Natural Language Engineering (2024), **30**, pp. 650–659
doi:10.1017/S1351324924000032

CAMBRIDGE UNIVERSITY PRESS

INDUSTRY WATCH

Start-up activity in the LLM ecosystem

Robert Dale

Language Technology Group
Email: rdale@language-technology.com

Abstract

The technical and mainstream media's headline coverage of AI invariably centers around the often astounding abilities demonstrated by large language models. That's hardly surprising, since to all intents and purposes that's where the newsworthy magic of generative AI lies. But it takes a village to raise a child: behind the scenes, there's an entire ecosystem that supports the development and deployment of these models and the applications that are built on top of them. Some parts of that ecosystem are dominated by the Big Tech incumbents, but there are also many niches where start-ups are aiming to gain a foothold. We take a look at some components of that ecosystem, with a particular focus on ideas that have led to investment in start-ups over the last year or so.

1. Introduction

Hardly a week goes by without the announcement of a large language model that exhibits either some new capability or a leapfrogging in performance over what's already out there. Many of these advances inevitably come from the major players battling it out in the space, principally OpenAI/Microsoft and Google, with Meta's open-sourced offerings close behind. There are also many smaller but still significant players who contribute to pushing forward the frontiers of generative AI, with Anthropic, Cohere, Mistral, Perplexity, and xAI currently being amongst the most visible. Model development is undoubtedly where the cutting edge of activity in generative AI lies.

But there's a lot more to developing and deploying an LLM than the language model itself. A complex supporting infrastructure has grown up around LLM development and deployment, constituting what we might think of as an ecosystem.

At the center of this ecosystem is the technology stack that brings LLMs to life, with the GPUs that provide the processing power for LLM training and inference at the bottom, the LLMs themselves in the middle, and the user-facing generative AI applications built using these models at the top. Access to LLMs for the bulk of developers who can't afford to, or don't want to, buy their own GPUs is mediated by an additional layer provided by the sellers of cloud compute, principally Amazon Web Services, Microsoft Azure, and Google Cloud, who will rent you the computational resource you need on demand. And interwoven with this technical infrastructure are the ecosystem's human resource elements, with a growing body of training and education assets and services, and every major management consultancy aiming to offer advice and assistance.

There's another technical element, though, and that's the one I want to focus on here. As in other areas of complex and sophisticated software development, there are a range of supporting functionalities you need in order to manage LLMs throughout their lifecycle. If you're just a solopreneur experimenting with generative AI, you might be able to cobble together much of the support you need via a few Python scripts and some version control software. But if you want to productionize your experiments, you need to look seriously at generative AI's analogue of conventional software's DevOps functionalities, widely referred to as LLMOps: a set of technology-supported workflow practices that streamline the process of developing, deploying

and maintaining LLMs, providing efficiency and reliability. Tools and jigs for AI makers, if you like.

It's difficult for a new start-up to compete when it comes to the core generative AI activities of chip development, model development, and compute provision; those elements of the ecosystem are generally the domain of the Big Tech incumbents with their deep pockets, offering relatively little room for smaller players to disrupt. But generative AI is at a sufficiently early stage that we're still working out what LLMOps needs to encompass, and that opens up opportunities for new players to contribute, with lots of space for new ideas and innovation. In what follows, I provide a structured overview of the start-ups that have ventured into this part of the ecosystem over the last year or so.

2. Start-up activity

What follows is an attempt at organizing the space of start-up LLM ecosystem offerings that have received funding over the last year. The categories I've adopted are as follows:

- Data management
- Vector search and databases
- Access to training and inference engines
- Testing and evaluation
- Risk management
- Security
- Customizing LLMs
- End-to-end platforms

The companies described in each section below are listed in order of when they received their most recent funding. I've restricted the review to companies that have received either seed or Series A funding; if a company has achieved a Series B round, that usually means it has already found product-market fit, established a customer base, and is focused on growth, so I no longer consider it a start-up for present purposes.

