chain of Thought Data," which is not a normal kind of text to be available. What this data contains is not clear, but it would be a highly plausible place to obtain example CoTs from existing effective LLMs. While examples have been shared online of getting DeepSeek to claim itself as another LLM, it is also the case that LLM outputs are becoming more ubiquitous online and could easily be captured inadvertently or intentionally. There are not yet any reliable mechanisms to distinguish the degree or intent of such interactions with another LLM.

Distillation and Smaller Models

Distillation is an AI technique and not in the realm of computation, per se. It was redeveloped for LLMs in 2023 a paper called "Distilling Step-by-Step", but the idea has been around since 2015, and it is a common technique for producing smaller and more compute-efficient models from larger, advanced models. Given the focus on moving to smaller models, DeepSeek uses distillation as a common approach. In addition to DeepSeek-R1, DeepSeek has open-sourced six dense models resulting from distillation of DeepSeek-R1. Distillation enables the training of a student model using the probability distributions of a larger model. The outcome is a smaller parameter LLM that, despite its size, is powerful on specific tasks.

DeepSeek demonstrates this by showing that DeepSeek-R1 (671B) can be distilled to smaller models such as Qwen (1.5B/7B/14B/32B) and Llama (8B/70B), outperforming reasoning tasks like math and live-coding comparably to OpenAI-o1-mini. For example, DeepSeek-R1-Distill-Llama-70B outperforms Claude-3.5-Sonnett, o1-mini, and GPT-4o-0513 on the AIME-2024 cons@64 (86.7 vs. o1-mini 80.0), MATH-500 pass@1 (94.5 vs. o1-mini 90.0), GPQA Diamond pass@1 (65.2 vs. 60.0 o1-mini), and LiveCodeBench (57.5 vs. 53.8 o1-mini).

[10] Hsieh, Cheng-Yu, et al. "Distilling step-by-step! Outperforming larger language models with less training data and smaller model sizes." arXiv preprint arXiv:2305.02301 (2023).

[11] Geoffrey Hinton, Oriol Vinyals, Jeff Dean. "Distilling the knowledge in a neural network." https://arxiv.org/abs/1503.02531

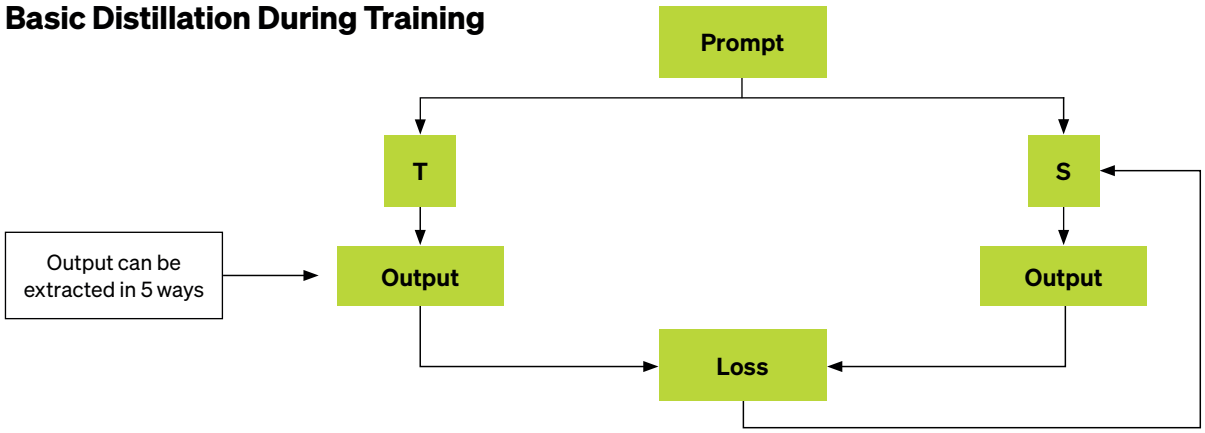


Figure 3: demonstrates how distillation works during training. This general procedure is how DeepSeek distilled Llama and Qwen models. Distillation relies on matching the outputs of a teacher and student model. These outputs can come in five forms. Then, the student model can match these outputs in four different ways. More details are provided in Table 2.

