JAN 24, 2024

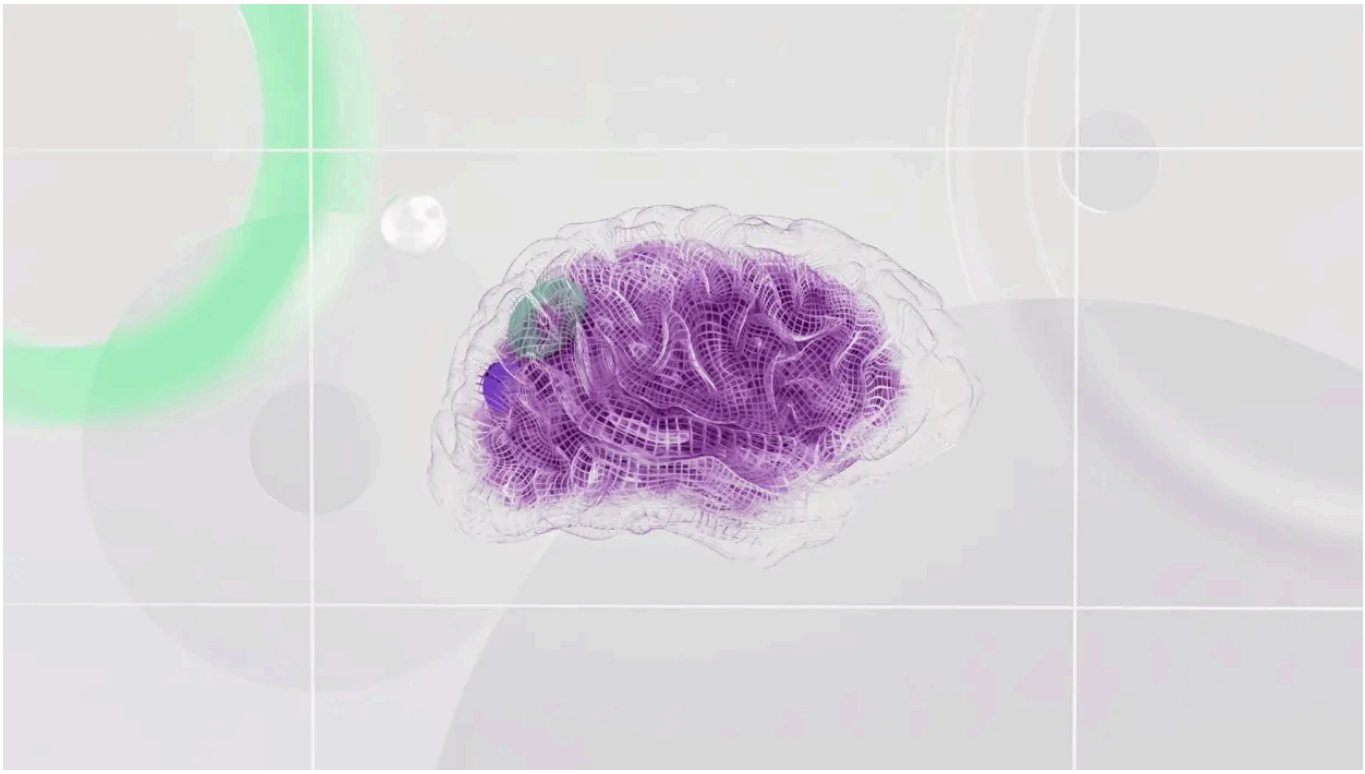
The State of AI in 2024

Our review of AI progress in 2023 and predictions for 2024.