Perhaps inevitably, the borders between the categories I've chosen are leaky, and there are a number of offerings that could just as easily fit into a category other than that which they are assigned to here. But I think this set of categories provides a reasonable snapshot of where the development and deployment pain points are perceived to be at this stage in the evolution of the generative AI industry. As the technology further develops and evolves, we might expect the contours of the landscape to shift, so things may look different in a year's time.

2.1 Data management

The training or fine-tuning of a generative AI model is completely dependent on data: as has been observed many times in the last few years, data is the new oil. But the wide variation in the specifics of data sources and formats, and concerns about the quality of that data, mean that data access, ingestion, and management are often messy processes that distract the AI researcher from getting on with the more interesting parts of the job. A fair number of start-ups have stepped in help.

- LlamaIndex (founded 2023; $8.5m seed, June 2023) positions itself as a data framework, providing a pipeline of data processing steps for ingestion, chunking, metadata extraction, embedding, and indexing; it supports 160+ data sources and formats and 40+ storage services.
- Cleanlab (founded 2021; $25m Series A, October 2023) targets automated data curation for reliability: the company claims to automatically find and fix errors for LLMs,

https://doi.org/10.1017/S1351324924000032 Published online by Cambridge University Press

identifying problematic instances and edge-cases that prevent an LLM from behaving reliably. Cleanlab's Trustworthy Language Model is an LLM that adds a trustworthiness/confidence score to each output, identifying which outputs are more reliable and which should be double-checked.

- Carbon (founded 2022; $1.3m seed, December 2023), like LlamaIndex, focuses on helping developers manage external data for LLMs; it positions itself as a universal retrieval engine for LLMs to access unstructured data from any source, providing 20+ pre-built connector APIs to ingest unstructured data from a variety of common data sources, and SDKs for a number of popular programming languages.
- DatalogyAI (founded 2023; $11.7m seed, February 2024) focuses on tools for the automated curation of AI training datasets. The platform can identify which data are most important for a given application; which concepts require redundancy in the data; how the dataset might be augmented with additional data; and how it should be batched during model training. The aim is to automatically optimize every step of the process.
- Metaplane (founded 2019; $13.8m Series A, March 2024) aims to improve and rectify data quality issues for enterprises via its end-to-end data observability platform. The toolset detects anomalies in data using machine learning, provides data usage analytics and visibility into data pipelines to identify the causes of problems, and integrates with notification channels to provide alerts when problems are detected.
- Foundational ($8m seed, March 2024) aims to find and fix data issues before any code is deployed. The company's platform automatically analyses data teams' source code to map data lineage and identify potential issues before deployment. The platform integrates with tools like GitHub to provide actionable suggestions and fixes directly within developers' existing workflows. Importantly, the tool doesn't require access to the underlying data itself, only the metadata expressed in the code, reducing data privacy and security concerns.

2.2 Vector search and databases

Vector databases, which index and store documents as vector embeddings for fast retrieval and similarity search, are central to the success of generative AI applications, and play a critical role, for example, in retrieval augmented generation solutions. You can build your own vector database, but in most cases, it makes more sense to use an off-the-shelf product. Again, a number of start-ups want to help.

- Marqo (founded 2022; $12.5m Series A, August 2023) provides a collection of vector search tools with a choice of hundreds of embedding models. The company aims to develop a new form of vector search technology that continuously improves based on user engagement, which differentiates it from existing vector databases.
- SuperDuperDB (founded 2023; $1.75m seed, December 2023) provides a Python package that aims to bridge the gap between existing data storage systems and AI, making it easier for organizations to build and manage AI applications by avoiding data migration, data duplication and ETL pipelines. Instead of requiring the user to move their data to a vector database, the product enables vector search in existing databases such as MongoDB or Snowflake.
- Qdrant (founded 2021; $28m Series A, January 2024) provides an open-source vector search engine and database; the company has developed an efficient compression technology called binary quantization which it says can reduce memory consumption by as much as 32 times and enhance retrieval speeds by around 40 times.

