PitchBook

PitchBook Data, Inc.

Nizar Tarhuni Vice President, Institutional Research and Editorial

Paul Condra Head of Emerging Technology Research

Institutional Research Group

Analysis

Brendan Burke
Senior Analyst, Emerging Technology
brendan.burke@pitchbook.com

Data

Matthew Nacionales
Senior Data Analyst

pbinstitutionalresearch@pitchbook.com

Publishing

Designed by **Jenna O'Malley**

Published on June 26, 2024

Contents

EMERGING TECH RESEARCH

High-Stakes Foundation Model Horse Race Out of the Gates

Handicapping secondary investment opportunities for startup AI research labs

PitchBook is a Morningstar company providing the most comprehensive, most accurate, and hard-to-find data for professionals doing business in the private markets.

Key takeaways

- We have tracked VC megadeals for 14 foundation model research labs since 2022. Financial engineering has aligned with investor demand to enable secondary trading in these companies' shares, creating liquid markets for OpenAI, Anthropic, Cohere, and xAI, along with opportunities for a broad spectrum of investors with other AI ecosystem leaders.

- Valuations of AI unicorns Anthropic, Databricks, and OpenAI trended up in Q2 2024 in the absence of primary rounds, when factoring in public comparables and secondary transactions.

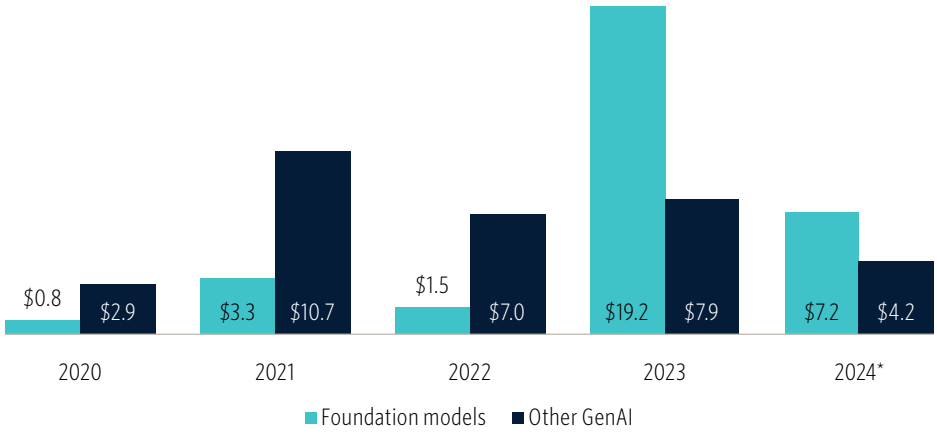
- Foundation model success will be dictated by reasoning capabilities, not the extent of training data nor performance on question-answering tasks. Research labs are developing new approaches to reasoning and tool usage that can yield further breakthroughs in model performance in successive iterations.

LLM race leaving the gates

The current wave of foundation model megadeals augurs a future of AI service providers similar to the internet service providers (ISPs) of the 1990s. In that decade, rapid content delivery innovation on top of the internet protocol suite yielded 4,500 ISPs by 1998 that threatened existing telecommunications vendors and forced M&A activity. The decade also was affected by seismic shifts from dial-up to broadband internet and partnerships with PC makers. Netscape emerged as a dominant leader in browsers while contributing basic innovations in web delivery architecture and merged with a larger service provider in AOL. One terminal result was a merger between AOL and Time Warner that created a company valued at a $638.4 billion in today's dollars. Yet even AOL was disrupted by the shift to broadband, thus limiting the success of its merger, and similar disruption in AI is likely over a 10-year horizon. Beyond this specific result for one of the fastest-scaling ISPs, a class of internet companies was created in a secondary wave around 10 years after the start of the democratization of the internet that rushed past the precedent telecom companies in value. While this period is associated with the dot-com crash, it did not lead to an "internet winter" in the way that some might assume AI will experience.

With history as a guide, startups are likely to define the tech utilities of the future. Cloud services created by post-dot-com-boom startups Amazon and Google surpassed incumbents with head starts including IBM and Oracle. Similarly, technical innovations needed to build companies around the latest AI models should create sustainable moats as the scope and scale of technology delivery expands. Examples might include new semiconductors or datacenter power-generation methods. We are beginning to see how technical leadership at startups can yield innovation in user interfaces and computing infrastructure. The rapid uptick in foundation model VC funding in 2023 has persisted thus far in 2024, with $7.2 billion raised by foundation model companies in 2024 YTD, on pace to exceed 2023's total outside of OpenAI's outlier $10 billion commitment.

Generative AI (GenAI) foundation model VC deal value ($B)



Source: PitchBook • Geography: Global • *As of April 30, 2024

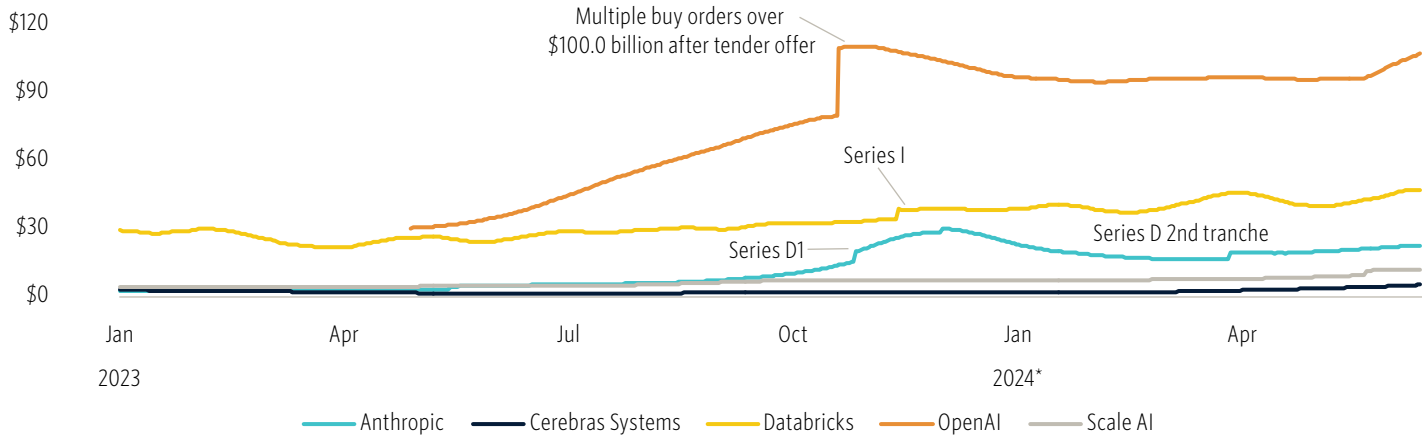
AI valuation estimates can vary with both public valuations and private funding rounds. Individual companies remain susceptible to market forces best measured in the aggregate. Our Morningstar PitchBook Global Artificial Intelligence Unicorn Index tracks how unicorn valuations in AI compare to other tech verticals. More information is publicly available here.

