**Module View**

Generative AI (GenAI) infrastructure covers the underlying technology and complex architecture used to develop, train, deploy, and monitor GenAI models.

Companies in this industry play a pivotal role in enabling the development and deployment of GenAI applications across various domains, including natural language processing, computer vision, and creative content generation. It encompasses diverse products and tools that support the lifecycle of GenAI models, including hardware, data storage, AI model management (development, training, deployment, monitoring), and prompt engineering.

\* Note: Additional sections such as market sizing can be provided on request.

**Detailed View**

**Behind the Scenes with ChatGPT, DALL-E, and Bard**

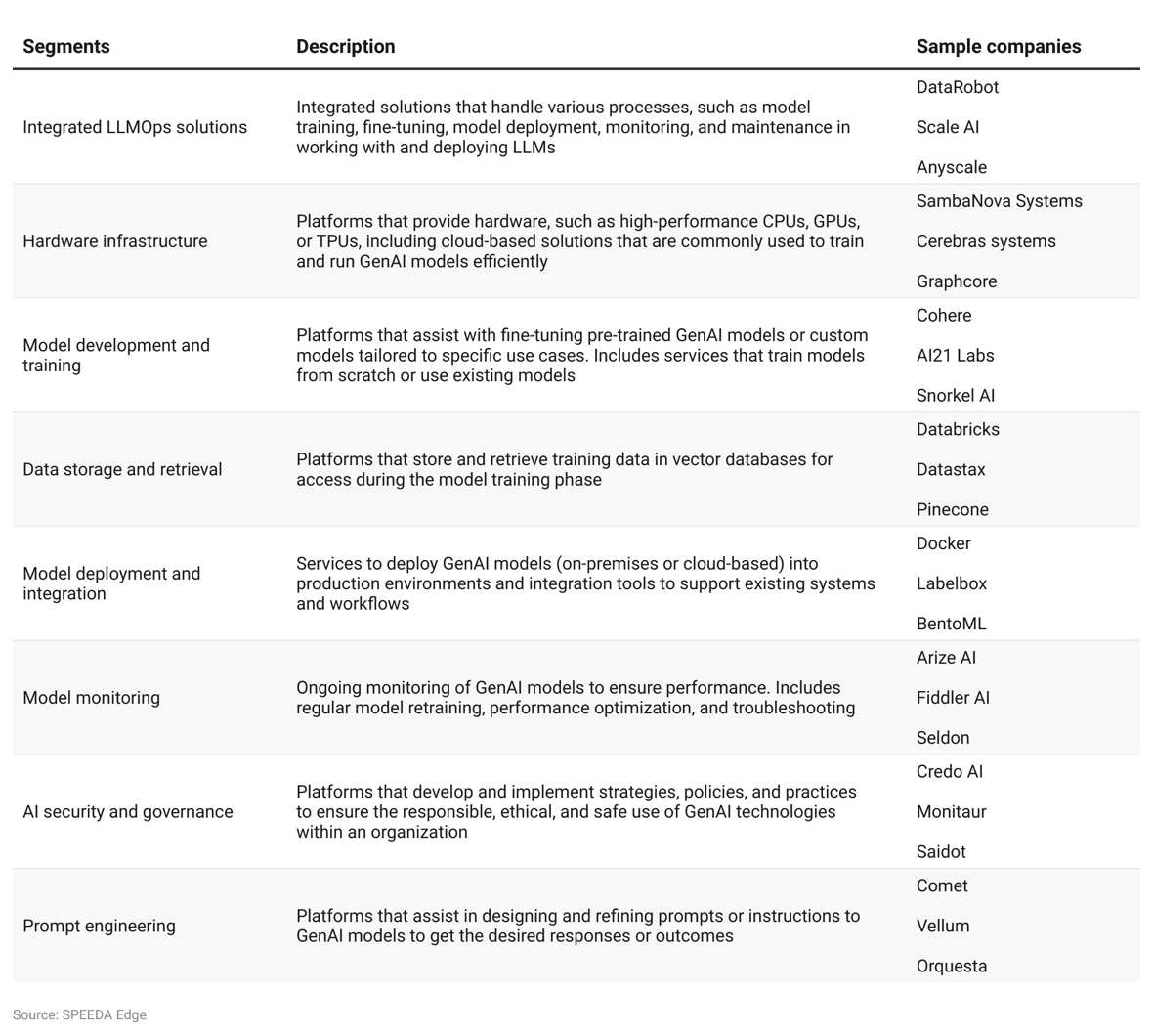
Since the launch of OpenAI’s ChatGPT in November 2022, “generative AI” (GenAI) is fast becoming a household term. 2024 has seen a sharp uptick in the usage of GenAI-powered solutions; one may even say we are spoiled for choice with the number of options currently available in the market. GenAI is now each person’s personal assistant, helping us with menial and creative tasks, such as answering questions (ChatGPT, Bard), writing (Jasper, Copy.ai), and creating images (DALL.E, Midjourney). GenAI can be found across a wide range of areas like healthcare diagnostics and software development and has expanded beyond technical fields to offer value across various domains, similar to the widespread impact of the internet.

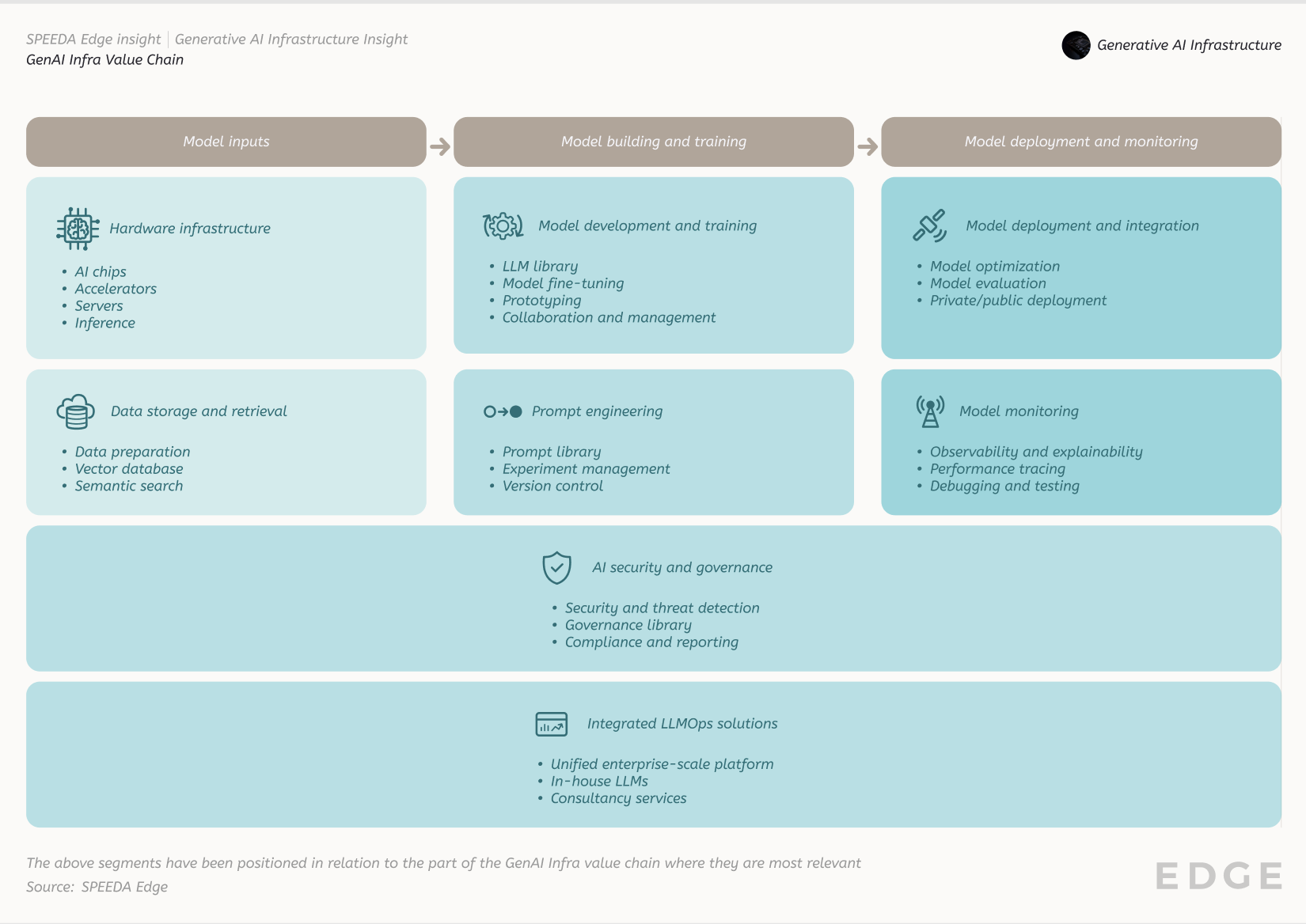
Because of this boom, a slew of firms have quickly jumped into the GenAI market with different use cases and solutions that can be applied across various industries, hoping to make them “firsts.” GenAI is now here to stay as it becomes increasingly embedded into our everyday lives, which is reason enough for companies to use it in their products and services and to augment their business processes.