Extraction – 5 Ways to get Outputs from the Teacher				
1	**2**	**3**	**4**	**5**
T labels unlabeled data, and these labels are used to train **S**	**T** makes new data and labels from some demonstrations, and this data is used to train **S**	**T** makes data based on a specific topic, and this data is used to train **S**	**T**'s internal knowledge on distributions and logits are extracted and **S** tries to match **T**	**T** generates feedback on **S**'s generations, which is used to update **S**
Distillation – 4 Ways to Put Knowledge Into Student				
1	**2**	**3**	**4**	
Supervised Fine Tunning	Divergence and Similarity	Reinforcement Learning – Rewarding Correctness	Ranking Optimization	

Table 1. Summary of Ways That Information Can Be Extracted from Teacher Models and Distilled into Student Models. DeepSeek used these methods to distill DeepSeek-R1 into Llama and Qwen.

The growing interest in and success researchers have had with distillation highlights a potential security vulnerability for U.S. organizations. Specifically, domestic entities serving AI models can have competitors steal intellectual property indirectly by matching their models' outputs to established models using teacher-student distillation. This is one example of a class of vulnerabilities that closed source model providers have been considering for the past two years, with various deployment-time interventions intended to mitigate its potential impact. Other examples of vulnerabilities include the potential to infer the contents of training data from model interactions or to learn non-public information about the model architecture. This type of vulnerability is particularly meaningful for U.S. government models trained on classified data. If foreign adversaries gain access to such models, they may be able to glean insights into secure government datasets. More machine learning security research is necessary to determine how acute these risks are and what the best means to combat them are.

Computation

No results of DeepSeek-R1 computation have been published so far. All computation information is based on speculation as published in the DeepSeek-V3 technical report.

Cost

Data from DeepSeek-V3, the base model used to train DeepSeek-R1, shows that the model cost less than $6 million to train. They assert some assumptions to justify the claim. First, DeepSeek-V3 was trained on a GPU cluster consisting of 2048 NVIDIA H800 GPUs. Specifically, the H800 is a card that was launched on 21 March 2023 in the Tesla Hopper generation. It has 80GB VRAM with 528 tensor cores and can perform 59.30 TFLOPS at floating-point-32 precision. At the rental price of a single H800 GPU card at $2/GPU/hour, DeepSeek-V3 costs 2.788M GPU hours (approximately 2 months) for its full training or $5.576 million on 14.8T tokens. In contrast, Llama 3 took almost 10 times longer to train, and ChatGPT-4 cost over $100 million to train on 25,000 A100 GPUs taking over 100 days. However, these metrics are also hard to compare, since DeepSeek used an FP8 type instead of FP16 or FP32. They also avoid using tensor parallelism, which has been necessary for all other pretrained models, instead opting to use their own low-level optimizations. It is possible that these two modifications could be critical in reducing the cost of pretraining. The cited $6 million value also only accounts for a single, end-to-end training run, neglecting any failed runs or ablations to arrive at the final model.

Lastly, DeepSeek's models were trained on their own GPU cluster. DeepSeek's parent company is a quantitative trading firm that presumably uses these GPUs for other tasks. As such, they can effectively train their LLMs during GPU downtime, reducing the effective cost of a training run. Put another way, the cost of purchasing their GPUs is spread between the primary trading business as well as the lesser DeepSeek models, allowing the developers to use GPUs "for free" when they aren't in use (relative to reserving them on Amazon Web Services [AWS]).

Scheduling

Each node in the 2048 cluster consisted of eight GPUs, indicating that there are 256 nodes. The authors refer to the "HAI LLM" framework as "an efficient and lightweight training framework crafted by our engineers from the ground up." This alludes to how the LLM is trained with careful engineering of distributed and parallel GPU processing across nodes and with some efficiencies in the way they use NVLink bandwidth and kernels across nodes. Most cutting-edge models do not use pipeline parallelism anymore and instead solely use tensor parallelism. U.S. export sanctions on GPUs make this less viable for DeepSeek, as efficient large-scale tensor parallelism requires having extremely high quality cross-node interconnect.

They describe DualPipe, which is a scheduling algorithm that ensures overlapping computation and communication during chunks of forward passes in the algorithm and backpropagation using parallelism. Each chunk of memory is divided into four components emulating a similar algorithm called "ZeroBubble" (30 November 2023) . ZeroBubble is a scheduling strategy published by the Chinese Sea AI Lab used for large-scale distributed training. The key idea is to split the backward computation (when gradients are updated in backpropagation) into two parts—one that computes the gradient for the input, and one that computes it for the parameters. In the figures below, each cell is a worker assigned with memory to hold tensors. The ZeroBubble paper shows many unused workers and siloed uses of applying the workers inefficiently. This contrasts to DualPipe to optimize the load and perform overlapping functions simultaneously across workers.

