Brought to you by Informa Tech

To keep up with GenAI innovation, enterprise buyers should think operationally

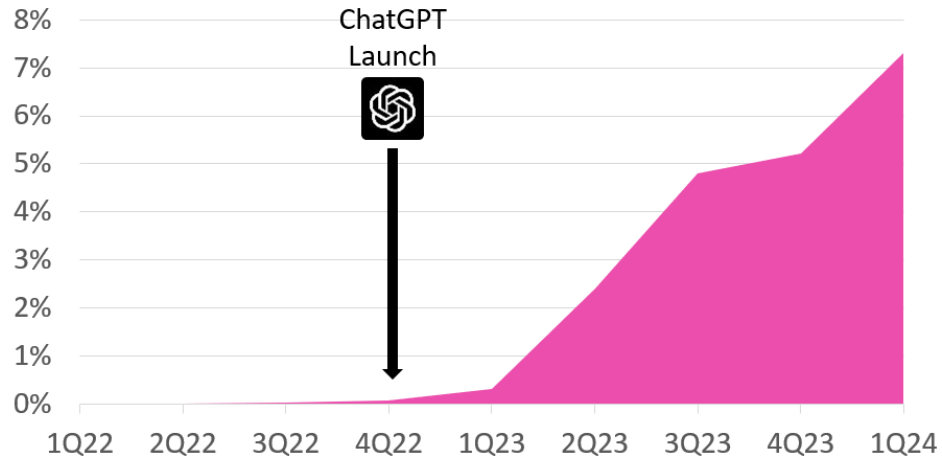
Publication date:
June 2024
Author(s):
Bradley Shimmin, Chief Analyst, AI and Data Analytics

The state of GenAI adoption

The era of generative AI (GenAI) has already changed the way consumers interact with their favorite brands, upended the way companies think about corporate data, and even changed the concept of work itself. Never before in the technology market has a single idea so quickly captured the attention of buyers and builders alike – Omdia research estimating an overall market valuation going from zero in late 2022 to $74 billion dollars by 2028.

Indeed the industry – from startup to hyperscaler – is scrambling to feed their talent engine too and deliver on the technological promise. According to Omdia ongoing research into the AI job market, GenAI jobs have grown by more than 743% over the past 18 months (see Figure 1).

Figure 1: The rising percentage of AI jobs focused on GenAI technologies



Source: Omdia AI Skills Tracker (% of job openings focused on GenAI technology)

With major investment and talent comes rapid innovation for the current state of the art GenAI offerings – with concepts from multi-modality, to nearly infinite context windows and semantic search tools now give GenAI large language models (LLMs) instant access to reams of contextual and timely information. And with rapidly maturing development frameworks and application programming interfaces (APIs), LLMs now live within everyday consumer applications and are quickly gaining autonomy, taking direct action on a user's behalf.

Perhaps most importantly, LLMs themselves are getting smaller thanks to improved training methodologies, more streamlined model architectures, and innovative deployment practices. These smaller models open new opportunities for companies to deploy GenAI at scale at either edge or on-device without compromising performance or quality. Both independent and major model manufacturers including Meta, Microsoft, Google, IBM, are doubling down on these smaller models, building highly adaptable LLMs capable of fitting into less than 8GB of virtual RAM (VRAM). Unsurprisingly, nearly 75% of GenAI job openings tracked by Omdia call for edge and internet of things (IoT) skills. The most sought-after LLM skills also revolve around this smaller footprint, led by open source models Microsoft Phi and Meta Llama.

Even amongst the cautious world of enterprise IT buyers, GenAI has almost overnight moved beyond experimentation and proof-of-concept (PoC) endeavors, to become a fully supported, mission-critical technology. According to Omdia data, nearly 48% of enterprises have already operationalized GenAI across numerous tasks ranging from simple text and image generation to complicated, agent-based workflows across several data modalities and industries.