What is GenAI Infrastructure?

GenAI Infrastructure (GenAI Infra) includes a range of tools and platforms used to develop, train, deploy, and monitor large language models (LLMs) as well as GenAI models and applications. Based on our research, we have identified a range of firms that make up a vast support system which supports the industry across the GenAI value chain

### **Key types of GenAI Infra solutions**



**

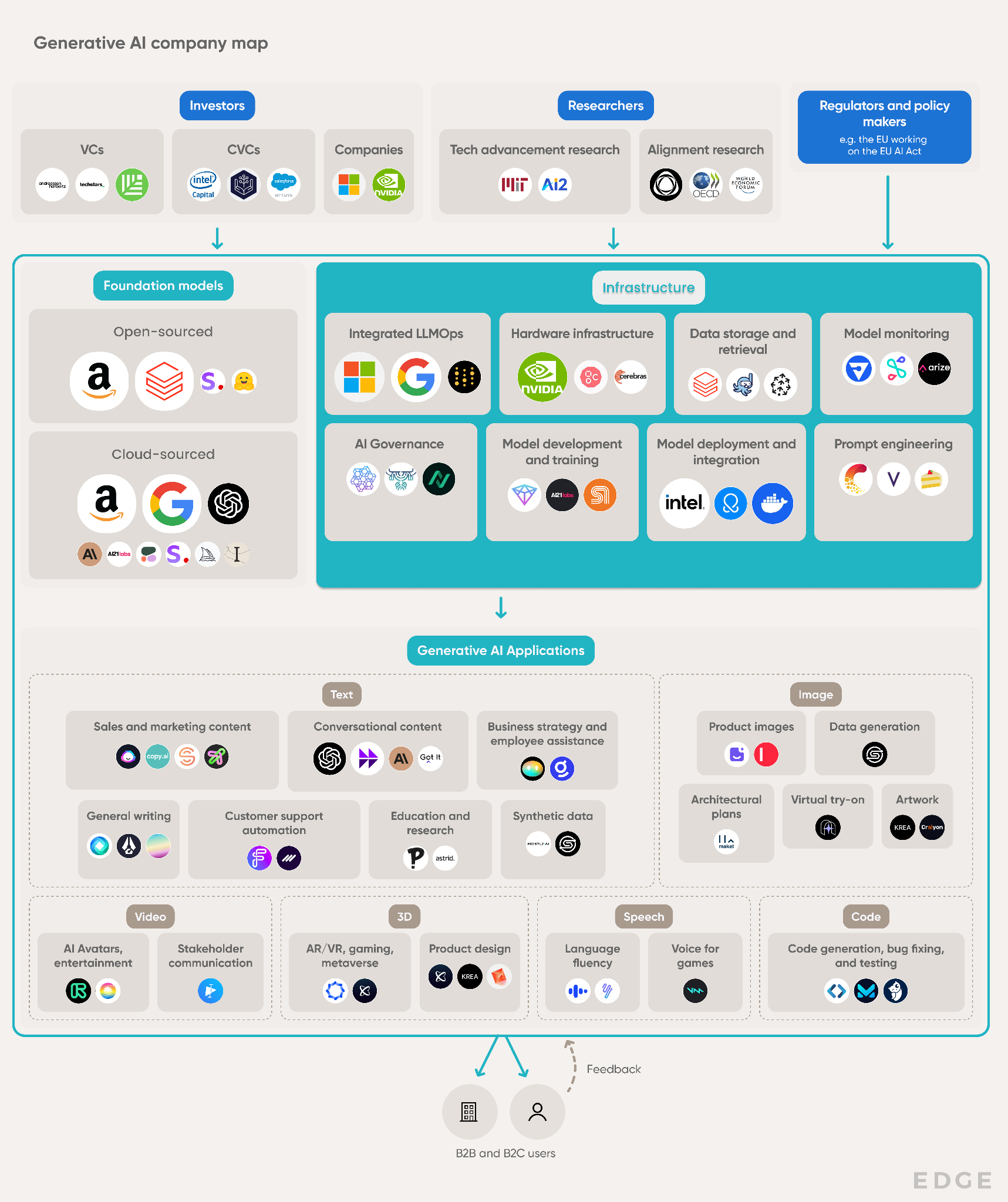
These firms provide a diverse range of tools and platforms that support the lifecycle of GenAI models across various domains, including natural language processing (NLP), creative content generation, and analytics and forecasting. Notably, some of the GenAI Infra players fall under the broader [Machine Learning Infrastructure](https://sp-edge.com/industry/160) industry, given their solutions are used by other AI-related uses (such as predictive modeling, autonomous vehicles, and robotics).

Most of these platforms and tools specialize in providing solutions for a particular segment, while a number of established players provide integrated solutions targeting the end-to-end LLMOps lifecycle. In addition to software-based solutions, GenAI Infra also includes hardware infrastructure that supports model development and covers application-specific processors, CPUs, GPUs, and inference that are used to train and run AI models. In addition, many enterprise-focused firms provide consultancy services to advise firms on complex, domain-specific model development and strategic guidance on AI safety and governance.

GenAI Infra solution providers support players within GenAI applications to develop end-user solutions based on GenAI, as well as to build domain-specific foundation models. For more details about [application](https://sp-edge.com/industry/163) and [foundation model](https://sp-edge.com/industry/169) layers, please refer to the respective industry hubs.

### 

### **The GenAI ecosystem**



Source: Created by SPEEDA Edge

The companies included in this diagram serve as representatives of their respective segments. For a comprehensive list of companies focused on GenAI Infra, please refer to the market map given below

**Shortfalls in off-the-shelf GenAI models lead firms to develop their own**

According to [McKinsey’s “State of AI” in early 2024](https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai), 65% of respondents reported regular use of GenAI in their organizations compared to [22% in 2023](https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-in-2023-generative-ais-breakout-year). It is therefore unsurprising that a host of off-the-shelf GenAI models have emerged, attempting to fit every use case from content creation and virtual assistants to analysis, research, and software coding tools. The potential use cases of GenAI are growing at a pace that is difficult to keep up with; many new use cases have already been identified, such as [content moderation](https://sp-edge.com/industry/177), [digital humans](https://sp-edge.com/insights/23419) and [Humanoid Robots](https://sp-edge.com/industry/185). See our insight on [The *State of GenAI Adoption*](https://sp-edge.com/insights/30429) *for the adoption of GenAI across emerging industries.*

However, integrating these solutions without a proper understanding of the business needs may result in inefficiencies and even hinder future adoption. At the same time, firms may unknowingly expose themselves to risks, such as data biases and security vulnerabilities, by force-fitting GenAI models into a firm's workflows without proper review and customization[[1]](#footnote-1).

