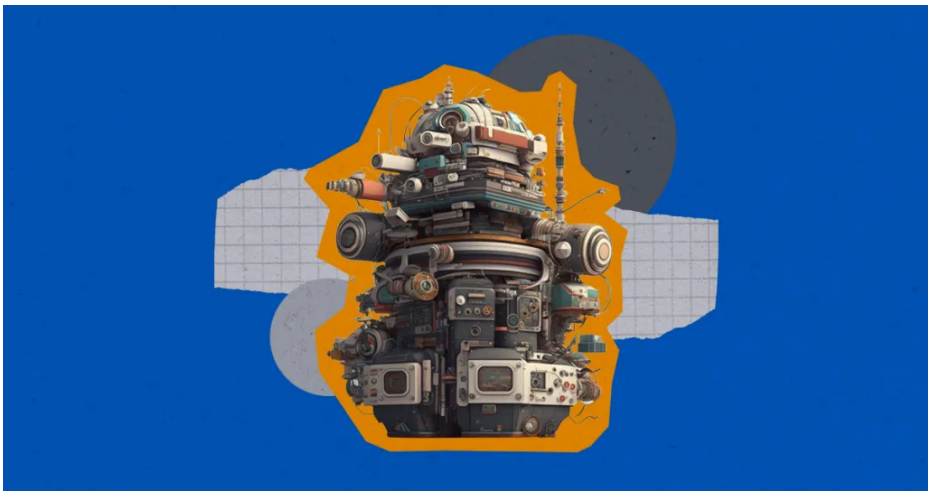
Who Owns the Generative AI Platform?

by Matt Bornstein, Guido Appenzeller, and Martin Casado

AI, machine & deep learning • enterprise & SaaS • Generative AI • machine learning

TABLE OF CONTENTS



We're starting to see the very early stages of a tech stack emerge in generative artificial intelligence (AI). Hundreds of new startups are rushing into the market to develop foundation models, build AI-native apps, and stand up infrastructure/tooling.

Many hot technology trends get over-hyped far before the market catches up. But the generative AI boom has been accompanied by real gains in real markets, and real traction from real companies. Models like Stable Diffusion and ChatGPT are setting historical records for user growth, and several applications have reached $100 million of annualized revenue less than a year after launch. Side-by-side comparisons show AI models outperforming humans in some tasks by multiple orders of magnitude.

So, there is enough early data to suggest massive transformation is taking place. What we don't know, and what has now become the critical question, is: **Where in this market will value accrue?**