By Jenny Xiao and Jay Zhao

Leonis Capital · leoniscap.com/research/the-state-of-ai-in-2024/

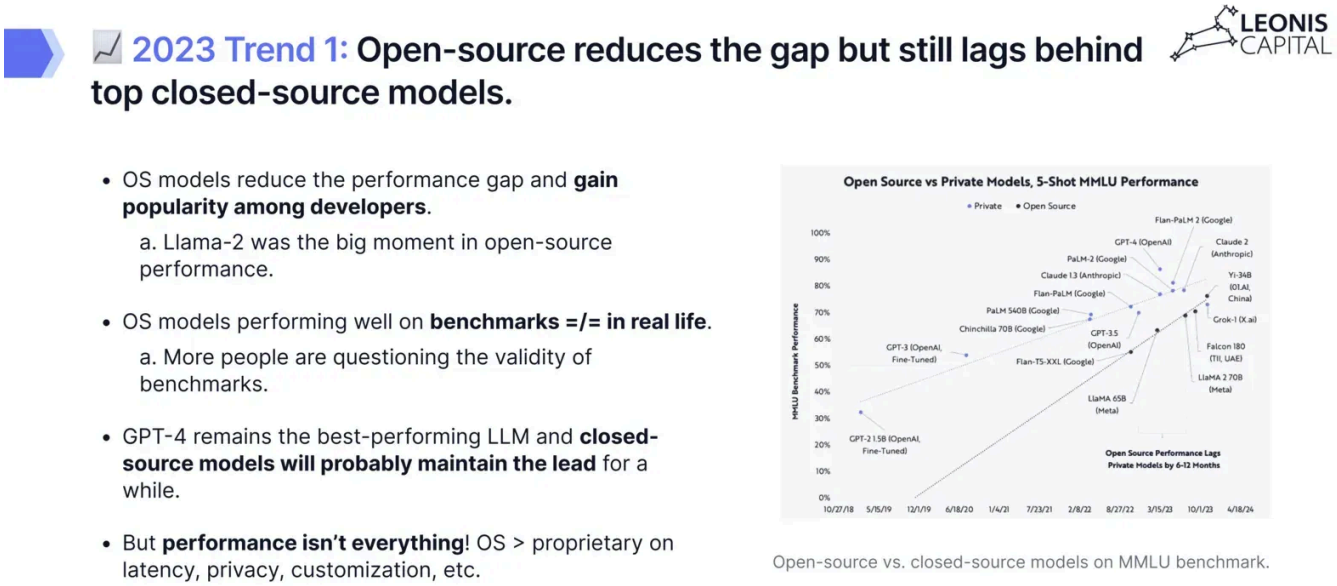


[Full webinar deck available here.]

AI has evolved rapidly over the past few years and 2023 was no exception. The past 12 months have been marked by significant breakthroughs, emerging trends, and heightened discussions on AI safety and governance. As we move into 2024, it is crucial to reflect on the key developments of the past year and anticipate what lies ahead. In this blog post, we will delve into the major AI trends observed in 2023 and offer predictions for 2024.

Trend 1: Open Source Models Gaining Ground But Still Behind

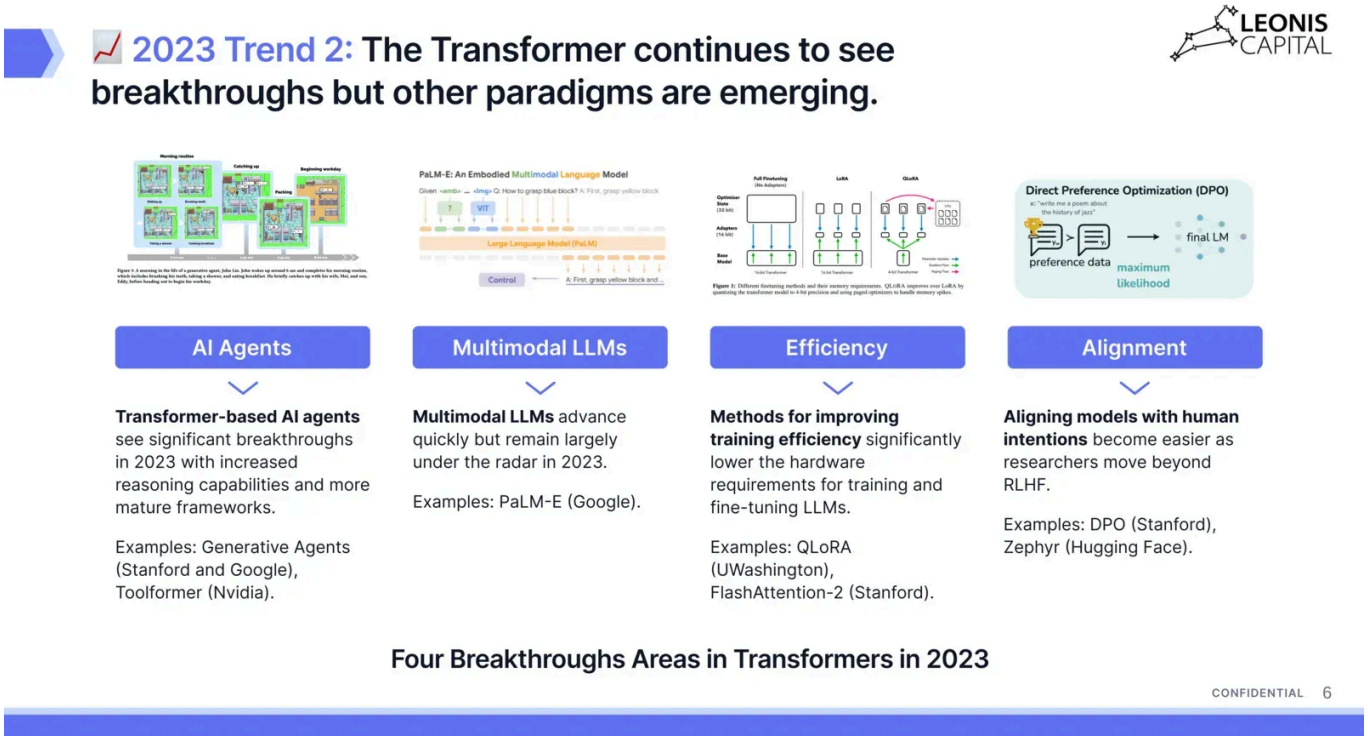
One of the most notable trends in 2023 was the closing performance gap between open-source and closed-source AI models. Historically, closed-source models, such as OpenAI's GPT-4, have led the way in natural language processing (NLP) benchmarks. However, open-source models have made significant strides, reducing this gap considerably. This improvement has made open-source models viable options for many startups and enterprises, offering advantages in data security and privacy. Despite this progress, open-source models still lag behind closed-source models by about six to twelve months in terms of real-world application performance.



