JAN 27, 2025

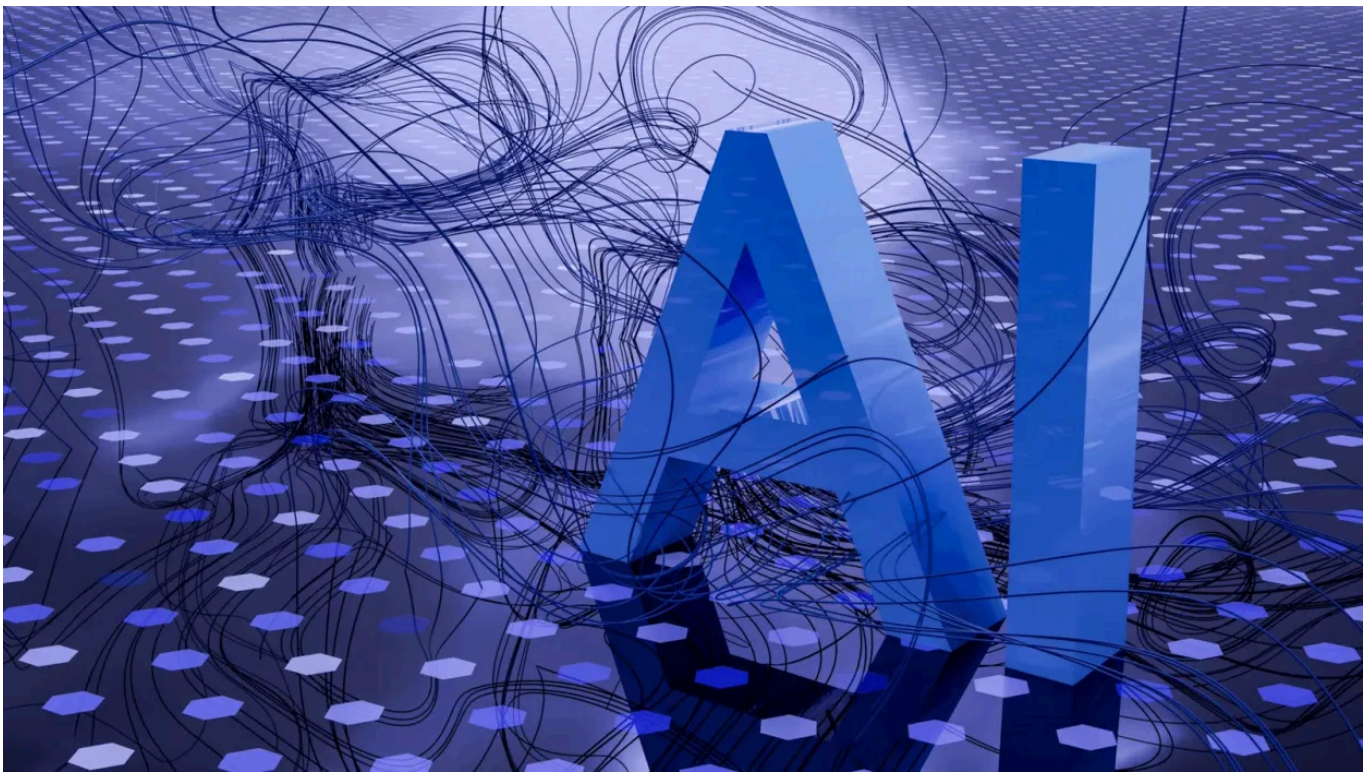
The State of AI in 2025

2024 Recap and 2025 Predictions: Scale vs. Efficiency at AI's Crossroads



By Jenny Xiao and Jay Zhao

Leonis Capital · leoniscap.com/research/the-state-of-ai-in-2025/



A year ago, we made bold predictions about AI's evolution. Some we got right—like the rise of vertical and the proliferation of AI agents. Others revealed how this technology continues to surprise even its close observers. Perhaps most unexpected was DeepSeek's demonstration that the future isn't just about massive compute: they achieved state-of-the-art performance with just 2,000 GPUs and $5 million, while industry giants were spending hundreds of millions. **This achievement exposed AI's key inflection point: will progress come from billion-dollar infrastructure or smarter, leaner approaches?**

As a Leonis Capital tradition, we begin the year by reviewing the past year and making some predictions for the next. This deep dive examines what we got right, what we got wrong, and

most importantly, what it all means for the future of AI in 2025 and beyond. From the end of scaling laws to the rise of reasoning-first architectures, 2024's developments have set the stage for even more dramatic transformations in 2025.

The 2024 Score Card

Our 2024 predictions turned out to be partially prescient and partially premature. One of our boldest predictions was that a new model architecture would replace the transformer as the scaling law hit a limit. This is partially right: 2024 was the year when researchers started questioning whether scaling laws would come to an end. However, we underestimated the ingenuity of researchers in extending the capabilities of existing model architectures (e.g., efficient scaling) and the determination of foundation model companies to overcome compute and data challenges (e.g., xAI's Colossus). While new architectures like Mamba emerged, they proved less reliable in production than transformers. Moreover, major tech companies like OpenAI and Anthropic were already heavily invested in transformer models, making architectural shifts particularly challenging.

1. A new AI paradigm outperforms the transformer.	🤔 Transformers remain dominant architecture but run into scaling issues. New architectures emerge but less viable in production.
2. AI agent-based software find initial applications.	✅ 2024 is the year of AI agents but agent applications are still in their infancy.
3. Smaller models become more popular as companies explore more use cases.	✅ Smaller, more specialized models see rise in deployment, e.g., real-time customer support, code, targeted coding assistance.
4. Model developers and users will face more regulatory challenges.	✅ More IP lawsuits. EU AI Act comes into effect.
5. The AI hype cools down but building continues.	🤔 Startups are generate less hype but are more customer-focused. But foundation model companies continue to be hyped up and hot. VCs continue to invest heavily in AI.
6. Vertical AI startups develop more mature applications.	✅ Vertical AI develop service-as-a-software model. Healthcare, law, and finance adopt AI.
7. The M&A market for AI booms as large companies look to acquire to innovate.	✅ Major acquisitions in 2024. Google acquiring Character.ai, Microsoft acquiring Inflection, etc.

Our 2024 predictions.

Another humbling miss? The “AI hype cooling” we’ve been predicting since 2023. Not only did the market stay hot, it reached peak fever. OpenAI’s $157B valuation and the near 3x increase in VC funding in AI companies from 2023 to 2024 proved that investor appetite for AI isn’t just sustaining—it’s growing. We underestimated how breakthroughs from foundation model companies would keep fueling investors' enthusiasm, despite the sky-high AI valuations.

On the brighter side, **our call on AI agents and vertical AI proved accurate, though their path of adoption remains rocky**. Particularly in highly specialized sectors like healthcare, finance, and law, we saw the emergence of what we call "service-as-a-software" agents—AI systems that don't just assist human work but actively perform it. This thesis led to our 2024 investments in Landscape (AI agents for venture capital and private equity analysis) and Sully AI (AI agents for doctors).

Another prediction that panned out is **the rise of smaller, specialized models**. We watched as our portfolio companies achieved remarkable results with focused, efficient small models that outperformed larger, general-purpose models in specific tasks. Our portfolio company Unify AI is helping enterprises intelligently route tasks between large and small models, while Spline recently demonstrated the power of specialized models with Spell, their efficient AI model for 3D generation.