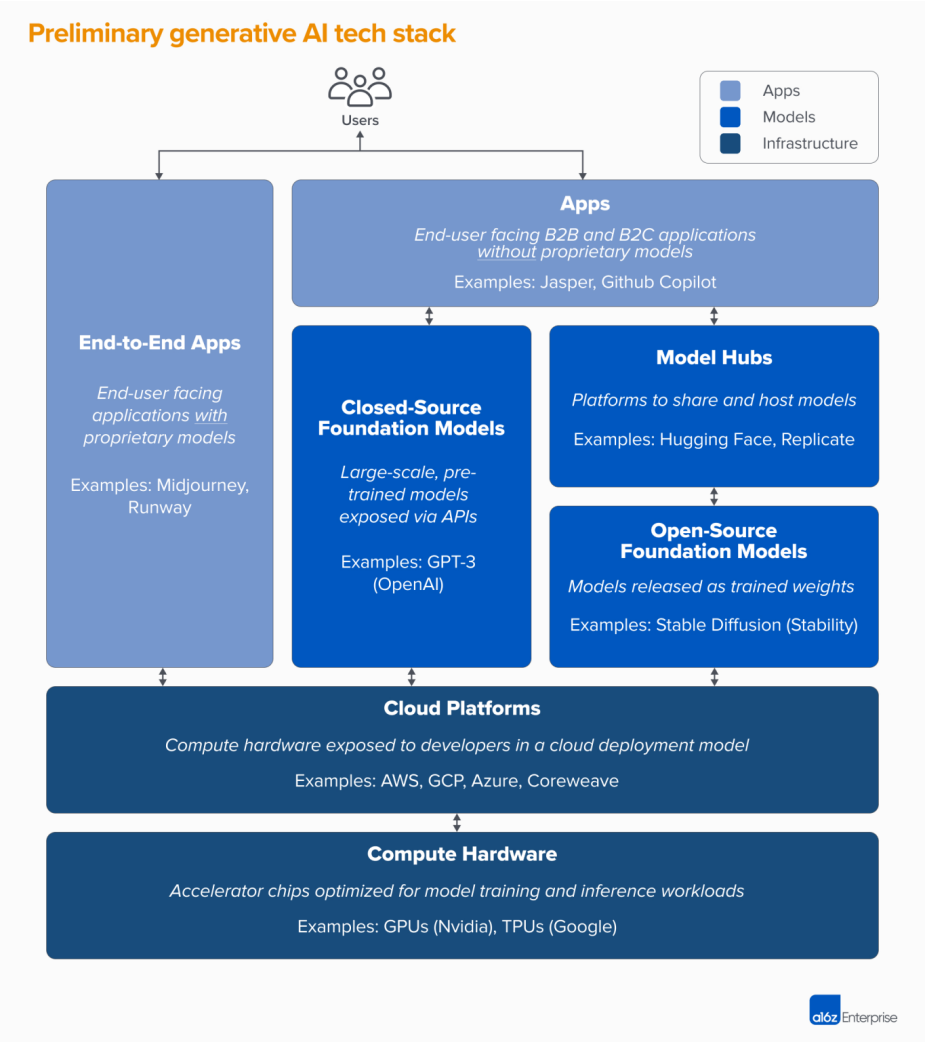
Over the last year, we've met with dozens of startup founders and operators in large companies who deal directly with generative AI. We've observed that **infrastructure vendors** are likely the biggest winners in this market so far, capturing the majority of dollars flowing through the stack. **Application companies** are growing topline revenues very quickly but often struggle with retention, product differentiation, and gross margins. And most **model providers**, though responsible for the very existence of this market, haven't yet achieved large commercial scale.

In other words, the companies creating the most value — i.e. training generative AI models and applying them in new apps — haven't captured most of it. Predicting what will happen next is much harder. But we think the key thing to understand is which parts of the stack are truly differentiated and defensible. This will have a major impact on market structure (i.e. horizontal vs. vertical company development) and the drivers of long-term value (e.g. margins and retention). So far, we've had a hard time finding structural defensibility *anywhere* in the stack, outside of traditional moats for incumbents.

We are incredibly bullish on generative AI and believe it will have a massive impact in the software industry and beyond. The goal of this post is to map out the dynamics of the market and start to answer the broader questions about generative AI business models.

High-level tech stack: Infrastructure, models, and apps

To understand how the generative AI market is taking shape, we first need to define how the stack looks today. Here's our preliminary view.



The stack can be divided into three layers:

- **Applications** that integrate generative AI models into a user-facing product, either running their own model pipelines ("end-to-end apps") or relying on a third-party API
- **Models** that power AI products, made available either as proprietary APIs or as open-source checkpoints (which, in turn, require a hosting solution)
- **Infrastructure** vendors (i.e. cloud platforms and hardware manufacturers) that run training and inference workloads for generative AI models

It's important to note: This is not a market map, but a framework to analyze the market. In each category, we've listed a few examples of well-known vendors. We haven't made any attempt to be comprehensive or list all the amazing generative AI applications that have been released. We're also not going deep here on MLops or LLMops tooling, which is not yet highly standardized and will be addressed in a future post.

The first wave of generative AI apps are starting to reach scale, but struggle with retention and differentiation

In prior technology cycles, the conventional wisdom was that to build a large, independent company, you must own the end-customer — whether that meant individual consumers or B2B buyers. It's tempting to believe that the biggest companies in generative AI will also be end-user applications. So far, it's not clear that's the case.

To be sure, the growth of generative AI applications has been staggering, propelled by sheer novelty and a plethora of use cases. In fact, we're aware of at least three product categories that have already exceeded $100 million of annualized revenue: image generation, copywriting, and code writing.