Open-source vs. closed-source models on MMLU benchmark.

Trend 2: The Transformer Sees Breakthroughs and But New Paradigms Emerge

Transformers have remained the dominant model architecture, with substantial improvements in efficiency, reasoning capabilities, and the development of AI agents. A Stanford paper on generative agents demonstrated how AI agents could exhibit social behaviors without human prompting, highlighting the potential for multi-agent collaboration. Additionally, multimodal large language models (LLMs), combining text, image, and video data, have shown promise in creating more robust and capable models.



New paradigms, such as Mamba and liquid neural networks (LNNs), have also emerged. Mamba, developed by Stanford, offers linear scaling, making it more efficient for longer sequences compared to the quadratic scaling of transformers. Liquid neural networks, introduced by MIT spinoff Liquid AI, present a novel approach with potential paradigm-shifting implications for deep learning.

Trend 3: The LLM Tech Stack Matures

The LLM tech stack has seen maturation with the adoption of retrieval-augmented generation (RAG) techniques. RAG, which integrates LLMs with vector databases, has evolved from a basic approach to a more modular and sophisticated system. This method addresses the issue of hallucination in LLMs by incorporating external information to improve accuracy. However, RAG still faces engineering challenges, particularly in handling large datasets and ensuring data security.

📈 2023 Trend 3: RAG becomes the standard tech stack for building LLM apps but significant pain points exist.

LEONIS CAPITAL

- **Retrieval Augmented Generation (RAG)** helps LLMs incorporate external knowledge and quickly became the standard way to build LLM applications.
- RAG has gone through **three generations**: 1) Naive RAG; 2) Advanced RAG; 3) Modular RAG.

➡ A sign that the LLM tech stack is maturing!

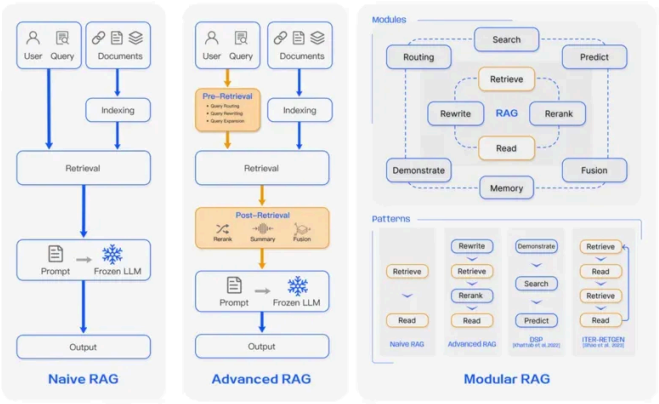
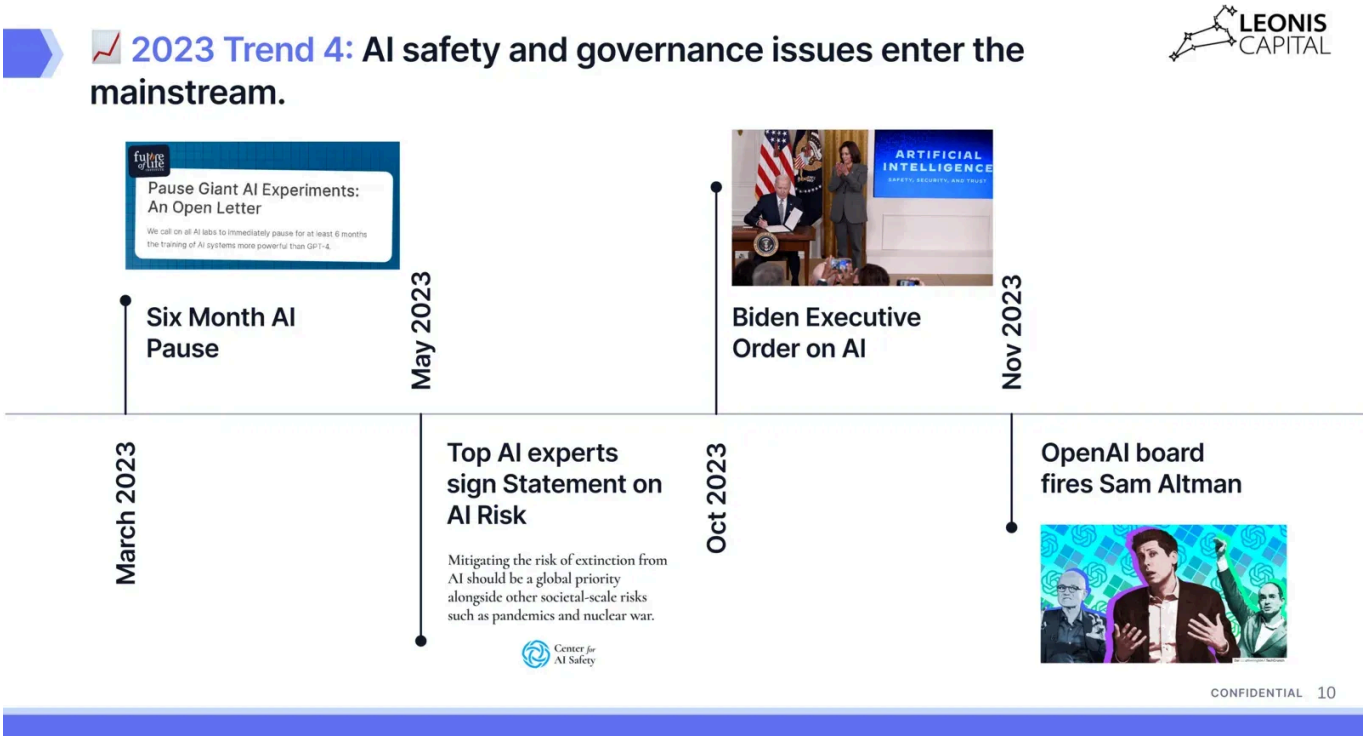