We also correctly anticipated the AI M&A boom that included Google’s acquisition of Character AI and Microsoft’s absorption of Inflection AI. In line with our prediction, the acquisitions weren’t about buying technology—they were about acquiring scarce AI talent and capabilities that can’t be built fast enough internally. Google paid $3 billion to acquire Character AI’s engineering team but had little interest in the company’s technology or platform. With lower interest rates and a more business-friendly, deregulatory environment ahead, we expect this trend of creative M&A structures to accelerate in 2025.”

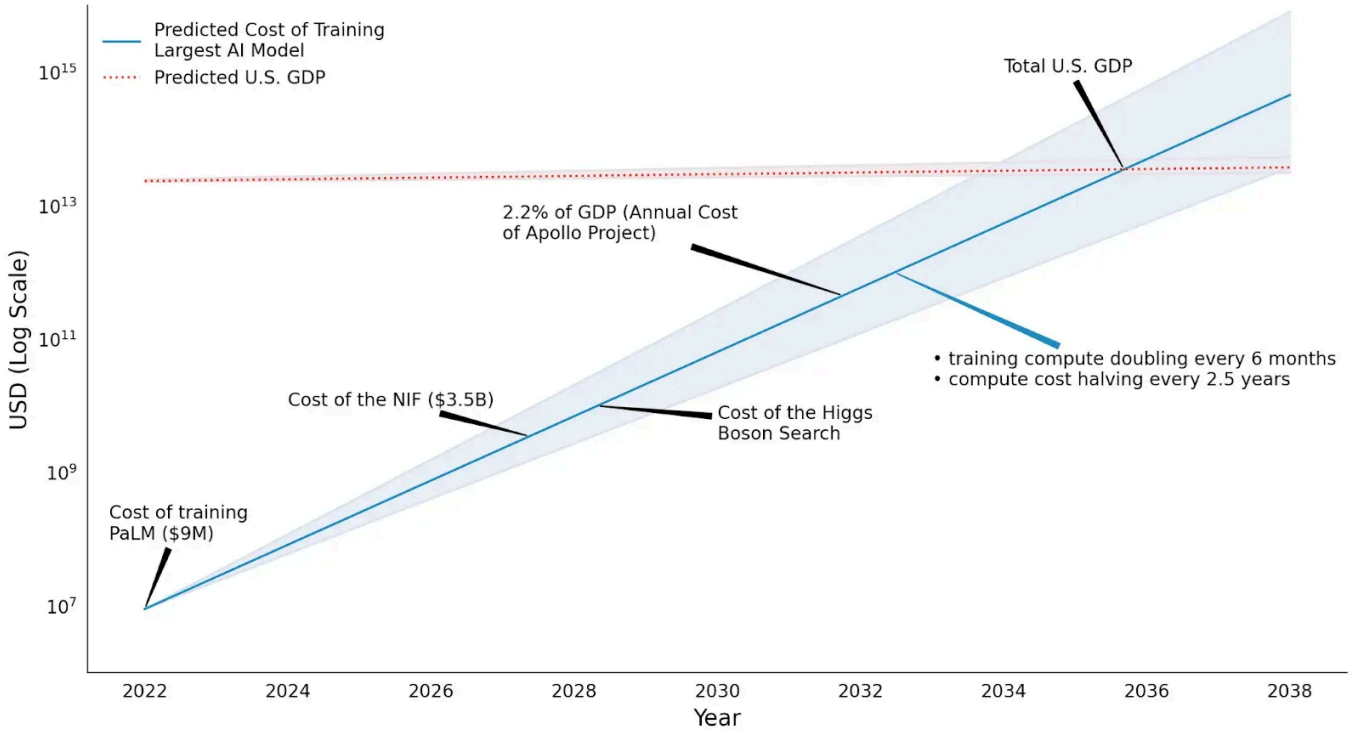
Going beyond our own predictions, here are five trends that have defined the AI industry in 2025. These shifts fundamentally changed how AI systems are built, deployed, and monetized, and more importantly, they signal a deeper transformation in how we think about software and services.

Five Trends That Defined 2024

Trend 1: The End of Scaling’s Golden Age

2024 marked the year when the AI industry’s obsession with model scaling started to hit a brick wall. In the last five years, the AI industry has operated on two fundamental assumptions: training compute doubles every six months (Scaling Law), and compute costs halve every 2.5 years

(Moore's Law). But few have considered their implications. AI researcher Lennart Heim's analysis shows that if current scaling trends continue, in the next few years, training a single large AI model would require resources equivalent to 2.2% of U.S. GDP—matching the scale of the Apollo Program. By 2035, the costs would exceed the entire U.S. economy—a mathematical impossibility. [1] Scaling Law is destined to end, but this reality was never clearer than in 2024.

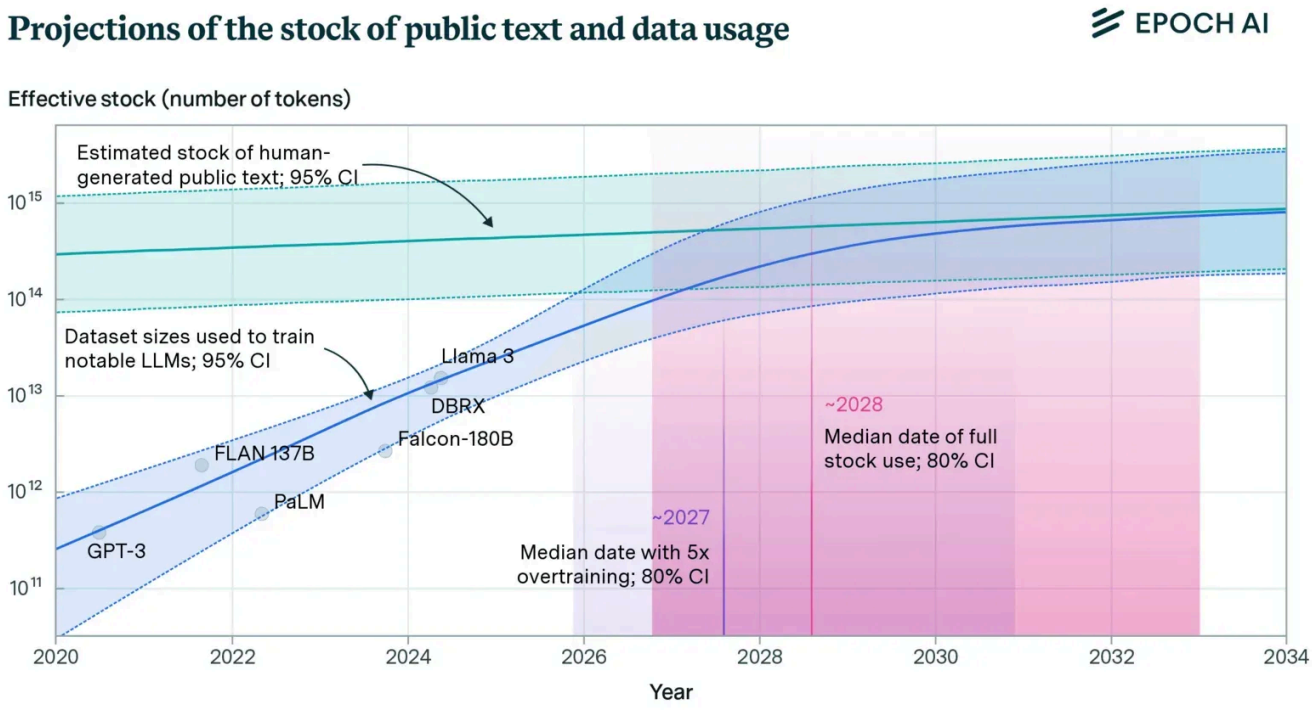


Source: Lennart Heim blog (2023).