Given these concerns over ill-fitting models, firms demand solutions for developing their own AI models. This is when GenAI Infra solutions enter the scene: They can help firms make informed choices and provide a range of services across the model development lifecycle.

The availability of user-friendly developer solutions has led many more firms across multiple industries to turn to GenAI Infra service providers in finding the best AI- and GenAI-powered solutions for their specific requirements. For instance, [A121 Labs](https://www.ai21.com/blog/harambee-case-study), the developer of the Jurassic series of LLMs, provides tools to build AI applications. Its users claim they are particularly drawn to its ability to select custom models to fit their unique requirements by interacting with them, exploring presets, and prototyping the different options before making a commitment. Further, startups are continuously innovating in the infrastructure space by improving the efficiency of the model development process. For instance, [Lightmatter](https://sp-edge.com/companies/596587) develops photonic chips that accelerate data processing and reduce energy consumption.

**Application of GenAI Infra across industries**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Customer** | **Sector** | **GenAI Infra partner** | **Use case** | **Description** | **Source** |
| University Medical Centre Mannheim | Healthcare | Datarobot | Predictive analytics | Used Datarobot’s platform to help physicians and researchers analyze clinical data for various applications, leading to 30% of patients experiencing improved quality of life within six months | [Link](https://www.datarobot.com/customers/how-a-medical-center-accelerates-clinical-research-with-ai/) |
| FifthEdge | Information technology | Accubits | Recruitment process automation | Accubits developed an AI-based recruitment system to streamline repetitive tasks in the hiring process, resulting in cost and time reductions of up to 90% | [Link](https://accubits.com/case_studies/recruitment-process-automation-solution-using-ai/) |
| Wayfair | Retail and manufacturing | Snorkel AI | Product tagging | Snorkel AI developed a data-centric AI workflow to improve automated catalog tagging and used computer vision to extract detailed product image information, resulting in 10x faster labeling of catalog data than manual methods | [Link](https://snorkel.ai/customer-story/wayfair/) |
| Airbnb | Transport and logistics | Labelbox | Object detection | Used Labelbox to integrate model-assisted labeling with human expertise to efficiently prepare high-quality data for training, testing, and validation | [Link](https://labelbox.com/solutions/object-detection/) |
| Blue River Technology | Information technology | Labelbox | Data labeling | Used Labelbox’s automation suite to standardize and manage its data from a single location, resulting in a 50% drop in labeling costs | [Link](https://labelbox.com/customers/brt-data-engine/) |
| Optimus | Transport and logistics | Scale AI | Data labeling | Selected Scale AI as its data annotation partner to speed up its GenAI development | [Link](https://scale.com/customers/optimus-ride-3d) |
| Pietra | Retail and manufacturing | Scale AI | Content generation | Used Scale Forge, an AI-powered marketing suite that utilizes fine-tuned models and humans in the loop to enhance its products on its ecommerce platform. It resulted in a 38% increase in clickthrough rates | [Link](https://scale.com/customers/pietra) |
| OYAK Cement | Power and energy | DataRobot | Process optimization | Used AI-assisted process controls to increase alternative fuel usage by 7x, reduce C02 emissions by 2%, and reduce costs by USD 39 million | [Link](https://www.datarobot.com/customers/oyak/) |
| Harris Farm | Retail and manufacturing | DataRobot | Demand forecasts | Implemented an AI-powered decision-making system for demand forecasting, which resulted in a 10x increase in resource capacity and resulting in more accurate purchasing of perishable inventory | [Link](https://www.datarobot.com/customers/harris-farm-markets/) |
| Flexiti | Financial services | DataRobot | Predictive analytics  Personalization | Used DataRobot’s solutions to speed up model creation time by 6x, resulting in faster fraud detection and increased collection rates | [Link](https://www.datarobot.com/customers/flexiti/) |
| Manz | Defense and professional services | DeepSet | Semantic search | Used DeepSet Cloud to speed up its research of case laws and to review documents with the ability to find 30 facets of a problem with a single query | [Link](https://www.deepset.ai/nlp-in-the-legal-industry) |
| Graphcore | Information technology | Weights & Biases | Model development and training | Enabled Graphcore to deliver 50x–100x more IPU experiments to speed up model development | [Link](https://wandb.ai/wandb_fc/case-studies/reports/Learn-How-Graphcore-is-Supporting-the-Next-Generation-of-Large-Models-with-the-Help-of-W-B--Vmlldzo3MjQ5Njc) |
| Socure | Information technology | Weights & Biases | Model development and training | Resulted in a 15% increase in model building efficiency while saving an additional 15% on hardware spend | [Link](https://wandb.ai/wandb_fc/use-cases/reports/How-Socure-Fights-Fraud-With-Machine-Learning--VmlldzozMDExNTAw) |
| US Air Force | Defense and professional services | C3.ai | Predictive analytics, Model development and training | To speed up the model development time to days compared to weeks or months with traditional data science tools, resulting in a 85% reduction in time taken to conduct alert analysis with a 92% accuracy rate | [Link](https://c3.ai/customers/enterprise-ai-for-aircraft-predictive-maintenance/) |
| Shell | Power and energy | C3.ai | Predictive analytics, Model development and training, Model management | Partnered with C3.ai to develop three AI-powered apps and automate predictive maintenance of 10,000 pieces of equipment | [Link](https://c3.ai/enterprise-ai-at-shell/) |
| Georgia Pacific | Retail and manufacturing | C3.ai | Predictive analytics | Partnered with C3.ai to deploy its Reliability application to improve its data science workflow, which reduced unplanned downtime for complex assets and other critical equipment | [Link](https://c3.ai/enterprise-ai-at-Georgia-Pacific/) |
| ReverseAds | Marketing and media | InData Labs | Model development and training | Developed an NLP semantics solution that enabled ReverseAds to effectively segment users to provide targeted ad campaigns, which increased ad campaign performance and reduced advertising costs by 54% | [Link](https://indatalabs.com/resources/targeted-advertising-system) |
| Apple | Information technology | Snorkel | Model development and training, Data labeling | Used an internal Snorkel-based system to create an AI-powered solution with 2.9x fewer errors and 32x increase in generating data labels | [Link](https://snorkel.ai/case-studies/#apple) |
| Google | Information technology | Snorkel | Data labeling | Used Snorkel to replace six months of hand-labeled data work with 30 minutes of automated data labeling, resulting in a 52% performance improvement | [Link](https://snorkel.ai/google-content-classification-models-case-study/) |
| Intel | Information technology | Snorkel | Data labeling | Deployed a proto version of Snorkel Osprey to replace six months of crowdsourced data-labeling work with programmatically generated labels that resulted in an 18.5-point performance improvement | [Link](https://snorkel.ai/case-studies/#apple) |
| Schllumberger | Power and energy | Snorkel | Model development and training | Partnered with Snorkel to build an AI/ML application for processing client reports within three days, resulting in a reduction in processing time from one to three hours per report to a matter of seconds | [Link](https://snorkel.ai/case-studies/#apple) |
| Stanford Medicine | Healthcare | Snorkel | Data labeling | Partnered with Snorkel to replace eight months of data labeling work with a 94% accuracy rate | [Link](https://snorkel.ai/case-studies/) |
| Tide | Financial services | Snorkel | Model development and training, Model deployment | Partnered with Snorkel to programmatically label invoices and transaction data, replacing 15 days worth of time spent hand-labeling data while achieving a 97% accuracy rate | [Link](https://snorkel.ai/case-studies/) |
| Harambee | Defense and professional services | A121 Labs | Model development and training | Used A121 Labs’ custom models to create an NLP chatbot app to onboard job seekers, resulting in a 20% sign-up increase | [Link](https://www.ai21.com/blog/harambee-case-study) |
| Gong | Marketing and media | Pinecone | Database management | Used Pinceone’s vector database, Gong, to run high-performance applications across 10+ billion records without any difficulty | [Link](https://www.pinecone.io/customers/) |
| Capital One | Financial services | Seldon | Model deployment | Used Seldon’s model-as-a-service platform to speed up model deployment and updates from months to minutes | [Link](https://www.seldon.io/customers-capital-one) |
| Comcast | Marketing and media | Datatron | Model monitoring | Used Dataron’s model monitoring module to save the work of four full-time employees | [Link](https://datatron.com/) |

