fractal

# Unlock the potential of generative AI with Amazon Web Services

aws PARTNER Advanced Tier Services

- Retail Software Competency
- Financial Software Competency
- Data & Analytics Software Competency

# Contents

# Introducing generative AI

# Introducing generative AI

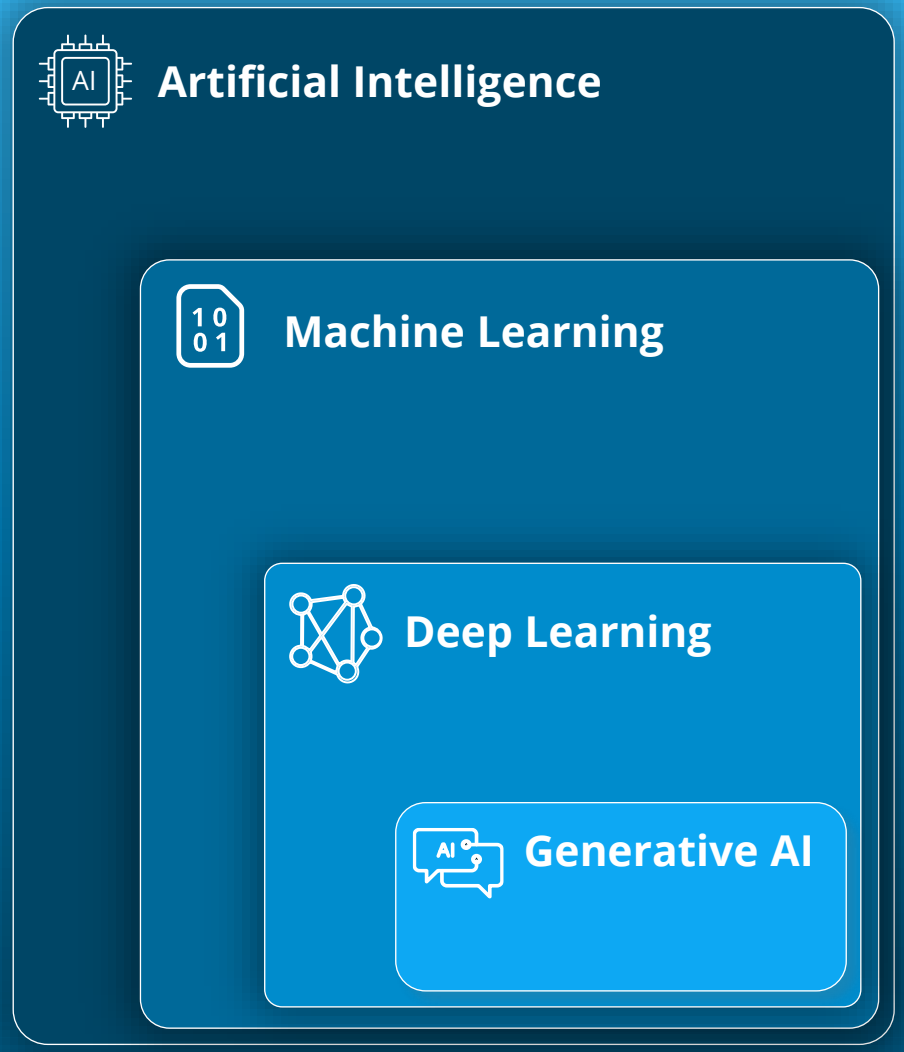
Generative AI refers to a class of artificial intelligence algorithms that focus on generating new, previously unseen data that conform to certain patterns or characteristics learned from real-world data.

It can generate original content, such as text, images, music, or videos, based on a set of inputs (“prompts”).

These algorithms are designed to learn patterns from data and then use those patterns to generate new content that is similar in style or form to the original data.

Recent advancements have opened new possibilities for using generative AI to solve complex real-life business problems, create art, and even assist in scientific research.

Those Deep Neural Networks (DNNs) based models use the latest “Transformer” (the “T” in ChatGPT) architectures to achieve those results.