Secondary courses for foundation model horses

Recent megadeals for large language model (LLM) companies resemble the internet IPO wave of the early 1990s without listings on public exchanges, given the newfound ability of tech giants to support startups to a greater extent than financial markets can. We have tracked VC megadeals for 14 foundation model research labs since 2022. Financial engineering has aligned with investor demand to enable secondary trading in these companies' shares. In part due to some special situations like FTX's bankruptcy after investing in Anthropic, shares in promising startups have become readily available, thereby creating opportunities for cross-asset investors to back companies without direct access. New startup-focused index products and secondary exchanges make this theme investable for retail investors and registered investment advisors. A word of caution is that excessive democratization of speculative securities can precede a decline if sellers are motivated to exit, yet we are seeing commercial traction and technical differentiation that can allow some of these companies to deliver outsized returns.

Companies vary in the availability of their shares. According to secondary market data broker Caplight, valuations of Anthropic, Databricks, and OpenAI are trending up in Q2 2024 in the absence of primary rounds, when factoring in public comparables and secondary transactions.[1] These increases mirror those seen in the Morningstar PitchBook Global Artificial Intelligence Unicorn Index, demonstrating the broad-based clustering in secondary valuation movements. Some market watchers perceive trades based on individual news events, like the Apple partnership with OpenAI. Among this sample, according to Caplight, only Databricks and Scale AI have active markets with close bid-ask spreads.[2] Cohere, Hugging Face, and Scale AI have also experienced significant secondary activity at the lower end of the spectrum, though volume can vary over time.

Valuation estimates ($B) for AI unicorns



Source: Data reproduced with permission from Caplight • Geography: Global • *As of June 12, 2024

1: "AI (Artificial Intelligence) Sector," Caplight, June 21, 2024.
2: "All Companies," Caplight, June 21, 2024.

Emerging capabilities can make models complementary to each other. Initial LLMs have been compared on standard tasks that make progress in the field appear to stem along a single axis of general intelligence. In the real world, experts specialize in diverse fields of intelligence with unique thought patterns and references. As an example, complete knowledge of physics is insufficient to design an aesthetically appealing building that is up to construction codes. When asked to do this, GPT-4o said, "I can't generate a detailed and code-compliant blueprint. I can certainly help you conceptualize a design and provide a basic layout."[3] We are starting to see how developers can find unique capabilities embedded in diverse models in the wild, whether on coding or creative tasks. Additionally, model success is context-specific and defies generalized benchmarks when selecting a model for production. While the ranking of models on benchmarks may vary over time, the underlying research capabilities of startups and adaptation to user needs will determine their ultimate success. This note focuses on underlying technical innovations and commercial traction that can encourage successive innovations in the future.

Handicapping the foundation model horses

OpenAI

OpenAI

- **Last Caplight valuation estimate:** $107.0 billion
- **Last primary round valuation:** $86.0 billion
- **Most recent employee count:** 1,200
- **Most recent revenue disclosure:** $3.4 billion in annual recurring revenue (ARR)[4]
- **Most recent revenue growth disclosure:** 225%

Overview

OpenAI has the potential to be a trillion-dollar company based on the novelty of its technology, revenue pace, and commercial partnerships. The company has been valued based on the revenue it has achieved, capital needs for additional computing resources, and expectations for future growth, surpassing $100.0 billion in secondary transaction valuations. The recent investment led by Thrive Capital was based on a tender offer, demonstrating the willingness of OpenAI to reprice its equity without further diluting shareholders. Recent employee tender offers have extended to $800.0 million, according to The Information.[5] Such tender offers can produce secondary investing opportunities, yet secondary trading activity has been limited since the round. Investors must evaluate the likelihood of future disruptive model releases, the economics of partnerships with Microsoft and publishers, and the sustainability of record-breaking revenue growth.

Technical differentiation

OpenAI has stood out for attracting elite talent and allowing its deep bench of experienced AI researchers to take risks on scaling promising innovations across massive training runs. While OpenAI has numerous research projects and user experiences under development, the company's technical differentiation relates primarily to three model families in general intelligence, image intelligence, and video generation.

3: "Can you generate a blueprint up to code for a multifamily building" prompt, OpenAI, ChatGPT 4o, generated June 20, 2024.
4: "OpenAI's Annualized Revenue Doubles to $3.4 Billion Since Late 2023," The Information, Stephanie Palazzolo and Erin Woo, June 12, 2024.
5: "OpenAI CEO Says Company Could Become Benefit Corporation Akin to Rivals Anthropic, xAI," The Information, Aaron Holmes, Natasha Mascarenhas, and Julia Hornstein, June 14, 2024.