Figure 3: Comparison between the three paradigms of RAG

Image from "RAG Survey" (Tongji University 2023).

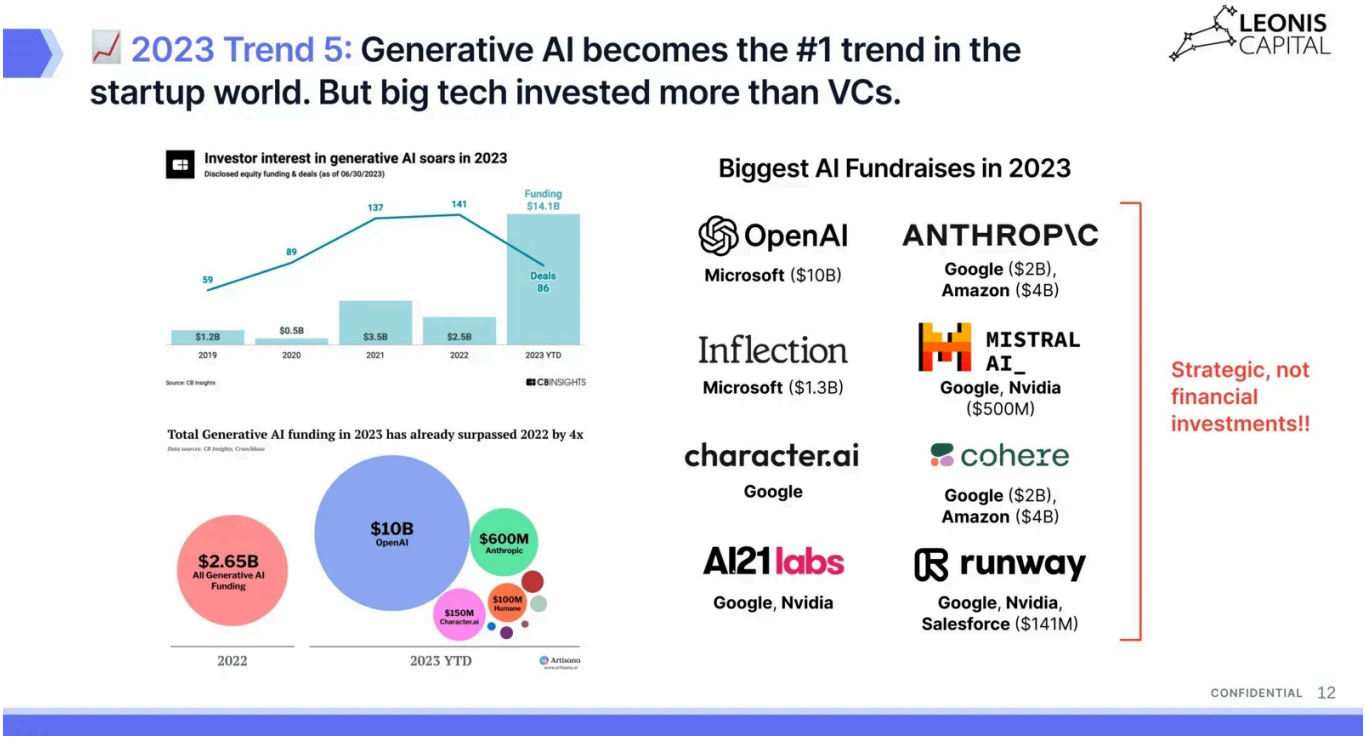
Trend 4: AI Safety and Governance Enter the Mainstream

AI safety and governance became mainstream topics in 2023, driven by high-profile events and regulatory initiatives. An open letter from the Future of Life Institute called for a six-month pause on LLM development, citing existential risks. This sentiment was echoed in a statement from AI experts and influenced the Biden administration's executive order on AI, which imposed model reporting requirements and other regulatory measures. The firing and rehiring of Sam Altman at OpenAI underscored the importance of governance structures for AI companies.



Trend 5: Generative AI Becomes the #1 Trend in the Startup and VC World

In 2023, generative AI was the leading trend in the VC and startup world, but what stood out was the significant investment from Big Tech, surpassing that from VCs. A total of $14.1 billion was invested in generative AI, with major players like OpenAI receiving $10 billion, Anthropic $600 million, and Character AI $150 million. These substantial investments primarily came from tech giants like Microsoft, Google, and Nvidia, rather than traditional venture capitalists. The strategic rationale behind these investments is not necessarily about financial returns but about building on top of these models for competitive advantage.