https://doi.org/10.1017/S1351324924000032 Published online by Cambridge University Press

- Upstash (founded 2021; $10m Series A, February 2024) offers a range of products that aim to simplify data management for developers. The most recent, Upstash Vector, is a high-performance vector database designed to maximize developer experience, cost efficiency, and scalability by adopting a serverless model, eliminating the need for developers to worry about deployment and maintenance.
- Superlinked (founded 2021; $9.5m seed, March 2024) provides a compute and data engineering framework to turn data into vector embeddings, optimizing retrieval control, quality, and efficiency in real time for analytics, RAG, search, and recommendation solutions. The toolset supports experimentation and easy deployment and auto-generates ingestion and query APIs.
- Activeloop (founded 2018; $11m Series A, March 2024) addresses the problem of dealing with petabyte-scale unstructured data covering modalities such as text, audio, and video, where that data is distributed across multiple idiosyncratic sources. Its Deep Lake database unifies the storage of these various data types as tensors and facilitates the streaming of these tensors to the SQL-like Tensor Query Language, an in-browser visualization engine, or deep learning frameworks like PyTorch and TensorFlow.

2.3 Access to training and inference engines

Most companies experimenting with generative AI are unlikely to host their own training and inference capabilities, given the high cost of GPU infrastructure, and the associated headaches of resource management; more likely, they will take advantage of the cloud-based model hosting services provided by Microsoft Azure, Google Cloud, or Amazon Bedrock. But even given the overwhelming presence of the heavyweights in this space, investors obviously still see room here for smaller players, with common selling points being efficiency, lower cost, and ease of use.

- Predibase (founded 2021; $12.2m Series A extension, June 2023) claims to be the fastest, most efficient way to fine-tune and serve open-source AI models in a private cloud, with up to a 50x improvement in training speed for task-specific models and a 15x reduction in deployment costs. A key capability is the use of low-rank adaptation (LoRA) to fine-tune large pretrained models to build smaller task-specific LLMs.
- Nomic (founded 2022; $17m Series A, July 2023) focuses on AI explainability and accessibility, providing tools that enable everyone to interact with AI scale datasets and run AI models on consumer computers: Atlas is a tool for interacting with massive datasets, providing visualizations of high-dimensional data to support the manipulation and curation of LLM training data, and GPT4All enables anyone to run open-source AI on any machine, offering a framework for local deployment and a collection of open-source models, requiring either 8 Gb or 16 Gb of local RAM.
- CentML (founded 2022; CA$37m seed, October 2023) focuses on compute efficiency, claiming up to 8x acceleration for model inference. The company's platform attempts to identify bottlenecks during model training and predict the total time and cost to deploy a model. CentML also provides a compiler that automatically optimizes model training workloads to perform best on the target hardware. The product focuses on optimizing inference workloads on NVIDIA GPUs specifically.
- TitanML (founded 2021; €2.6m pre-seed, October 2023) provides enterprise-ready software that claims to make LLM deployment faster, cheaper, and easier; targeting both cloud and self-hosted solutions, the Titan Takeoff Inference Server selects the best inference optimization techniques for your hardware and model and prepares it for deployment.
- DeepInfra (founded 2022; $8m seed, November 2023) offers a platform that aims to simplify AI model integration and execution by providing a simple pay-per-use API for over

https://doi.org/10.1017/S1351324924000032 Published online by Cambridge University Press

100 predominantly open-source models and their variants, relieving the customer of the heavy lifting related to running, scaling, and monitoring LLM inference. The company also provides dedicated instances and clusters and can deploy custom models.

