AUG 10, 2023

Demystifying the LLM Tech Stack (Part III: The Application Layer)

How to build an application-layer LLM company with a solid tech stack.



By Jenny Xiao, Jay Zhao, and Val Gui

Leonis Capital · leoniscap.com/research/demystifying-the-llm-tech-stack-(part-iii-the-application-layer)/

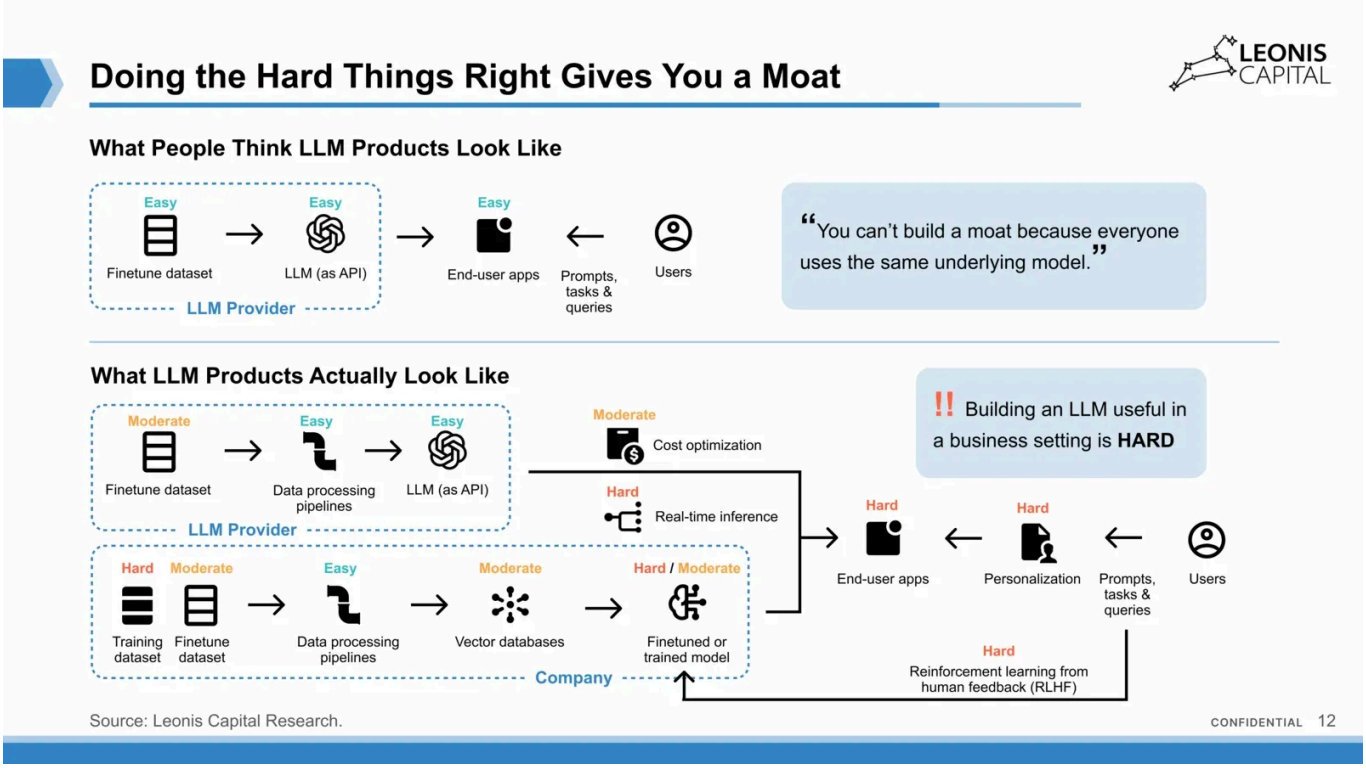


[Full webinar deck available here.]

In the rapidly evolving world of Large Language Models (LLMs), one question looms large for startups and investors alike: How can companies build defensible moats in this space? In our final session on the LLM tech stack, we tackle this question head-on.