The GenAI easy button

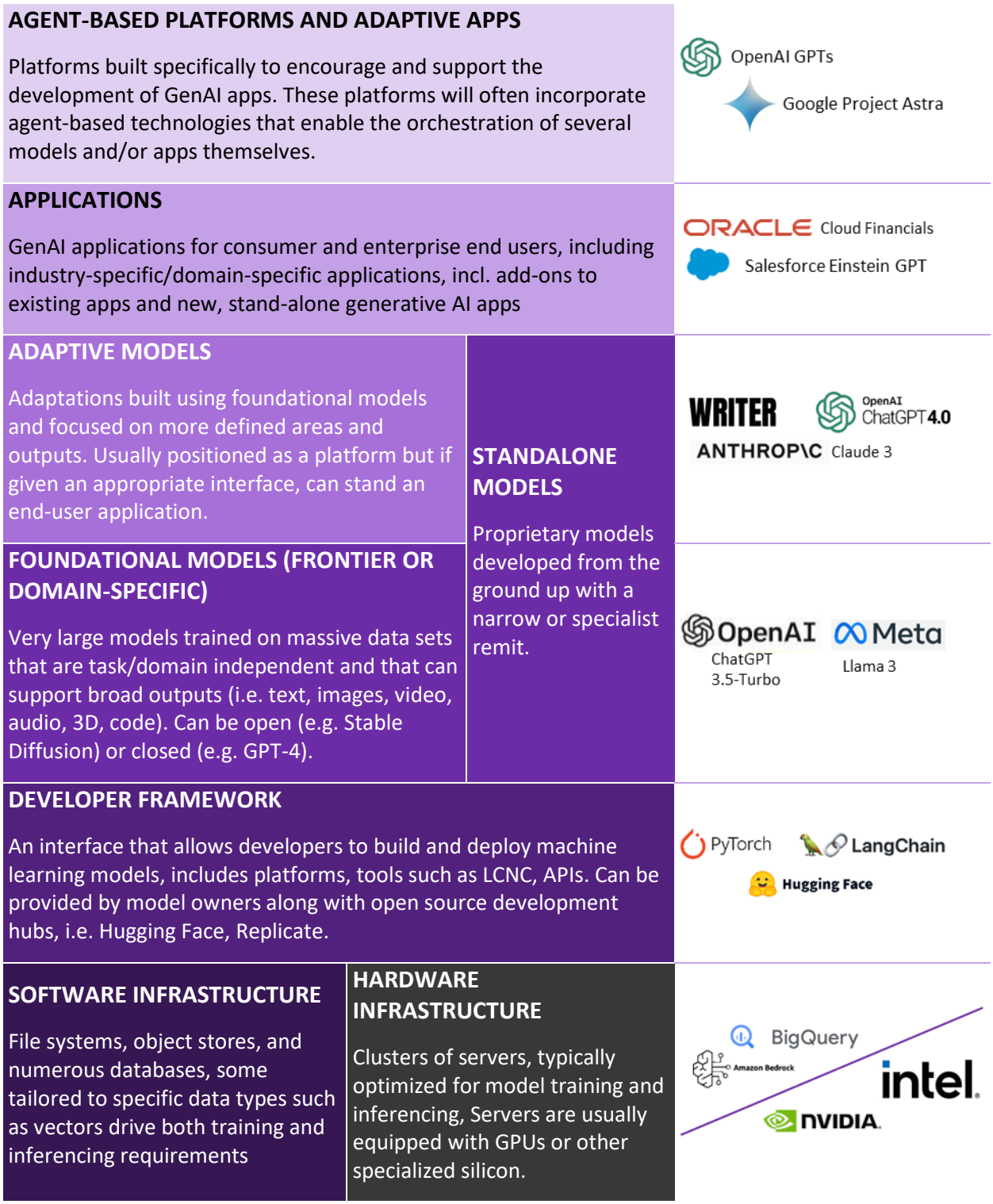
Three key factors stand out in driving this rapid development:

- The transformer model architecture, which underpins most LLMs, is a chameleon. **A single model and supporting architecture can tackle numerous use cases** ranging from customer support automation to predictive maintenance. For example, a search on AI community platform, HuggingFace, reveals more than 300 distinct LLMs have been built specifically to handle image to text tasks. This is why LLMs are often referred to as foundational models (FMs). Once trained, FMs are only partly finished, requiring further tuning or training to follow instructions, engage in effective turn-based conversation, and generate useful information based on domain-specific data.

- **Second, this adaptive quality of LLMs has shifted the necessary skills for enterprise AI away from traditional data science techniques to instead emphasize composability.** Rather than creating AI models and supportive training data from scratch, practitioners need only wire together the right model and the right supportive data using techniques such as retrieval augmented generation (RAG).

- Finally, the composable nature of GenAI coupled with its early reliance upon cloud-hosted API services for access to large-scale frontier models like Google Gemini 1.5 Pro and Anthropic Claude 3 Opus, has **enabled many companies to rapidly and affordably build simple GenAI solutions.** The equally quick rise of supportive tools such as development frameworks from model makers and third party developers (e.g., LangChain) along with RAG-supportive vector databases (e.g., Pinecone, Chroma, and Weaviate), has greatly enhanced the overall GenAI software stack (see Figure 2).