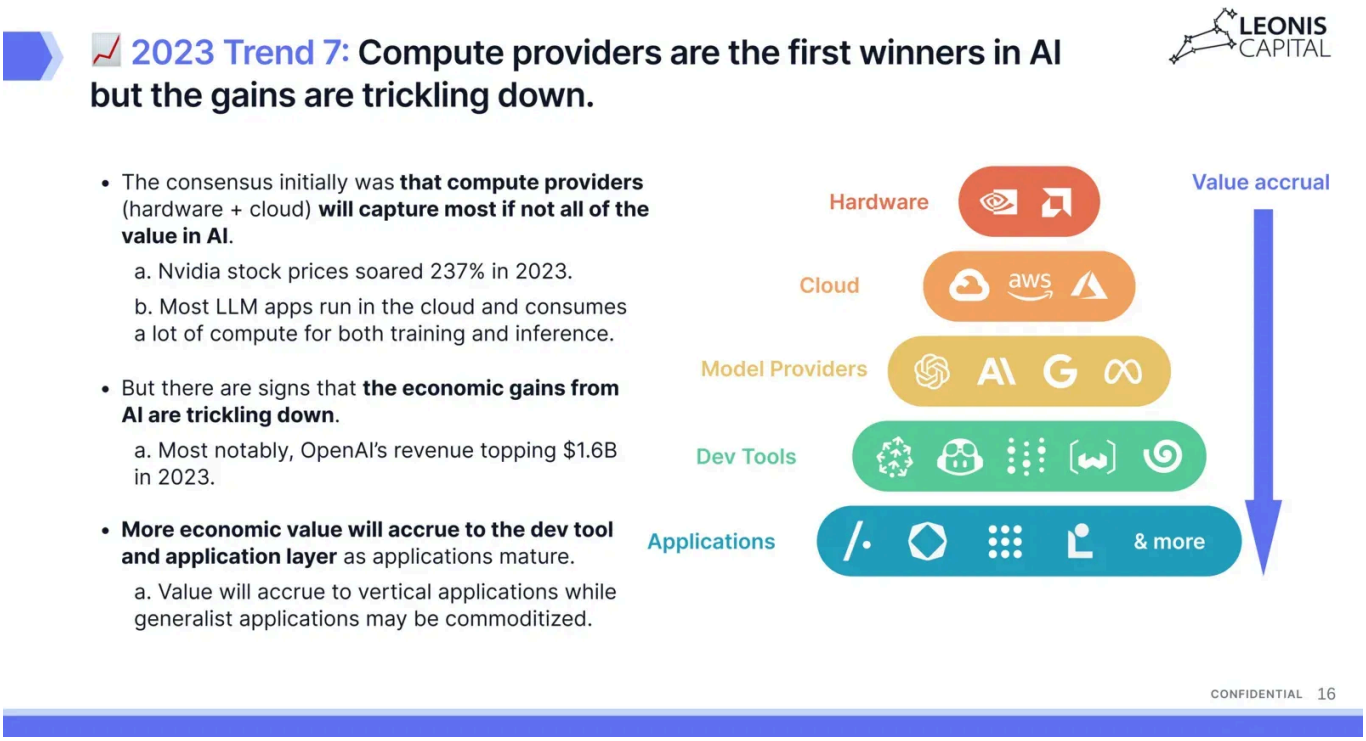
Trend 6: Enterprises Try but Struggle with AI Integration

In early 2023, many enterprises saw the potential of LLMs and aimed to create internal or customer-facing AI products. However, the complexity of integrating LLMs into their existing systems proved to be far greater than anticipated. The primary difficulties enterprises encounter are due to their existing tech stack and team structure. Traditional enterprise SaaS tech stacks are designed for stability and business applications, whereas LLM tech stacks focus on continuous AI model development and integration. This discrepancy requires enterprises to overhaul significant parts of their infrastructure, a process that is both time-consuming and resource-intensive. Additionally, traditional engineering teams often lack the specialized skills needed for AI-first development, further complicating efforts to implement LLMs effectively. We wrote an entire blog post on this topic and you can read it here.