- **GPT-4o:** OpenAI's newest transformer model continues the success of GPT-4 with a mixture of models to facilitate native multimodal experiences across voice, image, and text. This focus likely owes to the consumer success of OpenAI's models and desire to improve user experience for partnerships like Apple Intelligence, which uses ChatGPT's API to answer complex real-world questions that cannot be found in the planned on-device model or in Apple's cloud. The model also likely reduces the latency and operational costs of ChatGPT queries, given accompanying price declines for consumers. The development of GPT-4o came in lieu of a launch of a newly pretrained model, which would be called GPT-5 and remains in development. The ability to innovate around existing training runs shows OpenAI's sophistication in mixture-of-experts architectures, which can create a whole greater than the sum of the parts in terms of generalization and instruction following. It also means that each incremental base model can create different child models to accelerate inference and perform on diverse tasks.

- **DALL·E 3:** The image-generation model was the original diffusion model building on transformer architecture to generate images from text prompts. The model lags Midjourney, Stability AI, and Playground AI in Artificial Analysis' text-to-image leaderboard.[6]

- **Sora:** The video-generation model leverages diffusion research DALL·E 3. While the model will benefit from the quality of OpenAI's computer vision research team, its public release lags startups like Luma AI, Pika, and Runway, due in part to a priority on safety by collaborating with creators to ensure that outputs do not duplicate existing video.

Progress of these models will be accelerated by OpenAI's leading access to compute via its partnership with Microsoft. Microsoft has built the world's leading AI supercomputer, which is expanding at a linear rate. GPT-4o was trained in a Microsoft datacenter with 14,440 NVIDIA H100 graphics processing units (GPUs), and the company claims to be adding 5x that amount every month thus far in 2024, putting Microsoft ahead of Meta on GPU additions this year.[7] An internal forecast suggests that this pace may still accelerate, reaching 1.8 million GPUs by year-end.[8] OpenAI's ability to use this computing cluster is an independent contributor to its value, apart from its own financial resources. Even so, OpenAI realizes it will need to vertically integrate into hardware, and the company is unlikely to become a tech giant while depending on others for infrastructure.

GPT-5 remains in progress, in part due to the need for additional compute resources to be brought online. This model is likely to build significant reasoning capabilities after the hiring of Noam Brown from Meta in 2023 to drive reasoning research. We covered the potential for planning techniques in Brown's CICERO model in our 2023 Industry and Technology Outlook, though the idea is taking time to play out. There may be valid reasons for deferring the training of the model, including compute access, end-user product development, and shifting computing needs away from

6: "Text to Image Arena Leaderboard," Artificial Analysis, June 19, 2024.
7: "What Runs GPT-4o? | Inside the Biggest AI Supercomputer in the Cloud With Mark Russinovich," YouTube, Microsoft Mechanics, May 21, 2024.
8: "Microsoft Has a Target to Amass 1.8 Million AI Chips by the End of the Year, Internal Document Shows," Business Insider, Ashley Stewart, April 17, 2024.

the superalignment team. Even if the model development cycle takes two years, breakthroughs remain possible, which could extend the company's technical lead and create models that can be scaled across organizations, instead of the per-user subscriptions under the current primary business model.

Commercial traction

OpenAI stands out as the most widely adopted enterprise model provider even as much of its business comes from individual subscriptions. In surveys of both public chief information officers (CIOs) and startup chief financial officers (CFOs), OpenAI stands out as the most adopted.[9] To some extent, this is reflected in its revenue total yet is more likely a leading indicator of future revenue growth as pilot projects are scaled into deployments. In a Barclays CIO survey, 60% of a sample of 100 CIOs use OpenAI in Azure Cloud.[10] This result is validated by data from startup CFOs on model spending, with 60% of a sample of 800 startups paying for OpenAI.[11]

These partnerships allow for training data as well as retrieval of fresh content. The company's best data source is the human feedback its models receive. While the company's data sources remain opaque overall and may often cross copyright boundaries, the company's models perform well in integrity testing from HydroX AI, which measures models' tendencies to share private information, and these partnerships suggest that training data moats can be generated.

OpenAI has reached $3.4 billion in revenue yet reveals that direct-to-consumer and direct-to-developer applications present the easiest paths to scale. Only $200.0 million of the company's revenue comes via Microsoft Azure distribution, according to The Information,[12] thereby creating a strong incentive to create developer- and consumer-focused products along with better models. A large portion of the company's revenue stems from its API, while its ChatGPT user base has escalated into the millions, though these figures can be volatile.

Considerations

- The company's lack of governance controls may create instability among talent, which is the company's biggest risk. Recently, OpenAI has experienced high turnover given the disbanding of the superalignment team and recruitment from early-stage startups. The company is considering shifting from a nonprofit to a for-profit entity, which will enable employees to maximize the value of their equity. That equity is subject to restrictions that may encourage employees to seek positions elsewhere. The board continues to be a focus area for the company and may install some of the controls needed to alleviate employee concerns.

9: "Barclays CIO Survey Program," Barclays, April 2024.
10: Ibid.
11: "Kruze AI Spending Data," Kruze Consulting, April 30, 2024.
12: "OpenAI's Annualized Revenue Doubles to $3.4 Billion Since Late 2023," The Information, Stephanie Palazzolo and Erin Woo, June 12, 2024.

- OpenAI's safety and security efforts lag its peers. In independent testing from information security (infosec) startup HydroX AI, OpenAI's recent GPT-4o model received an 80/100 score for safety and a 92.7/100 score for security, lagging Anthropic and Meta.[13] The disbanding of the superalignment team intended to develop computational methods to protect future models from causing harm demonstrates that a near-term product focus will distract from the original mission to create long-term safety.

- OpenAI's web scrapers have been targeted by website hosts seeking to block them. Over 10% of the top 100,000 websites block OpenAI bots, according to Revealera data, the highest among all AI research labs.[14] This limitation compounds the lack of training data available to the company, thus encouraging costly synthetic data generation.