While companies like xAI still managed to pull off its 100,000 GPU Colossus cluster, the technical and financial challenges of scaling are piling up. The massive power demand, frequent hardware failures, and high bandwidth requirements made many question the feasibility of such a large GPU cluster. At the same time, foundation model companies like OpenAI and Anthropic face what Sequoia Capital's David Cahn calls "AI's $600B question"—their revenue and profit margins struggle to justify their high valuations and capital expenditures.

The second crisis lurking beneath the surface might be even more fundamental: we're running out of high-quality human data. According to Epoch AI's estimation, we will run out of publicly available human data sometime between 2026 and 2032. Prominent figures in AI echo this point: In 2024, Elon Musk and Ilya Sutskever both came out to warn that we're running out of training data. The situation dates back to DeepMind's 2022 Chinchilla model, which suggested most language models were actually undertrained relative to their size. This insight single-handedly encouraged foundation model companies to train on vastly more data, accelerating our path toward data scarcity. Synthetic data appears promising, but issues with accuracy and mode collapse continue to raise problems. Companies are trying to tap into larger data sources like

video data, physical data from IoT devices, and using privacy-preserving techniques to leverage user-generated content, but none of these approaches have yet proven to be a complete solution.



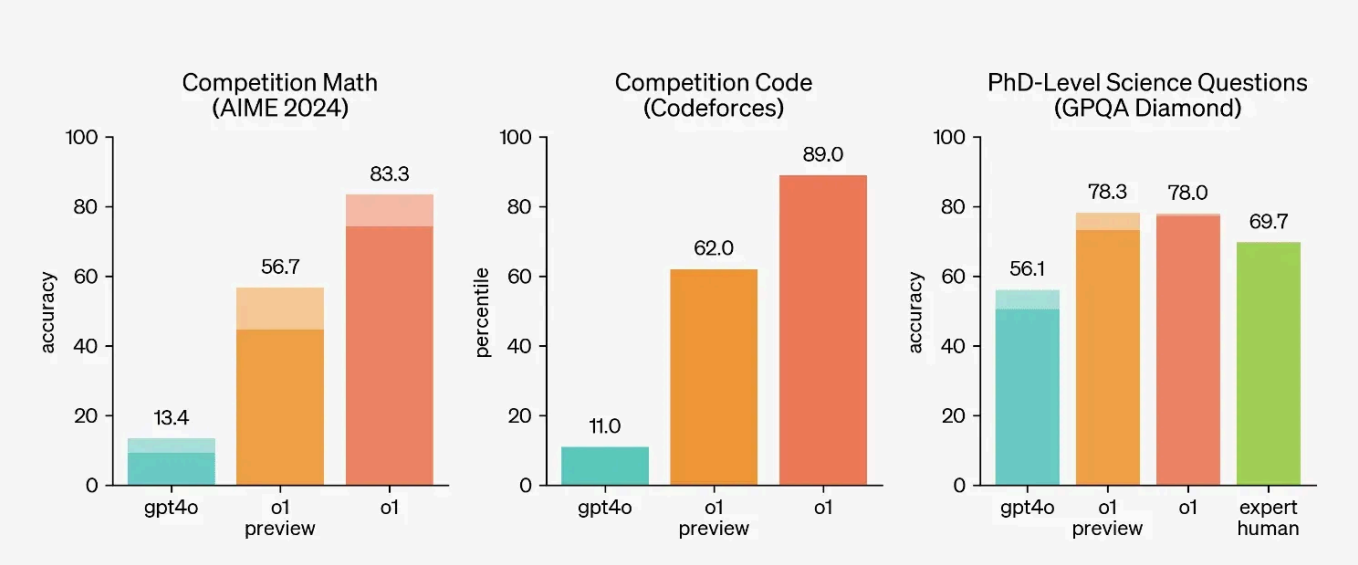
Necessity breeds innovation. The challenges of the scaling law have sparked a renaissance in interest in new architectures. Companies are also achieving remarkable results through model distillation, hybrid architectures, and more efficient training methodologies. The Chinese team at DeepSeek managed to train the remarkable DeepSeek-v3 and DeepSeek-R1 models with a fraction of the cost of typical LLMs while achieving state-of-the-art performance across various benchmarks. Their approach is elegant in its simplicity—using a technique called GRPO (Gradient Reward Policy Optimization), they optimize existing models directly for correctness without complex reward models. Most remarkably, their 1.5B parameter model outperforms much larger models like Claude and Llama 405B on advanced reasoning tasks, while their distilled 7B model competes with OpenAI's latest offerings.

Interestingly, funding and GPU constraints might be a blessing in disguise, forcing researchers to find new approaches in model development that achieve more with less. The future of AI might not be about building ever-larger models but about training them more intelligently.

Trend 2: The System 1 to System 2 Transition

The most profound technical breakthrough in 2024 wasn't about model scaling or training techniques—**it was about how to get AI to think deeper**. OpenAI's o-1 marked a turning point in this direction: By prompting the model to take more time to think through problems step-by-step, it achieved something remarkable—the ability to solve complex problems beyond simple pattern matching. This transition from System 1 (fast, intuitive thinking) to System 2 (slow, deliberate reasoning) represents a fundamental shift in how AI approaches problem-solving. Despite Anthropic CEO Dario Amodei's claims that the company wouldn't build specialized reasoning models, they quickly incorporated o-1's reasoning features in Claude—a pattern followed by Google's Gemini Deep Research, DeepSeek's R1, and others within just months.

What's particularly striking is the pace of improvement: in just two months, we went from college-level to PhD-level AI capabilities. The implications are most profound in domains that operate primarily in symbolic space—mathematics and coding being prime examples. These fields, relatively unbounded by physical world constraints, are seeing the most dramatic advances. This suggests a cascade effect into other symbolic-heavy fields: theoretical physics might be next, followed by theoretical chemistry and biology. The o-class models are particularly effective at optimizing for anything with a clear reward function—explaining their impressive performance in domains like mathematics and coding, where success criteria can be precisely defined. However, they still struggle in areas where defining success is more nuanced, like creative writing or emotional intelligence.



Source: Vellum AI (2024).

This transition represents more than just improved performance metrics. It signals a new approach to AI development, where the focus shifts from accumulating more training data to developing better reasoning capabilities. The tradeoff is clear in the model's behavior—o-1 is

notably slower than its predecessors, but this "thinking time" enables it to tackle problems that were previously out of reach. Interestingly, this shift has made inference compute more critical than training compute, potentially disrupting the current hardware landscape where training chips dominate. We also believe that this has **significant implications for agentic software and agents-networked companies**, which we will discuss in the prediction session below.

Trend 3: The Rise of AI Native Startups

The distinction between traditional software adding AI features and truly AI-native products grew even starker in 2024. The struggles of tech giants and established players have been particularly telling. Microsoft, despite its decabillion OpenAI partnership and endless fanfare around Copilot, has largely delivered an underwhelming suite of AI products that feel more like forced add-ons than transformative tools. Their approach of repackaging LLMs without fundamentally rethinking their products has led to what many users describe as an intrusive and even frustrating experience. Notion's AI journey tells an even more cautionary tale—their AI features have become almost a textbook example of how not to integrate AI. Despite Notion's reputation for excellent product design, their AI consistently stumbles with basic tasks, leading many users to dismiss it as more of a gimmick than a game-changer.