### Source: Compiled by SPEEDA Edge based on company websites

### **What’s driving demand?**

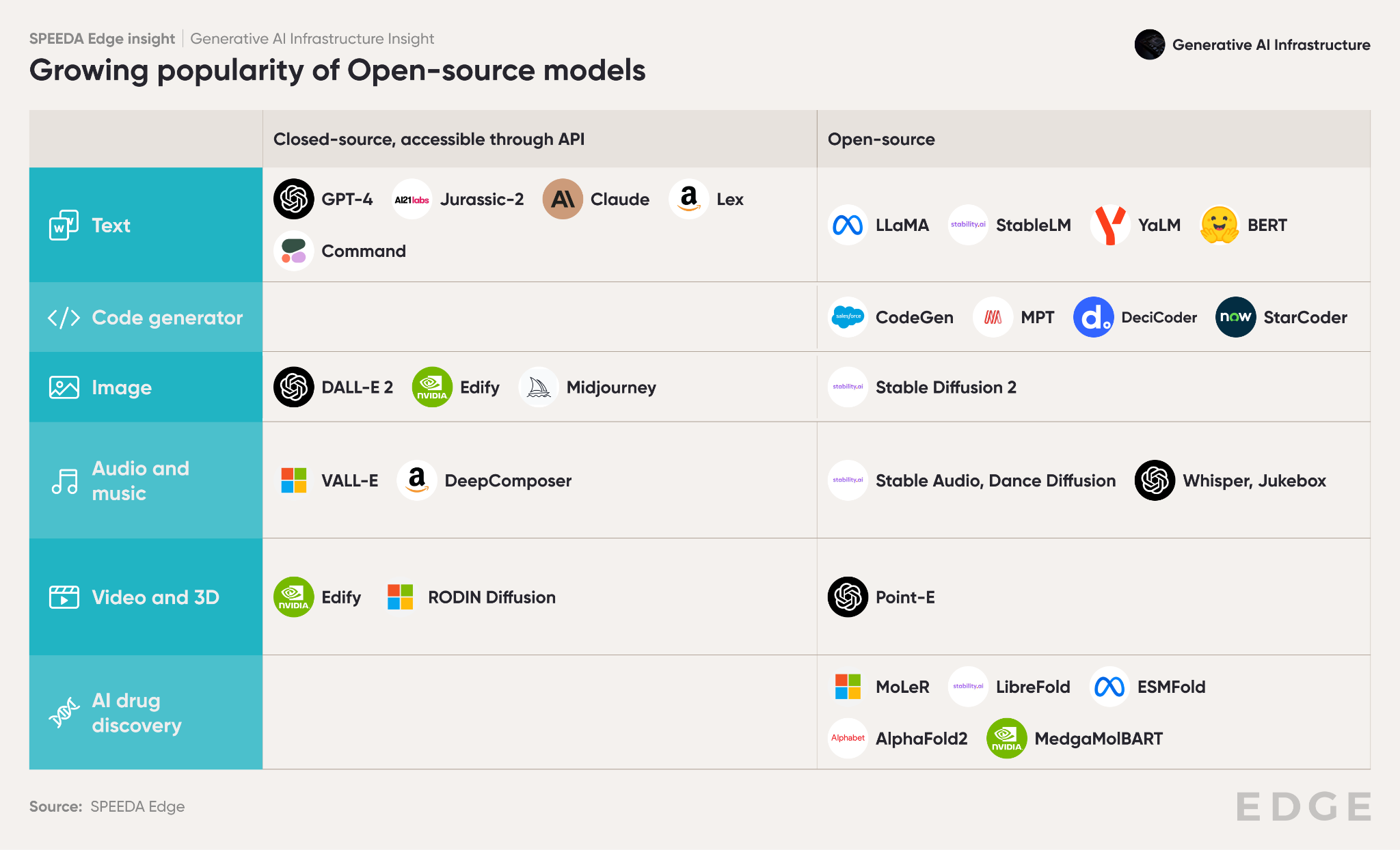
**1. The open-source movement expands opportunities for GenAI developers**

A number of prominent LLM developers, such as OpenAI (GPT-4), Alphabet (LaMDA), and A121 Labs (Jurassic-1)[[2]](#footnote-2), are closed-source models (the source code is kept secret). Concurrently, the open-source software movement has actively promoted open-source AI models. For example, Meta, the developer of the open-source LlaMA model, launched the [AI Startup Program](https://sp-edge.com/updates/23665)[[3]](#footnote-3) with Hugging Face and Scaleway to help French firms adopt open-source AI models. Further, Meta introduced [pre-trained models](https://sp-edge.com/updates/31479) using a multi-token prediction method to simplify the development and deployment of LLMs. In addition, many closed-source model developers, including OpenAI, Anthropic, and A121 Labs, have provided access to their models via API licensing agreements, facilitating firms to integrate these models into their respective business processes for various use cases. For instance, Anthropic offers models such as Claude, which are accessible through API licensing for tasks like conversational AI.

Studies have also shown that open-source models come with several benefits. Developers can harness more data and continuously tweak and make improvements, which has proven to produce more consistent and superior results. A case in point is image generation solutions provider DALL.E (a closed-source model developed by OpenAI) vs. Stability AI’s open-source Stable Diffusion model. While [app comparison studies](https://zapier.com/blog/stable-diffusion-vs-dalle/) show that both models offer impressive AI image generation solutions, Stable Diffusion offers a superior solution with more realistic output and better options and control[[4]](#footnote-4).

**Growing popularity of open-source models**

|  |  |  |
| --- | --- | --- |
| **Model type** | **Closed-source, accessible through API** | **Open-source** |
| Text | OpenAI (GPT-4)  Amazon (Lex)  Anthropic (Claude)  A121 (Jurassic-2)  Cohere (Command) | Meta (LLaMA)  Stability AI (StableLM)  Yandex (YaLM)  Hugging Face (BERT) |
| Code generator |  | Salesforce (CodeGen)  MosaicML (MPT)  Deci AI (DeciCoder)  ServiceNow (StarCoder) |
| Image | OpenAI (DALL.E 3)  Midjourney | Stability AI (Stable Diffusion 2) |
| Audio and music | Microsoft (VALL-E)  Meta (AudioGen)  Google (MusicLM)  Amazon (Deep Composer) | OpenAI (Whisper, JukeBox)  Stability AI (Stable Audio) |
| Video and 3D | Microsoft (RODIN Diffusion, Godiva)  NVIDIA (Edify) | OpenAI (Point-E) |
| Protein structures and DNA |  | Microsoft (MoLeR)  Meta (ESMFold)  Google (AlphaFold3)  Stability AI (LibreFold)  NVIDIA (MegaMoIBART) |



Firms have also exhibited a growing preference for open-source models. A [2024 survey by OpenLogic](https://www.openlogic.com/resources/state-of-open-source-report#form) found that ~95% of organizations have either continued or expanded their use of open-source software, where ~33% said their usage increased significantly. Many IT leaders expect to shift from proprietary, closed-source software to enterprise and community-based open-source solutions over the next two years.

While closed-source and private LLMs such as Odyssey (by [InteliGems Labs)](https://www.einnews.com/pr_news/661439665/inteligems-labs-unveils-odyssey-the-private-large-language-model-that-s-actually-private) and HealthGPT (by [Harman](https://www.pymnts.com/artificial-intelligence-2/2023/harman-launches-generative-ai-solution-for-healthcare-enterprises/)) will continue to play a role for specific use cases (e.g., in healthcare, where data privacy is a concern) the popularity of open-source solutions is anticipated to keep rising.