The Myth of "No Moats" in AI

You've probably heard it before: "There are no real moats in AI because everyone is using the same underlying models like GPT-4." At Leonis, we strongly disagree. While it's true that many companies leverage similar base models, building a truly useful and commercially viable AI product involves far more complexity than simply calling an API. There's a vast difference between cobbling together a minimum viable product (MVP) to show investors and creating a robust, enterprise-grade AI application.



The backend of production AI apps involves numerous challenging steps, such as finding and curating high-quality training data, cost optimization across different model types, real-time inference and low-latency responses, personalization and context management, implementing reinforcement learning from human feedback (RLHF), and potentially training custom models. By excelling in these difficult areas, startups can absolutely create technological moats that are not easily replicable.

A Staged Approach to Development

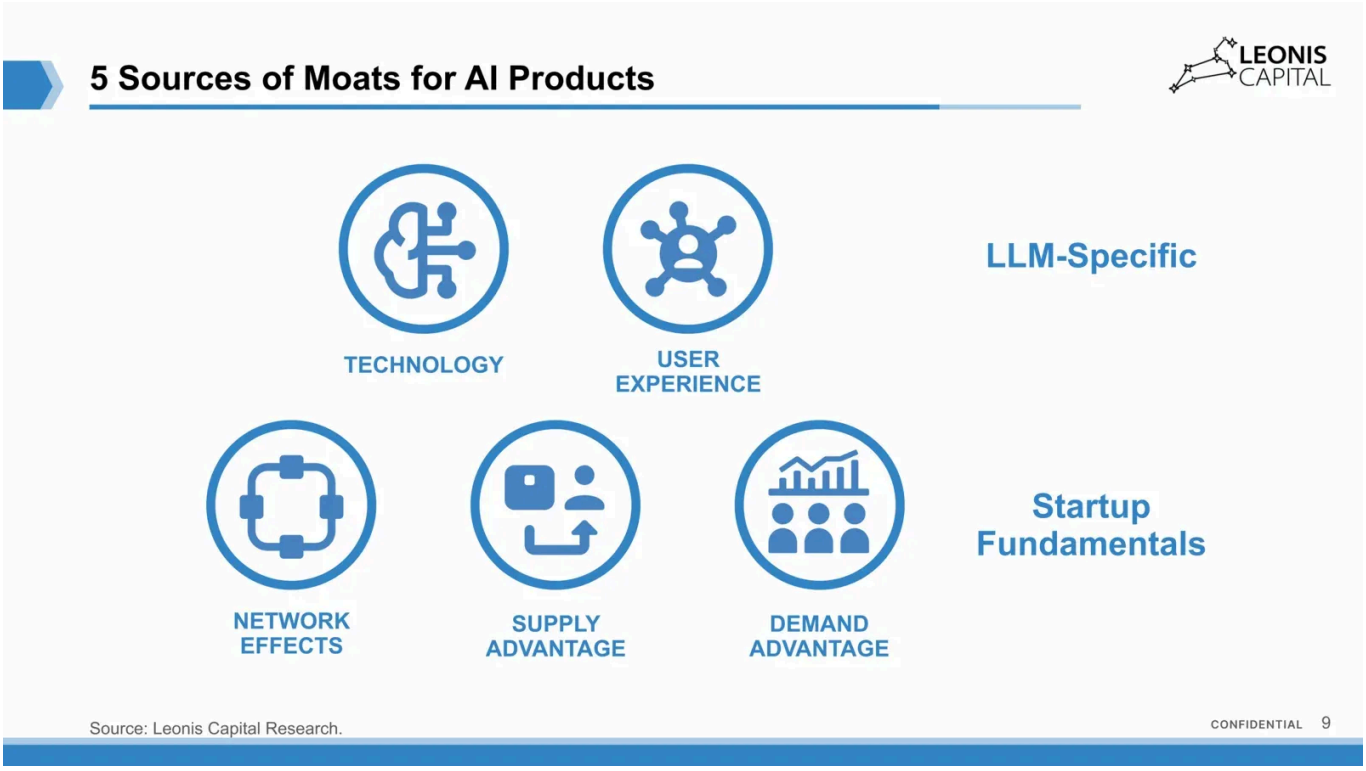
Startups should take a staged approach to developing their products:

1. **Start by building on top of proprietary models** like GPT-4 to test product-market fit quickly.
2. **Fine-tune open-source models** as you gather data and better understand your use case, dramatically reducing costs.
3. **Train custom models** from scratch once you hit performance bottlenecks with fine-tuned models for maximum control and performance.

