FEB 5, 2025

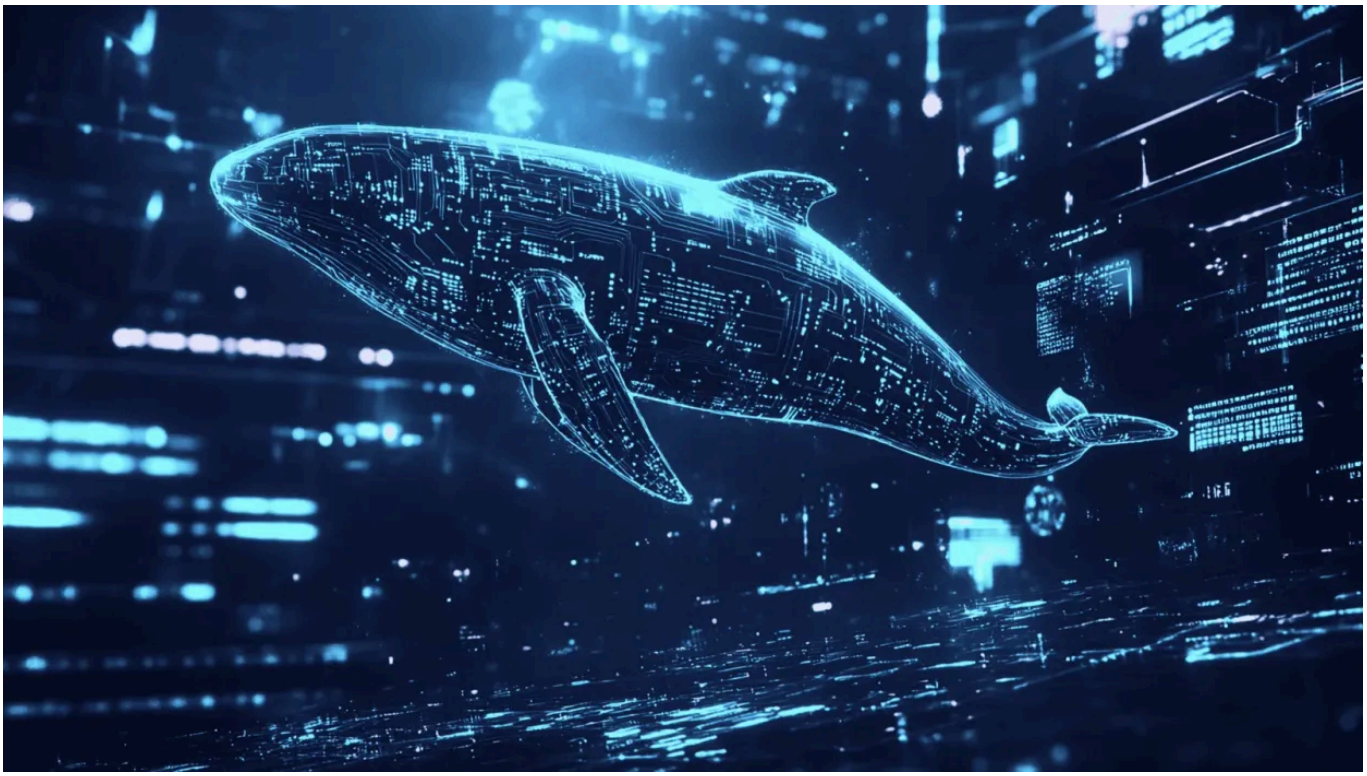
DeepSeek: A Technical and Strategic Analysis for VCs and Startups

A technical deep-dive into DeepSeek's AI breakthroughs and their implications for startup strategy in 2025.



By Jenny Xiao and Jay Zhao

Leonis Capital · leoniscap.com/research/deepseek-a-technical-and-strategic-analysis-for-vcs-and-startups/

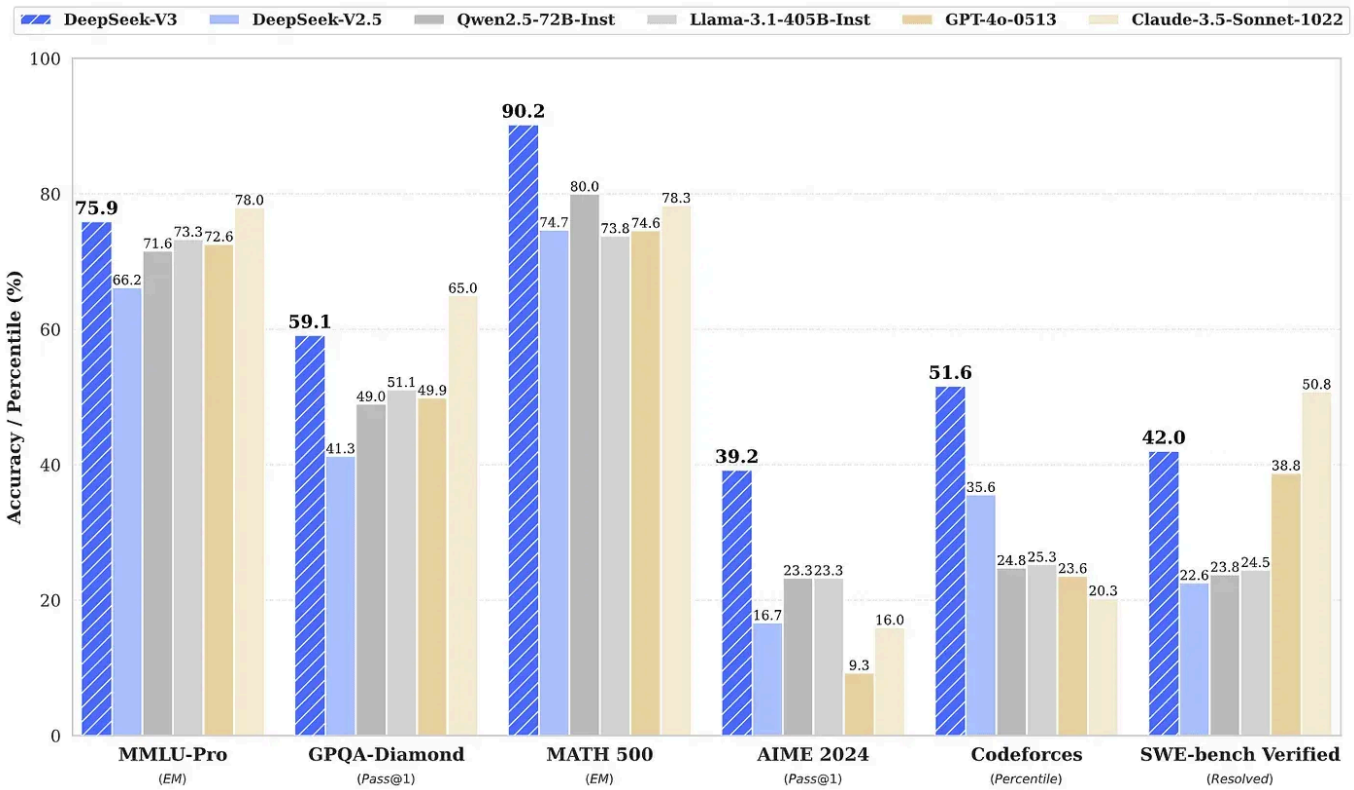


DeepSeek has fundamentally reshaped the economics of AI development, achieving state-of-the-art performance at a fraction of the cost of its U.S. competitors. While the "$5.5M to match OpenAI" narrative is totally overblown—the company's actual infrastructure spending is probably closer to $1B—their breakthrough remains remarkable. DeepSeek's innovations in model architecture and training efficiency demonstrate that AI development and deployment can be dramatically more cost-effective than the traditional approach taken by companies like OpenAI and Anthropic, even if not quite as revolutionary as the headlines suggest.