However, growth alone is not enough to build durable software companies. Critically, growth must be profitable — in the sense that users and customers, once they sign up, generate profits (high gross margins) and stick around for a long time (high retention). In the absence of strong technical differentiation, B2B and B2C apps drive long-term customer value through network effects, holding onto data, or building increasingly complex workflows.

In generative AI, those assumptions don't necessarily hold true. Across app companies we've spoken with, there's a wide range of gross margins — as high as 90% in a few cases but more often as low as 50-60%, driven largely by the cost of model inference. Top-of-funnel growth has been amazing, but it's unclear if current customer acquisition strategies will be scalable — we're already seeing paid acquisition efficacy and retention start to tail off. Many apps are also relatively undifferentiated, since they rely on similar underlying AI models and haven't discovered obvious network effects, or data/workflows, that are hard for competitors to duplicate.

So, it's not yet obvious that selling end-user apps is the only, or even the best, path to building a sustainable generative AI business. Margins should improve as competition and efficiency in language models increases (more on this below). Retention should increase as AI tourists leave the market. And there's a strong argument to be made that vertically integrated apps have an advantage in driving differentiation. But there's a lot still to prove out.

Looking ahead, some of the big questions facing generative AI app companies include:

- **Vertical integration ("model + app").** Consuming AI models as a service allows app developers to iterate quickly with a small team and swap model providers as technology advances. On the flip side, some devs argue that the product *is* the model, and that training from scratch is the only way to create defensibility — i.e. by continually re-training on proprietary product data. But it comes at the cost of much higher capital requirements and a less nimble product team.
- **Building features vs. apps.** Generative AI products take a number of different forms: desktop apps, mobile apps, Figma/Photoshop plugins, Chrome extensions, even Discord bots. It's easy to integrate AI products where users already work, since the UI is generally just a text box. Which of these will become standalone companies — and which will be absorbed by incumbents, like Microsoft or Google, already incorporating AI into their product lines?
- **Managing through the hype cycle.** It's not yet clear whether churn is inherent in the current batch of generative AI products, or if it's an artifact of an early market. Or if the surge of interest in generative AI will fall off as the hype subsides. These questions have important implications for app companies, including when to hit the gas pedal on fundraising; how aggressively to invest in customer acquisition; which user segments to prioritize; and when to declare product-market fit.

Model providers invented generative AI, but haven't reached large commercial scale

What we now call generative AI wouldn't exist without the brilliant research and engineering work done at places like Google, OpenAI, and Stability. Through novel model architectures and heroic efforts to scale training pipelines, we all benefit from the mind-blowing capabilities of current large language models (LLMs) and image-generation models.

Yet the revenue associated with these companies is still relatively small compared to the usage and buzz. In image generation, Stable Diffusion has seen explosive community growth, supported by an ecosystem of user interfaces, hosted offerings, and fine-tuning methods. But Stability gives their major checkpoints away for free as a core tenet of their business. In natural language models, OpenAI dominates with GPT-3/3.5 and ChatGPT. But *relatively* few killer apps built on OpenAI exist so far, and prices have already dropped once.

This may be just a temporary phenomenon. Stability is a new company that hasn't focused yet on monetization. OpenAI has the potential to become a massive business, earning a significant portion of all NLP category revenues as more killer apps are built — especially if their integration into Microsoft's product portfolio goes smoothly. Given the huge usage of these models, large-scale revenues may not be far behind.

But there are also countervailing forces. Models released as open source can be hosted by anyone, including outside companies that don't bear the costs associated with large-scale model training (up to tens or hundreds of millions of dollars). And it's not clear if any closed-source models can maintain their edge indefinitely. For example, we're starting to see LLMs built by companies like Anthropic, Cohere, and Character.ai come closer to OpenAI levels of performance, trained on similar datasets (i.e. the internet) and with similar model architectures. The example of Stable Diffusion suggests that *if* open source models reach a sufficient level of performance and community support, then proprietary alternatives may find it hard to compete.