Anthropic

ANTHROP\C

- **Last Caplight valuation estimate:** $22.5 billion
- **Last priced valuation:** $18.4 billion
- **Most recent employee count:** 375
- **Most recent revenue disclosure:** $100.0 million (as of Q3 2023)[15]
- **Most recent revenue growth disclosure:** 900%

Overview

Anthropic was founded in 2021 by a group of seven former senior members of OpenAI. Several of the founders worked at Google Research before OpenAI. The founders split off from OpenAI with a desire to train safe AI systems that could mitigate the worst risks of superintelligence, including human extinction. To serve this mission, the company was established as a public-benefit corporation. Since its founding, Anthropic has developed proprietary AI systems and has commercialized the Claude family of LLMs.

The company has produced a high volume of secondary transactions, making its valuation a moving target. The liquidation of FTX's stake made shares widely available in secondary markets. FTX was able to liquidate around two-thirds of its $500.0 million stake in Anthropic for $884.0 million in a series of transactions, achieving around 165.5% capital gains.[16] Bids for the company shares have valued the company as high as $50.0 billion, and shares have recently been offered for as low as $22.4 billion, according to Caplight.[17] Mutual fund Ark Invest has valued the company as high as $25.7 billion in public filings, also listed on Caplight.[18] All told, the company's valuation has increased an estimated 22.9% since its last primary round, demonstrating continued momentum as the company is encouraged to remain private.

Technical differentiation

Claude: Claude leverages transformer architecture and a reinforcement learning process referred to as Constitutional AI, which tests the alignment of model outputs against a list of ethical rules and principles. In this way, the model self-improves by selecting outputs that conform to the company's ethics. The Constitutional AI

13: "Enable AI Safety & Build Safe AI: LLM Safety Evaluation Leaderboard V0.2," HydroX AI, June 19, 2024.
14: "What's the Bear Case Against ChatGPT/OpenAI?" X, Hanley Weng (@AznWeng), April 9, 2024.
15: "Anthropic Forecasts More Than $850 Mln in Annualized Revenue Rate by 2024-End – Report," Reuters, Harshita Mary Varghese, December 26, 2023.
16: "FTX to Sell Majority of Stake in Anthropic for $884 Million," Bloomberg, Emily Nicolle, March 25, 2024.
17: "Anthropic," Caplight, June 21, 2024.
18: Ibid.

process improves the safety of responses without the extensive human intervention seen with other approaches, as in OpenAI's use of human feedback for outsourced reinforcement learning. Given the background of Anthropic's team at OpenAI, there is little evidence that training techniques vary greatly, though internal improvements on explainability can improve the selection of features to emphasize in training. The company's research into feature steering shows how individual calculations in the model can be prioritized to affect the model's outputs, along with simple user prompts, showing more clearly than ever how LLMs can be manipulated to produce reliable outcomes. Basic innovations like this can improve model performance in the future, and Anthropic has shown it can lead the field with the latest Claude 3.5 Sonnet model.

This approach yields more resilient models than others in the market in practice. While our view has long been that Anthropic takes a more procedural and organizational approach to security integration than other startups, AI security startup HydroX AI's testing shows that Anthropic outperforms other foundation models out of the box across safety, privacy, security, and integrity, reaching near-perfect scores in privacy and security.[19] Our review of HydroX AI's methodology shows thousands of dynamically generated tests for each model across over 30 categories including toxicity, hallucinations, vulnerability to model-specific cyberattacks, and copyright protection. Anthropic's Chief Information Security Officer (CISO), Jason Clinton, was a prominent figure at cybersecurity's RSA Conference in Q2. He gave cybersecurity teams a view into the second-order effects of AI deployment, including self-replication and autonomous hacking, which security teams can defend against today. Given enterprises' focus on privacy and security, we expect these features—and the promise of attack-proof AI—to be a differentiator.

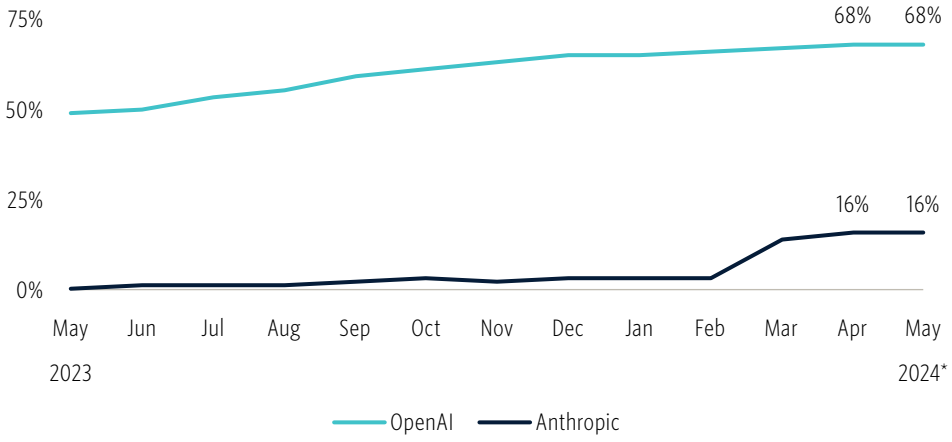
Commercial traction

Anthropic is positioned as a complement to OpenAI and a chosen partner for Amazon Web Services. In adoption data from accounting firm Kruze Consulting based on 125 revenue-generating startups in its client portfolio, multimodel adoption is perfectly correlated with Anthropic usage.[20] Usage of two paid LLMs is common, while three is rare. In part explaining this, Anthropic is a primary beneficiary of existing AWS customers leveraging Bedrock, because AWS does not support OpenAI, which competes via Microsoft. Additionally, significant AWS consulting relationships from Accenture, Deloitte, and Slalom will encourage the use of Anthropic models. Anthropic, AWS, and Accenture now have an official collaboration to deploy Claude across industries. Accenture is the leading AI services firm by market share and disclosed $2.0 billion in GenAI bookings in the first three quarters of its fiscal year.[21] Given that consultants present the first port of call for enterprise deployments, Anthropic can benefit from close relationships via AWS.

19: "Enable AI Safety & Build Safe AI: LLM Safety Evaluation Leaderboard V0.2," HydroX AI, June 19, 2024.
20: "Kruze AI Spending Data," Kruze Consulting, April 30, 2024.
21: "Accenture Reports Third-Quarter Fiscal 2024 Results," Accenture, June 20, 2024.