Combination of Efforts

The combination of efficiencies in algorithms, framework and hardware is a reason why training was faster. Some include numerical modifications such as FP8 (versus FP16 or 32), precision training, and quantization. Another reason is how DeepSeek-R1 uses MoE where the fully trained 671B parameter model only uses 37B "activated" parameters for each token. Yet another reason is that GRPO drops one of three traditional PPO models during RL training, thus saving cost and memory for yet another neural network.

The DeepSeek model seems to be focused on getting good performance on STEM/reasoning tasks. OpenAI and other LLM providers are focused on a much

[12] See, for example, Carlini et al. "Stealing Part of a Production Language Model." 2024.

[13] https://www.wired.com/story/openai-ceo-sam-altman-the-age-of-giant-ai-models-is-already-over/.

wider range of tasks, including essay-writing, poetry-writing, and more. It would be interesting to see if DeepSeek's MoE scales to working tasks that are not reasoning-based. It is possible that their architecture and training procedure has "just enough" experts to solve reasoning-based tasks but does not generalize well to creative tasks. In this vein, it is also worth thinking about how DeepSeek could use accuracy/format rewards to encourage creative responses. It is not clear how to measure the accuracy of a creative task. While a creative proxy-problem may be possible, it is unlikely the approach as-is enables efficacy on creative tasks alone.

Combination of Efforts

The combination of efficiencies in algorithms, framework and hardware is a reason why training was faster. Some include numerical modifications such as FP8 (versus FP16 or 32), precision training, and quantization. Another reason is how DeepSeek-R1 uses MoE where the fully trained 671B parameter model only uses 37B "activated" parameters for each token. Yet another reason is that GRPO drops one of three traditional PPO models during RL training, thus saving cost and memory for yet another neural network.

The DeepSeek model seems to be focused on getting good performance on STEM/reasoning tasks. OpenAI and other LLM providers are focused on a much wider range of tasks, including essay-writing, poetry-writing, and more. It would be interesting to see if DeepSeek's MoE scales to working tasks that are not reasoning-based. It is possible that their architecture and training procedure has "just enough" experts to solve reasoning-based tasks but does not generalize well to creative tasks. In this vein, it is also worth thinking about how DeepSeek could use accuracy/format rewards to encourage creative responses. It is not clear how to measure the accuracy of a creative task. While a creative proxy-problem may be possible, it is unlikely the approach as-is enables efficacy on creative tasks alone.

Performance

In the DeepSeek-R1 technical report, DeepSeek-V3 and DeepSeek-R1 are compared against Claude-3.5-Sonnet-1022, GPT-4o-0513, OpenAI-o1-mini, and OpenAI-o1-1217 for 21 different benchmarks. Benchmarks are published metrics and datasets that allow researchers to compare the performance of LLMs in a consistent, "apples-to-apples" method and are a requirement to demonstrate algorithmic improvements. But it is obvious that DeepSeek-R1 excels at math and some English reasoning tasks.

English, n=10	Code, n=5	Math, n=3	Chinese, n=3
MMLU (-Redux, -Pro), DROP, IF-Eval, GPQA Diamond, FRAMES, AlphaEval2.0, ArenaHard	LiveCodeBench, Codeforces (%, Rating), SWE Verified, Aider-Polyglot	AIME 2024, MATH-500, CNMO 2024	CLUEWSC, C-Eval, C-SimpleQA
• 6/10 benchmarks DeepSeek-R1 beat outcompetitors. • 3/10 benchmarks, OpenAI-o1-1217 won 1 benchmark, Claude won	• 1/5 DeepSeek-R1 won • 3/5 OpenAI-o1-1217 won • 6/10 benchmarks DeepSeek-R1 beat outcompetitors.	• 3/3 DeepSeek-R1 won	• 2/3 DeepSeek-R1 won • 1/3 DeepSeek-V3 won

Table 3: Asserted performance benchmarks in the DeepSeek-R1 technical report

[14] Qi, Penghui, et al. "Zero bubble pipeline parallelism." arXiv preprint arXiv:2401.10241 (2023).