Figure 2: The modern GenAI model software stack

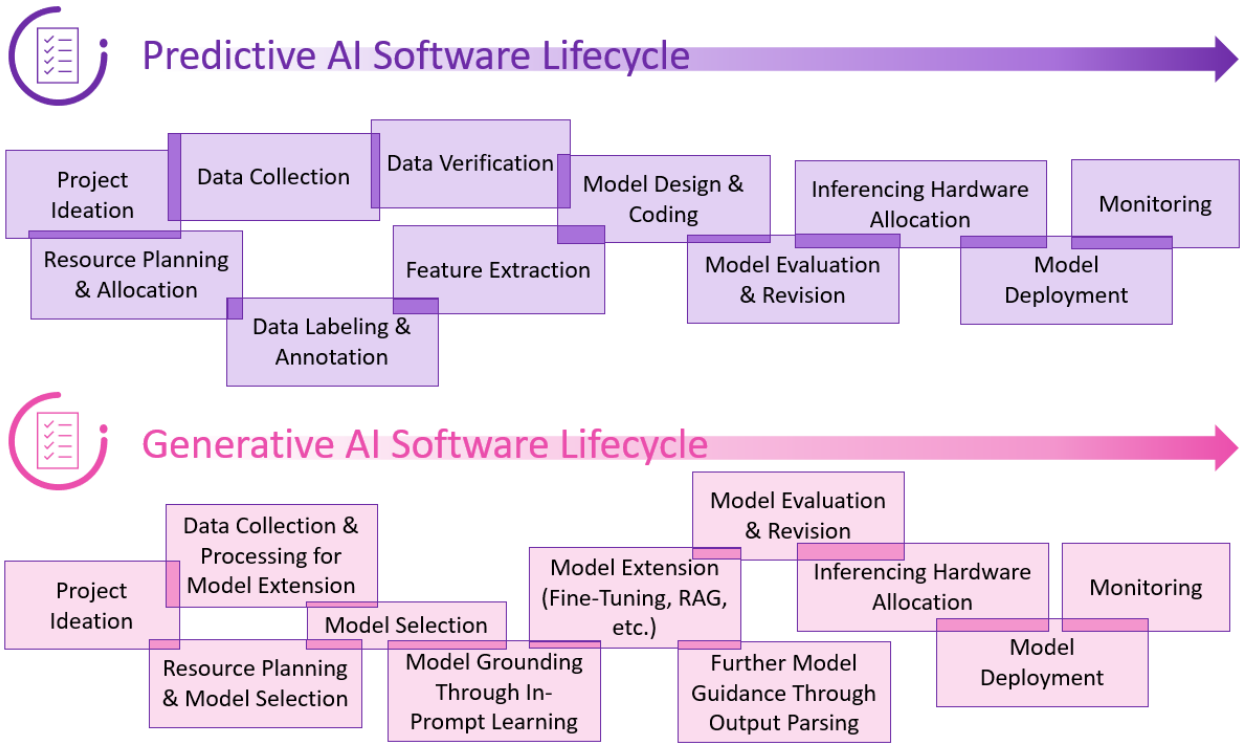


Source: Omdia

Here there be monsters

The ease of access to models and tooling along with the low-code nature of GenAI itself has unfortunately created a false sense of security and confidence among early adopters. While GenAI offers transformative potential, enterprises must still learn to mitigate a litany of both old and new risks around privacy, security, reliability, accountability, and more. Why old and new risks? At its core, a GenAI LLM is nothing more than a very large deep learning neural network that features a tremendous amount of training data -- old technology made new through access to more cost effective hardware and more inventive modeling techniques.

Figure 3: The many commonalities between predictive and generative AI software lifecycles



Source: Omdia

Adopters of GenAI hoping to build an enterprise-grade solution must first address the many operational complexities common to both traditional (predictive) AI and GenAI solutions (see Figure 3). Only then can they effectively respond to the many risks unique to GenAI that arise from the transformer model architecture itself. Those risks include

- **Repeatability** -- the stochastic nature of LLMs hampers their ability to deliver consistent results over time
- **Reliability** -- factors such as hidden bias can lead to unpredictable model behavior
- **Factuality** -- In the absence of generalizable training data and supportive contextual information, models are prone to fabricate inaccurate results

- **Transparency** -- The very nature of DL architectures means that companies can know the source but never the "reason" behind a given model decision

- **Accountability** -- The many dependencies intertwined across data, architecture, and partner decisions adds complexity to the already complex challenge of controlling exposure to risk

How do these challenges map to actual practitioner concerns? A survey of more than 350 enterprise GenAI adopters shows that data-related worries over privacy and security are at the forefront. Such worries generally center on preventing data leakage, identifying bias hidden in training data, and ensuring that models don't fabricate incorrect information.