Anthropic can stand out in enterprise deployments because of its integration with the leading public cloud, AWS, and its focus on baked-in safety without the need for volatile reinforcement learning from human feedback. The company disclosed revenue forecasts of $850.0 million in 2024. Our review of end-user spending data suggests that this hypergrowth is not well supported by market adoption of foundation models. Kruze Consulting data shows that payment for Anthropic has increased from a base of nearly zero, but only to 16%. This lack of adoption is backed up by CIO surveys at the high end of the market. Even if the company grows only 200% this year instead of a more outlandish total, it will be on a pathway to upsell research breakthroughs via existing customers.

Startup customers paying foundation model providers as a share of 125 SaaS and AI startups with revenues over $1 million



Source: Data reproduced with permission from Kruze Consulting • Geography: Global • *As of May 1, 2024
Note: Sample includes 125 SaaS and AI startups with revenues of over $1 million

Considerations

- Anthropic's Claude Pro subscriptions are not scaling to the extent of ChatGPT, even as API access remains popular for developers. Given the public statistics about Anthropic adoption and the neutrality of AWS in promoting different models, we find it unlikely that the company has multiplied the $100.0 million in ARR reported at year-end 2023, suggesting a maturing growth trajectory.

- Anthropic's pricing for its largest model falls well above OpenAI's. Compared with GPT-4o, Claude Opus costs 3x per input token and 5x per output token. Anthropic offers smaller and cheaper models that perform well, yet the higher pricing of its frontier model demonstrates a lack of flexibility in its commercial relationship and a need to outperform on model quality for high-value tasks. The company has focused on cost reduction with its next-generation Claude 3.5 Sonnet model.

databricks

- **Last Caplight valuation estimate:** $47.1 billion
- **Last priced valuation:** $43.2 billion
- **Most recent employee count:** 8,799
- **Most recent revenue disclosure:** $2.4 billion (as of June 2024)[22]
- **Most recent revenue growth disclosure:** 60%

Databricks

Overview

Founded in 2013, Databricks offers a data science platform that includes AI-as-a-service functionality with a suite of data science tools for data engineering, data warehousing, and machine learning (ML) algorithms. The company grew out of the open-source Apache Spark data science community and created an extensible product, the Unified Data Analytics Platform, that can ingest data from enterprise silos and prepare it for cluster-based computing. Once an effective open-source product was in place, the company moved to a closed-source model and rapidly increased revenue beginning in 2016. Databricks now offers a product suite on top of Apache Spark that includes an AI platform called Mosaic AI, which is optimized for GenAI with data storage, a model-training pipeline, and model-serving and monitoring; a data lake, which integrates unstructured data in a central database for analytics; and data security. The company has focused on the GenAI opportunity, training a custom LLM called DBRX that can run more efficiently than GPT-3.5 and comparable open-source models.

The company has deferred an IPO via readily available late-stage rounds and secondary investments. In addition to the Series I, at least 31 secondary transactions have closed over the past year, demonstrating robust demand and some supply of shares in the high-growth company. The company's founder-led and open-source values present some resistance to going public, which has played a role in some other open-source companies closing off their products, along with competitive pressures.

Technical differentiation

Databricks stands out in the AI & ML market by supporting customized AI model training using open-source models atop its data lake—functionality that distinguishes it from legacy vendors such as SAP, Microsoft, and AWS and allows it to be more AI oriented than next-generation data warehouse vendor Snowflake. The company used its acquisition of Mosaic ML to enable custom small-model training—a bold stance in contrast to closed-source model APIs from OpenAI and Anthropic. This approach builds on research into compound AI systems, which include adaptations to models including search and retrieval. By integrating high-performance vector search, the company can build performant AI applications, whereas foundation model labs tend to collaborate with consultants and open-source tooling partners to complete the last mile.

The data stack for AI applications is in flux, and Databricks has a claim to centralize the pipelines to LLMs in its platform. The company's Unity Catalog product stands out for data governance—a critical problem for data scientists new to AI. The company recently trained a custom LLM called DBRX. The model's training differs in the mixture-of-experts architecture used from other open-source models, though the model does not necessarily outperform Mistral AI in independent testing. While the model was only the "best model in the world for two whole weeks," according

22: "Databricks Tells Investors Annualized Revenue Will Reach $2.4 Billion at Midway Point of Year," CNBC, Jordan Novet, June 12, 2024.

to CEO Ali Ghodsi,[23] it offers a base model of open weights with fast inference for customer fine-tuning, and also helps Databricks enhance native code generation and knowledge retrieval with custom-built AI, which competitor Snowflake acquired startup Neeva in part to address. The research talent at the company should allow continued model innovation, yet will likely lag more specialized competitors.

Commercial traction

Databricks' AI solutions have become widespread. The company counts 10,000 corporations as users of its Data Intelligence platform, including 300 of the Fortune 500. This is broad definition that includes the data warehouse product and does mean that all of these customers are actively fine-tuning LLMs. The company disclosed that 76% of Data Intelligence users use LLMs, with 210% growth in the number of companies registering at least one model on the platform.[24] This high growth suggests there was a low base of customers building models natively within Databricks, yet a measurable increase exists in GenAI spending. The company's leading use cases are financial services and retail, which are also the industries that rank highest in GenAI spending in IDC data. Databricks disclosed continued 140% net retention, indicating that new AI solutions are effective in fueling spending growth.[25]

Considerations

- Databricks is not the primary choice for developing retrieval-augmented generation solutions, given the additional steps needed to load unstructured data into its data warehouse. The company is treated like a repository to be combined with structured databases for multidatabase retrieval using agnostic data loaders like LlamaIndex or Unstructured.

- Databricks faces competitive risks from the newly composable data stack. Data users may opt to build AI solutions on top of open data formats like Iceberg, which spurred the company's acquisition of Tabular, or PostgreSQL. Databricks can offer support for other open-source libraries yet remains an expensive platform.