These developments open up avenues for many new, independent GenAI Infra solution providers that offer model-agnostic solutions with access to various foundation models and custom-made models. For example, [Orquesta](https://sp-edge.com/companies/2277604), an Amsterdam-based prompt engineering platform, supports a large number of well-known LLM providers and models, such as Anthropic, Cohere, Google, and Hugging Face. This gives users the freedom to switch between different models, thus avoiding vendor lock-in.

**2. Transparent and flexible pricing makes AI development more accessible to a wider audience**

Subscription-based pricing and usage-based pricing are the most common forms of pricing for many GenAI Infra service providers. [LastMile AI](https://sp-edge.com/companies/3200989) goes further and offers a “free forever” feature for model development. Usage-based pricing that reflects the differing computing power and data needed for AI model training and development provides firms with greater visibility on the expected development cost.

In addition to paid pricing models, many firms also provide open-source solutions for model monitoring and prompt engineering on a free-to-use basis for individual developers, research, and other non-commercial purposes. For example, [Deepchecks](https://sp-edge.com/companies/928739), which has a focus on keeping ML observability and validation solutions open-source and community-focused, provides free access to its open-source models for individual developers while offering a “pay-per-model” pricing structure to firms.

**Common pricing models for GenAI Infra solutions**

|  |  |  |  |
| --- | --- | --- | --- |
| **Pricing model** | **Description** | **Firm segments using the pricing model** | **Example** |
| Free-to-use | Typically for open-source solutions that are offered to individual developers or small development teams | • Prompt engineering  • Data storage and retrieval | **Chroma**  Currently offered free of charge  **PromptLayer**  Free to use for individual developers  Custom pricing for higher usage limits |
| Subscription-based | Offers varied pricing for firms of different sizes. Many firms also offer a free tier for personal or non-commercial use with limited features  For enterprise users, custom pricing is offered when providing tailored solutions to fit their specific requirements | • Integrated LLMOps solutions  • Model development and training  • Data storage and retrieval  • Model monitoring  • AI security and governance | **Weights & Biases**  • Personal: Free of charge with limited features  •Teams: USD 50 per user/month (annual billing) • Enterprise: Custom pricing  **Portkey AI**  Developer: Free version with limited features  Production: USD 49/month  Enterprise: Custom pricing |
| Pay-per-use | Provides access to its computing systems or model fine-tuning training services, where the pricing is based on a time or resource usage | • Hardware Infrastructure  • Model development and training  • Model deployment and integration  • Model monitoring | **OctoML**  Small: USD 0.4/hour  Medium: USD 1.15/hour  Large: USD 5.2 /hour  **Cohere**  Price per token  1 million tokens: USD 0.4  100 million tokens: USD 40 |
| Licensing | Firms typically provide licensed access to a firm’s models and solutions via an API for commercial use | • Model development and training  • Data storage and retrieval | **Unstructured**  Licensed commercial API that supports a wider range of file formats |
| Outright sales | Many firms directly retail their AI chips, processors, and other hardware needed for AI model development | • Hardware infrastructure | **NVIDIA** Blackwell AI chip: USD 30,000–35,000  A100 AI chip: USD 10,000  H100: USD 30,000  **Flex Logix**  Small orders: USD 99–199 per chip  Large orders: USD 34–69 per chip |
| Managed Services | Add-on services, such as strategic planning and other consultancy services, for enterprises with complex use cases and firms lacking in-house AI expertise | • Integrated LLMOps solutions |  |

### Source: Compiled by SPEEDA Edge based on company websites

Access to open-source resources coupled with such transparent and flexible pricing structures make model development more affordable and accessible. A [survey](https://techaisle.com/blog/529-techaisle-survey-shows-the-rise-of-generative-ai-in-smbs-and-midmarket-firms) in 2023 indicated that 60% of SMBs and 84% of mid-market firms are either using or planning to use GenAI within the next six months. Additionally, 72% of mid-market firms have expressed interest in broadening their in-house hiring for GenAI; these efforts align with GenAI Infra service providers, which aim to complement and bolster these activities.

In addition, cost concerns have also spurred GenAI Infra solution providers to build their own smaller LLMs and more efficient model training infrastructure, which can make model development more efficient and less costly. For instance, [Giga ML](https://sp-edge.com/companies/3164885) (launched in 2023) provides a platform for deploying and fine-tuning on-premise LLMs. It has developed its own pre-trained, fine-tuned version of Meta’s Llama2 model, which claims to result in a 70% reduction in costs and a 300% increase in output delivery speed compared with GPT-4 models.

**3. GenAI Infra tools boost developer happiness and efficiency, expediting AI adoption**

Along with cost, firms need to invest a considerable amount of developer time and effort to build, train, and deploy GenAI models and applications. According to some estimates, depending on the complexity, data availability, and expertise, GenAI models could take around 5–7 months to develop[[5]](#footnote-5).

GenAI Infra solutions include tools and services for developers to speed up model development by automating time-consuming tasks. These solutions typically feature AI/GenAI-powered features, such as coding co-pilots for automating code writing, semantic search to quickly locate data for model fine-tuning, and performance tracing and observability solutions for automated model debugging and monitoring.

In addition, AI hardware providers are always looking to up their game to speed up model development. NVIDIA, a leader in the GenAI hardware space, claimed that its [EOS supercomputer](https://sp-edge.com/updates/23666) can train a 175-billion-parameter GPT-3 model on 1 billion tokens in under four minutes[[6]](#footnote-6).

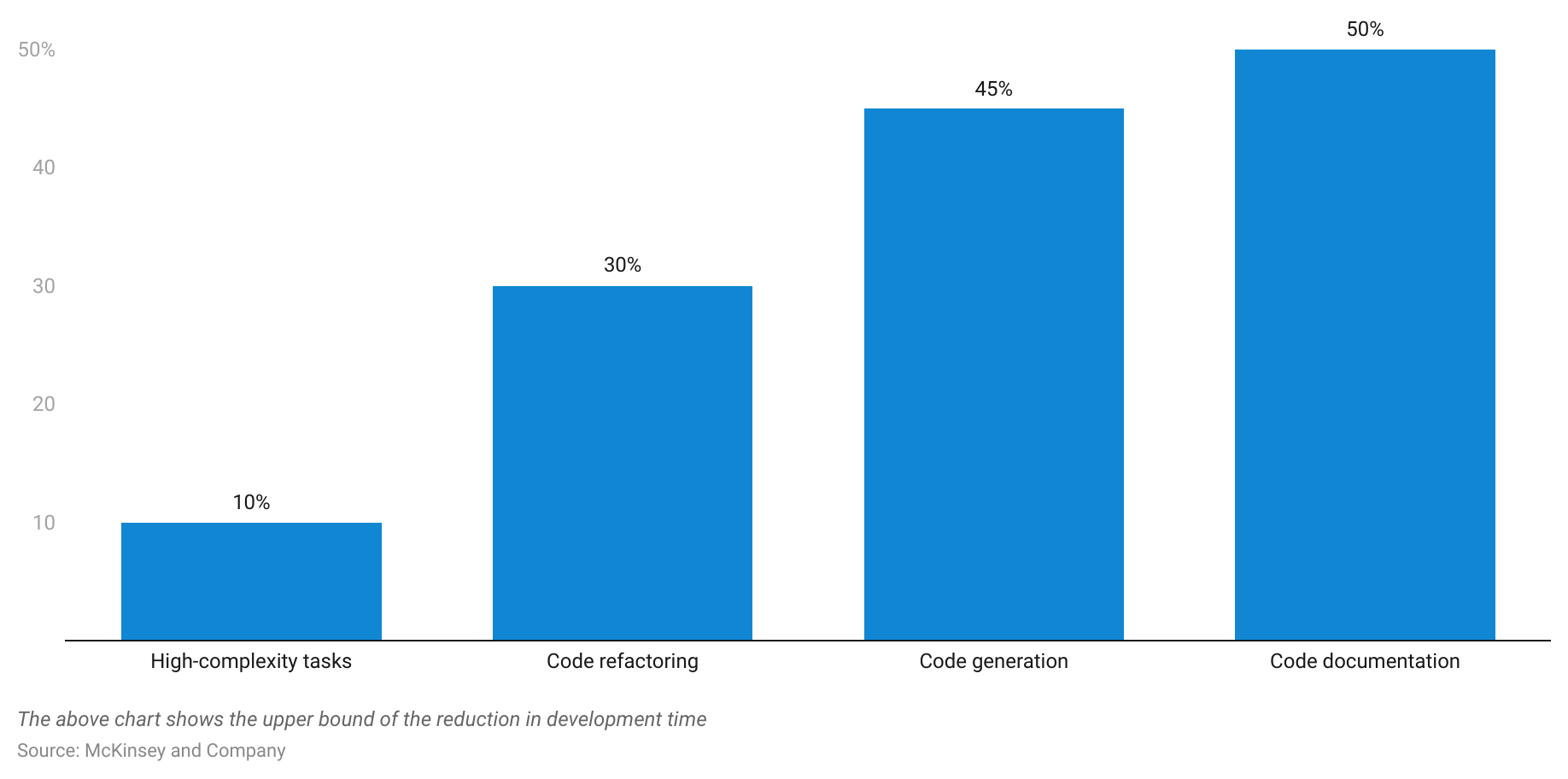
**Cost and efficiency advantages of using GenAI Infra solutions**

