JUL 25, 2023

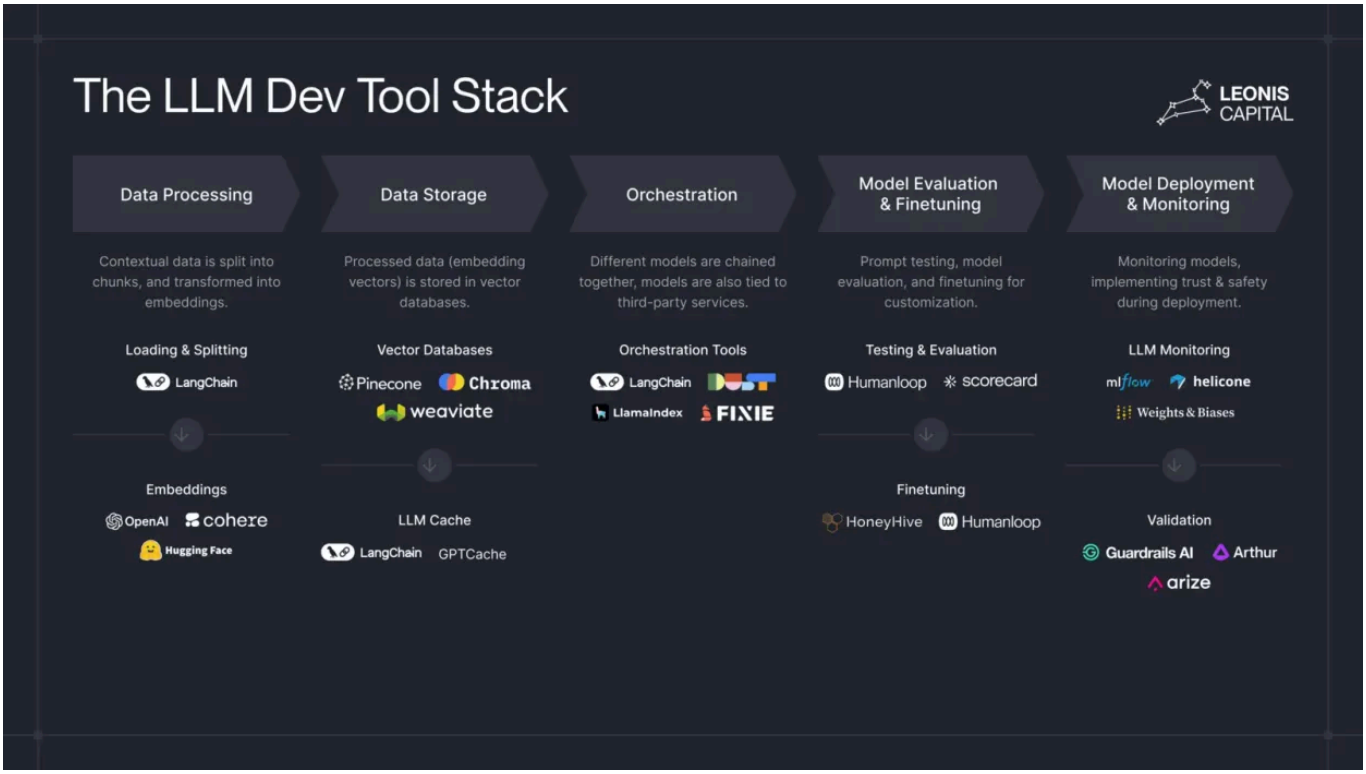
Demystifying the LLM Tech Stack (Part II: The Middle Layer)

What LLM developer tools will look like.