Future-proofing innovation

Keeping up with this cacophony of demands can seem an impossible task. To illustrate, in just 30 minutes' time, model creators, researchers, and hobbyists uploaded more than 120 new and/or derivative LLMs to the popular Hugging Face community site. Some are nothing more than volatile experiments, whilst others could pave the way for future innovation.

How can enterprise adopters keep up with such innovations, especially when they might help resolve GenAI challenges? How can they move forward quickly without succumbing to technical paralysis or worse incurring a high degree of technical debt necessary to integrate new technologies into production over time?

The key rests in abstraction through operationalization, in deploying an AI platform capable of both unifying and modularizing the entire GenAI stack from device to cloud. Fortunately, there are many LLM operational (LLMOps) platforms currently entering the market that bring together the following critical elements:

- An open yet opinionated repository of fully vetted GenAI models that are often optimized for specific tasks.

- A rich suite of frameworks and tools targeting the process of fine-tuning, optimizing, and evaluating LLMs within the context of corporate data.

- A model inference engine that lets developers write once and deploy across multiple processors, accelerators, and environments, all with a high degree of efficiency.

Companies so equipped can move rapidly without unnecessary risk, taking full advantage of new LLM model innovations while also maximizing the ever evolving landscape of underlying hardware assets.

This kind of unification will prove critical for companies looking to build more complicated solutions. Take for example, a GenAI-infused electric vehicle charging station such as that announced by DFI/Intel. A single solution may require the use of multiple models, a RAG pipeline, and an LLM-based orchestration layer to coordinate across several interdependent tasks including the charging unit itself alongside digital signage, a payment transaction system, and an interactive kiosk.

Each of those app workflows could demand different security/privacy, performance, and governance requirements, running together across several unique edge and cloud hardware footprints. Managing such complexity without locking everything down within a monolithic, brittle codebase and hardware stack takes a high degree of operationalized abstraction, which can only be found within an LLMOps platform.

In summary

Highly adaptive GenAI foundation models have powered explosive growth in just 18 months - yet the speed of innovation and ease of access to GenAI technologies have created a false sense of security among early enterprise adopters. Risks abound and the rapid rate of GenAI innovation can leave many practitioners either open to unseen risks or frozen in inaction.

To overcome the hype and rapid innovation cycles, enterprises must therefore operationalize GenAI through an LLMOps platform capable of creating a strategic layer of abstraction that still enables a more tactical response to market changes through access to curated models, tuning tools, and optimized deployment across diverse hardware. As models become progressively more capable and less dependent upon data center hardware, enterprise practitioners will find new, as yet unimagined opportunities to build, refine, extend, and deploy GenAI models not just in the cloud but anywhere where humans and computers come together.

Recommended next steps:

- **Start with a composable and modular mindset**, prioritizing the adoption of tools and frameworks that can easily incorporate model changes and readily move across use cases. Standards are emerging, but practitioners will continue to rely upon tools that can fill in the gaps.
- **Prioritize partners that combine deep GenAI expertise with intelligent operational tooling.** And invest in an LLMOps platform enabling model curation, optimization workflows, and unified cross-hardware deployment.
- **Never lose sight of the fundamentals:** the stringent application of privacy, security, and governance protocols by enacting regular audits, bias monitoring, etc. And don't forget to establish clear lines of accountability that cover the entire GenAI stack.
- **Above all, try to think of GenAI not as a single, monolithic LLM chatbot running in the cloud**, but rather as a set of capabilities made up of smaller LLMs, each tuned to meet specific needs (planning, chatting, coding, etc.), all working together to meet users where they are, on-device, at the edge, or in the cloud. This is the spirit of GenAI.

Author(s)

Bradley Shimmin, Chief Analyst, AI and Data Analytics

Bradley.Shimmin@Omdia.com

This piece of research was commissioned by Intel Corporation.

Request external citation and usage of Omdia research and data via citations@omdia.com.

We hope that this analysis will help you make informed and imaginative business decisions. If you have further requirements, Omdia's consulting team may be able to help you. For more information about Omdia's consulting capabilities, please contact us directly at consulting@omdia.com.

omdia.com

customersuccess@omdia.com