Trend 7: Value Capture Shifts from Compute to Applications

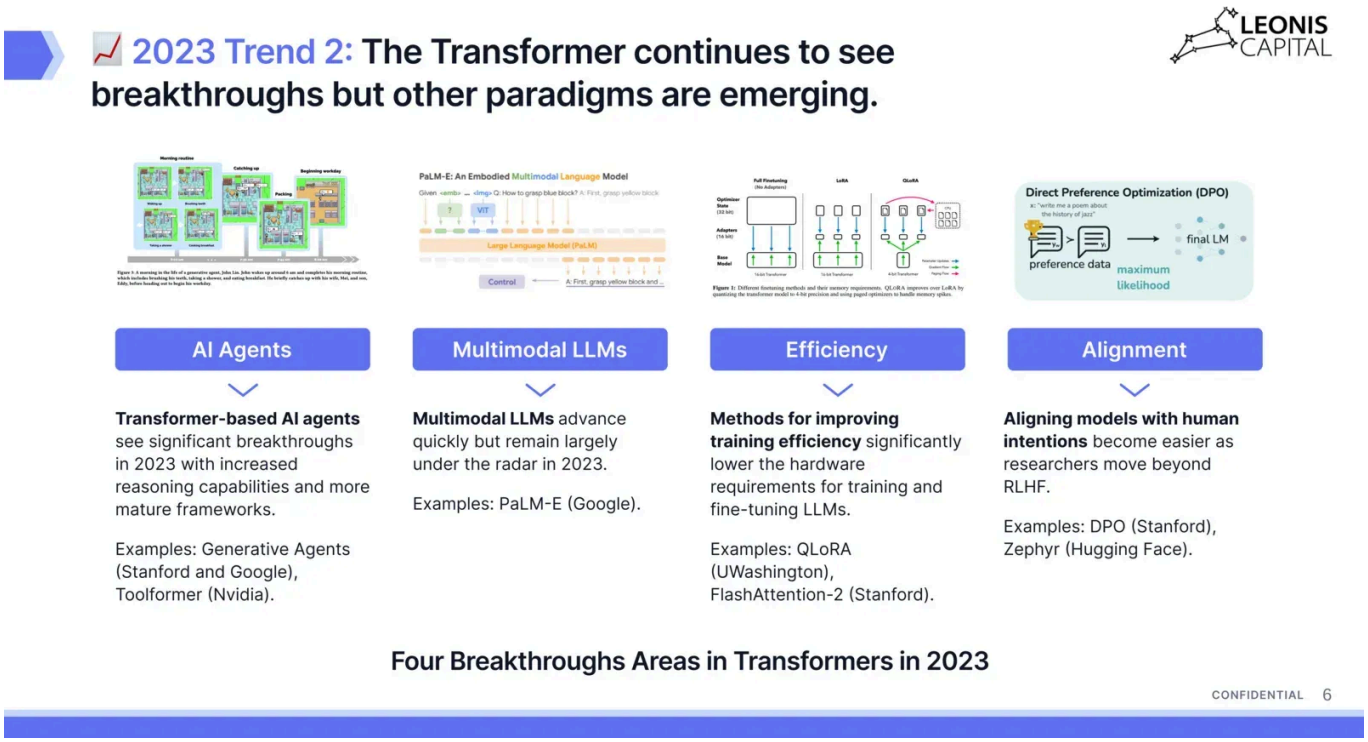
While compute providers initially reaped the most benefits from the AI boom, the value is now beginning to trickle down to the application and model provider layers. This shift is reminiscent of the early internet era, where the most value was initially created at the infrastructure level. Over time, as infrastructure became commoditized, the economic margins moved to software, which now enjoys high margins. Similarly, as AI infrastructure becomes more standardized, applications are poised to capture greater economic value. Although LLM applications are still in their nascent

stages and not yet highly profitable, they hold significant potential for the future, expected to realize substantial economic benefits over the next three to five years.



Prediction 1: A New AI Paradigm Will Emerge

The most exciting prediction for 2024 is the emergence of a new AI paradigm that will outperform transformers. Historically, AI paradigms have evolved rapidly, from CNNs and GANs in the early 2010s to reinforcement learning, transformers, and diffusion models in the late 2010s. We're now overdue for another shift. Potential candidates for the next big thing include Mamba, LNNs, or even a blend of smaller models. The limitations of transformer-based models, particularly regarding compute efficiency and their ability to achieve AGI, are driving this change. As the AI community explores these new paradigms, we expect significant breakthroughs in the coming year.