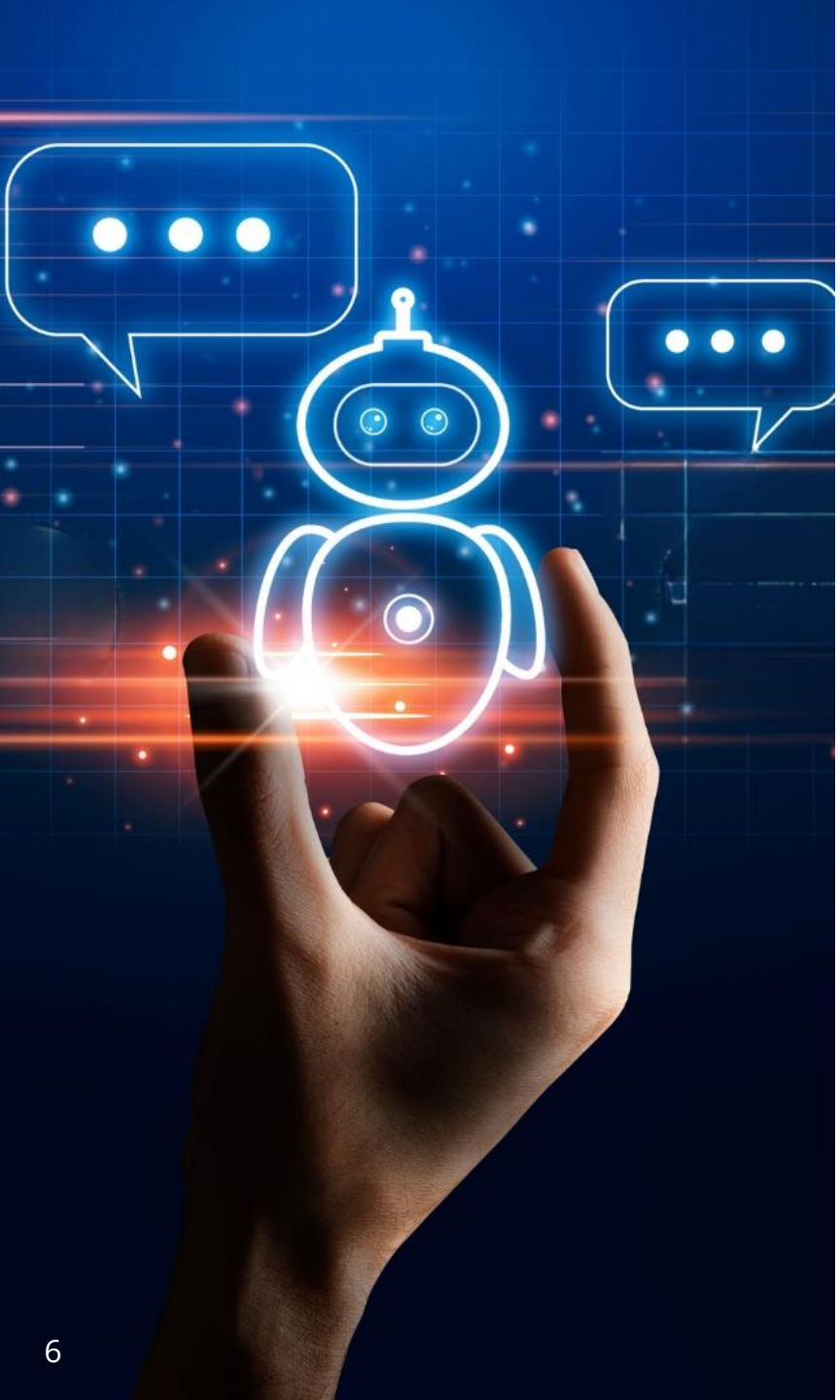

# History of generative AI

**Artificial Intelligence**
1956
The field of computer science that seeks to create intelligent machines that can replicate or exceed human intelligence

**Machine Learning**
1997
Subset of AI that enables machines to learn from existing data and improve upon that data to make decisions or predictions

**Deep Learning**
2017
A machine learning technique in which layers of neural networks are used to process data and make decisions

**Generative AI**
2021
Create new written, visual, and auditory content given prompts or existing data



# Key benefits of generative AI

**Increased efficiency**

Generative AI can free up humans from more low-value, repetitive, and boring activities.

It allows them to focus more on value-added activities requiring uniquely human skills, like creativity, lateral thinking, empathy, and many more.

**Improved quality**

Generative AI can enhance the quality of data or content by adding details, correcting errors, or removing noise.

For example, generative AI can improve the resolution of images, the accuracy of speech recognition, or the grammar of text.