Mistral AI

MISTRAL AI_

- **Last Caplight valuation estimate:** N/A
- **Last primary valuation:** $6.3 billion
- **Most recent employee count:** 55
- **Most recent revenue disclosure:** N/A
- **Most recent revenue growth disclosure:** N/A

Overview

Mistral AI presents an independent open-source foundation model startup that can compete with both Meta and OpenAI while responding directly to developer needs and developing its own computing resources. Mistral drew much of its talent from Meta, demonstrating a "brain drain" from the social media giant that enabled Mistral to make up ground in the LLM race. Since founding in 2023, the company has proven capable of training competitive foundation models that can be fine-tuned by enterprise customers.

23: "Data + AI Summit Keynote Day 1 - Ali Ghodsi, Co-Founder and CEO of Databricks," YouTube, Databricks, June 12, 2024.
24: "State of Data + AI: Data Intelligence and the Race to Customize LLMs," Databricks, n.d., accessed June 15, 2024.
25: "Databricks Tells Investors Annualized Revenue Will Reach $2.4 Billion at Midway Point of Year," CNBC, Jordan Novet, June 12, 2024.

Mistral receives financial support from open-source supporters and European sources. The company's last round, a Series B in Q2 2024, featured investment from IBM, which supports the open-source community partly through its Red Hat business, and French investors Bpifrance, BNP Paribas, and Korelya Capital, along with other European investors. The strategic importance of the company indicates continued support and possibly a reduced need for secondary investment. The company faces some political risk as a result, because a change in French government could alter the level of financial support.

Technical differentiation

Mixtral 8x22B: The name of the largest and most performant Mistral model owes to its mixture-of-experts architecture that uses eight expert models and can select from two for each inference, keeping computing costs low and resembling the architecture of fellow open-source model Grok-1. Mixtral has a larger context window than its primary open-source competitor, Llama 3, enabling retrieval applications. The model underperforms Llama in question answering but outperforms on the Abstraction and Reasoning Corpus (ARC) Challenge benchmark, which should ultimately be a more practical direction for future development.

Mixtral 8x7B: This medium-size model shrinks the size of expert models used by the overall LLM. In a surprising development, many developers find better results from Mistral's smaller models, especially when considering cost. We have seen this result in diverse use cases like cybersecurity threat detection and content generation. This result mirrors the heightened interest in Anthropic's medium and small models as well as Microsoft's Phi small model.

Commercial traction

Mistral is forming partnerships with open-source databases and model-serving frameworks to offer a low-cost alternative to OpenAI. By receiving investment from Cisco, Databricks, and IBM, the company gains access to model marketplaces and consulting firms that can encourage adoption. The company has also partnered with AI deployment leader OctoAI, encouraging OctoAI customers to experiment with Mistral's models. Mistral has become one of the most popular models on the platform, even though it also supports Llama 3.

Considerations

- Mixtral can be easily jailbroken out of the box. Model jailbreaks refer to prompting models to defy their instructions, which can lead to poisoned models being forked into black markets for usage by malicious actors. Developers can easily determine this in practice, and the company's models rank low in security testing by independent security testing company HydroX AI.[26] This may owe to a lack of costly reinforcement learning that can detect adversarial activities. The company may need to work with the security community to provide greater assurances to open-source users.

26: "Enable AI Safety & Build Safe AI: LLM Safety Evaluation Leaderboard V0.2," HydroX AI, June 19, 2024.

cohere

- **Last Caplight valuation estimate:** $3.7 billion
- **Last primary round valuation:** $3.0 billion
- **Most recent employee count:** 300
- **Most recent annualized recurring revenue disclosure:** $35 million
- **Most recent revenue growth disclosure:** 169.2%

Cohere

Overview

Cohere unites leading AI research talent with history at the University of Toronto, led by former students of AI pioneer Geoffrey Hinton. The founders have since done cutting-edge AI research and founded Cohere to train LLMs that can complete tasks for business professionals. The company's first funding round in 2020 featured AI luminaries that recognized the founders' abilities to move the field forward. The company has since aligned with the foundation model trend, partnering with Google Cloud in 2021 for training and commercialization, and began commercializing proprietary models via a platform tailored for business users in 2022. This product development can help gain traction with business users, yet basic innovation will drive the company's value going forward as buyers look for the most performant base models.

The company has been active on the secondary market from a relatively early stage. PitchBook tracked the sale of shares from Artisanal Ventures and Yenom Capital in August 2023, and Caplight tracked six other transactions. The company's valuation continued to rise after its last round, leading to a rumored recent round at a $5.0 billion valuation. The company receives strategic interest from Canadian investors looking to support local AI development, including the Canada Pension Plan Investment Board. Data provider ApeVue has seen secondary transactions valuing the company as high as $6.4 billion.[27] These investors and promising early revenue growth indicate that secondary deals will remain active for the company.

Technical differentiation

Command R+: In May 2024, Cohere released this dedicated model for retrieval-augmented generation tasks, which leads the market in latency and has become popular on Hugging Face. The model was fine-tuned in pre- and post-training with pairs of instructions and API-based tool usage. These pairs enable users to use a prompt template to allow the model to use tools like document retrievers and databases. The model also has a mode for grounded generation, which automatically pulls in documents to answer questions, rather than responding from pre-training data. OpenAI has a competitive product for this with function calling via its assistant API, yet the product remains unreliable in practice, and developers use open-source orchestration frameworks to connect GPT-4 to individual tools. OpenAI has put more resources into enhanced ChatGPT experiences, leaving a retrieval-oriented model to competitors like Cohere and Databricks.

Cohere's vision to transform AI from a passive tool into an active assistant sets it apart from competitors. By developing models capable of performing tasks like scheduling meetings or filing expense reports, the company addresses the growing demand for task-oriented AI solutions in the enterprise market. Demonstrating Command R's understanding and reasoning capabilities, the model outperforms fellow open-source model Mistral Small on the unpredictable LiveBench assessment and the larger Mixtral-

27: "Pre-IPO Stock Valuations | as of June 17, 2024," X, Aaron Dillon (@AaronGDillon), June 17, 2024.