- Foundry (founded 2022; $80m seed + Series A, March 2024) provides a public cloud platform capable of running AI models, aiming to take a bite out of the cloud computing business dominated by AWS, Microsoft, and Google. The company's goal is to make leveraging compute 'as simple as turning on the light', arguing that users of existing AI infrastructure have to spend time on capacity planning and related issues; it claims it can often provide computing power at an order of magnitude lower costs than existing providers.
- Together.ai (founded 2022; $106m Series A extension, March 2024) aims to create the fastest cloud platform for generative AI applications. The platform allows developers to quickly and easily integrate leading open-source models or create their own models through pretraining or fine-tuning; it offers serverless endpoints for 100+ models, an optimized inference stack that claims to provide the best performance at the lowest cost, playgrounds for testing, and dedicated GPU clusters for training and fine-tuning.
- Lumino ($2.8m pre-seed, March 2024) aims to provide an open, efficient, accessible, and cheaper infrastructure for AI workloads by building an integrated hardware and software compute protocol which uses economic incentives to bring distributed compute resources together, blockchain to ensure models are being trained correctly, and an SDK for developers. This sounds like a very interesting idea, but the company is very early stage, and its website is thin on detail.

2.4 Testing and evaluation

Generative AI models are famously nondeterministic: use the same prompt twice and you'll get different responses each time. This provides a dimension of the models' creativity, and in many contexts serves as an important feature of the application. But it makes testing rather more problematic than is the case for standard software development. That doesn't mean, though, that the testing and evaluation of generative AI models has to be ad hoc. A number of start-ups provide solutions that attempt to bring some structure to the task, both prior to the launch of a product and also once it's out there in the field.

- Context.ai (founded 2023; £2.8m seed, August 2023) provides an evaluation and analytics workflow that aims to help companies better understand how users are interacting with their LLMs, and how those models are performing. Support is provided both for before and after product launch. Prior to launch, the evaluation tools enable stress testing of changes to models via test cases whose responses are automatically evaluated. After launch, customers share their chat transcripts via an API, and Context analyses the results, classifying conversations based on topic and analyzing each to determine if the customer was satisfied with the response. Results are presented via a performance monitoring dashboard that makes it easy to identify topics that need attention.
- Raga AI (founded 2022; $4.7m seed, January 2024) aims to build an automated testing platform that can detect LLM issues, diagnose them, and fix them on the fly, using a battery of 300 tests that cover issues as diverse as bias in the training data, poor labeling, data drift, poor hyperparameter optimization during training, and a lack of model robustness to adversarial attacks. These tests are carried out using the company's foundation models, which are trained using AI testing-specific data and allow for domain-specific fine-tuning. Testing is supported both pre- and post-deployment.

https://doi.org/10.1017/S1351324924000032 Published online by Cambridge University Press

- Orq.ai (founded 2022; €1.5m pre-seed, March 2024) provides an all-in-one generative AI platform for enterprises to integrate with various LLMs; its latest products, Experiment and Optimize, complement its existing Run solution. Experiment makes it easy to set up and test thousands of generative AI use cases, prompts, and datasets across 75+ AI models; its Optimize product provides real-time analytics on interactions with generative AI models, supporting ongoing improvements and model fine-tuning.
- Adaptive ML (founded 2023; $20m Series A, March 2024), a start-up founded by the team behind open-source language model Falcon, develops technology that supports the continuous improvement of LLMs based on users' interactions. The platform abstracts away the technical details around fine-tuning and reinforcement learning in its Adaptive Engine product, which supports testing, serving, monitoring, and iteration on LLMs.

2.5 Risk management

Closely related to testing and evaluation is what we'll call here risk management, where the concern is to avoid undesirable outputs. These fall into two overlapping categories of problematic cases that are frequently considered: compliance failure, where a company's policies or other regulations may be violated by the generated output; and safety, where undesirable phenomena such bias, toxicity, or disinformation are present in the output. We also squeeze into this category the automatic detection of deepfakes and fraudulent output.