The root cause isn't just poor execution—it's architectural DNA. These companies built their foundations in the pre-AI era, optimizing for human workflows and traditional data structures. When they try to retrofit AI into these systems, it's like trying to attach a jet engine to a horse carriage. The mismatch goes far deeper than user interface issues—it reveals a fundamental disconnect between traditional software architecture and what AI-native applications require. Microsoft faces fundamental constraints on both the technical and business model side: their slower ARM hardware struggles to provide high memory sources for AI processing, and their traditional licensing model does not align well with the modern AI paradigm that requires data and ongoing improvements. Similarly, Notion’s rigid database structures and custom prompts significantly limit the functionality of its AI features. Even with unlimited resources and access to the best AI models, these companies find themselves constrained by their own architectural decisions made years ago.

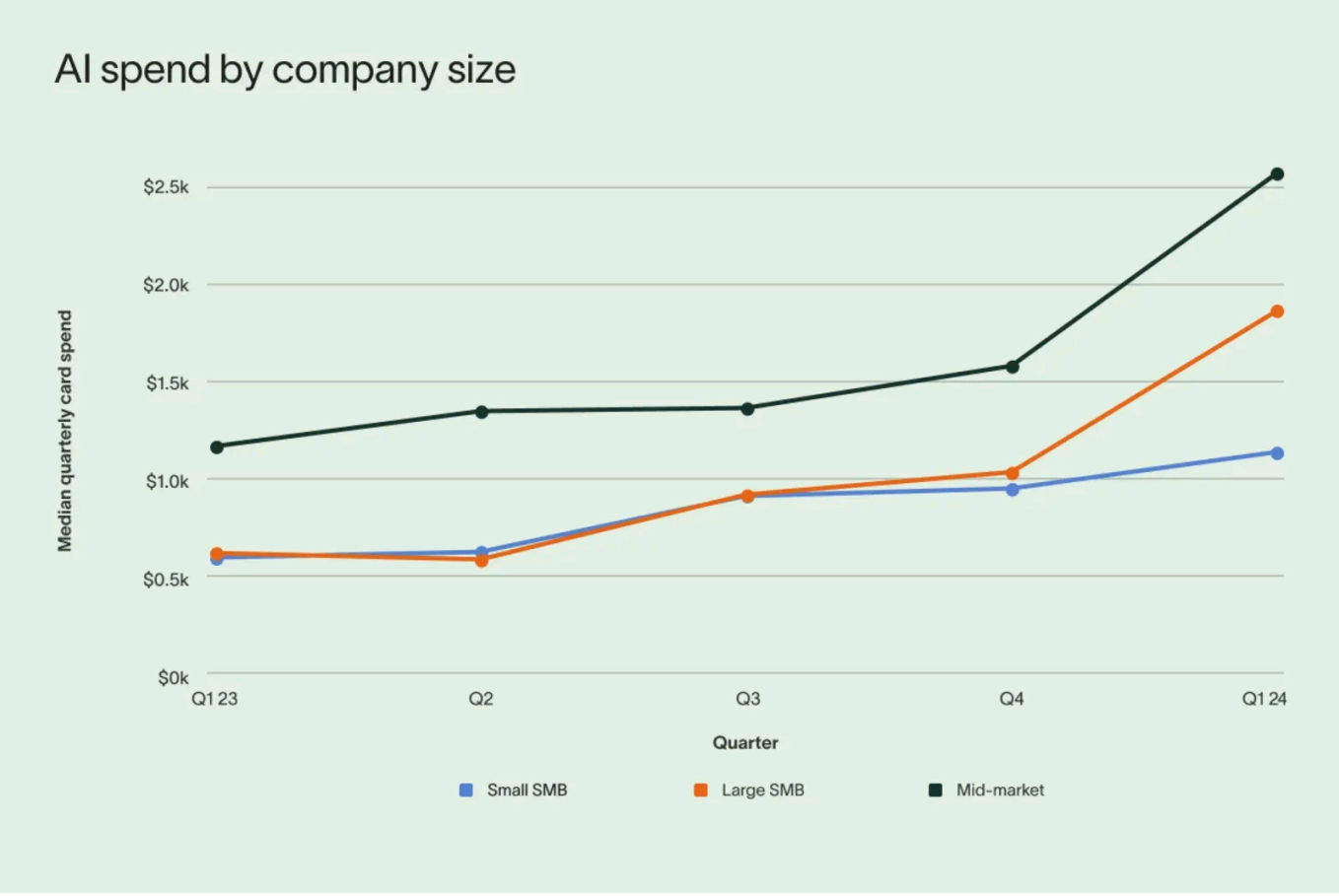
Meanwhile, a new breed of AI-native companies is rewriting the rules of software architecture. These companies aren't just using different tools—they're fundamentally rethinking how software works. Their architecture assumes AI is the computational core, not a feature: databases are designed for embedding and semantic search rather than just structured queries, APIs are built to handle streaming outputs and complex context windows, and infrastructure is optimized for both training and inference workloads from day one.

The difference becomes even more striking at the business model level. While traditional SaaS companies plod along with predictable-but-linear feature improvements, AI-native companies can hit exponential efficiency curves post-product-market fit. Every user interaction makes their

systems smarter, every piece of feedback tunes their model routing, and every new use case expands their capability surface. This creates a powerful triple flywheel: better data leads to better models, which attract more users, who then generate more data—a virtuous cycle that traditional software companies simply cannot match, no matter how many AI features they bolt on.

Trend 4: Enterprise AI Adoption Picks Up

Enterprise AI spending has shifted decisively from experimentation to execution. Based on data from Ramp, a corporate spend management platform, AI-related transactions increased by an astounding 293% year over year in early 2024, compared to just 6% growth in overall software spending. Over a third of companies now pay for at least one AI tool, spending an average of $1,500 per quarter—a 138% increase year over year. Within organizations, engineering departments are leading AI adoption, followed by a significant uptake in HR and Finance teams who are leveraging tools to automate repetitive back-office work. An even more encouraging sign is that once companies adopt AI tools, they tend to stick with them: 56% of businesses that started using AI vendors 12 months ago are still using them today.



Source: Ramp (2024).

In our conversations with enterprise clients, quality, not price, emerged as the decisive factor, revealing a stark new reality where enterprises would rather pay premium prices for reliable, accurate AI systems than risk the hidden costs of less capable alternatives—a single hallucination in a legal document or medical recommendation isn't just a bug, but a potential catastrophe that could cost millions or threaten human lives. The age of AI has dramatically elevated the bar for minimum viable products (MVPs), forcing startups to abandon the traditional "ship fast and iterate" playbook in favor of a more nuanced approach that demands extraordinary precision. Go-to-market (GTM) strategies must now operate like precision instruments, demonstrating an ability to solve hyper-specific problems with superhuman accuracy rather than appealing to the largest number of customers. AI demands not just a technological upgrade but fundamental strategic rethinking.

Trend 5: The Service-as-a-Software Revolution

Perhaps the most profound business model shift we've witnessed in the last year is the transformation of software from tools that help humans work to systems that actually do the work. The traditional SaaS model assumed white-collar workers would use software as a productivity tool, charging per seat and expecting users to navigate complex interfaces. In contrast, AI agents are now actively replacing human labor, charging not based on user headcount, but on concrete outcomes delivered.