8x22B in Hugging Face's Open LLM leaderboard.[28] Adjusting training samples can have unexpected effects on model performance, and Cohere can outperform as a result.

Commercial traction

Cohere had a blowout Q1, reportedly escalating its ARR from $13 million in 2023 to $35 million as of Q1 2024.[29] Cohere has high-end partnerships that support an escalating revenue profile. Strategic partnerships include Oracle, McKinsey, and Accenture, along with distribution via the three leading hyperscalers. Cohere offers Oracle a chance to maintain competitiveness in the foundation model ecosystem, along with xAI via model training services. The cloud provider is promoting Cohere via its Fusion Cloud Applications Suite, giving Cohere a similar position to what Anthropic has with AWS. McKinsey and Accenture lead consulting firms in GenAI deployment, though they partner across the ecosystem.

Considerations

- Cohere is not pursuing artificial general intelligence (AGI) and is generally pessimistic on the ability of LLMs to achieve AGI. While OpenAI and Anthropic have claims to develop human-level intelligence and are marshaling the computing resources they believe will do so, Cohere may not demand the funding total to keep up with their reasoning capabilities. While this constraint may result in a more profitable business, the company may not remain competitive with advanced models from successive AI launches.

- Cohere's Command R+ model is priced higher than its open-source competitors, at $3 per million input tokens and $15 for output. The models will have to compete on features and the creativity of developers in combining open-source components for retrieval augmented generation (RAG) applications.

xAI

- **Last secondary valuation:** N/A
- **Last primary valuation:** $24.0 billion
- **Most recent employee count:** 97
- **Most recent revenue disclosure:** N/A
- **Most recent revenue growth disclosure:** N/A

Overview

xAI is Elon Musk's entry into the foundation model horse race, founded in July 2023 and recently receiving $6.0 billion in outside investment. Musk has already proved adept at other high-stakes races and intends to develop best-in-class models via advanced hardware, exclusive data, and elite research talent. A fundamental motivation is the guardrails that have been placed on closed-source models that limit the reliability of the models' outcomes.

The company's Series B presented an opportunity for financial VCs to compete with tech giants, because xAI is unlikely to align with the existing hyperscalers that it intends to compete with. Given the 24 investors that participated in the round, we expect secondary shares to become available over time as they are for SpaceX. Early investors believe that an additional $6.0 billion will need to be raised in the coming year to keep pace with peers.[30]

28: "LiveBench: A Challenging, Contamination-Free LLM Benchmark," LiveBench, Abacus.AI, et al., Colin White, et al., June 20, 2024.
29: "Nvidia and Salesforce Double Down on AI Startup Cohere in $450 Million Round, Source Says," Reuters, Krystal Hu, June 4, 2024.
30: "Investing in xAI," Deepwater Asset Management, May 28, 2024.

Technical differentiation

xAI will compete based on the volume of exclusive training data and compute, more so than algorithm architecture. The company is building a supercomputer in partnership with NVIDIA server manufacturers Dell and Supermicro. The company's initial models Grok-1 and Grok-1.5 did not include social platform X posts in their training data, thus limiting the impact of their supposedly exclusive datasets. The company uses a mixture-of-experts architecture first developed at Google, owing to the at least 24 researchers the company has drawn from Google and OpenAI backgrounds. This approach is also used in other foundation model labs. The model can be differentiated in terms of experience, with the "fun mode" on X as an example of chat-tuned models that can be more engaging. The company intends to research AI reliability, envisioning scientific research applications in areas like aerospace and materials science to be built on top. This will require algorithmic breakthroughs that we have not seen evidence of yet and pose challenges to more experienced labs. These goals may encourage leading researchers from science-focused AI labs like DeepMind to take part.

The company is among the most aggressive in building proprietary AI datacenters. Legacy datacenter hosts face constraints in sourcing power, as we covered in our Q1 AI & ML Report. The company plans to leverage the latest semiconductors announced by NVIDIA along with custom Tesla silicon, giving a differentiated stack relative to what Meta and Microsoft are investing billions of dollars in. Given the accelerated pace of innovation at NVIDIA, future datacenters may enable far cheaper inference with more reliable networking. Construction of xAI's supercomputer is underway in Tennessee in partnership with Oracle, Dell, and Supermicro and will likely not yield cutting-edge models for several years. While any datacenter will include depreciating assets, incremental hardware innovations will encourage developers to gain access to leading compute clusters and will undergird the creation of category-defining companies.

Commercial traction

Grok has a built-in user base from X, which enables new features to receive feedback. Investors believe that the entire user base of 250 million daily active users will contribute to training data and will engage with the model.[31] A low percentage of X users have converted to premium usage, which enables conversations with the Grok chatbot. This revenue is important for X, so we do not expect a material contribution to xAI's business case, encouraging development of new business models.

Considerations

- **Training data limitations:** X posts are considered to be a leading source of training data yet may lack the reasoning steps and customer permissions to be leveraged for foundation model breakthroughs.

- **Capital scale:** Meta and Microsoft are likely investing upward of $30 billion in hardware this year, making competing datacenters highly capital-intensive. Musk plans to create something like the world's largest supercomputer, which will be a moving target.

31: "Investing in xAI," Deepwater Asset Management, May 28, 2024.

For more examples of research labs and their supporting tooling ecosystem, as well as those domain-specific foundation model developers in individual industries, PitchBook users can access:

- GenAI emerging space market map
- LLM operations market map
- LLM research lab list

Other contenders

We track 47 LLM research labs that work on both frontier models and secondary research topics that may not be well addressed by foundation models. Some also explore the constraints of open-source, including models with open weights and transparent datasets. Also, tech giant research labs continue to leverage their internal research teams for marginal breakthroughs, even as some of their leading researchers leave for startups. While each of these deserves consideration as future utilities in AI adoption, they lack current secondary investment opportunities and face significant competition for investment from strategic investors and public markets.