- DynamoFL (founded 2021; $15.1m Series A, August 2023) focuses on compliance with regulations. Using the company's DynamoGuard product, compliance teams can copy-and-paste their AI governance policies into the application, which then generates a series of example user-interaction edge cases that violate these policies; the compliance team then edits or rejects these edge-case examples to refine the product's understanding of nuanced edge-case violations, captured in a fine-tuned guard model based on the company's proprietary 8 billion parameter multilingual foundation model.
- Patronus AI (founded 2023; $3m seed, September 2023) is building an LLM evaluation tool oriented toward the needs of regulated industries, where there is little tolerance for errors or hallucinations. The product automatically generates adversarial test suites at scale with which it stress tests the models, and then scores model performance based on a proprietary taxonomy of criteria, determining which model is best for a given job. The company has also launched CopyrightCatcher, a copyright detection API for LLMs.
- Reality Defender (founded 2021; $15m Series A, October 2023) aims to detect deepfakes and other AI-generated content; it offers both an API and web app that use proprietary models trained on in-house datasets to analyze videos, audio, text, and images for signs of AI-driven modifications. At its core, and recognizing that there's no 'one-size-fits-all' solution, the product consists of a continuously updated ensemble of models that each focus on different methodologies.
- Intrinsic (founded 2022; $3.1m seed, December 2023) is a platform for building AI agents for user trust, allowing the user to design and implement custom policies beyond standard abuse categories. The company offers two products: content moderation and workflow management. Much like DynamoFL above, the content moderation product allows the user to define nuanced policies in plain English and uses a real-time detection classifier to determine which policies have been violated, along with a confidence score; human-in-the-loop retraining allows the model to learn from false positives. The workflow management platform is a collection of tools for managing trust and safety issues.
- Braintrust Data (founded 2023; $5.1m seed, December 2023) targets the problem of AI evaluation by providing a dedicated toolset that lets teams see how their AI model performs, allowing improvement well before the model reaches the production stage. The

https://doi.org/10.1017/S1351324924000032 Published online by Cambridge University Press

stack includes tools for dataset management; evaluating model performance and interrogating failures; a prompt playground for comparing prompts; easy access to a range of proprietary and open-source AI models via a single API; and support for continuous integration.

- Distributional (founded 2023; $11m seed, December 2023) aims to build a platform for robust and repeatable AI testing, enabling AI product teams to proactively and continuously identify, understand, and address AI risk before it harms their customers in production. The company is very early stage; its website points to the need for a systematic approach to testing, providing little detail on the proposed solution but pointing to a couple of research papers that appear to describe earlier iterations of the product.
- Guardrails AI (founded 2023; $7.5m seed, February 2024) provides a governance layer that aims to minimize the risks that come with the use of LLMs. Its Guardrails Hub lets developers build, contribute, share, and re-use advanced validation techniques known as validators. These validators can be used with Guardrails, the company's open-source product that acts as a critical reliability layer for building AI applications that adhere to specified guidelines and norms.
- Armilla AI (founded 2019; CA$6m seed, February 2024) provides an AI-powered assurance platform, Armilla Guaranteed, which evaluates AI models for compliance with global AI regulations, identifies risks such as bias, and offers performance guarantees to mitigate buyer risks, protecting enterprises against the failure of AI technologies by reimbursing license fees for underperforming AI models.

2.6 Security

If you're building a generative AI solution for an enterprise, you're sure to hear concerns expressed very early on about security, and about data security in particular. If you're using a cloud-hosted AI training and inference service, you'll need to be able to provide assurances that the enterprise's data won't find its way into the wider world. In the first few months following ChatGPT's release, many organizations forbade their employees from using the tool, lest important data provided in a prompt should leak or become part of the training data for a future version. Short of bringing the training and inference functions inside the corporate firewall, this is a hard problem, but some start-ups aim to help.