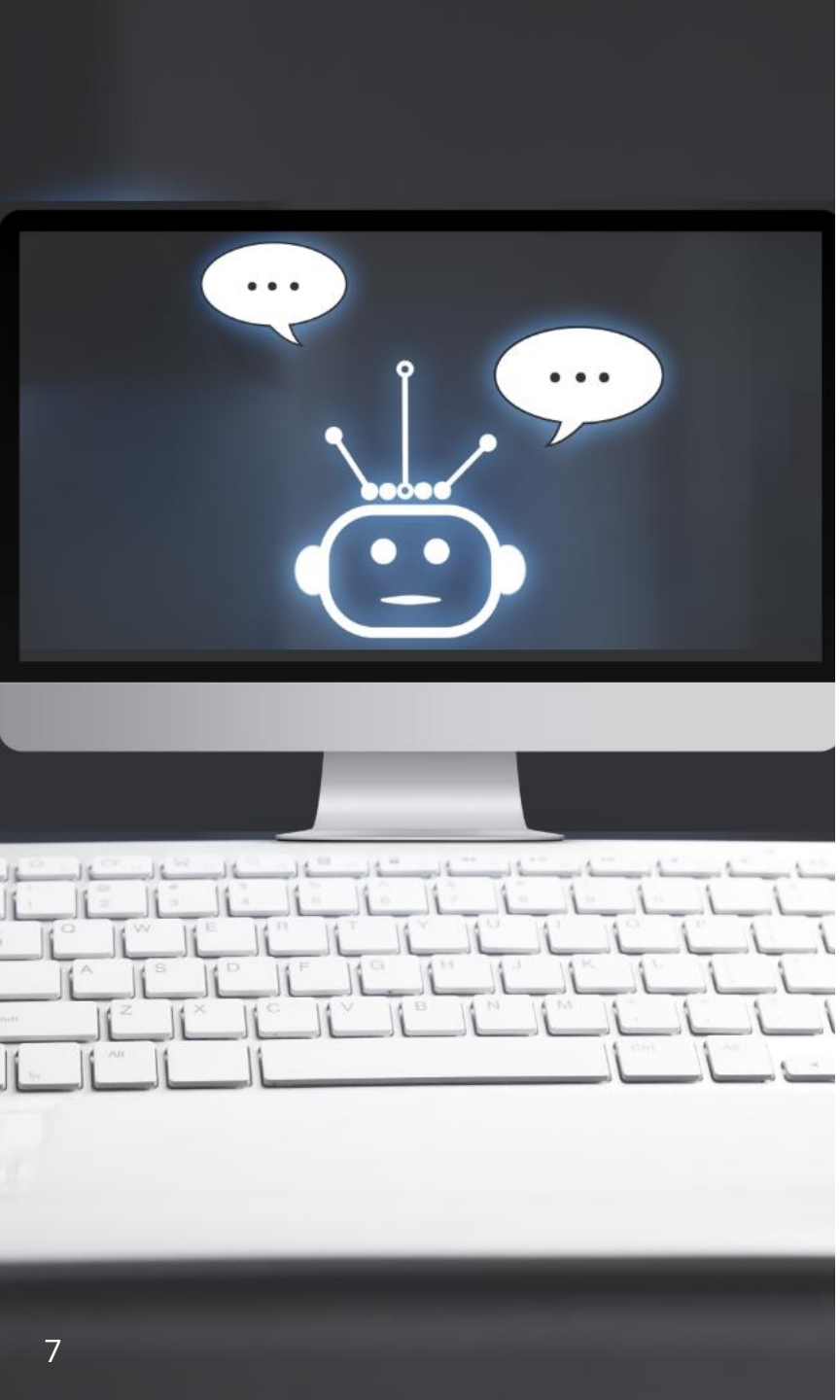
**Faster results**

Generative AI can produce data or content at a faster rate than human capabilities, especially for large-scale or complex projects.

Also, this can free humans from boring and repetitive tasks, such as writing the minutes of a meeting.

**Improved customer experience**

Generative AI can improve customer experience by providing customized content that is both relevant and engaging for a particular customer at a particular step in their buying or support journey.



# Key benefits of generative AI (Continued)

**Cost savings**

Generative AI can help reduce costs associated with scaling up business processes such as data processing or content creation.

**Improved decision-making**

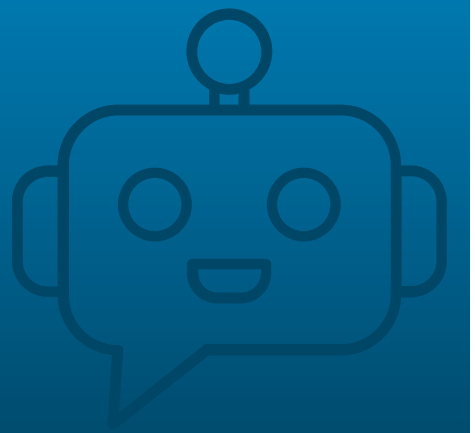
Generative AI can support decision-making by providing insights, predictions, or recommendations based on data or content analysis.

**Fostered creativity**

Generative AI can foster the creative process by generating initial concepts that can then trigger genuinely new products, entertainment, marketing, or educative ideas.

Generative AI has many applications across different domains, such as marketing, data analytics, software engineering, sales, and cross-functions.

Through its ability to create new content from various input types such as text, images, voice, and computer code, generative AI has applications across most business functions and processes.



# Generative AI applications

**Content marketing:** Create personalized and engaging content for your website and customers, such as web copies, blog posts, customer stories, and social media posts. Generative AI can also generate catchy slogans, headlines, and product descriptions.