|  |  |  |  |
| --- | --- | --- | --- |
| **Firm** | **Segment** | **Benefits claimed** | **Source** |
| [Pinecone](https://sp-edge.com/companies/1248787) | Data storage and retrieval | Users of Pinecone serverless architecture **reduced their costs by 60%** | [Press release](https://www.pinecone.io/blog/serverless-generally-available/) |
| [Seldon](https://sp-edge.com/companies/415200) | Model monitoring | Helped reduce its model monitoring time from days to hours  Users of Seldon Deploy report **productivity gains of up to 82%**  Covea Insurance claims its use of Seldon Deploy to serve models resulted in an **11x ROI** | [Company case study](https://www.seldon.io/how-covea-made-11x-roi-detecting-fraudulent-insurance-policies) |
| [Snorkel AI](https://sp-edge.com/companies/1079283) | Model development and training | Using its KYC documentation resulted in **~10,000 saved hours for investment managers** | [Company Website](https://www.datanami.com/this-just-in/snorkel-ai-announces-ga-of-its-data-centric-ai-platform-snorkel-flow/) |
| [Portkey.ai](https://sp-edge.com/companies/3004150) | Integrated LLMOps solutions | Its full-stack platform can speed up application **development and launches by 30%** | [Company website](https://portkey.ai/) |
| [FlexLogix](https://sp-edge.com/companies/271401) | Hardware infrastructure | Its eFPGA results in a **5–10x reduction in cost** and **power** of the FPGA while increasing computer density | [Press release](https://flex-logix.com/news/flex-logix-adoption-expands-into-datacenters/#:~:text=Flex%20Logix%20eFPGA%20enables%20volume,data%20centers%2C%20microcontrollers%20and%20others.) |
| [d-matrix](https://sp-edge.com/companies/1453946) | Hardware infrastructure | Provides a **10x increase in interactive speed** and a **3x increase in cost performance** | [d-matrix](https://www.d-matrix.ai/product/) |
| [Groq](https://sp-edge.com/companies/477143) | Hardware infrastructure | Set a new inference performance record with its language processing unit (LPU), achieving more than 736 tokens per second per user on Llama-3.1 8B, offering **faster delivery and lower cost** for model development and training | [Press Release](https://groq.com/news_press/groq-supercharges-fast-ai-inference-for-meta-llama-3-1/) |
| [Datatron](https://datatron.com/) | Model monitoring, Model deployment and integration | Deploys models securely and at scale in **90% less time and cost** compared to home-grown solutions | [Company website](https://datatron.com/) |
| [Giga ML](https://sp-edge.com/companies/3164885) | Model deployment and integration | Its innovative approach leads to a **70% reduction in costs** compared to the GPT-4 API and a **300% increase in output delivery speed** | [Company press release](https://www.usemano.com/post/mano-x-gigaml) |
| [Datatron](https://datatron.com/) | Model monitoring, Model deployment and integration | Users have experienced a **10x faster model development** and a **2x reduction** **in costs.** | [Company case study](https://datatron.com/whitepaper-download/success-case-dominos/) |
| [Robust Intelligence](https://sp-edge.com/companies/835022) | AI security and governance | Users of its AI risk intelligence platform experience a **3x decrease in time**-**to-market** for GenAI applications | [Company website](https://www.robustintelligence.com/) |

### Source: Compiled by SPEEDA Edge based on company websites

GenAI Infra solutions are a win for developers too. An internal study done by [McKinsey’s](https://futureskillsacademy.com/blog/generative-ai-for-developers/) for developers shows that GenAI-powered tools can reduce the time required for documentation of code functionality for maintenance by 50%. The [same study](https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/unleashing-developer-productivity-with-generative-ai) also showed that efficiency fosters happier employees; 50% strongly agreed that they felt happier when working with GenAI tools that automate repetitive, mundane tasks and establish more efficient workflows.

**Developers experience a significant productivity boost with GenAI tools**



**4. Popularity of domain-specific GenAI models drives demand for specialized knowledge**

As the GenAI space matures, industry commentators envision the current market landscape for LLMs evolving into [smaller, more focused models](https://techcrunch.com/2023/04/01/generative-ai-focused-language-models/) that are trained on a company's own data to meet its specific requirements. This is reflected by the growing interest in “domain-specific” GenAI models, which has seen the development of numerous models across a range of industries, such as [Bloomberg GPT](https://www.bloomberg.com/company/press/bloomberggpt-50-billion-parameter-llm-tuned-finance/) (financial services), [Hippocratic AI](https://sp-edge.com/companies/2894051) (healthcare), CoCounsel (legal) that leverage additional industry-specified data and use customized parameters.

Developing such models requires specialized domain knowledge and access to specific training datasets. By leveraging its in-house domain expertise and access to industry-specific datasets, GenAI Infra firms can speed up the development process by assisting developers through consultancy services, data access, and outsourced development services, thus bridging a firm’s knowledge and data gaps. For example, [InData Labs](https://sp-edge.com/companies/342254) claims to have experience building 150+ purpose-built AI solutions across a wide range of industries, while [AI21 Labs](https://sp-edge.com/companies/600017) has developed task-specific APIs leveraging its Jurassic series of LLMs, which can be used for specific model applications.