- Protect AI (founded 2022; $35m Series A, July 2023) aims to strengthen ML systems and AI applications against security vulnerabilities and data breaches—particularly those that arise from the use of open-source software—via its AI Radar security platform, which provides users with real-time visibility, detection, and management capabilities. The platform uses a 'machine learning bill of materials' to track all the components used in an AI application, running continuous security checks to find and remediate vulnerabilities. Security policy violations are detected using integrated model scanning tools for LLMs and other ML inference workloads, and a visualization layer provides real-time insights into an AI application's attack surface.
- Credal AI (founded 2022; $4.8m seed, October 2023) provides enterprises with a secure way to connect their internal data to cloud-hosted generative AI models. Using a collection of pre-built data connectors, the platform mirrors the permissions of the source data systems it connects to, automatically redacts and anonymizes PII and provides warnings when sensitive data is about to be shared with external LLMs. The company also provides tools for building generative AI applications using a range of popular LLMs.
- Enkrypt AI (founded 2023; $2.4m seed, February 2024) offers Sentry, a secure enterprise gateway that sits between users and models, delivering visibility and oversight of LLM

https://doi.org/10.1017/S1351324924000032 Published online by Cambridge University Press

usage and performance across business functions, protecting sensitive information while guarding against security threats, and managing compliance with automated monitoring and strict access controls.

2.7 Customizing LLMs

The diversity of training data used to develop most generative AI models means that they are necessarily general purpose in nature. It's often observed that better results in a specific domain are likely to be achieved by a model trained on data from that domain. A common requirement, then, is the fine-tuning of an existing LLM using a company's proprietary data. Of course, security is a key issue here too.

- Contextual.ai (founded 2021; $20m seed, June 2023) focuses on what it calls 'artificial specialized intelligence': their contention is that current 'first-generation' LLMs are not really appropriate for real-life applications, and that what's needed are contextual language models that take account of an enterprise's data. It's unclear from their website whether this is achieved via a refined version of RAG, domain-tailored LLMs, or some combination of both.
- Glaive (founded 2023; $3.5m seed, August 2023) provides a platform to build, measure, and improve use-case specific language models with the help of synthetic data, driven by the recognition that a significant blocker to having specialist models today is the availability of high-quality use-case specific data.
- Gradient (founded 2022; $10m seed, October 2023) offers a platform for fine-tuning and deploying LLMs with proprietary data. The platform hosts a number of open-source LLMs which users can fine-tune to their needs; the company also provides pretrained models that target particular use cases (e.g., data reconciliation, context-gathering, and paperwork processing) and specific industries (like finance, law, and healthcare). Gradient claims to be one of the few platforms that supports productionising multiple models at once, allowing an organization to develop and integrate large numbers of LLMs into a single system.
- Rohirrim (founded 2022; $15m Series A, December 2023) positions itself as specializing in 'domain-aware generative AI for the enterprise', and emphasizes its ability to securely leverage proprietary data. Although the company currently focuses on supporting the proposal writing process, it claims the potential use cases for its patented technology are much broader.
- Pienso (founded 2016; $10m Series A, March 2024) provides a platform that targets non-technical domain experts, allowing them to construct, deploy, and manage LLMs without having to write any code. The platform offers access to a range of foundation models which can be switched in and out at will and fine-tuned using the customer's own data, which is never transmitted outside of the customer's secure environment.
- OpenPipe (founded 2023; $6.7m seed, March 2024) offers a fine-tuning platform that lets developers build LLM models tailored to their specific use case; the key proposition is that smaller models can result in better performance and lower cost compared to using larger general purpose models.

2.8 End-to-end platforms

Each of the companies and products described above generally focuses on a specific pain point in the process of developing a generative AI application. A number of companies, however, aim to provide an end-to-end platform that supports all of the stages that arise in developing and deploying generative AI applications.

https://doi.org/10.1017/S1351324924000032 Published online by Cambridge University Press