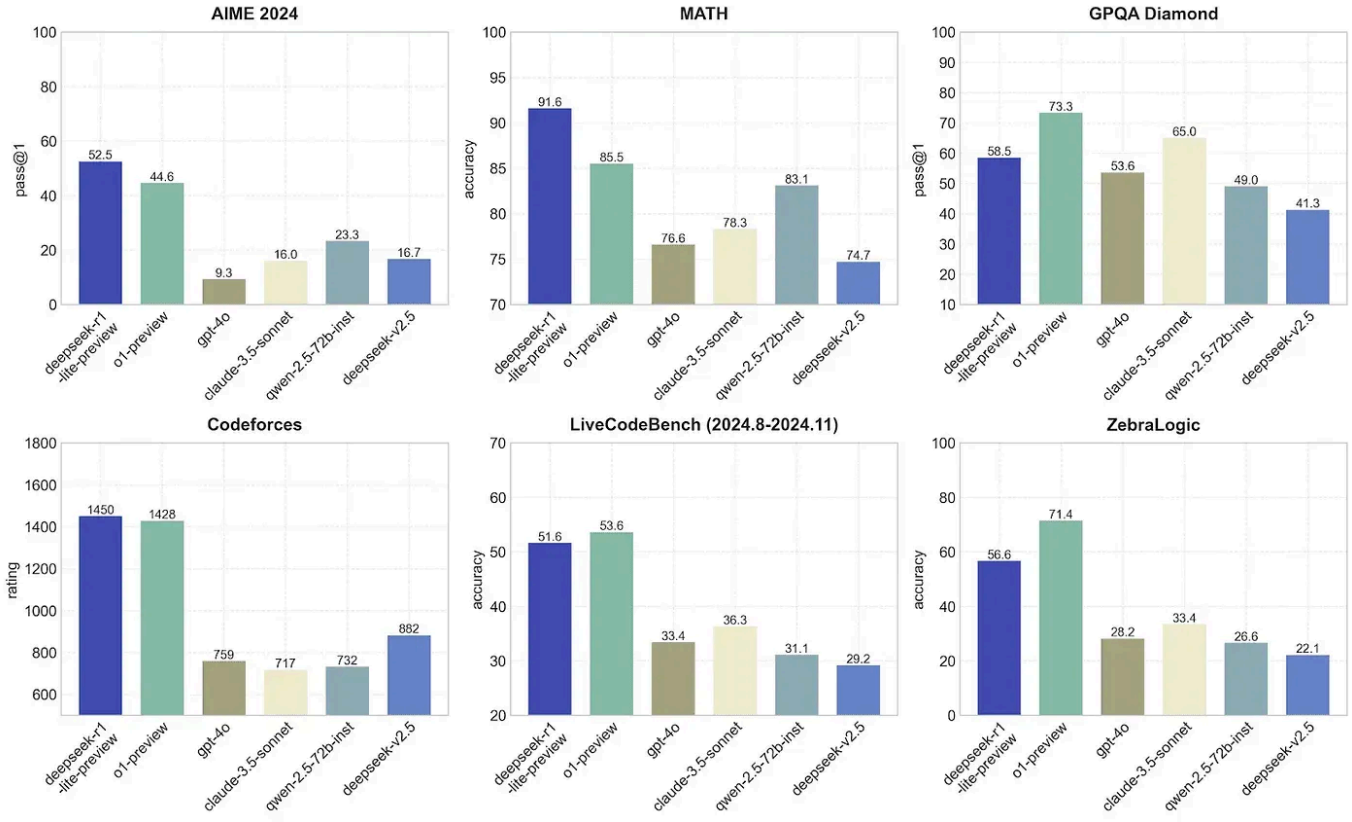
The implications extend far beyond a single company's achievements. For startups, investors, and technology leaders, DeepSeek's emergence signals a fundamental shift in how AI companies build competitive advantages. As base models become more efficient and accessible, we're seeing the rise of what we call **"Moat 2.0"**—a new paradigm where competitive advantages come not from raw compute power or massive datasets, but from how companies build, learn from, and deploy AI systems in sophisticated ways. This shift suggests that the next wave of AI leaders won't be determined by who has the most resources, but by who can most creatively deploy and optimize AI systems for specific use cases.

A Technical Primer

To understand DeepSeek's breakthrough, it's important first to recognize that we're actually looking at two distinct models: DeepSeek-V3, their base model, and DeepSeek-R1, their reasoning-focused model. This relationship mirrors OpenAI's GPT-4o and o1—in both cases, a powerful base model serves as the foundation for a more specialized reasoning model. While DeepSeek-R1 has captured recent headlines and triggered a 17% drop in NVIDIA's stock price, **it's actually DeepSeek-V3 that represents the more significant technical breakthrough**, achieving GPT-4o-level performance for its reported $5.5M training cost, while DeepSeek-R1 competes with OpenAI's o1 on reasoning tasks (see charts below).



DeepSeek-V3 on performance benchmarks. V3 matches or outperforms GPT-4o on many benchmarks.



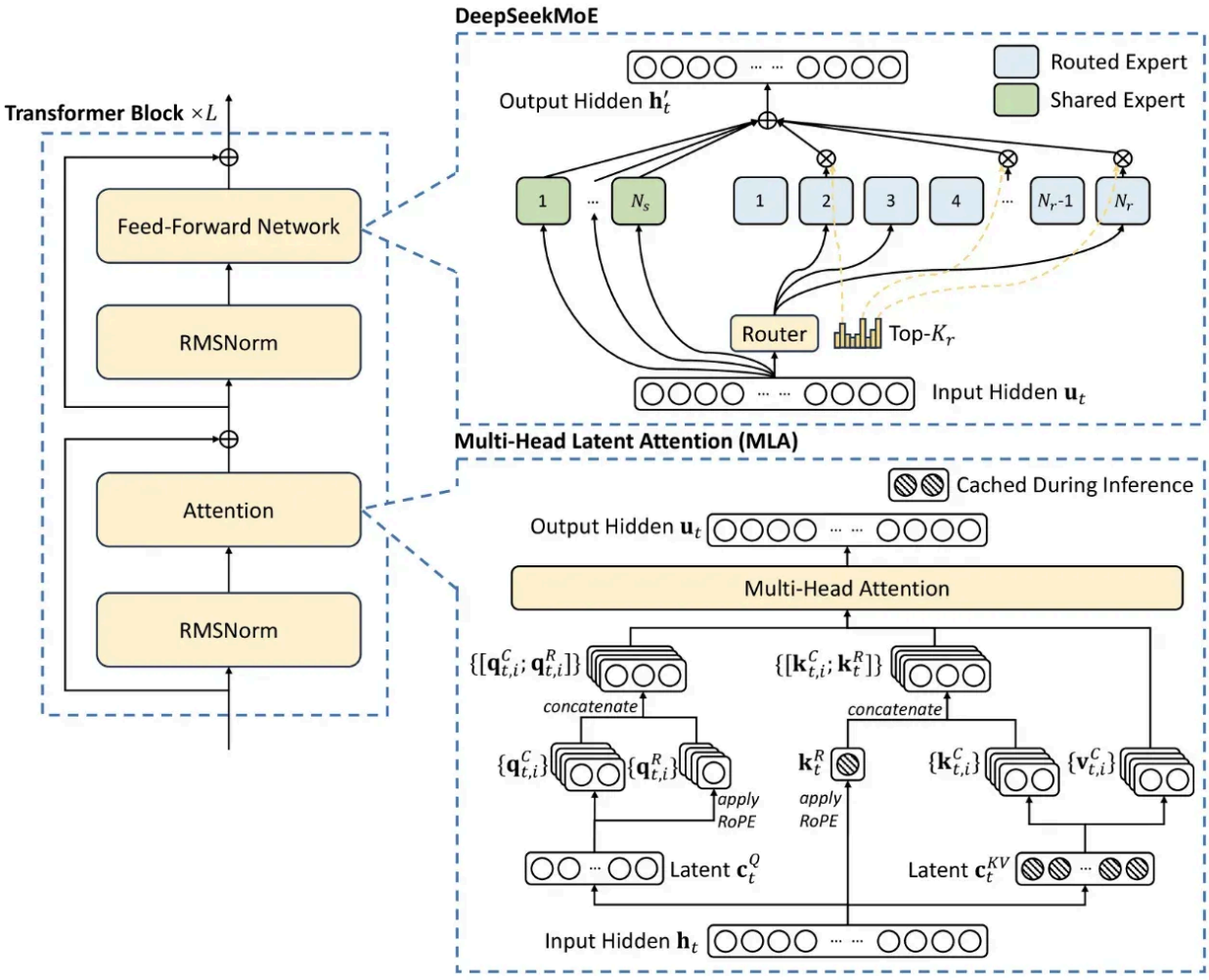
DeepSeek-R1 on performance benchmarks. R1 matches o1 on many reasoning tasks.

In the following sections, we'll break down DeepSeek's key innovations in plain English, making them accessible even if you don't have a technical background. From their novel approach to model architecture to their breakthroughs in memory efficiency, these advances help explain how DeepSeek achieved competitive performance at a fraction of the traditional cost.

> Note: For readers more interested in business implications than technical details, feel free to skip ahead to the analysis and impact sections. The key takeaway of the technical side is that DeepSeek achieved competitive performance at a fraction of traditional costs through innovative architectural design. Here is a short summary of the takeaways:
>
> 1. **Mixture of Experts (MoE)** → Specialized experts that are selectively activated.
> 2. **Multi-Head Latent Attention (MLA)** → Efficient memory storage
> 3. **GRPO (Group Relative Policy Optimization)** → Learning through comparative feedback on output quality
> 4. **Distillation** → Letting a small model absorb knowledge from a large model

Mixture of Experts: A New Paradigm for Model Architecture

The cornerstone of DeepSeek's approach is their innovative implementation of Mixture of Experts (MoE). Unlike traditional models that activate all parameters for every prediction, **DeepSeek-V3 activates only 37 billion parameters out of its total 671 billion for each token**. This selective activation is made possible through their auxiliary-loss-free load balancing system, which uses dynamic bias adjustments to improve training stability and efficiency. For founders and investors familiar with distributed systems, this is analogous to how modern cloud architectures dynamically allocate resources—but applied at the neural network level.



DeepSeek's architecture. This architecture cleverly combines selective expert activation (MoE) with efficient memory management (MLA) to achieve high performance at a lower cost.