**Data analytics:** Provide insights and predictions based on business data that can help companies make informed decisions, identify patterns and trends, and optimize business operations. Generative AI can also help generate data visualizations and reports for your data analysis.

**Software engineering:** Simplify and automate various software development tasks such as generating code from natural language descriptions or pseudocode, completing code snippets, fixing bugs, suggesting improvements, and explaining code logic.

**Cross-functional use cases:** Summarize meetings, answer questions, correct or translate text, design illustrations using text prompts, create videos from text or still images, and much more with generative AI.

**Sales:** Create personalized and persuasive sales pitches, proposals, and emails.

# The different types of generative AI models

1

## Language models

Large Language Models (LLMs) process and understand natural language in text (or through speech recognition, voice) format. These models are trained on massive amounts of text. The most popular type of generative AI models currently, LLMs can not only be used for creative, academic, and business writing, but also for translation, grammatical correction or analysis, computer code writing, and more. The most well-known LLM is **ChatGPT** from OpenAI.

2

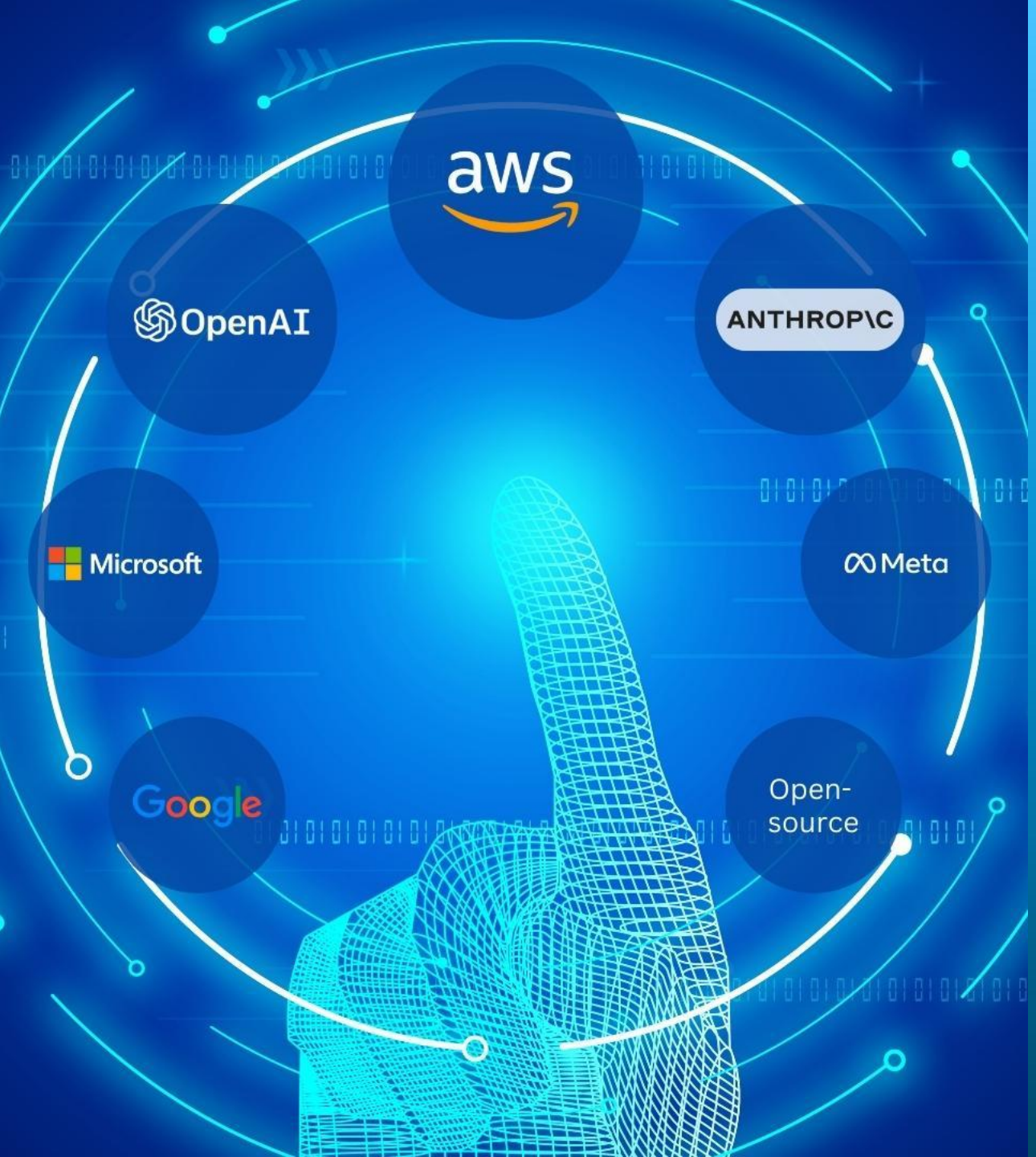
## Visual or imagery models

These models can generate realistic images or videos based on inputs such as a text description, a sketch, a pre-existing image, or a style. For instance, they can generate a photo of your dog that looks like a Kandinsky painting. Visual models can be used for creative design, 3D modeling, image editing, architectural rendering, CAD, and more. For example, **DALL·E** is a visual model that can create images from text descriptions (aka “prompts”).