- Shakudo (founded 2021; CA$9.5m Series A, July 2023) describes itself as 'the operating system for data and AI', aiming to create compatibility across the best-of-breed data and AI tools available. The company provides a large library of third-party components for all stages of the generative AI life cycle and a unified interface that connects these tools to a range of popular infrastructure providers.
- FEDML (founded 2022; $11.5m seed, July 2023) aims to help companies efficiently train and serve custom LLMs using proprietary data, while reducing costs through decentralized GPU cloud resources shared by the community. Its platform provides access to a wide range of open-source models, tools for managing training runs, and management of distributed compute resources for fine-tuning and deployment via federated learning, where multiple entities collaboratively train a model on decentralized GPUs, multiclouds, edge servers, and smartphones.
- Datasaur (founded 2019; $4m seed, August 2023) is primarily a customizable tool for labeling language data, but the company has also launched LLM Lab, a comprehensive one-stop shop to help teams build and train custom LLMs. The product provides a unified interface for handling different aspects of building an LLM application, covering internal data ingestion, data preparation, retrieval augmented generation, embedded model selection, similarity search optimization, and optimization of server costs.
- Portkey (founded 2023; $3m seed, August 2023) provides tools that allow businesses to monitor their language model operations, connect to multiple LLMs efficiently, and experiment, improve, and manage prompts effectively. Its Observability suite tracks key metrics and streamlines debugging; its AI Gateway supports connection to 100+ AI models using a single consistent API and provides load balancing; and its Prompt Management tool assists with prompt creation, versioning, and deployment.
- LastMile (founded 2023; $10m seed, September 2023) is motivated by the view that we need a new class of AI developer tools built for software engineers, not just ML research scientists. The company's goal is to provide a single developer platform that is model agnostic, encompasses the entire lifecycle of AI application development, and is accessible for app developers and engineers. Their first product is the AI workbook, a notebook interface for experimenting with generative AI models.
- LangChain (founded 2022; $25m Series A, February 2024) is a framework for constructing LLM-powered applications. It offers three products: LangChain, a library of interoperable and interchangeable components for building end-to-end applications; LangSmith, a unified platform for developing, collaborating, testing, deploying, and monitoring LLM applications; and LangServe, a platform for application deployment and maintenance.
- VectorShift (founded 2023; $3m seed, February 2024) aims to simplify LLM application development with a modular no-code approach: it provides an end-to-end AI platform that lets users drag and drop components to build, deploy and maintain production-grade LLM workflows. The product provides a library of components and a collection of pre-built application templates that can be customized.

3. Summing up

It's an oft-used aphorism that, when everyone is digging for gold, the best business to be in is one that sells picks and shovels. In the context of the AI gold rush, GPU chips are the most obvious picks and shovels, and there's no doubt that the current clear winner in the AI boom is Nvidia. Over the last year, Nvidia's market value has shot past US$2 trillion, making it the third most valuable company in the world by market cap after Microsoft and Apple.

https://doi.org/10.1017/S1351324924000032 Published online by Cambridge University Press

The cloud compute providers—AWS, Microsoft Azure, and Google Cloud—are the other big pick-and-shovel winners in the LLM ecosystem; in 2023Q3, the three had a 66 percent share of the $68 billion worldwide cloud market, with 31, 24, and 11%, respectively.

The start-ups described here are obviously tiny in comparison to the tech giants; they together account for a total of just over $600m invested in the last year. But they are nonetheless also firmly in the picks and shovels business. Of course, they will not all be winners. In the medium to longer term, it seems likely that end-to-end solutions will win out; fragmented point solutions multiply integration costs. For most of the pain points identified above, a quick analysis suggests that there are multiple competing solutions on offer, and the focus here on start-ups means that we haven't mentioned the more established companies already gaining traction addressing the same problems. And the more successful will inevitably find themselves competing with the ever-richer feature sets of the product offerings from the Big Tech players, unless of course they are acquired by those same players to become those features—surely the exit strategy of many.

As long as generative AI promises rich seams of gold, the ecosystem will only grow. Whether these start-ups get acquired or carve out their own niche, one thing's certain: in the AI gold rush, those selling the right picks and shovels are poised to strike it rich.

If you'd like to keep up to date with what's happening in the NLP industry, consider subscribing to the free *This Week in NLP* newsletter at https://www.language-technology.com/twin.

Cite this article: Dale R (2024). Start-up activity in the LLM ecosystem. *Natural Language Engineering* **30**, 650–659. https://doi.org/10.1017/S1351324924000032