What makes this particularly remarkable is their DualPipe system for pipeline parallelism. This innovation solved one of the most challenging aspects of distributed MoE models: managing the complex routing of information across different expert networks. DualPipe achieves this by overlapping computation and communication phases—imagine a highly sophisticated assembly line where products are being processed and transported simultaneously rather than sequentially. The result is a more efficient and scalable architecture that processes an astonishing 14.8 trillion tokens during pre-training with nearly zero communication overhead.

What's especially unique about DeepSeek's implementation is their use of custom PTX instructions—essentially assembly language for NVIDIA GPUs—to optimize these operations. This level of low-level hardware optimization is extremely rare in AI development, where most

researchers work with high-level frameworks like Nvidia's CUDA platform. **The capability likely stems from their parent company High Flyer's background in high-frequency trading, where writing such low-level code is crucial for competitive advantage.** This combination of AI expertise with deep hardware optimization skills represents a significant competitive advantage in the race for AI efficiency.

Breaking the Memory Wall

DeepSeek's **Multi-Head Latent Attention (MLA) mechanism** represents a breakthrough in memory efficiency, reducing memory overhead by an astounding 93.3% compared to standard attention mechanisms. Originally introduced in DeepSeek-v2, this innovation specifically targets the KV Cache—a memory-intensive component that stores conversation context. By dramatically reducing these memory requirements, MLA makes inference significantly more cost-effective, enabling longer conversations without proportional increases in computational costs. The innovation has proven so significant that it caught the attention of leading US labs, and DeepSeek has further optimized it for their H20 GPUs, achieving even better memory bandwidth and capacity utilization than on H100s.

Complementing this is their Multi-Token Prediction (MTP) system, which enables the model to predict multiple tokens simultaneously. **The DeepSeek team has implemented MTP at a scale previously unheard of.** This isn't just about speed—it fundamentally changes the efficiency equation for both training and inference. Combined with their implementation of FP8 mixed precision training, these innovations allowed DeepSeek to achieve their remarkable $5.5M training budget.

The Reinforcement Learning Revolution

Perhaps most intriguing is DeepSeek's approach to model improvement through pure reinforcement learning (RL). Their R1 model introduces **GRPO (Group Relative Policy Optimization), which enables the model to optimize its behavior without explicit correct answers.** Compared to other RL models, GRPO optimizes directly for correctness without complex reward models. The efficiency of this approach is evident in the model's learning trajectory: as shown in the graph below, DeepSeek's R1 demonstrates a steady upward curve in performance over time. The rate of improvement is pretty remarkable, eventually overtaking OpenAI's o1.

Model	AIME 2024		MATH-500	GPQA Diamond	LiveCode Bench	CodeForces
	pass@1	cons@64	pass@1	pass@1	pass@1	rating
OpenAI-o1-mini	63.6	80.0	90.0	60.0	53.8	1820
OpenAI-o1-0912	74.4	83.3	94.8	77.3	63.4	1843
DeepSeek-R1-Zero	71.0	86.7	95.9	73.3	50.0	1444

Table 2 | Comparison of DeepSeek-R1-Zero and OpenAI o1 models on reasoning-related benchmarks.

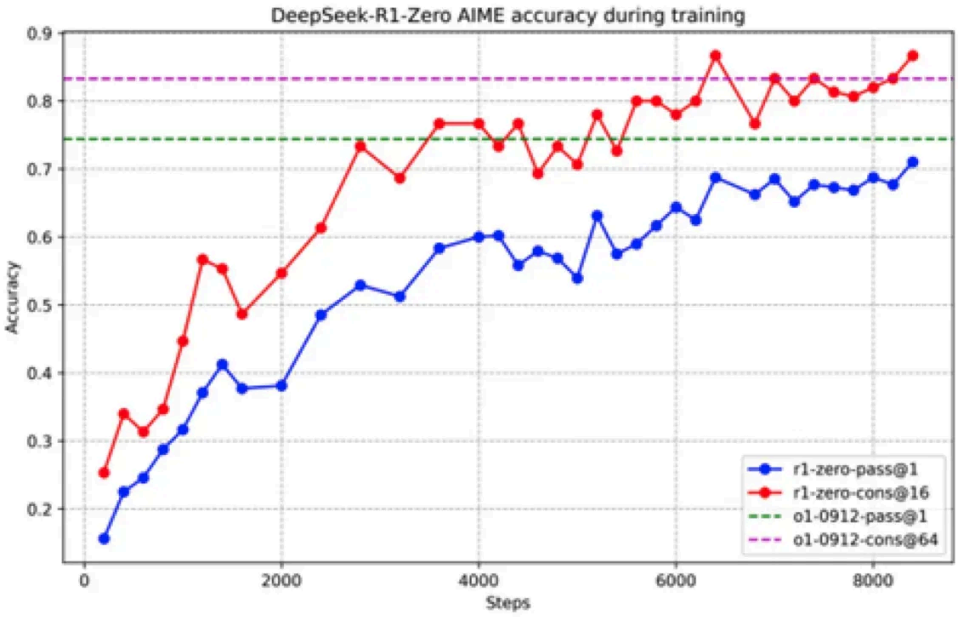
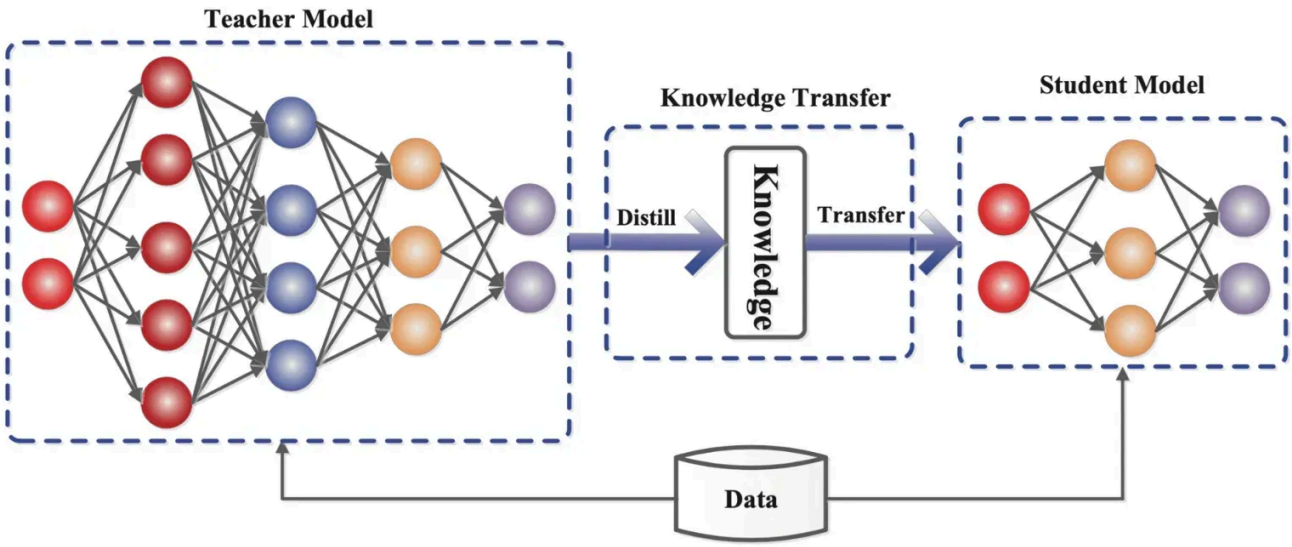


Figure 2 | AIME accuracy of DeepSeek-R1-Zero during training. For each question, we sample 16 responses and calculate the overall average accuracy to ensure a stable evaluation.

The GRPO implementation focuses on two key areas: formatting (ensuring coherent output) and helpfulness/harmlessness optimization. Reasoning capabilities emerged during fine-tuning on synthetic datasets, similar to o1's development path. However, what's particularly noteworthy is that while competing approaches often require extensive compute for reward modeling, GRPO achieves similar results with a more streamlined approach. This self-improvement capability is particularly effective in domains requiring complex reasoning, such as mathematical problem-solving and coding.

Democratizing AI Through Distillation

DeepSeek's impressiveness isn't just their large model's performance——it's their ability to create smaller, more accessible versions through distillation. In the distillation process, a large "teacher" model provides training data for a small "student model." DeepSeek's 7B parameter model—roughly 100 times smaller than their full system—achieves remarkable performance, outperforming many larger models including QwQ-32B-Preview on key benchmarks. Their approach suggests that reasoning patterns from larger models can transfer effectively to smaller ones. For startups and developers, this means access to near state-of-the-art capabilities without the massive infrastructure costs typically associated with running large AI models.



A simple illustration of how distillation works. Source: Neptune AI.

However, DeepSeek's perfection of the distillation technique has also sparked controversy. OpenAI, along with Microsoft, has raised concerns that DeepSeek may have used OpenAI's API outputs for distillation—essentially having their smaller model learn from GPT-4o and o1's responses. While this practice (known as black-box distillation) doesn't require access to the original model's internal workings, it is in direct violation of OpenAI's terms of service. **Ironically, OpenAI has also faced lawsuits over using copyrighted content to train its models.**

Bringing It All Together

DeepSeek's architectural innovations work together harmoniously, each component enhancing and complementing the others to create a highly efficient and powerful system. The **MoE** efficiently directs information to specialized expert networks, activating only the most relevant parameters for each task, while **MLA** compresses and stores key information, reducing memory overhead and enabling the processing of longer sequences. **GRPO** continuously refines the

model's decision-making by comparing output quality and making targeted updates, allowing for performance improvement without explicitly labeled data. Finally, **distillation** transfers the capabilities of the large model to smaller, more practical versions, democratizing access to state-of-the-art AI and making it more accessible and affordable for a wider range of applications.