This iterative process allows startups to balance speed, cost, and performance as they scale. And remember, disruption from new models is inevitable - successful companies will need to adapt their tech stack as cutting-edge models emerge.

Five Sources of Moats in AI Applications

During our research, we identified at least five key areas where AI companies can build defensible moats. While many focus solely on the technology side, business fundamentals remain crucial for AI startups. Let's dive deeper into each of these potential moat sources.



Technology Moats

There are many technical challenges in productionizing AI beyond simply calling an API. Companies that excel in areas like efficient fine-tuning, low-latency inference, and effective use of reinforcement learning can create significant technological barriers to entry. Additionally, gathering large amounts of high-quality, domain-specific data over time can lead to models that dramatically outperform general-purpose alternatives in niche applications.

User Experience (UX) Moats

User experience improvements can be a powerful differentiator in AI products. Take ChatGPT, for example, compared to the raw GPT-3 API. Key UX enhancements in ChatGPT included a more intuitive, chat-like interface, streaming tokens for a more natural reading experience, and maintaining context across multiple messages.

For AI startups, reducing user workload and frustration can be a quick win. Some strategies include providing preset tasks and commands, simplifying data input processes, and offering user-specific settings and preferences. Balancing model performance with low latency is another crucial UX challenge. Users have little patience for long wait times, so optimizing response speed is essential.

Network Effects

While not often associated with AI products, network effects can play a significant role, especially for open-source AI tools. As more users contribute to and improve an open-source project, it becomes increasingly valuable and difficult to replicate. Companies can monetize successful open-source projects through "open core" business models, offering advanced features or hosted versions for paying customers.

Supply Advantages

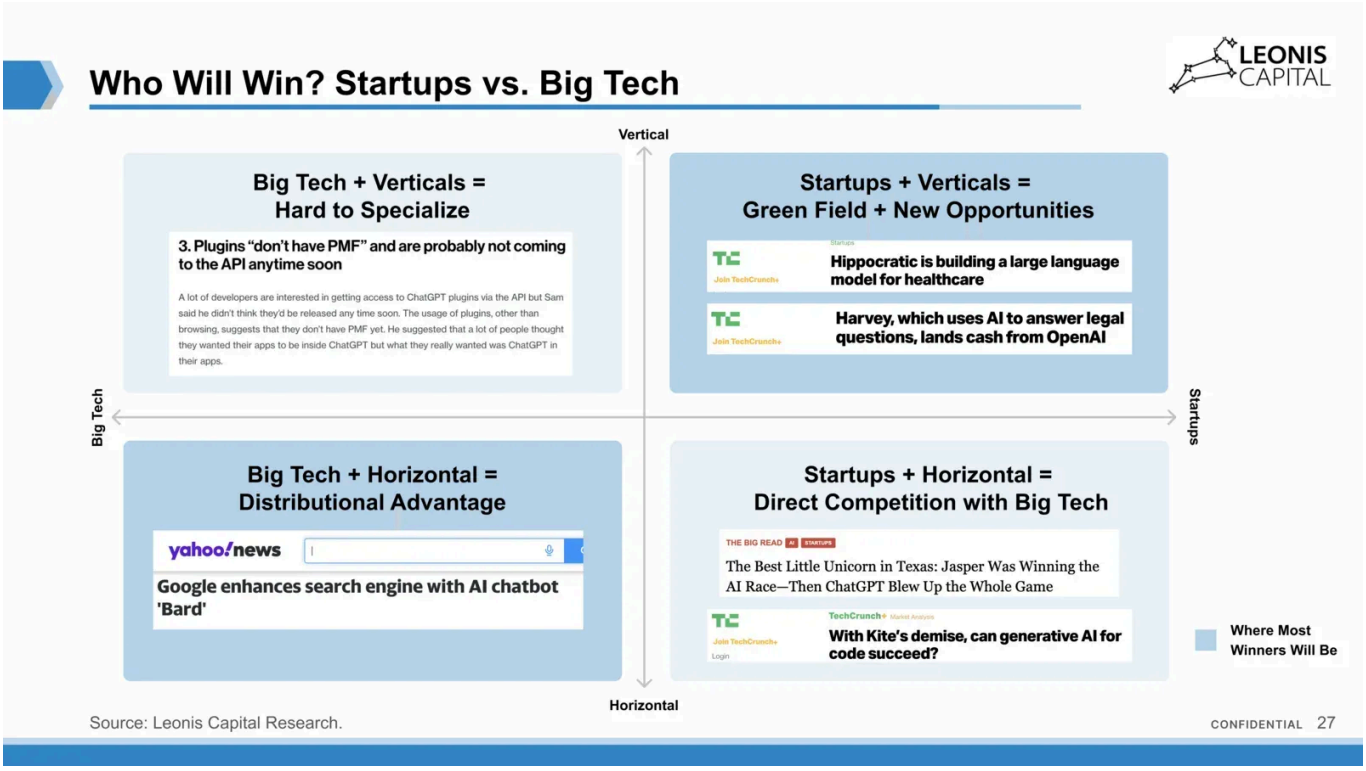
Controlling the supply of an AI product or service can create a strong moat. This is particularly relevant in highly regulated industries like healthcare, finance, and law. Factors that can lead to supply advantages include regulatory hurdles and compliance requirements, stringent data security and privacy measures, and exclusive partnerships with key enterprises or data providers. High-fidelity verticals like healthcare, legal services, and finance offer some of the best opportunities for startups to build defensible positions against big tech incumbents.