Perhaps the clearest takeaway for model providers, so far, is that commercialization is likely tied to hosting. Demand for proprietary APIs (e.g. from OpenAI) is growing rapidly. Hosting services for open-source models (e.g. Hugging Face and Replicate) are emerging as useful hubs to easily share and integrate models — and even have some indirect network effects between model producers and consumers. There's also a strong hypothesis that it's possible to monetize through fine-tuning and hosting agreements with enterprise customers.

Beyond that, though, there are a number of big questions facing model providers:

- **Commoditization.** There's a common belief that AI models will converge in performance over time. Talking to app developers, it's clear that hasn't happened yet, with strong leaders in both text and image models. Their advantages are based not on unique model architectures, but on high capital requirements, proprietary product interaction data, and scarce AI talent. Will this serve as a durable advantage?
- **Graduation risk.** Relying on model providers is a great way for app companies to get started, and even to grow their businesses. But there's incentive for them to build and/or host their own models once they reach scale. And many model providers have highly skewed customer distributions, with a few apps representing the majority of revenue. What happens if/when these customers switch to in-house AI development?
- **Is money important?** The promise of generative AI is so great — and also potentially so harmful — that many model providers have organized as public benefit corporations (B corps), issued capped profit shares, or otherwise incorporated the public good explicitly into their mission. This has not at all hindered their fundraising efforts. But there's a reasonable discussion to have around whether most model providers actually *want* to capture value, and if they should.

Infrastructure vendors touch everything, and reap the rewards

Nearly everything in generative AI passes through a cloud-hosted GPU (or TPU) at some point. Whether for model providers / research labs running training workloads, hosting companies running inference/fine-tuning, or application companies doing some combination of both — FLOPS are the lifeblood of generative AI. For the first time in a very long time, progress on the most disruptive computing technology is massively compute bound.

As a result, a lot of the money in the generative AI market ultimately flows through to infrastructure companies. To put some *very* rough numbers around it: We estimate that, on average, app companies spend around 20-40% of revenue on inference and per-customer fine-tuning. This is typically paid either directly to cloud providers for compute instances or to third-party model providers — who, in turn, spend about half their revenue on cloud infrastructure. So, it's reasonable to guess that 10-20% of *total revenue* in generative AI today goes to cloud providers.

On top of this, startups training their own models have raised billions of dollars in venture capital — the majority of which (up to 80-90% in early rounds) is typically also spent with the cloud providers. Many public tech companies spend hundreds of millions per year on model training, either with external cloud providers or directly with hardware manufacturers.

This is what we'd call, in technical terms, "a lot of money" — especially for a nascent market. Most of it is spent at the *Big 3* clouds: Amazon Web Services (AWS), Google Cloud Platform (GCP), and Microsoft Azure. These cloud providers collectively spend more than *$100 billion per year* in capex to ensure they have the most comprehensive, reliable, and cost-competitive platforms. In generative AI, in particular, they also benefit from supply constraints because they have preferential access to scarce hardware (e.g. Nvidia A100 and H100 GPUs).

Interestingly, though, we are starting to see credible competition emerge. Challengers like Oracle have made inroads with big capex expenditures and sales incentives. And a few startups, like Coreweave and Lambda Labs, have grown rapidly with solutions targeted specifically at large model developers. They compete on cost, availability, and personalized support. They also expose more granular resource abstractions (i.e. containers), while the large clouds offer only VM instances due to GPU virtualization limits.

Behind the scenes, running the vast majority of AI workloads, is perhaps the biggest winner in generative AI so far: Nvidia. The company reported $3.8 *billion* of data center GPU revenue in the third quarter of its fiscal year 2023, including a meaningful portion for generative AI use cases. And they've built strong moats around this business via decades of investment in the GPU architecture, a robust software ecosystem, and deep usage in the academic community. One recent analysis found that Nvidia GPUs are cited in research papers 90 times more than the top AI chip startups combined.