In essence, DeepSeek has not only pushed the boundaries of what is possible with AI but also redefined the way we think about the relationship between model size, computational cost, and performance. **This rare combination of engineering excellence and technological innovation positions DeepSeek as a true trailblazer in AI.**

The Real Economics of DeepSeek

While headlines tout DeepSeek's $5.5M training cost, the reality of their investment is far more substantial. Indeed, their own technical paper acknowledges that **the $5.5M figure only accounted for the costs of running the GPUs for training DeepSeek-V3, not prior research, infrastructure, or operating costs**. DeepSeek also has many more GPUs than the 2,048 H800s that they mention in their paper, as AI labs need many additional GPUs to run experiments and research before committing to a full training run. Notably, DeepSeek has been selective in their disclosures, releasing cost figures only for their V3 model while remaining silent about R1's training costs—possibly to avoid drawing attention to their true GPU infrastructure scale, particularly given export control sensitivities.

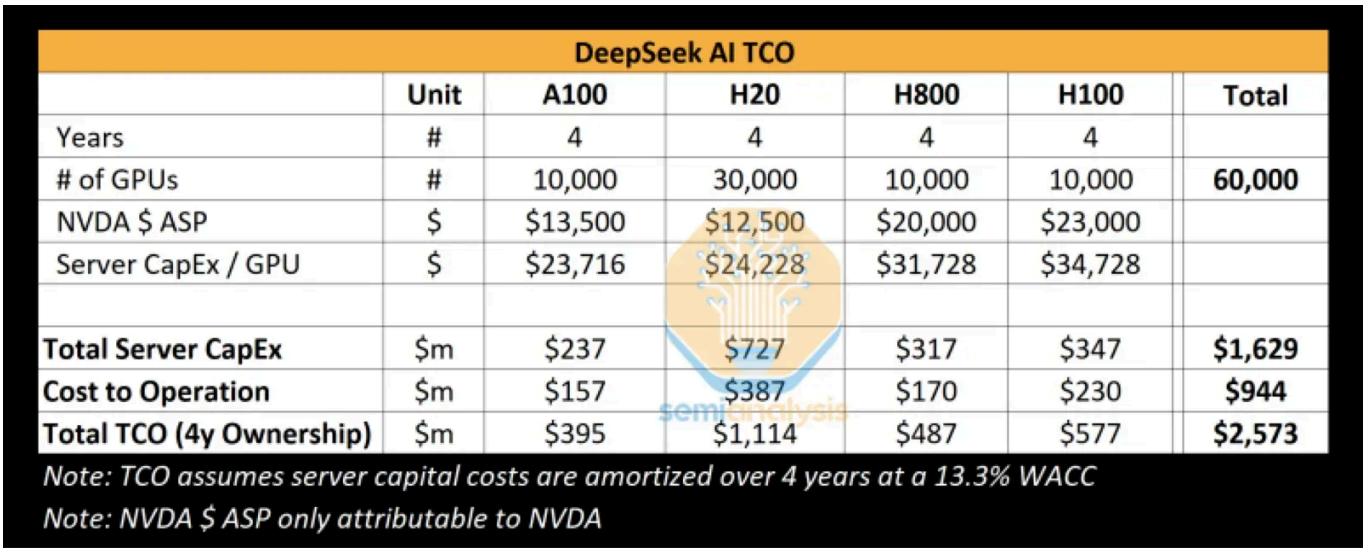
Training Costs	Pre-Training	Context Extension	Post-Training	Total
in H800 GPU Hours	2664K	119K	5K	2788K
in USD	$5.328M	$0.238M	$0.01M	$5.576M

Table 1: Training costs of DeepSeek-V3, assuming the rental price of H800 is $2 per GPU hour.

Lastly, we emphasize again the economical training costs of DeepSeek-V3, summarized in Table 1, achieved through our optimized co-design of algorithms, frameworks, and hardware. During the pre-training stage, training DeepSeek-V3 on each trillion tokens requires only 180K H800 GPU hours, i.e., 3.7 days on our cluster with 2048 H800 GPUs. Consequently, our pre-training stage is completed in less than two months and costs 2664K GPU hours. Combined with 119K GPU hours for the context length extension and 5K GPU hours for post-training, DeepSeek-V3 costs only 2.788M GPU hours for its full training. Assuming the rental price of the H800 GPU is $2 per GPU hour, our total training costs amount to only $5.576M. Note that the aforementioned costs include only the official training of DeepSeek-V3, excluding the costs associated with prior research and ablation experiments on architectures, algorithms, or data.

DeepSeek-V3's technical report on the cost of training the model.

According to SemiAnalysis's detailed investigation, **DeepSeek's total server infrastructure investment approaches $1.6B**, with approximately $500M dedicated specifically to GPU hardware. This breaks down into their extensive GPU fleet: roughly 50,000 Hopper GPUs distributed across different variants—10,000 H800s, 10,000 H100s, and the remainder being H20s (the China-specific version of Hopper GPUs that Nvidia produced in compliance with export regulations). The $1.6B figure encompasses not just the GPUs but the complete server systems, networking equipment, and data center facilities needed to operate at this scale. Beyond the capital expenditure, they incur around $944M in operating costs to maintain these systems, including offering up to $1.3M in compensation packages for top talent, a number that dwarfs the typical compensation for AI researchers in top labs like China like Zhipu or Moonshot.



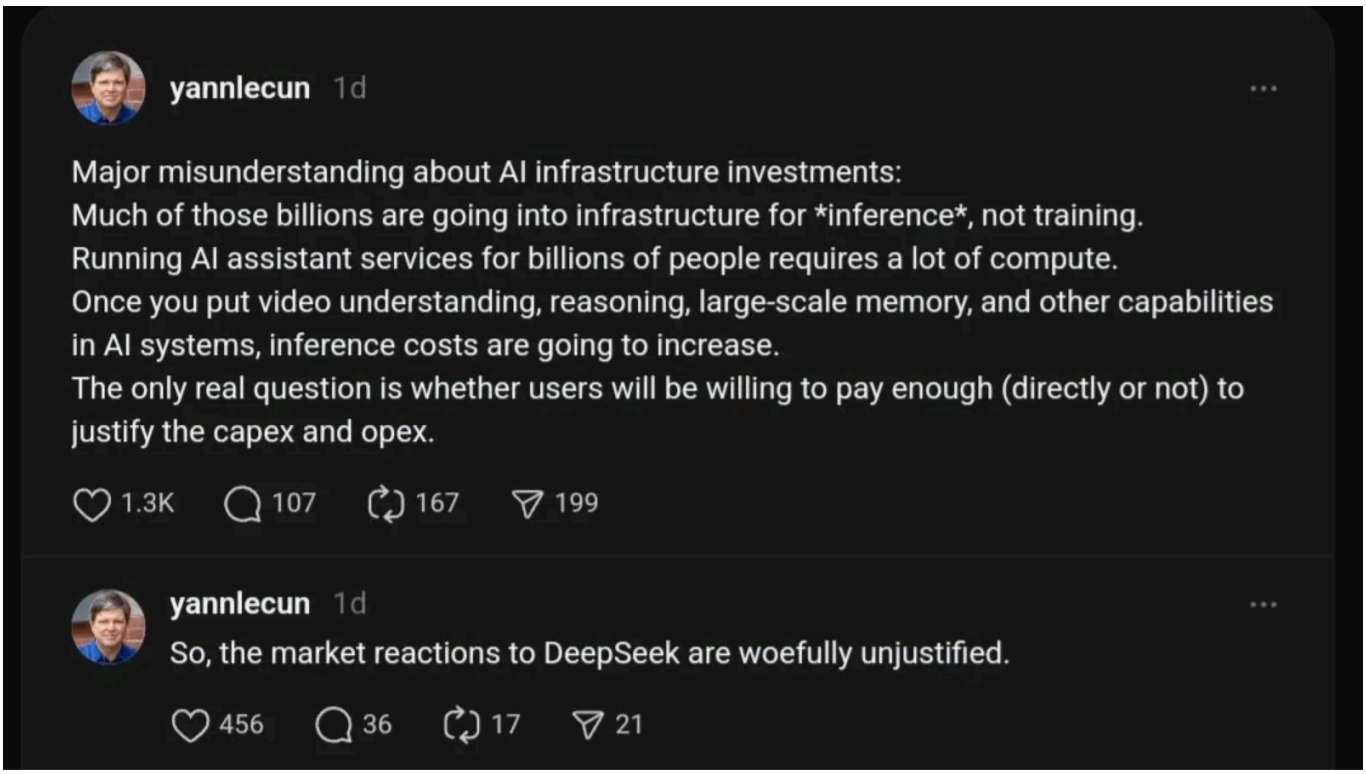
DeepSeek AI TCO						
	Unit	A100	H20	H800	H100	Total
Years	#	4	4	4	4	
# of GPUs	#	10,000	30,000	10,000	10,000	**60,000**
NVDA $ ASP	$	$13,500	$12,500	$20,000	$23,000	
Server CapEx / GPU	$	$23,716	$24,228	$31,728	$34,728	
Total Server CapEx	$m	$237	$727	$317	$347	**$1,629**
Cost to Operation	$m	$157	$387	$170	$230	**$944**
Total TCO (4y Ownership)	$m	$395	$1,114	$487	$577	**$2,573**

Note: TCO assumes server capital costs are amortized over 4 years at a 13.3% WACC

Note: NVDA $ ASP only attributable to NVDA

> *SemiAnalysis's estimations of DeepSeek's GPU resources.*

The company's ability to invest significant CapEx stem from their unique relationship with High Flyer, one of China's leading quant hedge funds. This partnership, which began when DeepSeek spun off from High Flyer in May 2023, enables shared resource utilization between AI development and quantitative trading operations. Notably, High Flyer purchased 10,000 A100 GPUs in 2021, a major AI infrastructure investment ahead of other Chinese AI labs and before any GPU export controls.