By Jenny Xiao, Jay Zhao, and Val Gui

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

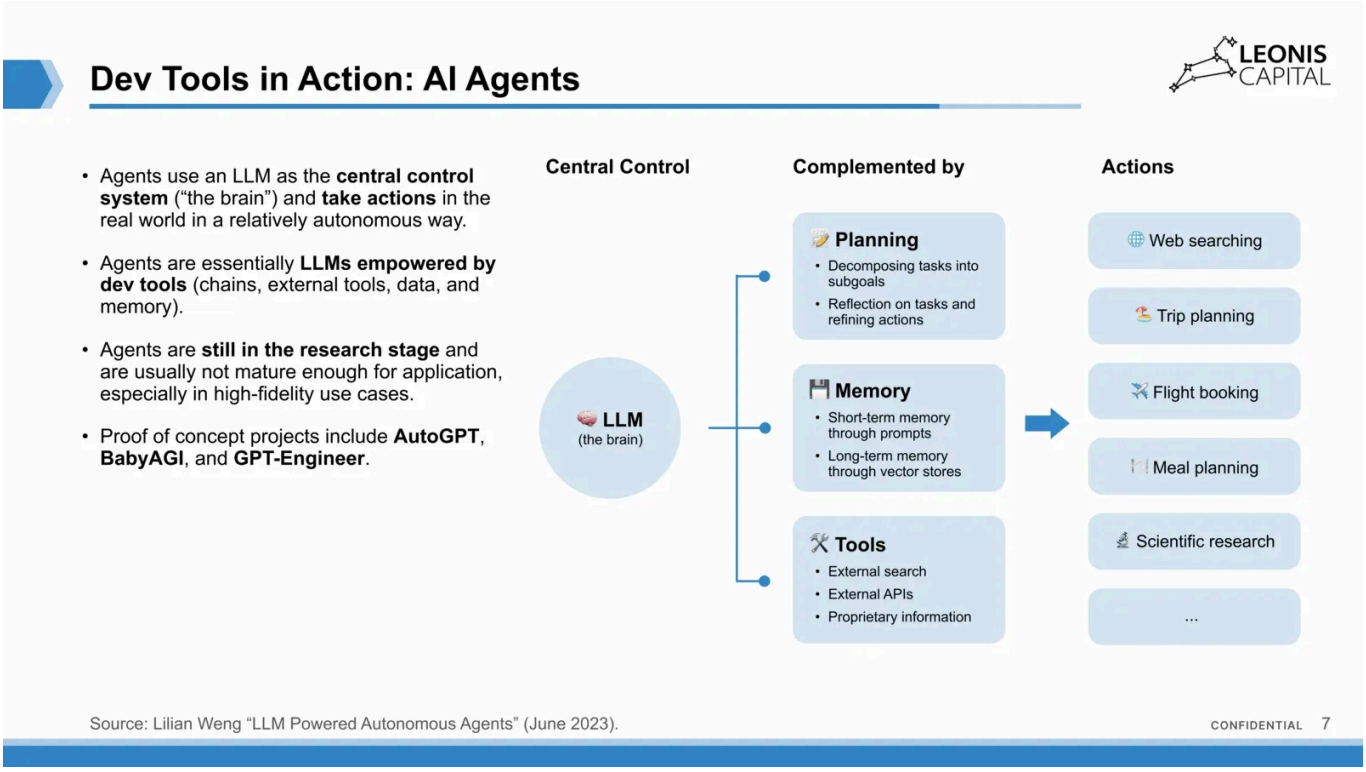


[Full webinar deck available here.]

As investors focused on AI, we've been closely following the rapidly evolving landscape of LLMs and the developer tools being built around them. We recently conducted a deep dive on the LLM tech stack that covers all three layers of the LLM stack — from the foundation layer to the middle layer and the application layer. In a time of AI boom, investing in the middle layer — developer tools — is akin to investing in shovels during a gold rush. This session sheds light on where we think the most exciting opportunities lie in this space.

The Rise of AI Agents

One of the most fascinating developments enabled by LLM developer tools is the emergence of AI agents. These are systems that use LLMs as a central "brain" to take relatively autonomous actions in the real world. Unlike a standard chatbot, an agent can plan, decompose tasks into sub-goals, reflect on its performance, and interact with external tools and data sources.