[15] DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning, 22 January 2025.

DeepSeek-R1 Evaluation

	Benchmark (Metric)	Claude-3.5-Sonnet-1022	GPT-4o 0513	DeepSeek V3	OpenAI o1-mini	OpenAI o1-1217	DeepSeek R1
	Architecture	-	-	MoE	-	-	MoE
	#Activated Params	-	-	37B	-	-	37B
	#Total Params	-	-	671B	-	-	671B
English	**MMLU** (Pass@1)	88.3	87.2	88.5	85.2	**91.8**	90.8
	MMLU-Redux (EM)	88.9	88.0	89.1	86.7	-	**92.9**
	MMLU-Pro (EM)	78.0	72.6	75.9	80.3	-	**84.0**
	DROP (3-shot F1)	88.3	83.7	91.6	83.9	90.2	**92.2**
	IF-Eval (Prompt Strict)	**86.5**	84.3	86.1	84.8	-	83.3
	GPQA Diamond (pass@1)	65.0	49.9	59.1	60.0	**75.7**	71.5
	SimpleQA (Correct)	28.4	38.2	24.9	7.0	**47.0**	30.1
	FRAMES (Acc.)	72.5	80.5	73.3	76.9	-	**82.5**
	AlPacaEval2.0 (LC-winrate)	52.0	51.1	70.0	57.8	-	**87.6**
	ArenaHard (GPT-4-1106)	85.2	80.4	85.5	92.0	-	**92.3**
Code	**LiveCodeBench** (Pass@1-COT)	38.9	32.9	36.2	53.8	63.4	**65.9**
	Codeforces (Percentile)	20.3	23.6	58.7	93.4	**96.6**	96.3
	Codeforces (Rating)	717	759	1134	1820	**2061**	2029
	SWE Verified (Resolved)	**50.8**	38.8	42.0	41.6	48.9	49.2
	Aider-Polyglot (Acc.)	45.3	16.0	49.6	32.9	**61.7**	53.3
Math	**AIME 2024** (Pass@1)	16.0	9.3	39.2	63.6	79.2	**79.8**
	MATH-500 (Pass@1)	78.3	74.6	90.2	90.0	96.4	**97.3**
	CNMO 2024 (Pass@1)	13.1	10.8	43.2	67.6	-	**78.8**
Chinese	**CLUEWSC**(EM)	85.4	87.9	90.9	89.9	-	**92.8**
	C-Eval (EM)	76.7	76.0	86.5	68.9	-	**91.8**
	C-SimpleQA (Correct)	55.4	58.7	**68.0**	40.3	-	63.7

Figure 4: Evaluation Benchmark Results from the DeepSeek-R1 GitHub Repository Technical Documentation

For the smaller models, the distilled versions of Qwen (1.5, 7, 14, 32B) models and Llama (8B, 70B), there are only five benchmarks: AIME 2024 and MATH 500 (Math tasks), GPQA Diamond (English tasks), and LiveCodeBench and CodeForces (live coding tasks). For the three scores pertaining to math, the distilled models outperformed GPT-4o-0513, Claude-3.5-Sonnet-1022, OpenAI-o1-mini, and QwQ-32B-Preview. The distilled models also outperformed the comparable models on GPQA and LiveCode Bench, but OpenAI-o1-mini achieved the highest ranking on CodeForces.

Model	AIME 2024		MATH-500	GPQA Diamond	LiveCode Bench	CodeForces
	pass@1	cons@64	pass@1	pass@1	pass@1	rating
GPT-4o-0513	9.3	13.4	74.6	49.9	32.9	759
Claude-3.5-Sonnet-1022	16.0	26.7	78.3	65.0	38.9	717
OpenAI-o1-mini	63.6	80.0	90.0	60.0	53.8	**1820**
QwQ-32B-Preview	50.0	60.0	90.6	54.5	41.9	1316
DeepSeek-R1-Distill-Qwen-1.5B	28.9	52.7	83.9	33.8	16.9	954
DeepSeek-R1-Distill-Qwen-7B	55.5	83.3	92.8	49.1	37.6	1189
DeepSeek-R1-Distill-Qwen-14B	69.7	80.0	93.9	59.1	53.1	1481
DeepSeek-R1-Distill-Qwen-32B	**72.6**	83.3	94.3	62.1	57.2	1691
DeepSeek-R1-Distill-Llama-8B	50.4	80.0	89.1	49.0	39.6	1205
DeepSeek-R1-Distill-Llama-70B	70.0	**86.7**	**94.5**	**65.2**	**57.5**	1633