Meta's Chief AI Scientist Yann LeCun points out the training vs. inference cost difference in a Tweet (or more accurately, Thread).

Despite these substantial investments, DeepSeek's approach still represents a marked efficiency improvement over U.S. competitors. To put DeepSeek's $1.6B number in perspective, OpenAI has raised a total of $17.9B as of February 2025, a number similar to Anthropic's total funding raised so far. OpenAI is rumored to be targeting another $40B in new funding this year. This means that DeepSeek is achieving competitive performance at perhaps **one-tenth the infrastructure cost of major U.S. AI labs** (though not one-thousandth as some interpretations of the $5.5M figure might suggest).

Nevertheless, it is noteworthy to point out that **the vast majority of U.S. AI labs' infrastructure investments actually go towards inference**—preparing for future deployment and scaling of their models—rather than training costs. DeepSeek's $5.5M figure only accounts for the training costs, but the total inference and training cost is closer to the $1B mark. **Comparing the cost of training a single model to building an entire future-ready AI infrastructure is extremely misleading.** Understanding this full economic picture is crucial for grasping both the scale of DeepSeek's achievement and its implications for the AI industry.

The End of Scaling Laws?

The widespread narrative around DeepSeek's achievements has fueled speculation about the end of AI scaling laws. The industry rightly questions whether bigger is always better for model performance. But this view misses a crucial transformation: **we're not seeing the end of scaling, but rather its evolution into new dimensions**. As we mentioned in our 2025 State of AI report, the industry is moving from an era dominated by pre-training scaling to one where post-training optimization offers immense untapped potential.

Anthropic's CEO Dario Amodei also notes that **the opportunities for scaling in the post-training phase could yield even greater improvements than traditional pre-training scaling**. From 2020 to 2023, the main thing being scaled was pre-trained models trained on internet text. As the performance improvements of this brute-force approach came to an end, AI companies like OpenAI and Anthropic turned to using RL and chain-of-thought as the new focus of scaling. This approach proved especially effective in improving performance on measurable tasks like math, coding, and scientific reasoning. In 2025, we can expect much more from this post-training-centered scaling, including more sophisticated RL approaches, better synthetic data generation, and more efficient inference-time computation—areas where we're just beginning to scratch the surface.

Yet DeepSeek's achievements do demonstrate that the relationship between spending and performance isn't linear. **Through architectural innovations like MoE and MLA, they're achieving more with less—similar to how the semiconductor industry evolved when it hit the CPU clock speed wall in the late 2000s**. Just as chip makers found new vectors for improvement beyond raw clock speeds—multi-core architectures, advanced packaging, and parallel computing—AI development isn't hitting a wall so much as diversifying its scaling vectors.

The actions of major AI labs support this more nuanced view. Major AI labs are continuing massive infrastructure investments. OpenAI just announced its $500B Stargate infrastructure program, xAI rolled out its 100k GPU Colossus supercomputer, Meta is planning 2GW data centers for 2026, and other companies are also projecting multi-billion dollar investments in AI development. These investments suggest that scaling isn't dead—it's evolving. The future likely holds both: continuing efficiency gains through architectural innovation alongside massive infrastructure investments to push the boundaries of what's possible.

DeepSeek's Impact on Different Market Segments

DeepSeek's emergence represents a pivotal moment in AI's evolution, but its impact varies dramatically across different segments of the market. Rather than simply challenging U.S. AI dominance, DeepSeek's innovations are reshaping the competitive dynamics of the entire

industry in complex and often unexpected ways. Here, we will dive into DeepSeek's impact on different players in the AI landscape.

Closed-Source Model Providers: The Valuation and Pricing Challenge

DeepSeek's breakthrough fundamentally challenges the narrative that has supported massive valuations in AI. OpenAI's $157B valuation and Anthropic's fundraises were built on the assumption that competitive AI development requires tens of billions in investment, creating natural barriers to entry. However, DeepSeek's achievement—even accounting for their true $1.6B infrastructure cost—suggests **these AI companies' moats are less defensible than investors assumed**, representing roughly one-tenth of what U.S. competitors have spent to achieve similar capabilities.

The pressure extends beyond valuations to core business models. While DeepSeek's models might not match OpenAI and Anthropic's capabilities exactly, they're reaching a threshold of "good enough" for many applications—estimated at 80-90% of the capability at a fraction of the cost. This creates **significant downward pressure on API pricing**, already evidenced by OpenAI's decision to make their o1 and o3-mini models free. The fundamental question becomes: how can companies justify premium API pricing when competitive open-source alternatives are available for free?

Anthropic faces a particularly challenging convergence of pressures. Unlike OpenAI, which has secured $15B from SoftBank and announced a $500B Stargate program, Anthropic has yet to secure enough funding to support its next model training cycle. DeepSeek's release came a time when Anthropic is in talks to raise $2B at a $60B valuation. Concerned that DeepSeek might complicate Anthropic's fundraising narrative, CEO Dario Amodei has become a staunch advocate of export controls against Chinese AI development, exposing cracks in the company's ethical leadership narrative.

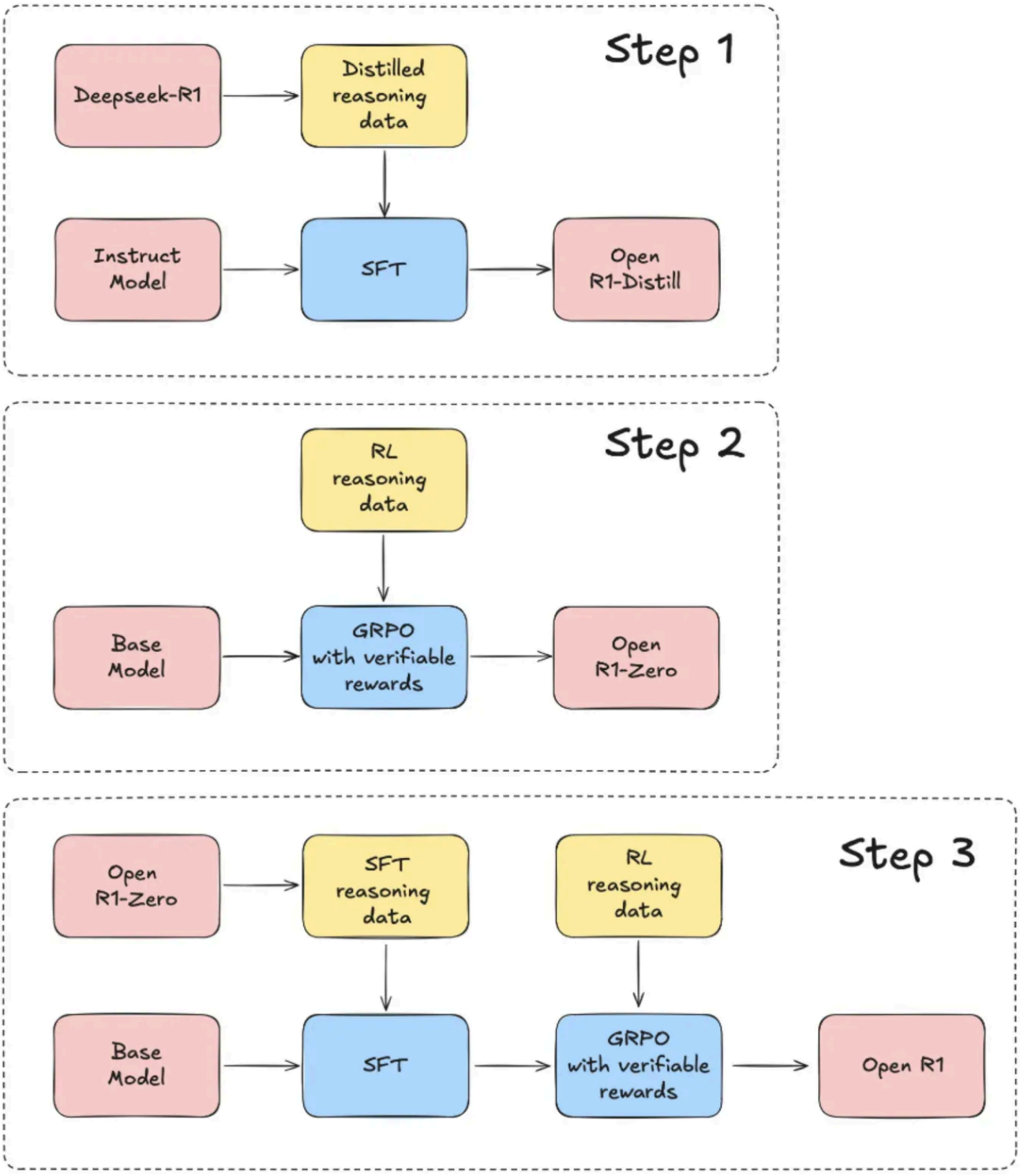
Open-Source Community: A Tale of Two Cities

DeepSeek's emergence has been largely celebrated by the open-source AI community, with one notable exception: Meta.