3

## Voice models

These models can generate natural sounding speech or music based on input such as a text, a melody, or a voice. Voice models can be used for composing, songwriting, dubbing, speech recognition, sound editing, and more. For example, **Jukebox** is a voice model that can generate music, including rudimentary singing, as raw audio in a variety of genres and artist styles. Jukebox can also generate lyrics conditioned on a genre, an artist, or a user prompt.



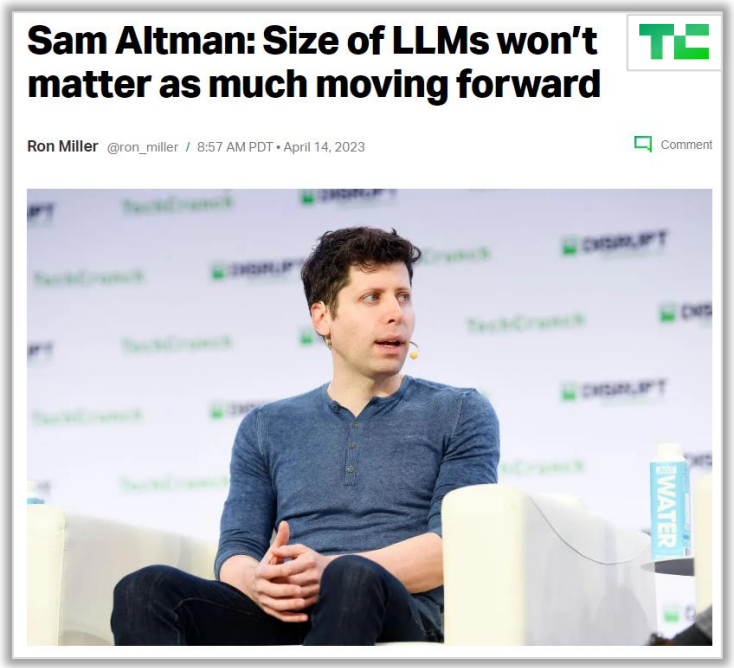

# Key Large Language Model (LLM) generative AI providers

# Key LLM platforms and models providers: summary table

| Company | Generative AI models strategy | LLM access through… |
|---|---|---|
| **Amazon** | • AWS Bedrock platform enables access to third-party LLMs such as Anthropic's Claude | • Bedrock APIs |
| **Anthropic** | • In-house model: Claude 2, Claude Instant | • Anthropic APIs<br>• Bedrock APIs<br>• Claude chat webapp |
| **OpenAI** | • In house-developed models: GPT 3.5 and 4<br>• Also offering image creation with DALL·E | • OpenAI APIs<br>• ChatGPT and ChatGPT webapp |
| **Microsoft** | • Platform approach in Azure Machine Learning<br>• Supports OpenAI GPT modes, Meta's Llama & other open-source models | • Azure OpenAI APIs, Azure ML<br>• Integrated in Bing Chat, Microsoft 365, Windows 11, and GitHub "Copilots" |
| **Google** | • Platform approach with Vertex AI<br>• In-house models with PaLM 2, LaMDA, BERT | • Vertex AI APIs<br>• Bard chat webapp, Google Workspace |
| **Meta** | • In-house model with Llama 2 (various sizes) | • Stand-alone open-source models |
| **Open-source & proprietary** | • Based on custom (native) open-source models<br>• Leveraging Llama's open-source model<br>• Proprietary internally developed and dedicated models | • Varied |

*Note: This table represents an August 2023 technology snapshot of a rapidly evolving ecosystem*