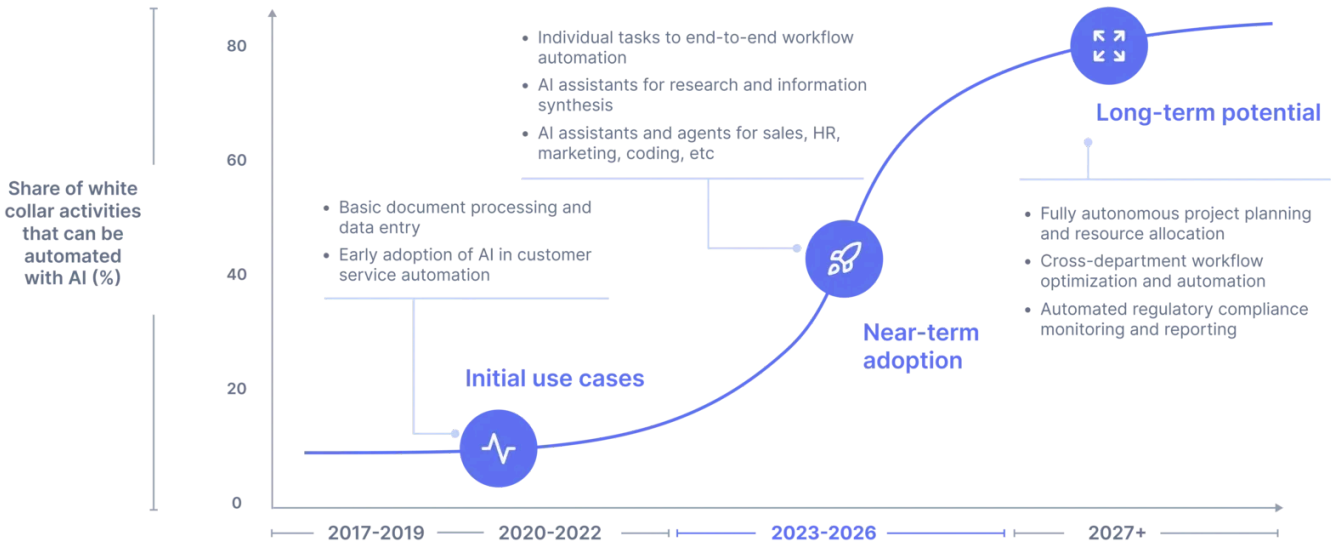
The economics tell a compelling story. Consider a business analyst role with an annual salary of $150,000 producing around 200 reports per year—effectively $750 per report. An AI system costing $100,000 annually can generate 20,000 reports, reducing the cost per report to just $5. This 150x efficiency gain is unlocking budgets previously allocated to payroll rather than IT spending, fundamentally reshaping how companies think about workforce and technological investment.

	Human Analyst	AI Agent	Ratio
💰 **Annual Cost**	$150,000	$100,000	1.5:1
📋 **Output (Reports/Year)**	200	20,000	1:100
👥 **Cost per Report**	$750	$5	150:1
⏰ **Operating Hours**	2,000h/year	8,760h/year	1:4.4

Our back-of-envelope math on AI vs. human labor.

This shift is particularly evident in sectors that traditionally haven't been big software spenders, such as healthcare and finance. These industries are now willing to allocate significant budgets to AI solutions because they can directly replace or augment human labor costs. According to Ramp's 2024 AI spending data, healthcare saw a 131% year-over-year increase in companies adopting AI tools, while financial services grew their AI spending by an impressive 331%. This surge isn't just about following trends—these industries are finding concrete return-on-investment (ROI) in specific use cases, from automating healthcare workflows to enhancing financial analysis.

The transition from software-as-a-service to service-as-a-software will create existential threats to traditional SaaS platforms that depend on human users. LinkedIn's revenue model, which primarily comes from selling seats to recruiting agencies, and Salesforce's entire ecosystem built around human sales teams, cannot simply pivot to AI agents without fundamentally dismantling their core business model. We've seen the same dynamic play out in Google, whose lucrative ad business has slowed down its adoption of LLMs up until recently. Traditional SaaS companies will struggle to embrace the AI agent revolution because their revenue depends on perpetuating existing human-centric workflows.



> *We're entering the rapid adoption phase for AI.*

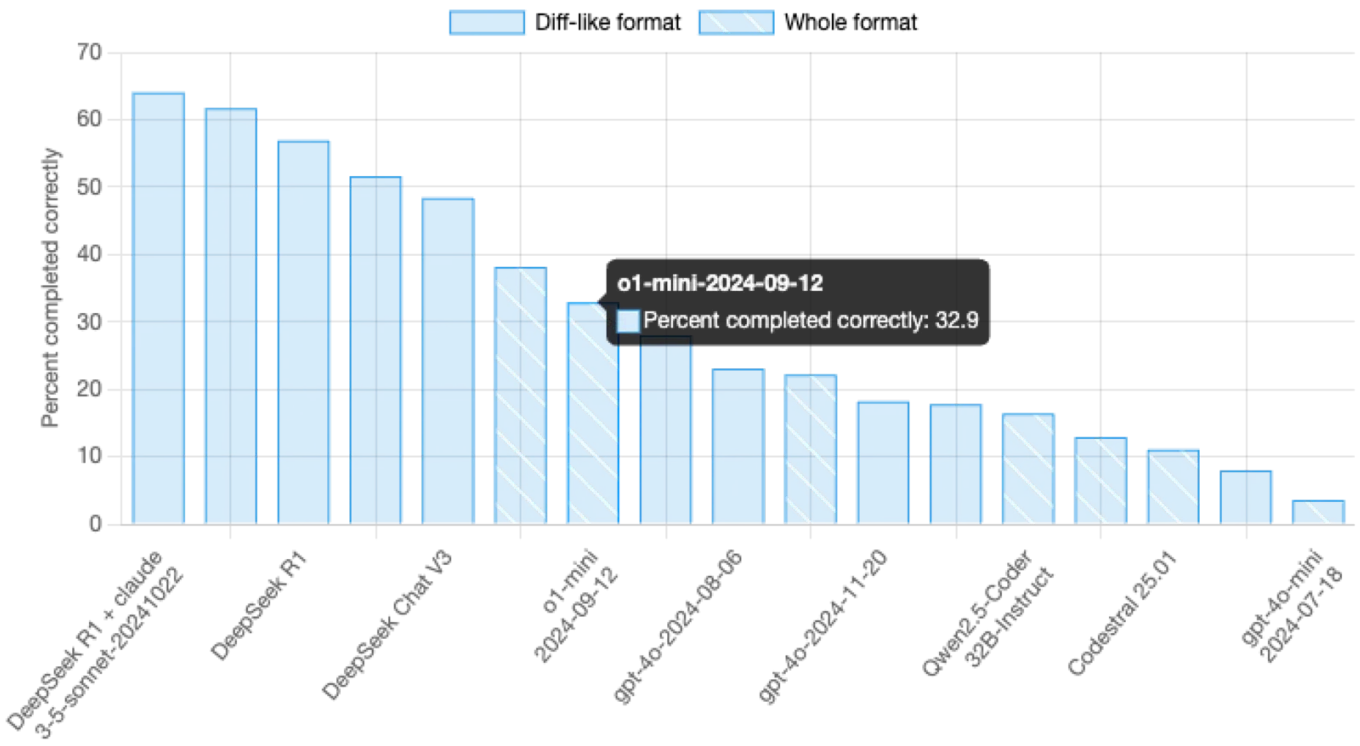
Our key insight for 2024 is that SaaS has always been a fundamentally flawed business model—a construct unique to high-labor-cost economies like the United States and Western Europe. In regions like India, Southeast Asia, and China, enterprise software simply doesn't exist because labor costs are too low to justify complex software purchases. Silicon Valley's SaaS obsession is essentially a workaround for expensive labor. Most non-tech professionals don't actually want to buy software—they want to buy services and pay for outcomes. A sales manager wants to sell more products, not navigate Gong; a marketing director wants leads, not a complex CRM; a

doctor wants efficient patient care, not another EHR. The traditional SaaS model forces users to learn to operate complex tools, whereas AI agents promise to actually solve problems.