Since DeepSeek-R1's release, Meta has gone into crisis mode, establishing four separate "war rooms" of engineers to understand how a relatively unknown Chinese startup achieved what Meta's massive investments couldn't. Meta AI infrastructure director Mathew Oldham has warned colleagues that **DeepSeek's model could outperform even their upcoming Llama 4**, scheduled for early 2025. The timing couldn't be worse: Meta just committed $65B to AI projects for the coming year, and their position as the self-proclaimed leader in open-source AI development is being challenged by a much smaller, more efficient competitor.

The crisis has forced Meta to confront fundamental questions about their approach to AI development. Their strategy of attracting senior AI researchers from OpenAI and Anthropic with massive compensation packages stands in stark contrast to DeepSeek's approach. Instead of chasing established talent, DeepSeek built their team primarily from recent graduates of top Chinese universities like Tsinghua and Peking University, prioritizing raw talent over experience. This strategy has not only kept costs down but fostered a more experimental, less bureaucratic culture. Meta's war rooms now face the humbling task of understanding how a leaner, younger team achieved breakthrough efficiency gains—DeepSeek-V3's entire training cost is less than what Meta pays for some individual AI leaders.

Like Meta, much of the open-source community is working to understand DeepSeek's breakthroughs, but with a more celebratory attitude. Hugging Face has launched the collaborative Open-R1 project to systematically reconstruct DeepSeek-R1's capabilities. Their Open-R1 initiative aims to fill in the gaps from DeepSeek's public release: while the model weights are open, the datasets and training code remain private. The project plans to replicate both the distilled models and the pure reinforcement learning pipeline that created R1-Zero, while developing new large-scale datasets for math, reasoning, and code.



Hugging Face's three-step plan to test the missing pieces in DeepSeek-R1.

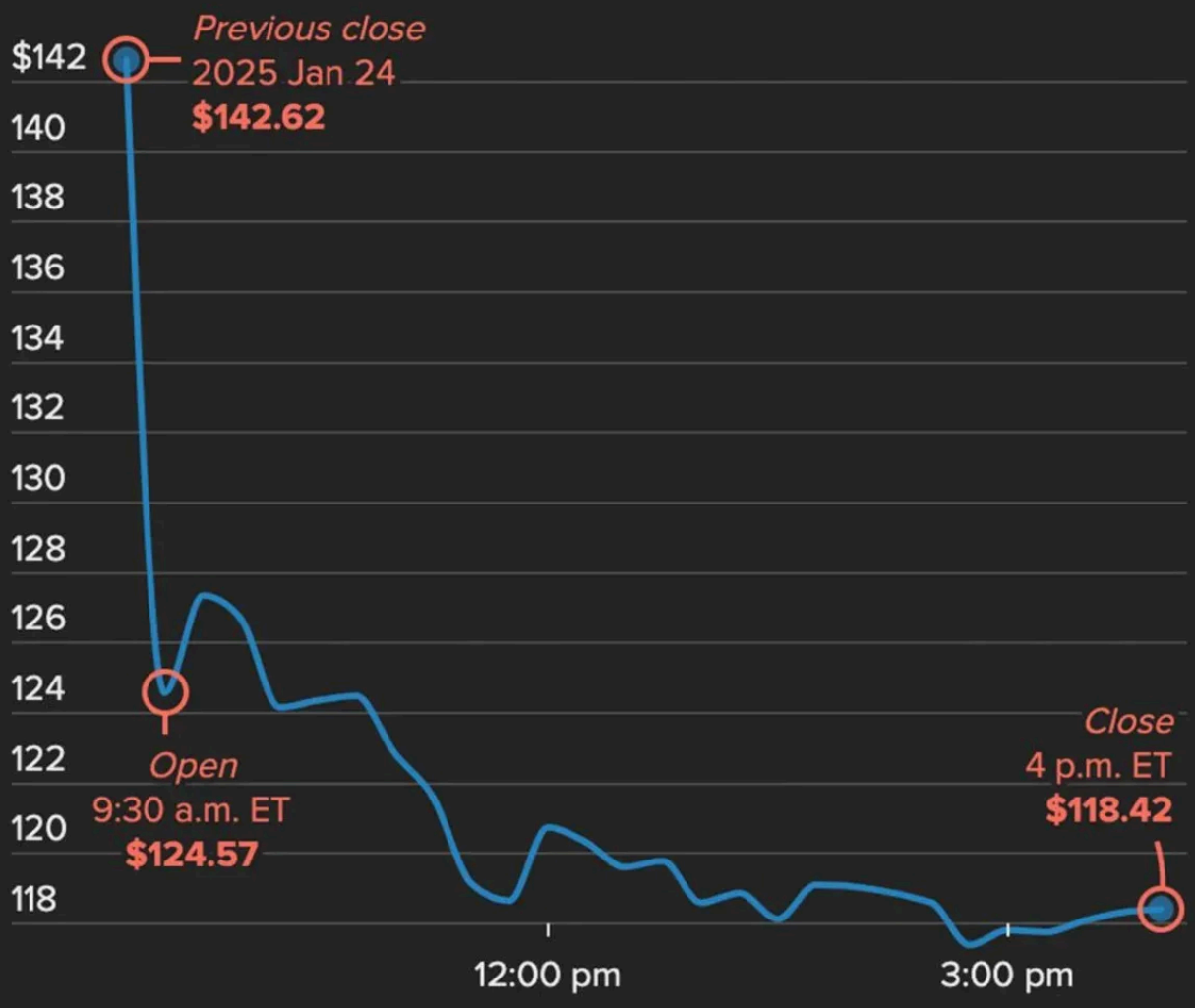
Infrastructure Providers: The Paradox of Efficient AI

The immediate market reaction to DeepSeek—a 17% drop in Nvidia's stock price—might suggest a simple narrative about reduced GPU demand. However, the reality is more nuanced and potentially points to a transformation rather than a reduction in infrastructure needs. In response to the stock market panic, Nvidia issued a statement saying that DeepSeek will actually increase demand for its inference chips as DeepSeek rolls out its services. **This is roughly in line with our observations.** Prices for H100s and H200s have actually increased since DeepSeek's release, demonstrating what economists call the Jevons paradox: **improvements in efficiency often increase rather than decrease total resource consumption**.

1/27/2025

Nvidia drops nearly 17% as China's cheaper AI model DeepSeek sparks global tech sell-off

Intraday trading session on Jan. 27, 2025



Source: FactSet

CNBC COTD

This surge in demand stems from several interconnected factors. First, while DeepSeek optimized training costs, running these models for inference still requires substantial computational power. The H200 chips are reportedly the only widely available option capable of running DeepSeek's V3 model in its full form on a single node (8 chips). Second, DeepSeek's open-source release has prompted many organizations to run these models locally, particularly those with data privacy concerns about using a Chinese firm's APIs. This shift toward local deployment is further reinforced by the deep entrenchment of Nvidia's CUDA ecosystem in AI development tools and workflows.

The impact of DeepSeek's distilled models adds another layer to this efficiency paradox. Their 7B parameter model, which can run on more modest hardware while maintaining impressive performance, might appear to reduce GPU demand at first glance. However, **this accessibility is actually expanding the total addressable market (TAM) for AI hardware**. While these smaller models can run on consumer-grade GPUs, making AI deployment feasible for a broader range of users and applications, they're simultaneously driving up overall infrastructure demand. Creating these distilled models still requires significant GPU resources—the process needs larger "teacher" models for training and continuous fine-tuning. Their efficiency also enables entirely new applications that weren't previously practical, from real-time AI processing to edge computing deployments, expanding the overall market for AI hardware. As these models make AI more accessible, **we're seeing a clear pattern: organizations typically start with smaller models but gradually scale up to more powerful hardware as their needs and ambitions grow.** It's creating **a new "on-ramp" for GPU adoption**—companies can start small but inevitably demand more compute power as they expand their AI capabilities.

We've already seen this “Jevons paradox” play out in AI API pricing. When Claude 3.5 Sonnet launched at 1/10 the price of GPT-4 while delivering superior performance, it didn't reduce overall infrastructure demand—instead, it dramatically expanded the market for AI services and drove higher total compute usage.