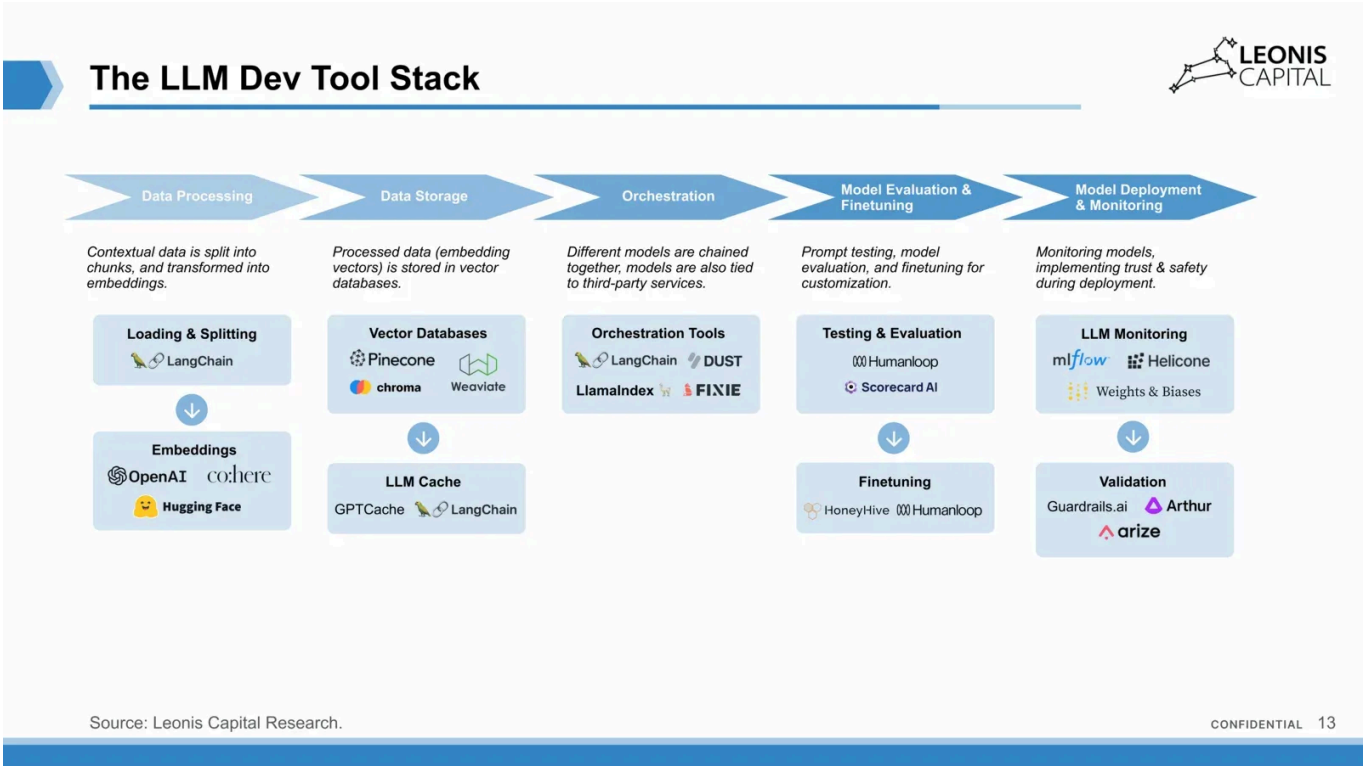
We've already seen some impressive proof-of-concept projects like AutoGPT, which can autonomously perform research and writing tasks with minimal human input. While still early, I believe agents represent the future of LLMs — moving from co-pilots that work alongside humans to more autonomous systems that can handle complex workflows.

That said, current agents still face significant limitations around performance, stability, and cost. The failed experiment with ChatGPT plugins demonstrated that we're not quite ready for fully autonomous AI assistants in consumer applications. But we're keeping a close eye on this space, as we expect rapid improvements over the next 3-5 years that unlock transformative use cases.

The LLM Developer Tools Landscape

To understand where the opportunities lie, it's helpful to break down the key components of the LLM developer stack:

1. **Data Processing:** Splitting and transforming contextual data into machine-readable embeddings.
2. **Data Storage:** Vector databases for long-term storage and semantic caches for short-term memory.
3. **LLM Orchestration:** Chaining together prompts, managing memory, and connecting to external tools/APIs.
4. **Evaluation & Fine-Tuning:** Customizing models and improving performance.
5. **Deployment & Monitoring:** Ensuring models are secure, compliant, and performing as expected.