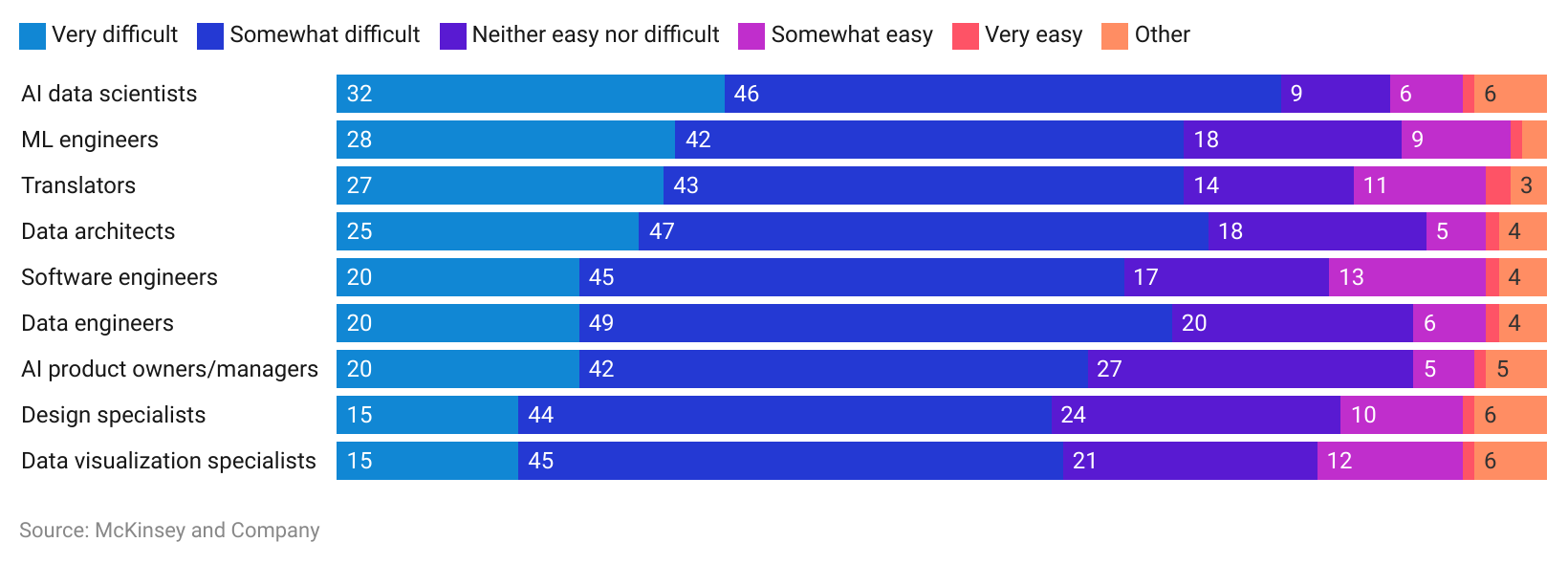
**What are the risks to growth?**

1. **AI talent shortage: A scarce and costly resource**

Numerous firms have shown interest in integrating GenAI into their business activities though there is a disparity in their willingness to pay for it. According to a [Gartner survey](https://www.gartner.com/en/newsroom/press-releases/2024-05-07-gartner-survey-finds-generative-ai-is-now-the-most-frequently-deployed-ai-solution-in-organizations), just 29% of 644 respondents from organizations in the US, Germany, and the UK reported that they have invested for and deployed GenAI models. A key reason for this lack of budget prioritization is the AI talent crunch—IBM’s AI Adoption Index 2022 identified a lack of AI skills as the biggest barrier to AI adoption. However, [Gartner](https://www.gartner.com/en/newsroom/press-releases/2024-10-03-gartner-says-generative-ai-will-require-80-percent-of-engineering-workforce-to-upskill-through-2027) identified that through 2027, GenAI will create new roles in software engineering and operations, requiring 80% of the engineering workforce to improve skills. Further, [global AI adoption](https://explodingtopics.com/blog/ai-statistics) by organizations is set to expand at a CAGR of 36.6% from 2024 to 2030.

A survey by [SAS](https://www.prnewswire.com/news-releases/ai-skills-crisis-may-lead-to-wasted-investments-and-stifled-innovation-research-shows-301630769.html) reflected that63% of respondents’ largest skills shortages are in AI and machine learning. It is, therefore, no surprise that AI talent and relevant roles, such as AI data scientists, remain in short supply and pose a significant challenge for organizations, particularly within SMBs that lack the brand recognition of large enterprises.

**AI data scientists are among the most difficult roles to fill**

****

This shortage has also led to a wage spiral in AI-related roles. The US Bureau of Labor Statistics reported that the median pay for computer and information research scientists in 2023 was USD 145,080 and is expected to grow much faster on average (26% vs. 3% overall average) in the next decade[[7]](#footnote-7).

Although many firms are opting to train and upskill their in-house teams, the AI talent shortage and wage inflation remain factors that may lead firms to think twice about AI model development.

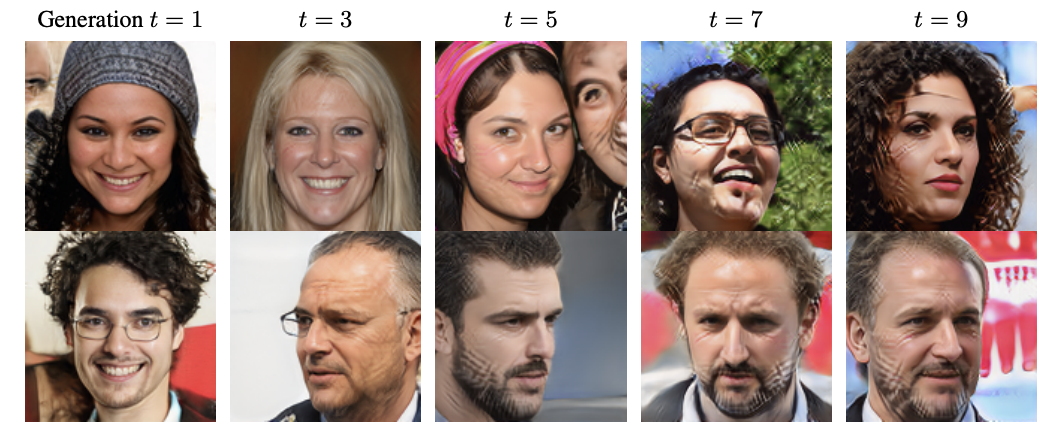
1. **Ensuing data deserts and cost escalations from data gathering challenges**

AI models are only as good as the data used to train its models. This is especially true when developing GenAI models, as a high quantum of diverse and relevant data is essential to ensure their effectiveness and reliability while avoiding bias. However, the availability of high-quality data is increasingly threatened by a host of personal privacy laws (GDPR, CCPA, HIPAA) and initiatives by firms such as Apple’s consumer privacy features and Google phasing out third-party cookies by the end of 2024.

Parallely, gathering quality data is anticipated to rise in cost. Content creators are uneasy about sharing their data to train AI models, which not only restricts model developers' access to such data but makes AI deployment more expensive. These concerns were at the heart of recent strike actions like the one conducted by the Writers and Actors Guilds of America, which resulted in guardrails on the use of AI in film and television projects[[8]](#footnote-8).

While synthetic data (artificially generated data) may help address these issues to a certain degree, studies done at Standford and Rice Universities have shown that without enough fresh, real data, generative models would experience a progressive decrease in the quality and diversity of their data—a condition known as MAD (model autophagy disorder)[[9]](#footnote-9).

**Impact of a lack of real training data**



Source: Rice University

This lack of quality data has been identified as a key inhibitor of GenAI development and restricts firms from taking full advantage of the benefits of GenAI.[[10]](#footnote-10)

1. **Stringent AI regulations add to AI model development cost**

Given the concerns over copyright violations, the use of GenAI for unethical purposes, and data and output biases, a number of regulations are on the horizon. The upcoming [EU AI Act](https://sp-edge.com/updates/23832) requires disclosure of AI-generated content and GenAI systems to be subject to review before commercial release and US President Joe Biden’s [executive order regulating AI](https://sp-edge.com/updates/23271) plans to implement standards for watermarking AI-generated content for copyright protection. *See our Insight, “*[*The evolving rulebook of AI governance,*](https://sp-edge.com/insights/21255)*” for more details on the regulatory landscape.*

According to a study by [Harvard Business Review](https://hbr.org/2021/09/ai-regulation-is-coming), such regulations could reduce the potential for scale advantages from AI, while the additional layers of algorithms and other process developments needed to remain in compliance with emerging regulations would raise development costs[[11]](#footnote-11). For instance, numerous studies have proven that watermarking AI-generated content is complicated to implement and can be sidestepped. This has also necessitated the development of more advanced (and expensive) solutions, such as [SynthID](https://sp-edge.com/updates/21420), an invisible watermarking tool developed by DeepMind in partnership with Google Cloud, which uses patented steganographic techniques that it claims aren’t affected by attempts to manipulate watermarks, such as resizing or cropping, among others.[[12]](#footnote-12)

**Toward a safe and secure GenAI future**

AI safety and AI governance have taken center-stage in dominating the conversations on GenAI. Several nations and organizations, such as the [UN](https://sp-edge.com/updates/23177), [G7](https://sp-edge.com/updates/23315), and the [Bletchley Declaration](https://sp-edge.com/updates/23426)—which brought together 28 nations, including China and the US—have unanimously acknowledged the importance of establishing global standards to address potential AI risks.