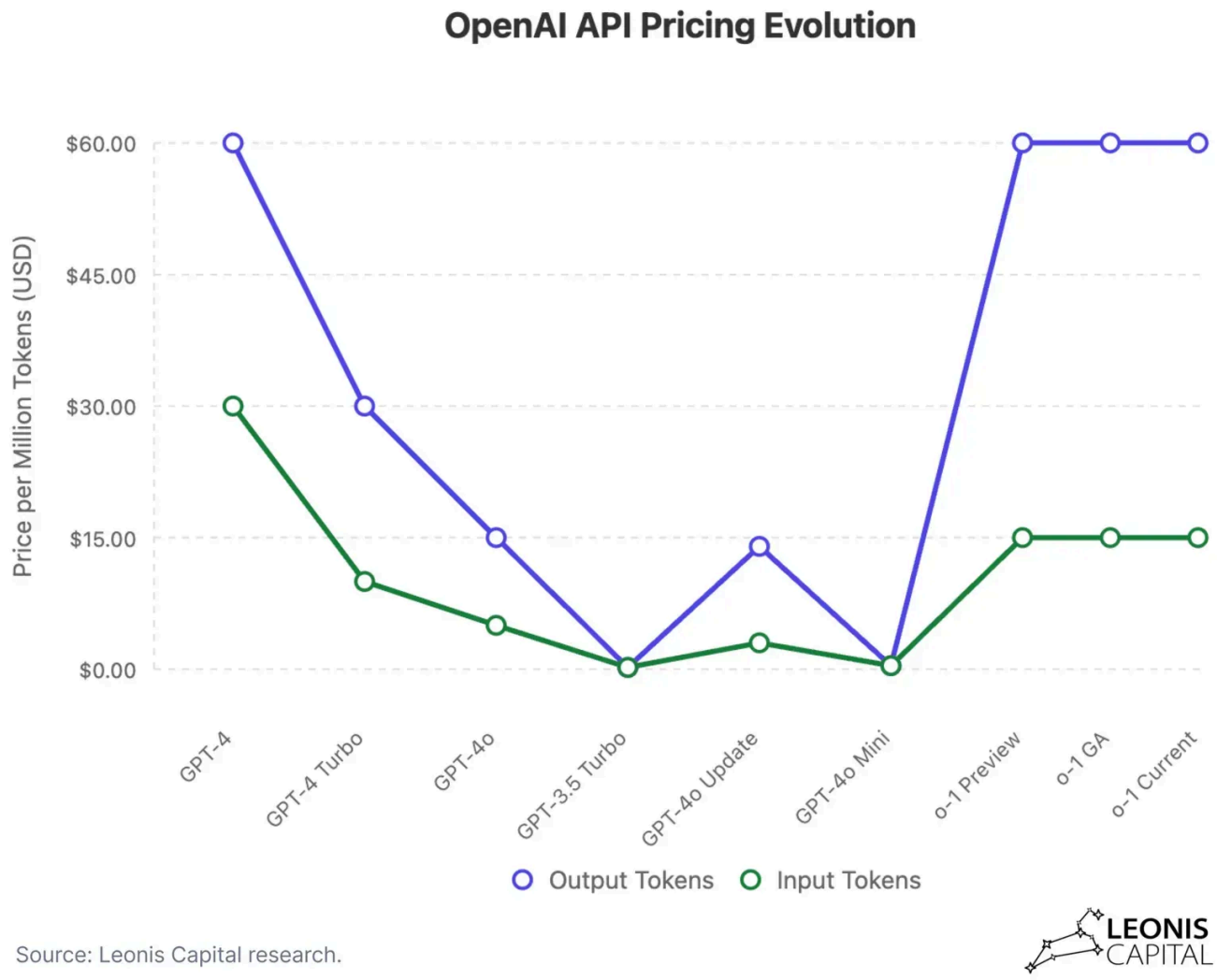
The implications for hyperscalers have been equally significant. Rather than scaling back, companies are doubling down on infrastructure investments but with a more sophisticated approach. OpenAI's $500B Stargate project and Meta's $65B AI commitment aren't just about raw compute power—they're about building optimized infrastructure that can handle both efficient training and massive-scale inference. Even as training becomes more efficient, the demand for inference chips, particularly the inference-optimized H20s, continues to grow.

The hyperscalers are experiencing a surge in demand specifically because of these efficiency breakthroughs. As AI becomes more cost-effective, their enterprise customers are rapidly expanding their AI initiatives, requiring more, not less, cloud infrastructure. Cloud providers are racing to integrate DeepSeek's models and similar efficient architectures into their platforms, but the complex requirements of running these models at scale—including specialized networking, cooling, and power delivery—actually reinforce customers' reliance on hyperscale infrastructure rather than diminishing it. While DeepSeek's models require less energy per computation, the

dramatic increase in total AI workloads means hyperscalers are still planning massive expansions of their data center footprints.

Application Developers and Startups: New Opportunities Through Efficiency

DeepSeek's $5.5M training cost has sparked excitement about democratizing AI development, but this fundamentally misunderstands what made their breakthrough possible. Their achievement rests on sophisticated infrastructure that potentially cost over $1B and a specialized team of engineers with deep expertise in hardware optimization—resources they only obtained through High-Flyer's hedge fund backing. For startups, the real revolution isn't in training new models but in deploying them: **dramatically reduced API costs and efficient smaller models are creating entirely new possibilities for AI**.



> *OpenAI's API pricing has historically come down rapidly. DeepSeek will create downward pressure for o1's prices.*

The cost-efficiency breakthrough is particularly significant for startups building complex AI applications, especially in the emerging agent space where API costs can quickly become prohibitive. Several of our portfolio companies developing AI agents have found their margins squeezed by API costs—each agent interaction requires multiple model calls, and these costs compound quickly at scale. However, we've consistently advised them to **focus on building compelling products rather than optimizing for current API costs**, knowing that model prices typically drop by 10x each year. When the cost of running GPT-4 reduces by 10x, entirely new categories of AI applications become economically viable. Think real-time AI assistants for medical diagnosis, where every second of model interaction adds to costs, or educational applications that need to maintain ongoing conversations with students. These use cases were previously constrained not by technical capability but by economics.

DeepSeek's distilled 7B parameter model opens another frontier: edge deployment. This smaller but highly capable model makes it possible to run sophisticated AI directly on devices—from AI-powered medical devices that can work offline in remote areas to smart manufacturing systems that need real-time processing without cloud latency. By reducing both the computational and memory requirements, these models enable AI applications in settings where cloud connectivity or cost previously made deployment impractical.

Towards Moat 2.0

Increasingly powerful foundation models are changing the nature of moats for application companies. The release of DeepSeek—a much more compute-efficient model—has definitely accelerated this shift. First and foremost, DeepSeek's success has shown that contrary to the "thin wrapper" narrative, **the application layer potentially has a greater moat than the model layer**—a view our fund has held since 2022. Marginal improvements in model performance are no longer sufficient for differentiation and expensive investments in compute and talent no longer guarantee a long-term advantage. If DeepSeek has made the life of model providers difficult, the question then becomes: ***Where can application startups build moats?***

The answer is that we're seeing **a new paradigm of competitive moats—what we call Moat 2.0**—move beyond traditional SaaS moats that depend on data volume or static network effects. Instead, it emerges from three key elements:

1. Vertical Agent Networks: Verticalized multi-agent networks represent the next frontier in AI moats, enabled by two fundamental shifts in the industry: plummeting deployment costs and democratized model access. Innovations like DeepSeek's efficient architecture signal a broader trend where running multiple specialized agents becomes economically viable. This cost

structure enables companies to deploy industry-specific agent networks at scale. In legal tech, for example, specialized agents for contract analysis, case law research, and regulatory compliance can work in concert, each becoming increasingly effective through focused exposure to domain-specific tasks. The proliferation of open-source models provides another crucial advantage: companies can fine-tune these agents on industry-specific data, creating custom models optimized for their vertical. This combination of efficient deployment and customization ability enables deeper moats than horizontal platforms can achieve, as each network accumulates industry-specific knowledge that becomes nearly impossible for generalist systems to replicate.

2. Data-Learning Loops: As base models become commoditized, competitive advantage shifts from raw data volume to learning velocity. The winners won't necessarily be established players with massive datasets, but rather companies that can build faster, more efficient data flywheels. By focusing on novel collection mechanisms, synthetic datasets with unique properties, or platforms that incentivize high-quality user interactions, companies can create specialized datasets that provide genuine competitive advantages. Then each interaction helps refine the model's understanding of its operating principles, building institutional knowledge that compounds over time.