Figure 5: Evaluation Benchmark Results from the DeepSeek-R1 GitHub Repository Technical Documentation

Assessment of Technical Claims

Training Costs

There is no real way to validate the budget of their hardware. But here are some comparisons. Currently on AWS, an H100 card has 80GB VRAM similar to the H800. As of late December 2024, a p5.48xlarge AWS cloud GPU server (H100 80GB x8) costs $98.32 per instance as of the time of this writing, which is approximately $12/hour/GPU. Using this basis as an example, the pre-training costs would be six times greater than $5.328M. The DeepSeek-R1 paper contemplates $2/hour/GPU, which is on par, not with an 80GB VRAM card, but with a U.S. AWS p3.2xlarge server, which translates into a V100 16GB x1—a single 16GB VRAM card that costs $3.06. However, realizing that calculating exact GPU costs is extremely difficult, and multiple factors are involved in pricing.

Training Costs	Pre-Training	Context Extension	Post-Training	Total
In H800 GPU Hours	2664K	119K	5K	2788K
In USD	$5.328M	$0.238M	$0.01M	$5.576M

Table 4: Asserted GPU Training Time and Costs based on data presented in the DeepSeek-V3 Paper

Lastly, they have manpower. When we counted the number of contributors mentioned on the last three pages of the DeepSeek-R1 paper, we arrived at somewhere over 200 contributors. This is hardly a bootstrapped startup; rather, it is an industrial lab. The cost that matters in most contexts includes headcount and the full cluster, which was used to train many models, built by many expensive experts, over a long time.

Benchmarks

Benchmarks are widely accepted in AI academia as tests to measure the performance of a claimed, novel algorithm. Furthermore, any researcher, student, or citizen is free to access the models (all listed in the timeline) through Hugging Face, by downloading the model directly and also viewing the code used to train the models. For example, the entire training script written in Python for DeepSeek-MoE is available for free on Github. Although it takes an AI expert to fully understand and run the code, the release of these models and their training scripts and utilities provides an obvious, unfiltered way to test every benchmark. However, DeepSeek has not released the training code specific to DeepSeek-R1, the reinforcement learning fine-tuning. For this reason, it may be difficult to replicate the fine-tuning process in DeepSeek-R1, but, since the models have been released, any researcher can verify the benchmark metrics DeepSeek claims in their papers.

However, careful examination of benchmarks reveals that DeepSeek-R1 excels at math, logic, and code related tasks. This makes sense given the fundamental nature of existing RL algorithms. Most RL algorithms require a well-defined problem, with a well-defined target. Tasks such as generating poetry are unfit for RL because the subjective nature of what is “right” is inappropriate. As a result, when examining their multi-stage pipeline, they purposely curate 600,000 samples that are “reasoning”-based and retrain their RL model against these samples. Compare this to a paltry 200,000 samples used for “non-reasoning.” It is fair to assume that DeepSeek-R1 is not genuinely a rival to OpenAI-mini, but a very good task-oriented STEM reasoner.

Allegations of Data Theft and the Bigger Picture

OpenAI and Microsoft are claiming that DeepSeek was constructed via the theft of their data. While they have not provided any public evidence of this, there are various reports online of DeepSeek responding that it is "ChatGPT" when asked. The accusations underscore the importance of a larger discussion and context around the difficulties and risks in making LLM outputs freely available. There are many ways in which models can have been benignly influenced or been intentionally copied.

First, DeepSeek's model claiming that it is "ChatGPT" is not unique to DeepSeek. This same phenomenon has been observed with most LLMs released since ChatGPT. This is in large part because there is a massive amount of ChatGPT logs, verbatim examples, human written and ChatGPT augmented, and ChatGPT written and human augmented, text released onto the internet since OpenAI made their model public. We wrote almost a year ago about how this "poisons the well" for unadulterated "purely human" text, biasing all subsequent models created to that of OpenAI's. Even with curated filters, it will be challenging to avoid some amount of bias due to people using such systems in editing and copy work.