Looking Forward: Five Predictions for 2025

Prediction 1: The First AI-Generated Killer App

The technical foundations for truly AI-generated applications have reached a critical inflection point. While AI coding assistants could only improve developer productivity by about 30% six months ago, today's models like Claude 3 can solve complex engineering challenges that would take experienced developers several hours. The progression in AI coding capabilities has been exponential rather than linear—previous improvements focused on helping developers with discrete tasks, today's models can engage in sophisticated software design discussions, understand business requirements, and translate them into working applications.



Rapid improvement in just six months (July-December 2024). Source: Aider polygot benchmark results (Dec 26, 2024).

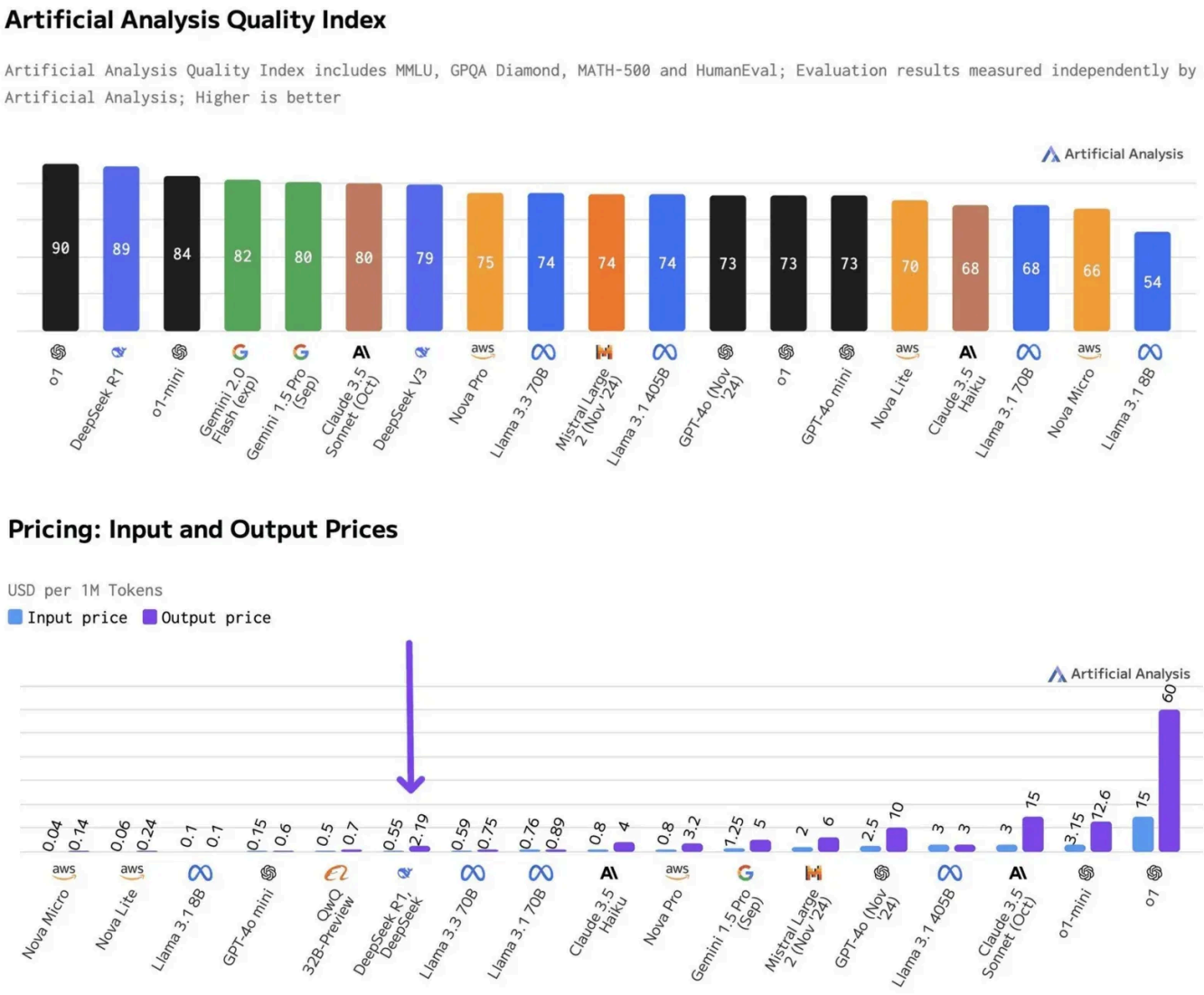
Three key technological breakthroughs make this moment different. First, code generation accuracy has improved dramatically. AI can now understand entire codebases contextually, reason about complex software architectures, and generate production-quality code across multiple languages. Second, context windows have expanded enough to handle complete codebases, allowing AI to maintain consistency across an entire project. Third, we're seeing sophisticated multi-language support expand beyond Python and JavaScript to include modern frameworks and statically-typed languages like Java and C++.

This confluence of advances suggests something bigger than just improved developer tools. We believe 2025 will be the year when non-developers can use AI to generate and deploy production-ready applications. Of course, engineers who can move up and down the stack—understanding everything from binary to high-level architecture—will maintain a crucial advantage. Nevertheless, advances in AI code generation will democratize software development and change who can build software. In 2025, we expect to see the first AI-generated application that's not just a technical curiosity but a legitimate competitor in its market space, potentially coming from someone with no traditional coding background.

Prediction 2: The AI Cost Reckoning

The economics of AI is overdue for a fundamental reset. While OpenAI and SoftBank announce the ambitious $500B Stargate project to double down on AI infrastructure, their Chinese counterparts like DeepSeek are focusing on how to achieve state-of-the-art results with a much less capital investment—just $5.5M for training their v3 model. Critically, DeepSeek's open-source models now perform alarmingly

Meta appears to be the first major U.S. tech company to confront this unsustainable reality. Their GenAI team is reportedly in "panic mode" after DeepSeek's announcements, as they struggle to justify massive investments when a relatively unknown Chinese company has achieved comparable results with a fraction of the budget—less than what Meta pays individual AI leaders. This cost crisis extends beyond just infrastructure: OpenAI's o-1 costs $15 per million input tokens and $60 per million output tokens—three to four times more expensive than GPT-4o. Even at $200 per month, ChatGPT Pro subscriptions are losing money for OpenAI.



Source: Artificial Analysis (2025).

Adding to this complexity, the traditional notion of AI model moats appears increasingly tenuous—researchers frequently move between labs and share information. This fluid exchange of talent and knowledge suggests that no single company can maintain a lasting technical advantage for more than a year. Despite the U.S. government's attempts to regulate AI through the AI diffusion framework, the field remains remarkably open. Researchers continue to share breakthroughs through open-source papers and code, talent flows freely across borders, and companies rapidly replicate each other's innovations. The reality is that there are essentially no secrets in AI—any significant breakthrough, whether from Silicon Valley or Beijing, quickly becomes common knowledge in the global AI community.