# Avoiding the halo effect





**The halo effect of ChatGPT (and its competitors)**

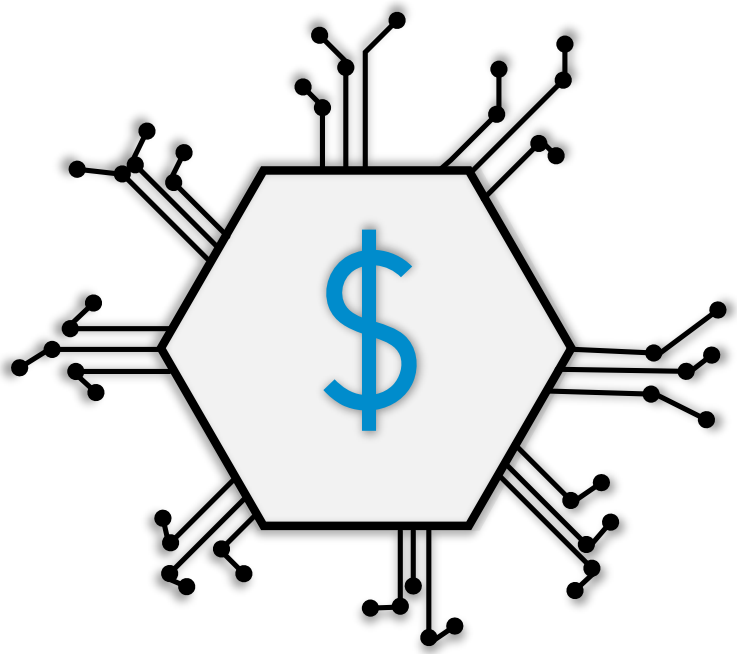
In December 2022, ChatGPT shook the world. Initially, it was mostly the technology world that understood its potential impact, especially for developers. Then, in early 2023, ChatGPT spread quickly to education and to many knowledge worker roles.

It also showed its limitations when a **lawyer went to court with a ChatGPT-generated argument** that was full of so-called "hallucinations" (i.e., made-up content not rooted in reality).

Soon after ChatGPT came out, the race for better models picked up with LLMs from Anthropic (Claude), Meta (Llama), Google (PaLM model through the Bard service), and OpenAI's own GPT-4.

The cost associated with training and operating those larger and larger models, however, is quickly becoming prohibitive.

So much so that even OpenAI's CEO, Sam Altman, was **quoted in April 2023** saying that "*I think we're at the end of the era where it's going to be these giant models, and we'll make them better in other ways.*"



# Avoiding the halo effect (Continued)

As the battle was being fought over which one of those giant models was better, Facebook open-sourced its Llama model.

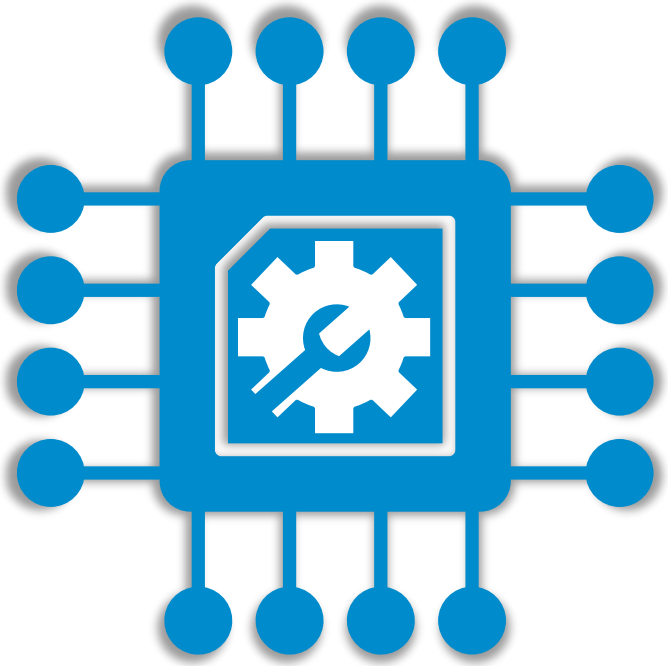
This model was quickly embraced by the open-source movement, especially given that Llama made models of multiple sizes available, some of them small enough to run on a smartphone.

This changed the dynamic again, and the focus switched from creating ever-larger generic models to creating smaller or customized ones.

Additionally, the cost associated with LLMs started to become a reality. In most cases, “traditional” AI models were 10 to 1,000 times cheaper to operate than LLMs.

For instance, even if text translation generated from LLMs was marginally better than from “traditional” AI machine translation models, customers quickly realized that the cost differential didn’t justify switching to LLMs for their large-scale machine translation processes.

# Avoiding the halo effect (Continued)



**What does it mean for companies wanting to integrate Generative AI into their business processes?**

As always, for any new ground-breaking technology, it's easy to fall into the "hammer in search of a nail" trap. Generative AI is and will remain transformational across most business processes for many years to come, no doubt about it.

LLMs' capabilities will continue to improve in the coming years. However, once the initial momentum of rapid successive quality improvements recedes (as it always does with technology) and that progress becomes more gradual, there will be two likely outcomes:

- One (or a few) players will dominate the field with a quality level way above all the rest of the LLM pack.
- All the main players will converge within a small distance of each-others quality-wise.