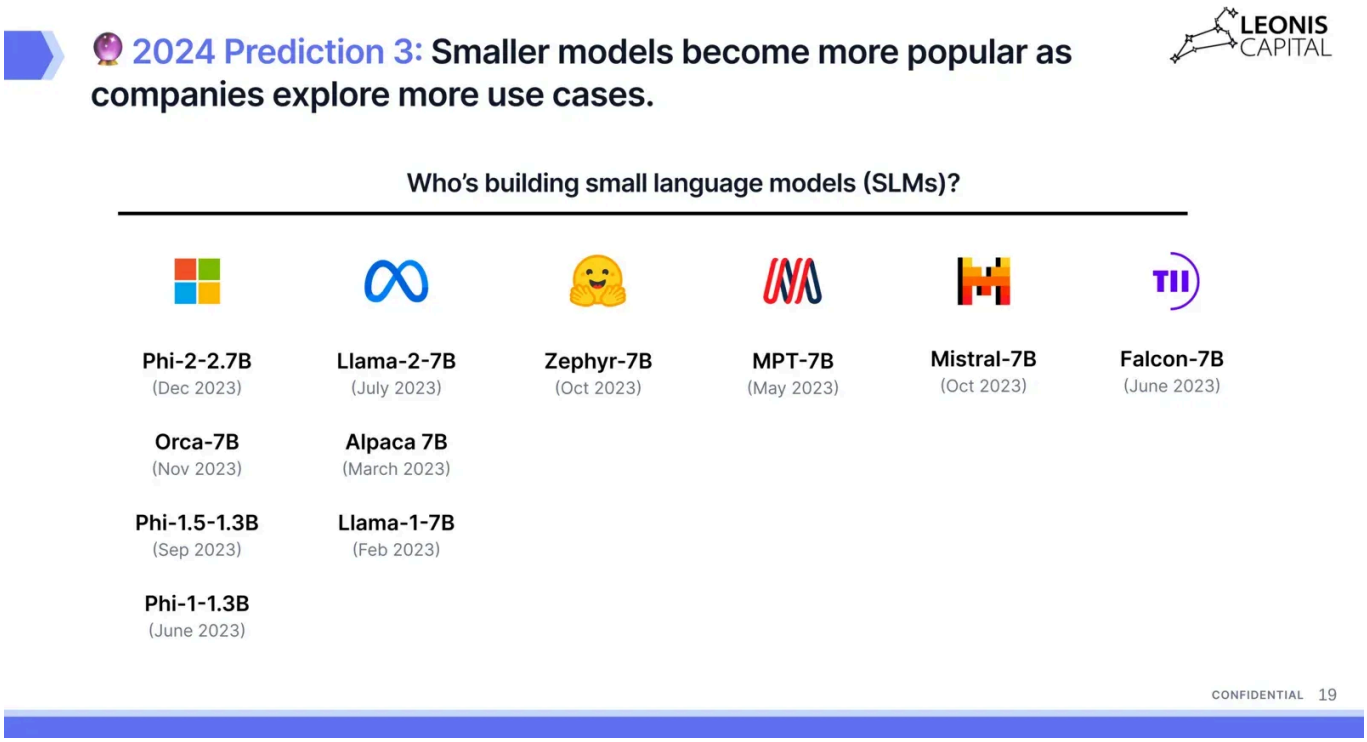


Prediction 2: The Rise of AI Agent-Based Software

AI agents represent a major evolution from traditional co-pilots and chatbots. Unlike their predecessors, AI agents can take autonomous actions in the real world, thanks to advancements in reducing hallucinations and improving reasoning capabilities. Major players like Meta are integrating AI agents into their products, while academic researchers are developing frameworks to enhance these models' efficiency. We predict that 2024 will be the year AI agents find real-world applications, significantly impacting various industries. If you're interested in AI agents, check out our previous research on AI agents and the future of software.

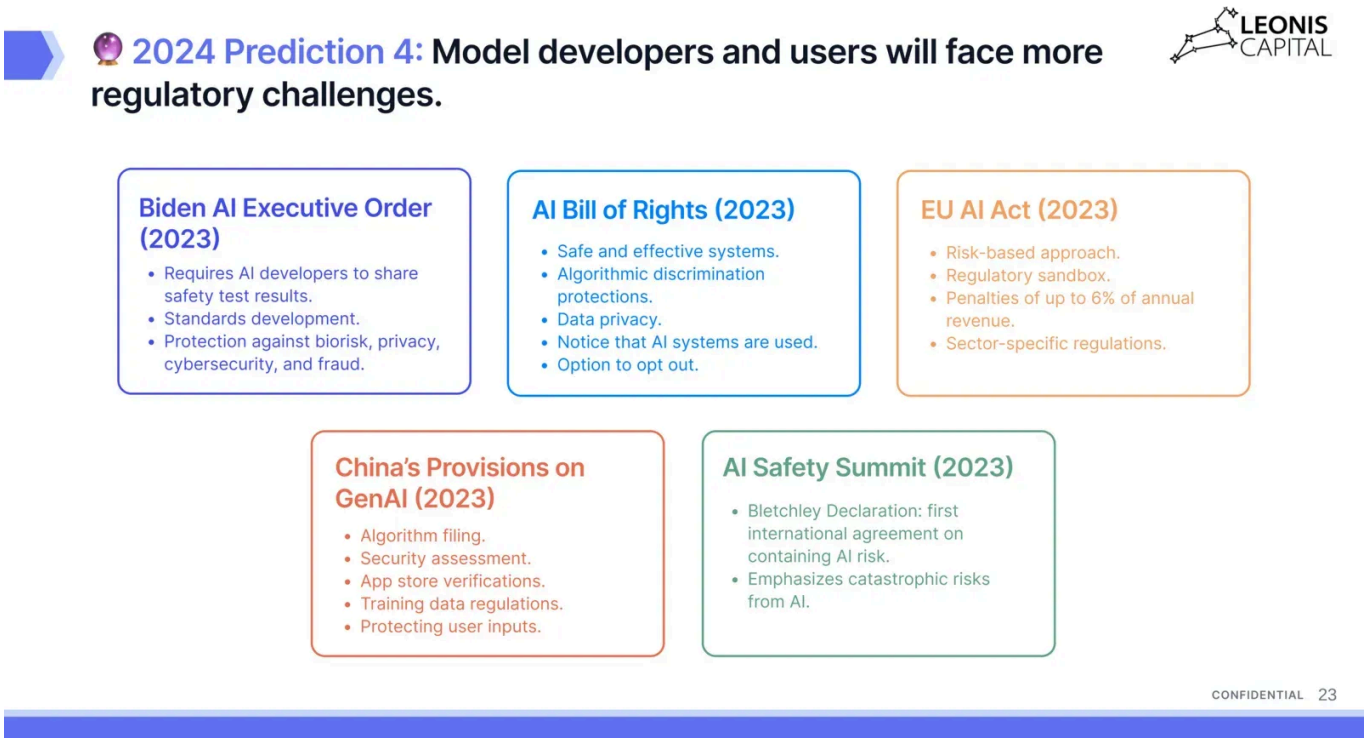
Prediction 3: Smaller Models Gaining Popularity

In 2024, smaller AI models (under 10 billion parameters) will become more prevalent, especially for edge devices. Companies like Meta and Microsoft are investing heavily in these compact models, which are faster to run and cheaper to train. Techniques like knowledge distillation and quantization are enabling this shift. Apple, in particular, is focused on making AI models suitable for edge deployment, which offers advantages like reduced latency, improved data privacy, and lower operational costs. These small models, while not as performant as larger ones, are ideal for real-time applications on personal devices.