Demand Advantages

On the demand side, companies can create stickiness through superior products that consistently outperform alternatives, high switching costs by deeply integrating into enterprise workflows, and network effects that increase value as more users join the platform. While cutting-edge AI technology is exciting, founders shouldn't overlook these traditional business moats that have proven effective in the software industry for decades.

Vertical vs. Horizontal AI Applications

An important strategic consideration for AI startups is whether to pursue vertical (industry-specific) or horizontal (general-purpose) applications. **We believe that startups are more likely to succeed in vertical domains, while big tech companies will dominate horizontal applications.**



Vertical AI applications in areas like legal tech, healthcare, and specialized industrial use cases offer several advantages for startups, including less direct competition from tech giants, the ability to build deep domain expertise, more opportunities for data and workflow lock-in, and a higher willingness to pay from enterprise customers. Meanwhile, horizontal applications like

general-purpose chatbots or search tools play more to the strengths of big tech companies with their massive distribution advantages and existing user bases.

Key Takeaways for Founders and Investors

To wrap things up, here are some key takeaways for founders and investors looking to build or invest in AI applications:

1. **Don't underestimate the complexity** of building production-ready AI applications. There's much more involved than simply calling an API.
2. **Take a staged approach to model development:** start with proprietary APIs, move to fine-tuned open-source models, and only build custom models when truly necessary.
3. **Focus on user experience early.** Many current AI products have significant UX issues that create opportunities for differentiation.
4. **Look beyond pure technology.** Network effects, supply advantages, and demand advantages can be powerful moats in AI businesses.
5. **For startups, vertical AI applications** often offer better opportunities to build defensible positions compared to horizontal tools.
6. **Be prepared for ongoing disruption** from new model releases. Build flexibility into your tech stack and business model.
7. **Traditional startup fundamentals still matter** immensely in AI. Don't neglect things like unit economics, go-to-market strategy, and workflow integration.

In the end, building a defensible moat in the AI space is absolutely possible. It requires a deep understanding of both the technology and business landscape, strategic planning, and relentless execution. By following these guidelines, startups can position themselves for long-term success in this exciting and rapidly evolving field.