DeepSeek also used a "judge" LLM, where a pretrained LLM is used as an arbiter in deciding if an input text does/does not satisfy some criterion. It is entirely possible, and even reasonable, for them to have used a different LLM other than their own as the judge. This may be done in an attempt to mitigate bias risks from their own systems using and creating a negative feedback cycle. However, it could certainly look like heavy API usage that would not necessarily appear benign from looking at API accesses. Broadly, there may be many data-processing tasks that fall along a spectrum of benign and acceptable use that would be easy to construe as malicious due to heavy API use.

To date, no public evidence is available to make any special informed judgment. We are using these claims to talk about the bigger picture, in that many behaviors that are likely benign will be hard to trace for any new LLM generated. In this same vein, preventing intentional "theft" is also fraught with challenges.

The underlying weights that control the model are not accessible, nor is the model provider (e.g., OpenAI) deprived of their own model. The question from a malicious use perspective is how much information one can intuitionally leak from a victim model. This could be achieved by prompting the victim model, copying its outputs, and using that as SFT data for your new competitor. Information leakage is an open question, and the more effective the attacker is, the more they can "steal" with a limited budget. There are also legal questions around the copyright, terms of service, and ownership of LLM-generated output in this context that are matters of the court that are currently being litigated, beyond the scope of this primer.

In the context of the current allegations, it is important to note that DeepSeek's primary improvement is in generated reasoning-like capabilities similar to OpenAI's o1 model. However, OpenAI does not disclose these "reasoning" tokens from the end user.

Finally, while some work on watermarking exists, it has been in the context of detecting text from an LLM, not in detecting if another LLM was trained on such text. A similar risk is in protecting the original data used for training to avoid leaking specific and sensitive information. The theoretical tool needed for such protection is known as "Differential Privacy" (DP), although it often requires further development to make it practical to real-world use rather than just theoretically useful. In both cases, there are tools and methods that could be developed for simpler models like linear regression, but the scale of LLMs makes them non-trivial to apply for these current needs.

[16] https://www.ft.com/content/a0dfedd1-5255-4fa9-8ccc-1fe01de87ea6

[17] https://www.yahoo.com/tech/deepseek-just-insisted-chatgpt-think-161449975.html?guccounter=1

[18] https://www.manning.com/books/how-gpt-works

[19] John Kirchenbauer et. al., "On the reliability of watermarks for large language models." https://openreview.net/forum?id=DEJIDCmWOz.

Conclusion

DeepSeek's latest release is a conflation of DeepSeek-R1-Zero, which has been purely trained by RL, and DeepSeek-R1, which has used a combination of SFT and two rounds of RL to train a 671B LLM. DeepSeek has been developing and releasing models since January 2024 at a fast clip. DeepSeek-R1 claims to be comparable in performance to LLMs from OpenAI, Meta, and Anthropic, as it excels in math, logic, and code benchmarks. There are many AI engineering functions they have optimized from the algorithm, training pipeline, and hardware, so that they could train with fewer GPUs (2048 H800s). Many of these methods are not novel to the AI community, but rather a clever application of existing research tools. Their claim to not use SFT and rely only on RL is true for DeepSeek-R1-Zero, but DeepSeek-R1 (the 671B model) uses SFT and RL, combined. Their claim that the cost is $6 million is misleading because the figure is derived from old numbers in the DeepSeek-V3 paper—the base model of DeepSeek-R1. Further, there is no way to easily validate the cost, which can vary wildly. Lastly, there are technical details that are missing from the DeepSeek-R1 paper and would provide additional insight into training. Some include the nature of the reward model, the computational scheduler used to optimize training, the manner and model from which they collected 800K SFT samples, and how they exactly evaluated for "helpfulness" and "harmlessness."

Authors

—Catherine Ordun, Ph.D.

—Edward Raff, Ph.D.

—Stella Biderman

—John Larson

—Alison Smith

—Amol Khanna

—Ryan Swope

—Tyler Nivin

Booz Allen®

About Booz Allen

Booz Allen is the advanced technology company delivering outcomes with speed for America's most critical defense, civil, and national security priorities. We build technology solutions using AI, cyber, and other cutting-edge technologies to advance and protect the nation and its citizens. By focusing on outcomes, we enable our people, clients, and their missions to succeed—accelerating the nation to realize our purpose: Empower People to Change the World®.