Prediction 4: AI Regulations are Coming

As AI technology advances, so does the regulatory landscape. In 2024, we expect to see more stringent regulations affecting model developers and users. The Biden administration's executive order and the EU AI Act are leading examples. These regulations require developers to share safety test results and comply with strict guidelines, which could pose challenges for startups with limited resources. However, these measures are crucial for ensuring the safe and ethical deployment of AI technologies.

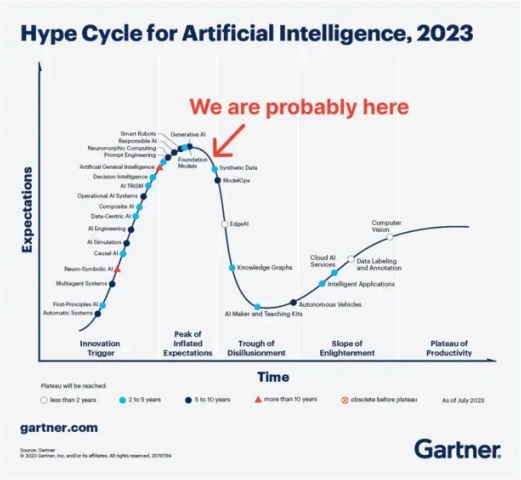


Prediction 5: The Decline of AI Hype

The AI hype that peaked in 2023 is expected to wane in 2024. While this may seem negative at first glance, it's actually a positive development for the industry. As the hype fades, AI valuations will return to more rational levels, allowing serious founders and investors to focus on building and investing in sustainable technologies. The reduction in hype-driven investments will also weed out companies that fail to deliver, ultimately strengthening the overall AI ecosystem.

🔮 2024 Prediction 5: The AI hype cools down but building continues.

- **Enterprises will find it harder** to build LLM products or integrate LLMs than they thought.
- Startups that are **thin wrappers will increasingly struggle**. "Tourist founders" will gradually leave AI.
- AI startups that **raised massive rounds but deliver little** will struggle.
- AI **valuations will go back to a rational level**.
- **Fewer VCs** will claim to be "AI experts" and raise AI funds.
- However, the best time to build and invest is when there is little noise and more action.



Prediction 6: Vertical AI Startups Become More Mature

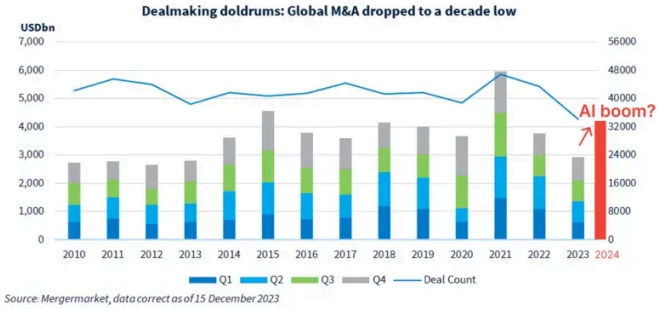
Vertical AI startups will come into their own in 2024, developing more mature applications. These startups have an advantage over big tech in terms of data access and agility, particularly in highly regulated domains like finance, healthcare, and legal tech. With a focus on specific industries, these companies can move faster and innovate in ways that large enterprises cannot. The ability to harness structured data for fine-tuning and training models will be a key factor in their success.

Prediction 7: Booming AI M&A Market

Finally, we predict a surge in AI-related mergers and acquisitions (M&A) in 2024. Enterprises struggling with AI adoption, big tech companies looking to augment their AI infrastructure, and the ongoing talent war will drive this trend. Companies like ServiceNow, NVIDIA, and Thomson Reuters have already made strategic acquisitions to enhance their AI capabilities. Additionally, some startups may seek M&A opportunities due to commercialization challenges or the need for additional resources. This M&A activity will further fuel innovation and growth in the AI sector.

🔮 2024 Prediction 7: The M&A market for AI booms as large companies look to acquire to innovate.

LEONIS CAPITAL



Source: Mergermarket, data correct as of 15 December 2023

AI M&A in 2023

OMNI ML → NVIDIA

mosaicML → databricks

splunk> → cisco

photomath → Google

casetext → THOMSON REUTERS

As we stand at the beginning of 2024, reflecting on the transformative year that was 2023 in AI, we can't help but anticipate the exciting developments yet to come. Our predictions for the year ahead are fueled by the momentum of progress and innovation. We'll return in 12 months to review our predictions, evaluating what we got right and learning from what we missed. It will serve as a reminder that we are still in the early days of AI, with boundless potential and discoveries awaiting us in the future.