Foundation model startups

01.AI: The LLM research lab of AI pioneer Kai-Fu Lee has developed a leading open-source LLM that concentrates on multilingual operation between English and Chinese. Alibaba and Xiaomi invested in the company, suggesting strategic interest for the Chinese internet industry and a complementary role to the cutting-edge research being done within Alibaba.

AI21 Labs: AI21 Labs addresses the limitations of LLMs via development of the MRKL system, a compound system that incorporates deterministic reasoning along with predictive generation, resulting in a base foundation model Jurassic-X. This model has been fine-tuned for domain-specific applications including financial document chat, customer services, healthcare chat, academic research summarization, product documentation chat, and insurance documentation summarization. The company continues to push beyond LLMs, recently releasing an advanced structured state space model based on academic research to enhance computational efficiency.

Adept AI: Adept has innovated in user interface understanding using multimodal transformers trained on examples of web browsing. The company's newest model, Fuyu-Heavy, opens the company's research into browser manipulation.

DeepSeek: The research lab was founded in 2023 but lacks the financing history to have a PitchBook profile yet. The company's open-source models include popular coding models. Its latest DeepSeek-Coder-V2 supports 338 programming languages and became the open-source leader for Aider's code-editing leaderboard while succeeding in other tests against GPT-4o.[32]

Nous Research: The startup research lab's Hermes 2 Pro model excels at coding, including function calling based on enriched training data. This makes the model adept at information retrieval via API and generating JavaScript Object Notation (JSON) outputs that can carry out action steps.

32: "Aider LLM Leaderboards," Aider, n.d., accessed June 20, 2024.

Reka AI: The startup co-founded by former Google and Meta researchers launched a natively multimodal model called Core across image, video, and audio before OpenAI did, which led the company to receive acquisition interest from its partner Snowflake. The model excels at understanding art and design content, leading all multimodal models in the Multidiscipline Multimodal Understanding (MMMU) benchmark.[33]

Runway: The video-generation startup recently leapt forward with Gen-3 Alpha, a model that uses video understanding to create physically realistic videos from user prompts. This is a goal of OpenAI's, yet the Runway team has consistently shipped ahead of time while working closely with creatives on content safety and quality requirements.

Safe Superintelligence: This startup was recently founded by former OpenAI co-founder Ilya Sutskever, along with leading AI investor Daniel Gross and OpenAI researcher Daniel Levy. Sutskever clashed with Sam Altman over AI safety efforts and definitions of general intelligence, likely inspiring a company focused primarily on safety. With an office in Tel Aviv, Israel, the company may be positioned to receive strategic funding for further domestic development.

Together AI: Mixture of Agents layers multiple LLMs to form a compound system, similar to mixture of experts yet leveraging diverse open-source models that have been trained with different techniques. The system aims for high performance on the newer AlpacaEval benchmark instead of conventional question answering.

Tech giant research labs

Alibaba Cloud: Qwen stands out as a leading open-source model based on quality-assurance steps taken during the pretraining phase of the model. The company evaluates the quality of its training data for key use cases like Python generation to make sure the model will not be biased.

Google Gemini: Gemini is popular given the quality of the training data, research behind the foundation model, and application tools built on top. The model ranks third in adoption behind OpenAI and Anthropic in Barclays' CIO survey.[34]

Meta: Meta competes with Microsoft for the largest GPU cluster in the market, resulting in groundbreaking open-source models. Llama 3 70B is the most-adopted open-source LLM in both the Databricks environment and in CIO surveys.[35]

33: "MMMU: A Massive Multi-Discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI," MMMU, IN.AI Research, et al., Xiang Yue, et al., n.d., accessed June 21, 2024.
34: "Barclays CIO Survey Program," Barclays, April 2024.
35: "State of Data + AI: Data Intelligence and the Race to Customize LLMs," Databricks, n.d., accessed June 15, 2024.

For primers on AI agents and coding automation innovations that are continuing to play out, see our Q3 2023 AI & ML Report.

Outlook

Foundation model success will be dictated by reasoning capabilities, not the extent of training data nor performance on question-answering tasks. Reasoning techniques can adapt AI models to new real-world situations and respond to circumstances as an intelligent human would, including action steps and communication. These approaches will require advanced model breakthroughs and do not necessarily present themselves in the existing capabilities of LLMs. New evaluation datasets including AgentBench, ARC-AGI, AlpacaEval, and LiveBench can test the abilities of LLMs to act similarly to humans, instead of purely recalling information. While recall tasks are approaching human-level performance, these benchmarks show that LLMs have higher goals to reach and are still far from them. We are still in the dial-up phase of AI, including manual prompting, latency reduction, and vibes-based evaluation. Broadband made internet more accessible for more high-value tasks, like advanced reasoning may for AI applications.

The limitations of LLMs present both constraints and opportunities for startups. For LLMs, hallucinations are a feature, not a bug, which means that production systems must use creative engineering with deterministic guardrails. Additional research directions include agent frameworks and coding models, which can supplement LLMs with improved multimodel orchestration and logic engineering. Individual research labs will make progress on these tasks and must encourage progress beyond current techniques to keep researchers engaged and move the field forward. Tension within teams can lead to researchers splitting off to work independently, which can ultimately lead to new startups taking leading positions on new research tasks.

Research labs will benefit from remaining independent and rewarding employees via well-timed tender offers. The financial competition for AI researchers is intense, and equity packages will present an important incentive, encouraged by secondary markets to a greater extent than other tech companies, which may be overstaffed. Emerging nation-state interests will discourage widespread consolidation while working on tasks that are culturally and linguistically specific. OpenAI's leadership in this area will encourage challengers to follow suit and create tender offer programs for their researchers. There is a flurry of reports of acquisition interest for some startup research labs, and we expect that mergers will happen in the long run, yet the field remains early enough that diversity will be fruitful. Given the trend toward private market value creation, generalist investors will need to grow comfortable with evaluating secondary investment opportunities instead of expecting public companies to capture all the prizes.