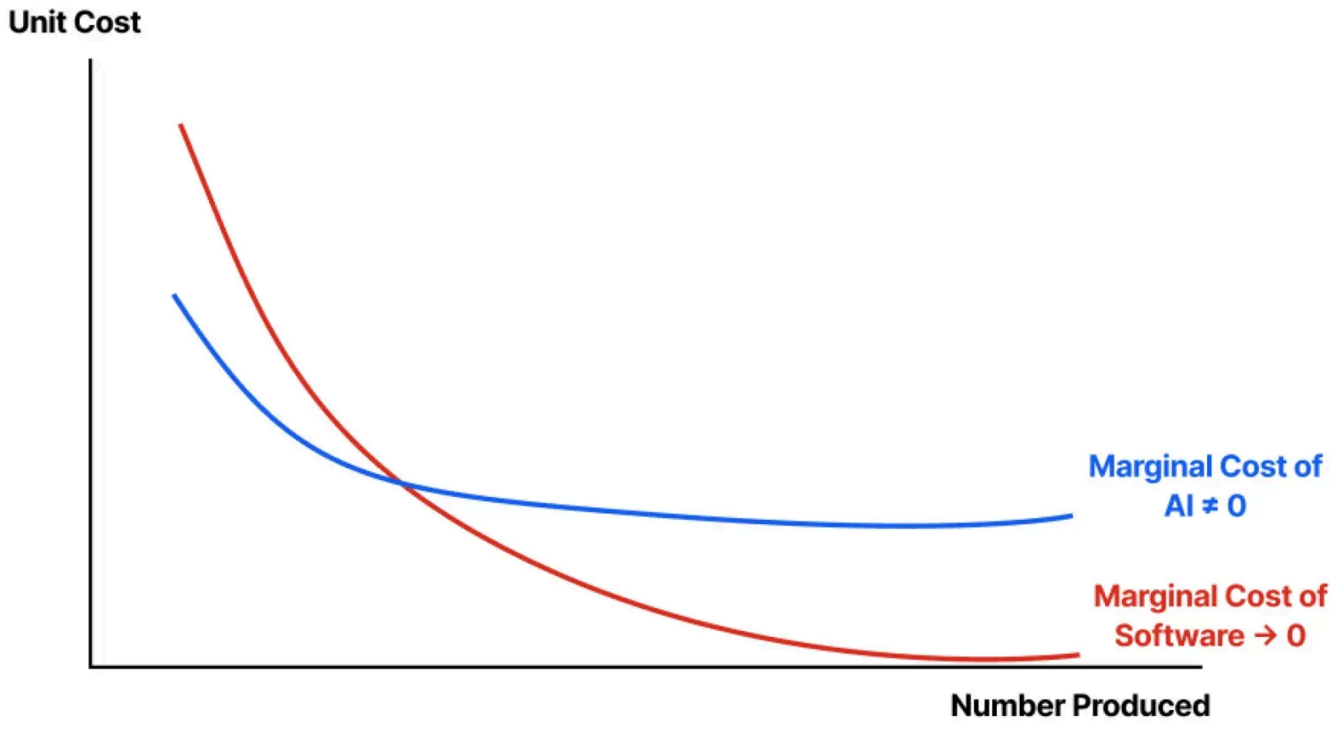
DeepSeek's achievements raise an uncomfortable question: how long can U.S. companies continue to justify billion-dollar AI investments when similar results can be achieved for orders of magnitude less and the moats less sustainable than investors thought? We believe 2025 will be

the year when this cost reckoning finally hits home, forcing a fundamental rethinking of AI development economics.

Prediction 3: The Evolution of AI Business Models

Following our last prediction, we believe that a serious AI cost reckoning will force a complete rethinking of how AI services are valued and monetized. Today, we live in an era of heavily subsidized AI. While inference costs are dropping 50-100x every couple of years, current AI pricing still doesn't reflect true costs—it's artificially suppressed by big tech companies, AI labs, and their investors. But this subsidy era will ultimately come to an end.

The stark reality is that AI economics differ fundamentally from traditional SaaS. Traditional SaaS has maintained exceptional profit margins because the marginal cost of serving each new customer approaches zero. AI services, however, face persistent marginal costs for every inference, every token, every task. As AI companies begin to price their models to reflect true costs rather than subsidized rates, this distinction will become even more apparent. Even today, with heavily subsidized AI models, we have heard founders complain that their API costs eat into their margins. But what will happen if their API costs force them to price their apps at $200 or $2,000 a month instead of $20?

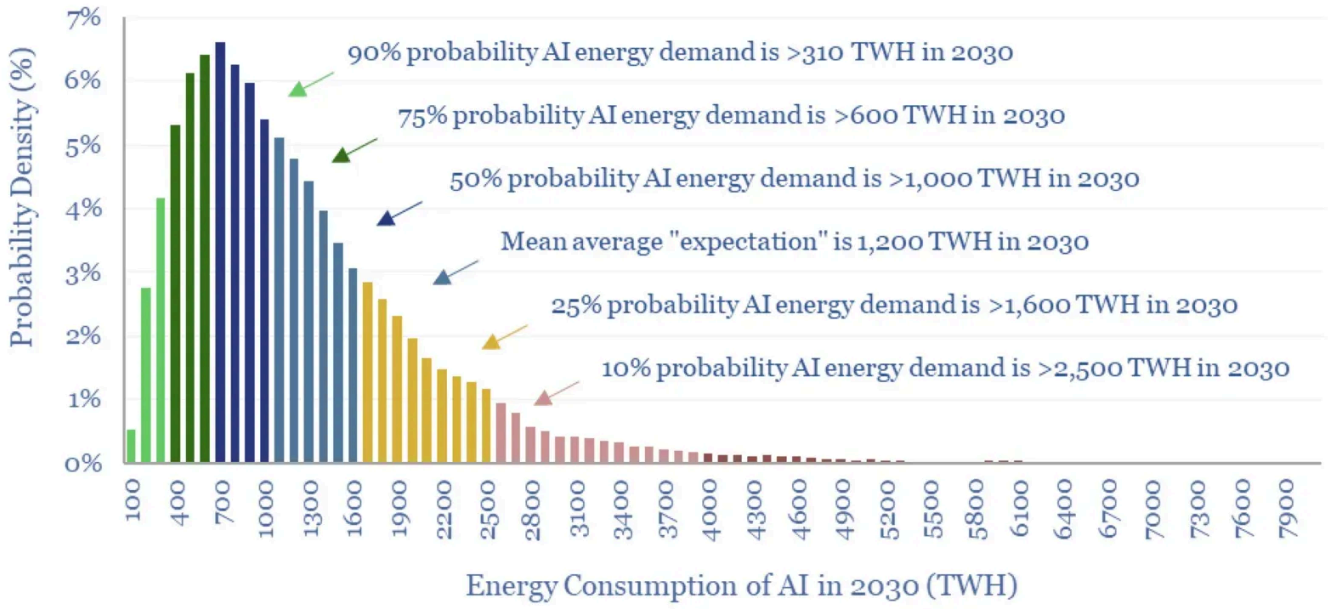


Why the traditional SaaS business model does not work for AI.

A changing cost structure will force a shift in pricing models. We anticipate this economic transformation will unfold in three distinct evolutionary phases. In 2024, we're witnessing the "Premium AI" phase, where advanced capabilities command high-end pricing—think $20 ChatGPT Plus subscriptions. But the next two years will usher in the "Enterprise Scale" phase, where AI companies will price their models based on tangible actions and results, more accurately reflecting the actual value delivered to customers. By 2027-2029, we'll enter the "AI Economy" phase—a radical reimagining of economic interactions. Here, autonomous agent economies will operate as independent economic units, with AI systems not just providing services, but becoming economic actors in their own right. Imagine AI agents negotiating, trading, and creating value with minimal human intervention. In this new economic landscape, the line between technology and economic activity will blur, creating a future where AI doesn't just serve the economy—it becomes the economy.

Prediction 4: The Power Crisis

Data center power consumption has already become a critical challenge for the AI industry but few outside of infrastructure management fully grasp the scale of the impending crisis. Current trajectories suggest that AI-related electricity consumption will grow up to 50% annually through 2030, with data centers potentially requiring 400 terawatt-hours of electricity—a figure that would exceed the entire United Kingdom's total electricity output in 2022. These demands will push against the physical limits of our existing infrastructure in ways that are both technically and politically complex.