Whatever the end game might be, it's too early to bet your company's generative AI multi-year strategy on one specific LLM.

So, what is the most logical strategy to move forward today given those uncertainties?

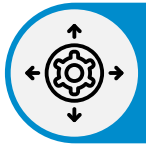
# Future-proof your Generative AI strategy

Knowing which LLM will win, if any, is not a bet you want to take. So, what's the best approach to strategically position your company to take this generative AI tectonic shift seriously while still not getting bogged down by (ever-changing) LLM-level details? To future-proof your generative AI strategy, it's crucial to avoid locking yourself in a specific LLM, or to assume a single approach could be used for all your business needs.

## Select the best tool for your challenge

First, you need to clearly define the most appropriate LLM per business process. Should you go with the one that offers the largest context window to be able to consume long text, or should it be the one that has the best mathematical or data analytics capabilities? Or maybe it's the model's coding, debugging, and software language translation that's the most important?

Also, you need to ensure that the model can be customized with your data to better align with your company's unique situation. Depending on the business process that you will redesign and augment with generative AI, the best LLM to select can change dramatically. You need to keep the flexibility of which LLM to use now and in the future.

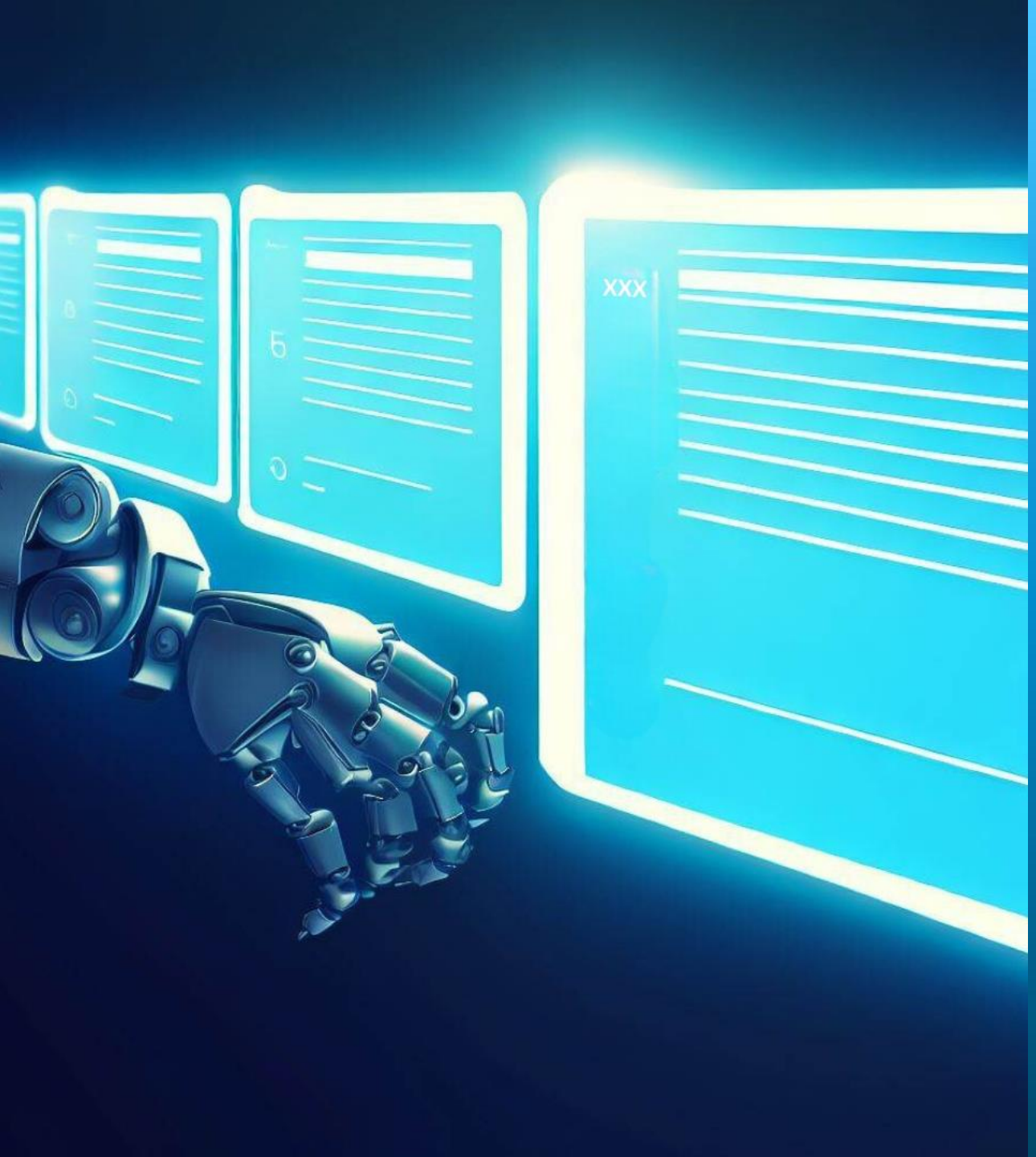


## Adopt a model-agnostic platform approach

Regardless of the LLM selected, you also want to ensure you can swap it out seamlessly if better options become available.