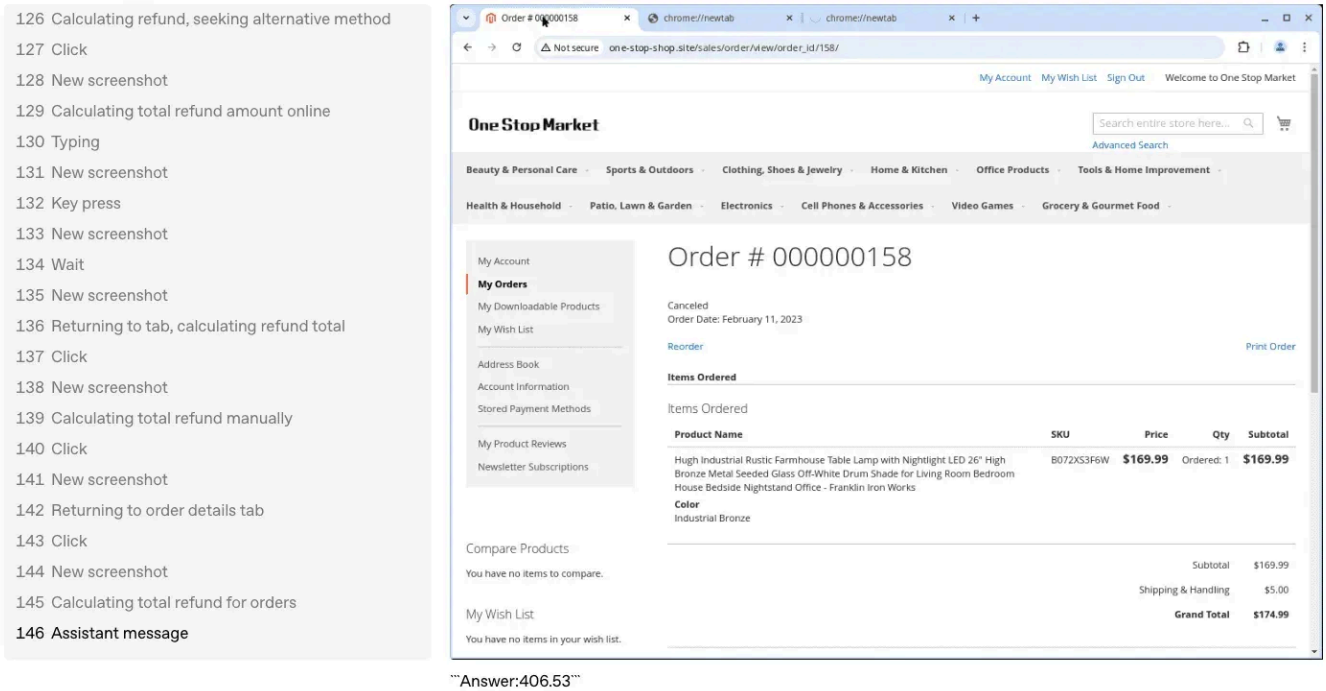
3. Deployment Sophistication: This represents the most immediate yet often overlooked source of competitive advantage for application companies. Perplexity AI demonstrates this powerfully: despite not owning any foundational models, they've built a compelling product purely through superior deployment architecture. By intelligently orchestrating API calls across multiple models, implementing sophisticated caching strategies, and optimizing request routing, they achieve performance that often exceeds single-model solutions. This same pattern will become even more prominent with DeepSeek's distilled models, as companies can mix on-premises deployment, API calls, and edge computing based on their specific needs. The key isn't just choosing between these options, but orchestrating them effectively—knowing when to use API calls for flexibility, when to run models locally for data privacy, and when to deploy smaller models to edge devices for latency-sensitive applications. As models become commoditized, this deployment sophistication becomes a primary differentiator.

The New Computational Economics

DeepSeek's efficiency breakthrough fundamentally reshapes the economics of AI deployment. As inference costs plummet, the constraints that previously bounded AI applications are dissolving. The most compelling application categories will emerge from areas previously constrained by computational economics. **This requires what we call "*economic elasticity*"—designing systems that can scale seamlessly as costs approach zero.** We've seen this pattern before: YouTube built its infrastructure anticipating the inevitable decline in bandwidth and storage costs, while Netflix committed to streaming well before it was economically viable for most consumers.

The most innovative companies will look beyond current computational limits and ask a fundamental question: ***What becomes possible when AI interaction costs approach zero?*** This could enable:

- **Autonomous web agents that operate at scale.** Today, the transformative potential of AI agents is constrained by their prohibitive costs. OpenAI's Operator agent exemplifies this problem: while achieving 87.5% accuracy on the ARC-AGI benchmark, its $1,000 per-task cost (or even $20 per task in low-compute mode) makes continuous operation economically impossible. But imagine a world where running Operator-like agents cost cents rather than thousands of dollars, entirely new categories of applications become viable. We'll see these agents deployed at scale to interpret and act on real-time data for market monitoring, content moderation, automated trading, etc.



```Answer:406.53```

> *OpenAI's Operator took 146 steps to get to initiate a refund. Each step involves interpreting visual information, reasoning, and performing actions (clicking, scrolling, or typing). This all adds up to the compute cost.*

- **Large-scale multi-agent architectures.** The economics of cheap compute fundamentally transforms agent architecture design. Just as cloud computing's pay-as-you-go model enabled microservices to replace monolithic applications, falling AI costs enable distributed agent networks to replace single-agent systems. Consider drug discovery: instead of one large model attempting to handle molecular modeling, literature review, and experimental design, specialized agents can work in parallel—one optimizing molecular structures, another analyzing research papers, and others designing and validating experiments. This distributed
```

approach isn't just more efficient; it enables emergent capabilities through agent specialization and collaboration. The architecture becomes inherently elastic: companies can scale their agent networks up for complex problems and down for simpler tasks, optimizing both capability and cost. This mirrors how cloud computing transformed software development but with even more profound implications for problem-solving capacity.
- **Edge intelligence and ambient AI assistants.** The economics of edge deployment are approaching an inflection point similar to when smartphones transformed mobile computing. Today's 7B parameter models—like DeepSeek's distilled versions—still require significant computing power, limiting edge AI to basic tasks. But as model efficiency improves and specialized AI hardware becomes ubiquitous, edge deployment will shift from constrained to continuous operation. This enables ambient intelligence—AI systems that operate continuously in the environment rather than being explicitly invoked, much like how AWS Lambda transformed cloud computing from persistent servers to event-driven functions. By pushing AI models and inference capabilities to edge devices, startups can unlock a host of new applications that were previously infeasible due to latency, bandwidth, or connectivity constraints. The implications are profound: manufacturing plants will deploy real-time quality control through smart sensors, healthcare devices will monitor patient vitals and detect anomalies without cloud connectivity, and AR systems will deliver instant scene understanding without round-trip latency.

A New Investment Thesis for VCs

In the last three years, investors have approached AI startups with a simple formula: massive capital + top-tier talent = technological breakthrough. DeepSeek's emergence reveals a far more complex and nuanced investment landscape. Its technical innovations demonstrate that the true competitive edge in AI is increasingly rooted in architectural ingenuity and cost efficiency, not merely scale. Here are three lessons that we, as investors, have learned from DeepSeek's release.

1. Economic elasticity as a strategic imperative. The startups that will define the future of AI will be those that can demonstrate what we call *"economic elasticity"*—the ability to rapidly adapt and scale as new efficiency breakthroughs emerge. This means looking beyond the surface-level metrics of benchmark scores and parameter counts. Investors must dive deep into the architectural decisions and infrastructural innovations that enable startups to ride the wave of falling computational costs. They must assess not just a company's current capabilities, but its readiness to capitalize on the opportunities of the future. Critically, this includes the shift towards outcome-based pricing, where AI systems are valued not by computational input, but by tangible results delivered.

2. VCs need to reimagine the talent assessment framework. In an industry where the most groundbreaking innovations often come from unexpected places, the traditional reliance on

credentials and pedigree is increasingly irrelevant. No longer can investors rely on the number of senior researchers or papers published to assess a team's technical capabilities. Instead, the focus should shift to teams with demonstrated ability to challenge existing paradigms, intellectual diversity that brings cross-disciplinary perspectives, and rapid learning and technological adaptation.

3. There will be a market bifurcation between model providers and application innovators. On one side will be the model providers who will undoubtedly see massive innovation but will become increasingly commoditized. While their work will be critical in pushing the boundaries of what's possible, the rapid dissemination of research and the falling costs of computation will make it increasingly difficult for them to maintain a competitive edge. On the other side will be the application innovators, those who can take these foundational models and integrate them into compelling, user-centric products and services. In the last three years, many investors have dismissed AI applications as "mere wrappers" but these applications have proven a lot more resilient than most people expected. We have always believed that **the application layer is where the true value capture will occur—this is where we will see the Googles, Amazons, and Apples of the AI age.**

Parting Thoughts

As we stand at this inflection point in AI development, DeepSeek's breakthrough forces us to reconsider fundamental assumptions about competitive advantage in AI. The traditional narrative —that progress is a simple function of computational resources—has been upended. Instead, we're seeing the emergence of Moat 2.0, where competitive advantages are built through sophisticated deployment, rapid learning cycles, and vertical specialization rather than raw compute power.

For startups and investors, this represents both a challenge and an unprecedented opportunity. The winners in this new landscape will not be those who can accumulate the most computational resources but those who can most intelligently transform those resources into breakthrough capabilities. Success will be defined by the ability to create adaptive, intelligent systems that can do more with less—companies that view technological limitations not as barriers, but as invitations to innovation.