Source: Thunder Said Energy (2024).

Through conversations with data center operators in Arizona, Texas, and California's Central Valley, we've discovered that AI usage is already straining local power grids in unprecedented ways. This isn't just a technical problem—it's becoming a profound political and social challenge. Communities in water-scarce regions are actively resisting the construction of power-hungry AI data centers, recognizing the significant local impact of these energy-intensive facilities.

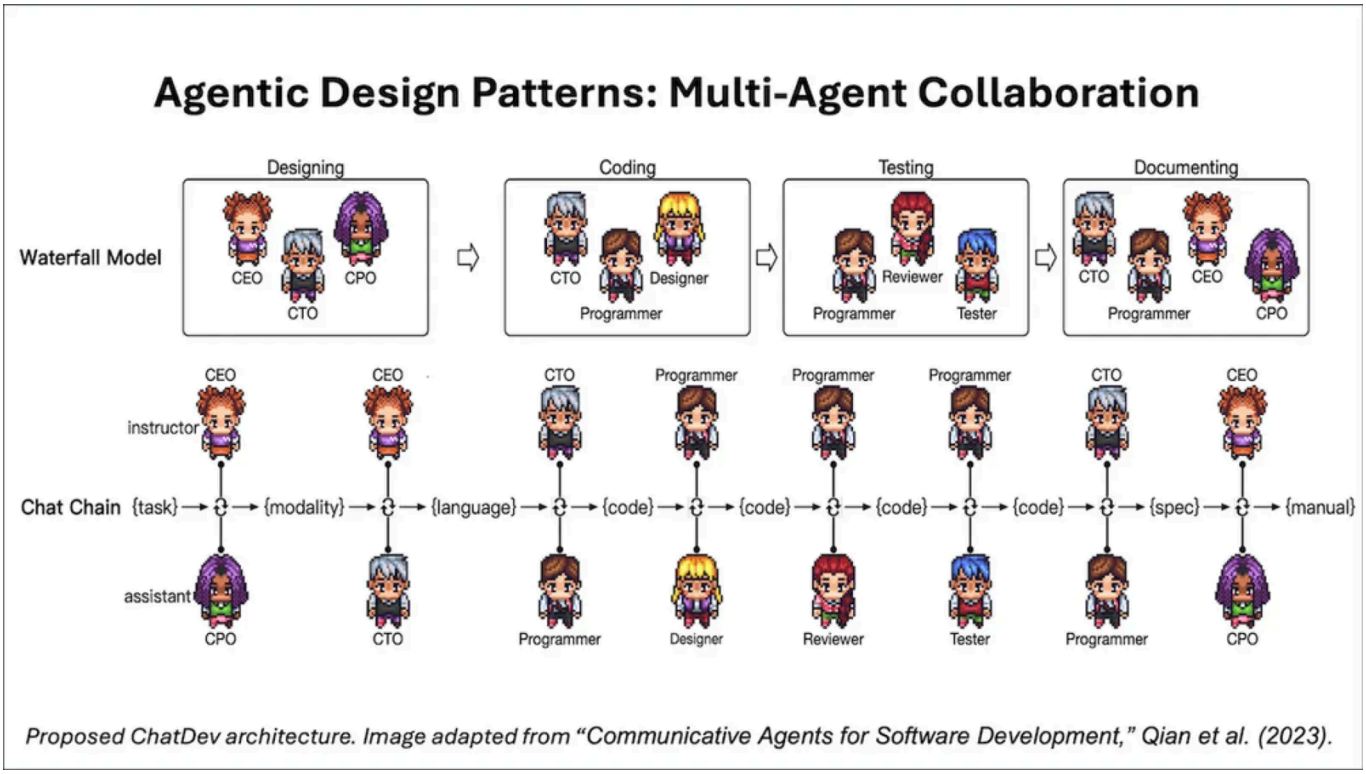
We anticipate 2025 will be the critical inflection point where AI's power grid strain becomes impossible to ignore, driven by the convergence of massive infrastructure projects like Stargate, the rapid scaling of enterprise AI adoption, the increasing demands of inference-time compute, and utility companies' growing recognition that current power generation capabilities are fundamentally misaligned with emerging computational demands.

Interestingly, this challenge is driving innovation across multiple domains. Major tech companies like Microsoft, Amazon, and Google are proactively investing in their own power generation, including experimental nuclear plants, to secure future energy supply. On a positive note, the power challenge is also accelerating innovation in both hardware architectures and software efficiency. We anticipate seeing breakthrough approaches in computational design, energy-efficient computing, and distributed computing models that can dramatically reduce the energy footprint of AI systems.

Prediction 5: The Rise of Multi-Agent Systems

The next frontier in AI isn't about building bigger models—it's about creating networks of specialized agents that can work together intelligently. By 2025, we expect to see a fundamental shift from monolithic AI models to multi-agent systems that collaborate to solve complex problems. We're already seeing early signs of this transition with frameworks like Microsoft's AutoGen, which enables sophisticated agent-to-agent conversations and human-in-loop interactions.

We envision a clear hierarchy in these multi-agent systems. At the foundation are research agents optimized for fast information retrieval, designed to rapidly process and find relevant information. These feed into analysis agents that focus on pattern recognition and synthesis, working with more specialized decision agents that handle complex reasoning tasks. Finally, implementation agents execute specific actions with robust error handling capabilities. Each agent is optimized for its specific role, making the entire system more efficient than a single large model trying to do everything.



Proposed ChatDev architecture. Image adapted from "Communicative Agents for Software Development," Qian et al. (2023).

> *Different design patterns for multi-agent collaboration.*

The advantages of this approach extend beyond mere specialization. These systems can adapt and improve through self-directed learning and interaction, similar to how Virtual Scientists (VirSci) mimics teamwork in scientific research. We're seeing breakthroughs in self-learning capabilities, like Meta FAIR's Self-Taught Evaluator that can generate synthetic preference data for continuous improvement.

By 2025, we expect these multi-agent systems to operate 'under the hood' of most AI applications. Users will interact with a single copilot interface, while behind the scenes, multiple specialized agents collaborate to complete complex tasks. The real innovation isn't just in the individual agents but in how they communicate, share context, and work together to solve problems that would be intractable for any single model.

Beyond Brute Force: AI's Next Chapter

2025 promises to be yet another transformative year in AI but not in the ways many expect. The trends we've outlined—from the evolution of System 2 reasoning to the rise of multi-agent systems, from the cost reckoning sparked by DeepSeek to the emergence of service-as-software business models—all point to a fundamental shift in how AI systems will be built, deployed, and monetized. **We're seeing a maturation of the industry that favors intelligence over brute force**,

where success comes from smarter architectures and efficient deployment rather than just larger models and more compute.

For investors and founders alike, this evolution demands a fundamental rethinking of how we evaluate and build AI companies. While traditional metrics like revenue growth remain important, the winners will be those who can achieve computational efficiency while maintaining sustainable cost structures. This means considering not just technical sophistication—like reasoning capabilities and multi-agent architectures—but also innovative business models that can capture AI's value in entirely new ways.

As we enter 2025, the industry stands at a crossroads between two visions of AI's future. One follows the traditional path of massive infrastructure investments and scaling; the other embraces a more efficient, focused approach that emphasizes intelligence over raw compute power. For those building in this space, success will come not from chasing the latest technical trends, but from solving real business problems in ways that weren't possible before—and doing so sustainably.

FOOTNOTES

1. Unless you assume that the U.S. economy will also grow exponentially at the same rate, which is unprecedented in human history and very unlikely given the scale of disruption.