Global firms have also begun to recognize the need of the hour, with firms like Alphabet and Microsoft developing formal AI governance policies while collaborating with others like OpenAI and Anthropic to form the [Frontier Model Forum](https://sp-edge.com/updates/23096). This resulted in a USD 10 million AI safety fund to advance AI governance[[13]](#footnote-13).

Although this fast-paced, ever-evolving regulatory environment is a challenge to keep up with, it has bred a new category of GenAI Infra solution providers. Many of these firms offer platforms and tools for securing AI models from external threats, implementing governance policies, ensuring compliance with regulations, and detecting problematic content and model deterioration.

**Examples of emerging AI safety and governance solutions**

|  |  |  |
| --- | --- | --- |
| **Product feature** | **Company** | **Product description** |
| Model visibility | Credo AI | Offers an integration hub that links with third-party AI Ops and business tools to improve oversight of AI policies. It gathers metadata on safety and security metrics, uploads models for compliance checks, and integrates datasets for governance (October 2024) |
| Model security | Patronus AI | Launched a self-serve API platform to detect and prevent AI failures in real time. The platform operates on a pay-as-you-go model, starting at USD 10 per 1,000 API calls for smaller evaluators and USD 20 per 1,000 API calls for larger ones (October 2024) |
| Model security | Patronus AI | Launched Lynx, an open-source hallucination detection model designed to address hallucinations in LLMs. Patronus AI claims its Lynx model outperforms GPT-4 by 8.3% in detecting medical inaccuracies (July 2024) |
| Model security | [Galileo](https://sp-edge.com/companies/1762847) | Launched the Galileo Luna suite, which includes evaluation foundation models (EFMs) designed to assess the performance of LLMs such as GPT-4 and Gemini Pro. These models detect issues like hallucinations, data leaks, and malicious prompts. Benchmark tests showed they are up to 20% more accurate than existing models and offer enterprise-scale evaluations that are 97% cheaper, 11x faster, and 18% more accurate than GPT-3.5 (June 2024) |
| Model security | RagaAI | Unveiled RagaAI LLM Hub to evaluate and establish guardrails for LLMs. The RagaAI LLM Hub includes over 100 curated metrics, assisting users in effectively assessing and rating LLMs. The platform's key features include evaluating relevance and understanding, content quality, hallucination, safety and bias, and vulnerability scanning (March 2024) |
| AI content detection | [Sensity AI](https://sp-edge.com/companies/706080) | Offers deepfake detection tools generated by generative adversarial networks (GANs). The company claims that it can detect AI-generated images from models like DALL.E, Stable Diffusion, and Midjourney, as well as AI-generated voices |
| Model security | [Arthur AI](https://sp-edge.com/companies/869868) | Launched Arthur Shield, reportedly the “first firewall” for LLMs. Arthur Shield enables enterprises to deploy LLM applications safely and protects against vulnerabilities like data leakage, toxic or problematic language generation, hallucinations, malicious prompts, and prompt injections |
| Model security | [Datatron](https://sp-edge.com/companies/381299) | Offers an AI model monitoring and AI governance solution focused on monitoring deployed models for drift (data drift and concept drift), bias, performance, and anomalies, enabling developers to identify issues and correct them on time |
| Model security | [Robust Intelligence](https://sp-edge.com/companies/835022) | Offers an AI firewall that inspects every input to an LLM and automatically blocks malicious input that can poison the data. It also validates model output in real-time to ensure they are absent of sensitive information, hallucinations, and other harmful content |
| Model security | [Copyleaks](https://sp-edge.com/companies/285007) | Its GenAI governance platform provides military-grade security and surfaces potential risks by providing comprehensive data on keyword searches and user conversation history with AI generators |
| Model compliance and reporting | [Enzai](https://sp-edge.com/companies/3186980) | Enables users to choose relevant governance policies from its policy center or create custom policies and measure GenAI models against it to assess compliance |
| Model compliance and reporting | [Saidot](https://sp-edge.com/companies/800386) | The Saidot library includes a collection of up-to-date AI policy templates, with the option to publish compliance reports and procure compliance reports from third parties |
| Model security | [Saidot](https://sp-edge.com/companies/800386) | Provides up-to-date risk assessment metrics for foundation models to make informed decisions |
| Model compliance and Reporting | [Fairly AI](https://sp-edge.com/companies/1097344) | Its Fairly control center automatically maps AI regulations and policies to monitor AI models for compliance. Also provides Auditors with access to real-time reports |
| AI content detection | [Copyleaks](https://sp-edge.com/companies/285007) | Claims to be the most accurate AI detector for ensuring cyber compliance for preventing copyright infringement. Claims to have a 99.1% detection accuracy and covers GPT-4 and Bard |
| Model compliance and reporting | [Cranium](https://sp-edge.com/companies/2791878) | Cranium’s “AI cards” can be used by organizations to demonstrate AI/ML compliance and to showcase the security and trustworthiness of one AI pipeline, one asset, or an entire organization’s AI systems |
| Model compliance and reporting | [Holistic AI](https://sp-edge.com/companies/2199328) | Offers an independent bias auditing solution to be in compliance with laws such as the NYC local Law 144, EEOC guidelines, and other emerging legislation |
| Model compliance and reporting | [Scale AI](https://sp-edge.com/companies/408174) | Launched the AI Safety, Evaluations, and Analysis Lab (SEAL) to help enterprises and governments comply with forthcoming AI standards and regulations |

### Source: Compiled by SPEEDA Edge based on company websites

While the proliferation of AI raises valid concerns, the right guardrails and the aid of GenAI Infra solutions can promote ethical and safe AI development, motivating firms, developers, and customers to shape a safe and secure AI future.

***Last updated: December 2024***

1. https://www.linkedin.com/pulse/dangers-force-fitting-generative-ai-your-business-syed-q-ahmed [↑](#footnote-ref-1)
2. https://analyticsindiamag.com/16-largest-closed-source-llms-you-must-know-about/ [↑](#footnote-ref-2)
3. https://about.fb.com/news/2023/11/meta-partners-with-hugging-face-scaleway-to-support-open-source/ [↑](#footnote-ref-3)
4. https://zapier.com/blog/stable-diffusion-vs-dalle/ [↑](#footnote-ref-4)
5. https://www.sparxitsolutions.com/blog/everything-you-need-to-know-about-generative-ai-development/ [↑](#footnote-ref-5)
6. https://www.engadget.com/nvidias-eos-supercomputer-just-broke-its-own-ai-training-benchmark-record-170042546.html?guccounter=1 [↑](#footnote-ref-6)
7. https://www.bls.gov/ooh/computer-and-information-technology/computer-and-information-research-scientists.htm [↑](#footnote-ref-7)
8. https://www.theguardian.com/culture/2023/oct/01/hollywood-writers-strike-artificial-intelligence [↑](#footnote-ref-8)
9. https://arxiv.org/pdf/2307.01850.pdf [↑](#footnote-ref-9)
10. https://www.forbes.com/sites/joemckendrick/2023/01/22/a-data-gap-continues-to-inhibit-artificial-intelligence/?sh=22b1546c5a84 [↑](#footnote-ref-10)
11. https://hbr.org/2021/09/ai-regulation-is-coming [↑](#footnote-ref-11)
12. https://www.imatag.com/ [↑](#footnote-ref-12)
13. https://techcrunch.com/2023/07/26/openai-google-microsoft-and-anthropic-form-body-to-oversee-safe-frontier-ai-development/ [↑](#footnote-ref-13)