Each of these areas has seen a flurry of startup activity. But as an investor, we are wary of following hype cycles. It's critical to look at the long-term value creation potential and understand how the landscape may evolve.

For example, vector databases have been red hot, with VCs clamoring to invest. But we are skeptical of the investment opportunity here. The market is already saturated with well-funded

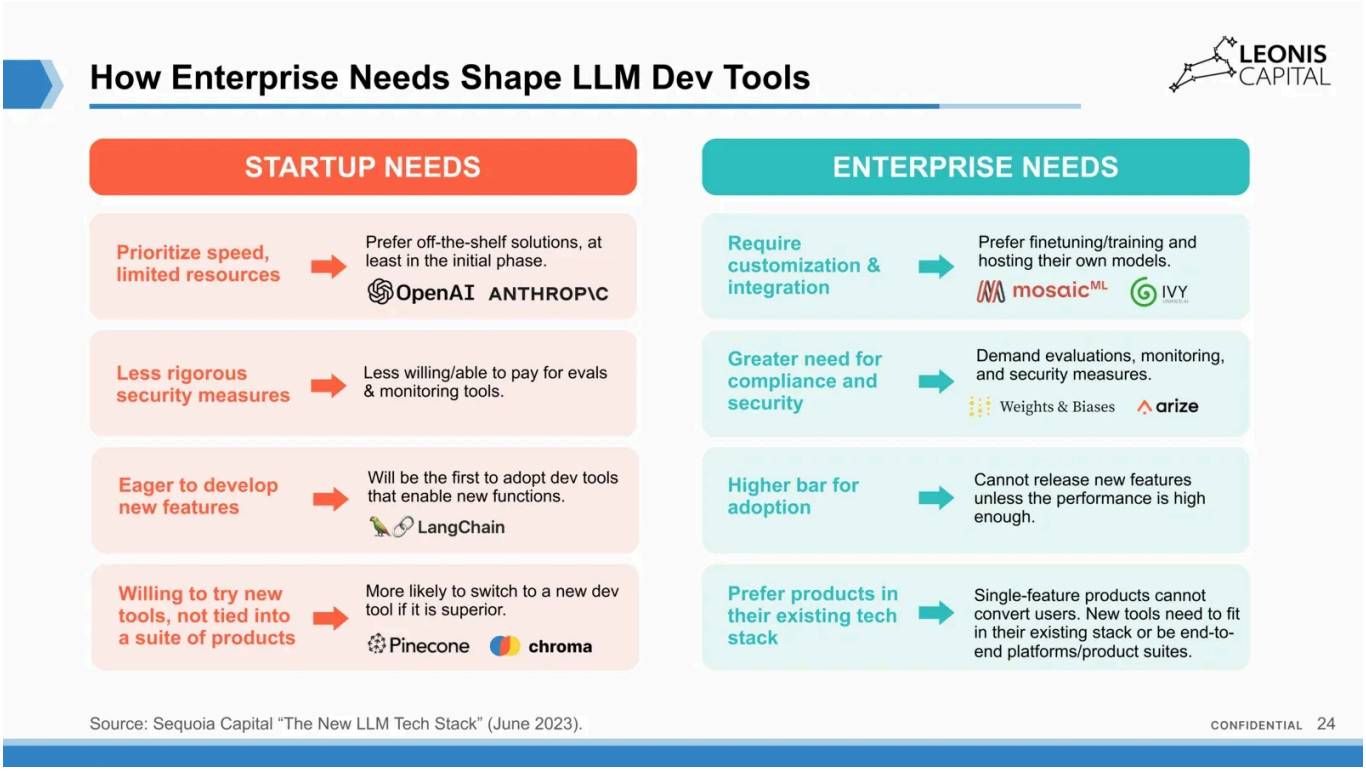
players, and existing database providers are quickly adding vector search capabilities. Plus, this functionality is relatively straightforward to build, limiting the potential for strong defensibility.

Similarly, while orchestration tools like LangChain are currently essential for building advanced LLM applications, we expect much of this functionality to eventually be absorbed by the core LLM providers. OpenAI and others have a strong incentive to make their models easier to use and more powerful out of the box.