Other hardware options do exist, including Google Tensor Processing Units (TPUs); AMD Instinct GPUs; AWS Inferentia and Trainium chips; and AI accelerators from startups like Cerebras, Sambanova, and Graphcore. Intel, late to the game, is also entering the market with their high-end Habana chips and Ponte Vecchio GPUs. But so far, few of these new chips have taken significant market share. The two exceptions to watch are Google, whose TPUs have gained traction in the Stable Diffusion community and in some large GCP deals, and TSMC, who is believed to manufacture *all* of the chips listed here, including Nvidia GPUs (Intel uses a mix of its own fabs and TSMC to make its chips).

Infrastructure is, in other words, a lucrative, durable, and seemingly defensible layer in the stack. The big questions to answer for infra companies include:

- **Holding onto stateless workloads.** Nvidia GPUs are the same wherever you rent them. Most AI workloads are stateless, in the sense that model inference does not require attached databases or storage (other than for the model weights themselves). This means that AI workloads may be more portable across clouds than traditional application workloads. How, in this context, can cloud providers create stickiness and prevent customers from jumping to the cheapest option?
- **Surviving the end of chip scarcity.** Pricing for cloud providers, and for Nvidia itself, has been supported by scarce supplies of the most desirable GPUs. One provider told us that the list price for A100s has actually *increased* since launch, which is highly unusual for compute hardware. When this supply constraint is eventually removed, through increased production and/or adoption of new hardware platforms, how will this impact cloud providers?
- **Can a challenger cloud break through?** We are strong believers that vertical clouds will take market share from the Big 3 with more specialized offerings. In AI so far, challengers have carved out meaningful traction through moderate technical differentiation and the support of Nvidia — for whom the incumbent cloud providers are both the biggest customers and emerging competitors. The long term question is, will this be enough to overcome the scale advantages of the Big 3?

So... where will value accrue?

Of course, we don't know yet. But based on the early data we have for generative AI, combined with our experience with earlier AI/ML companies, our intuition is the following.

There don't appear, today, to be any systemic moats in generative AI. As a first-order approximation, applications lack strong product differentiation because they use similar models; models face unclear long-term differentiation because they are trained on similar datasets with similar architectures; cloud providers lack deep technical differentiation because they run the same GPUs; and even the hardware companies manufacture their chips at the same fabs.

There are, of course, the standard moats: scale moats ("I have or can raise more money than you!"), supply-chain moats ("I have the GPUs, you don't!"), ecosystem moats ("Everyone uses my software already!"), algorithmic moats ("We're more clever than you!"), distribution moats ("I already have a sales team and more customers than you!") and data pipeline moats ("I've crawled more of the internet than you!"). But none of these moats tend to be durable over the long term. And it's too early to tell if strong, direct network effects are taking hold in any layer of the stack.

Based on the available data, it's just not clear if there will be a long-term, winner-take-all dynamic in generative AI.

This is weird. But to us, it's good news. The potential size of this market is hard to grasp — somewhere between *all software* and *all human endeavors* — so we expect many, many players and healthy competition at all levels of the stack. We also expect both horizontal and vertical companies to succeed, with the best approach dictated by end-markets and end-users. For example, if the primary differentiation in the end-product is the AI itself, it's likely that verticalization (i.e. tightly coupling the user-facing app to the home-grown model) will win out. Whereas if the AI is part of a larger, long-tail feature set, then it's more likely horizontalization will occur. Of course, we should also see the building of more traditional moats over time — and we may even see new types of moats take hold.

Whatever the case, one thing we're certain about is that generative AI changes the game. We're all learning the rules in real time, there is a tremendous amount of value that will be unlocked, and the tech landscape is going to look much, much different as a result. And we're here for it!

All images in this post were created using Midjourney.

***

January 19, 2023

Related Stories

AI + a16z

by a16z editorial

Emerging Architectures for LLM Applications

by Matt Bornstein and Rajko Radovanovic

The Getting Started with AI Stack for JavaScript

by Yoko Li, Jennifer Li, and Martin Casado

It's Not a Computer, It's a Companion!

by Justine Moore, Bryan Kim, Yoko Li, and Martin Casado

Navigating the High Cost of AI Compute

by Guido Appenzeller, Matt Bornstein, and Martin Casado

The enterprise is changing

Sign up for our enterprise newsletter to get the a16z take on the trends reshaping B2B and enterprise tech.

Email Address* Sign Up