Where we see the most compelling opportunities is in tools that address enterprise needs as LLM adoption moves beyond early-stage startups. **Two areas stand out: customization and integration, and compliance and security.** Enterprises need ways to fine-tune models on their proprietary data and integrate LLMs into existing tech stacks. We've invested in a company called Ivy that helps translate between different machine learning frameworks, making it easier for enterprises to adopt open-source models. As regulated industries start adopting LLMs, there will be a massive demand for tools that ensure models are secure, explainable, and compliant with relevant laws and policies.

The Shift from Startup to Enterprise Adoption

A key trend we're watching is the transition from startup to enterprise LLM adoption. The needs and constraints of these two groups are quite different. Startups prioritize speed and often prefer off-the-shelf solutions in the early stages. They're willing to experiment with cutting-edge tools and aren't as constrained by security and compliance requirements. Enterprises, on the other hand, require much more customization, integration with existing systems, and robust security/compliance measures. They have a higher bar for adoption and often can't just rip and replace existing tech stacks.



This shift presents both challenges and opportunities. The initial wave of developer tools has focused on satisfying startup needs — allowing teams to quickly prototype and launch LLM-powered applications. But we believe the next wave of winners will be those that can fulfill enterprise requirements.

Interestingly, we expect enterprise adoption of LLMs to happen much faster than previous waves of AI/ML. With traditional machine learning, enterprises faced friction at every step - from data collection to hiring ML experts to model deployment. But with LLMs, there's now an easier path. Companies can start by fine-tuning existing models on their proprietary data, dramatically lowering the barriers to entry.

This faster adoption cycle means there's a narrow window of opportunity for startups building enterprise-focused LLM tools. Those who can establish themselves as trusted providers in the next 2-3 years will be well-positioned as the market matures.

The Road Ahead

We're still in the early innings of the LLM revolution. The pace of innovation is breathtaking, with new models and capabilities emerging almost weekly. As an investor, it's an incredibly exciting time — but also one that requires careful analysis and a long-term perspective.

In the coming years, we anticipate several key developments that will significantly shape the landscape of LLM technology.

First, the lines between different layers of the tech stack will start to blur. The stack, currently divided into the foundation, middle, and application layers, will see base models absorbing more functionality, leading to a convergence. Core LLM providers will integrate advanced features directly into their models, simplifying the development of sophisticated applications without extensive middleware. Meanwhile, application-layer companies will need deeper, specialized technologies to stand out in a competitive market.

Second, enterprises across various industries will reshape their tech stacks around LLMs, creating massive opportunities for startups facilitating this transition. LLM adoption will spread beyond tech giants and early adopters to sectors like finance, healthcare, and manufacturing. These enterprises will need tools that integrate LLM capabilities into existing systems, ensuring compatibility, security, and scalability. Startups offering robust, enterprise-grade solutions, from customized model training to compliance tools, will be in high demand, bridging the gap between cutting-edge LLM technology and traditional enterprise infrastructure.

Third, AI agents will evolve from research projects to practical tools, creating new categories of products and services. AI agents like AutoGPT, while currently experimental, will become more reliable and capable, handling complex, multi-step workflows autonomously. This evolution will enable new applications in customer service, autonomous research, personalized education, and beyond, transitioning AI agents from supportive co-pilots to autonomous systems driving efficiency and innovation.

Finally, as these systems become more powerful and prevalent, we will grapple with thorny questions around AI safety, security, and governance. The increasing capabilities of LLMs and AI agents pose risks that must be managed. Ensuring AI systems are safe, secure, and aligned with human values will be crucial. Issues like data privacy, algorithmic bias, and AI misuse will require robust regulatory frameworks and ethical guidelines. Companies must implement stringent measures to protect sensitive information, maintain transparency, and foster trust, addressing these challenges for the sustainable and responsible development of AI technology.

For founders building in this space, our advice is to stay laser-focused on creating tangible value for users while remaining adaptable as the landscape evolves. For fellow investors, it's crucial to look beyond the hype and develop deep, nuanced views on how this technology will reshape industries.

The companies being founded today have the potential to define the next era of computing. It's a privilege to play a small role in supporting that journey, and we can't wait to see what emerges in the months and years ahead.