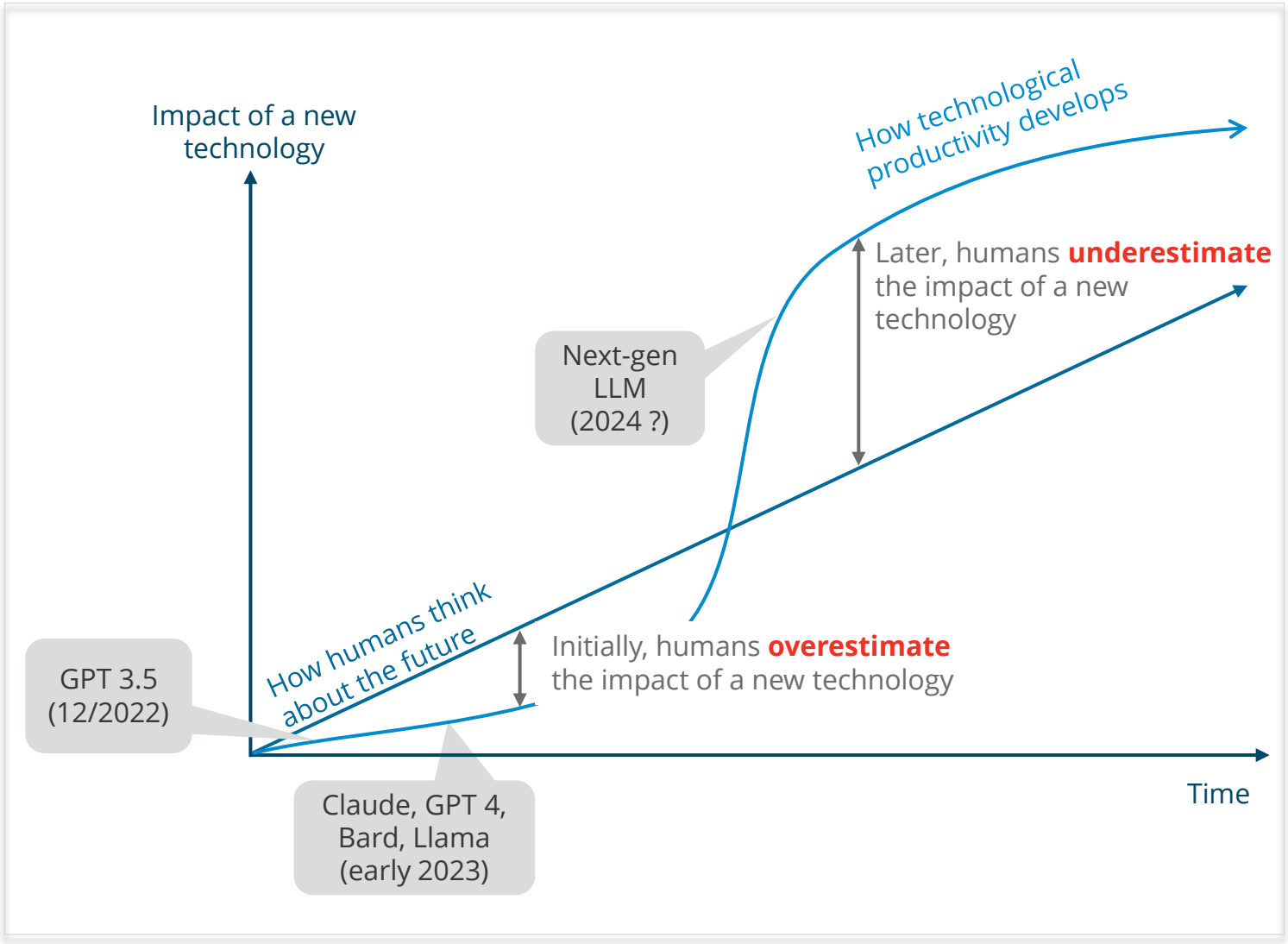
Therefore, your generative AI solution must be architected in a way that insulates the application from the LLM, like you would ensure that any business application would not be locked to a specific database.

Your overall generative AI platform, such as **Amazon Web Services' Bedrock platform**, must allow for selecting the right model (LLM, visual, code, voice, etc.) for the right use case. It should enable not only customization capabilities but, at least as importantly, guarantee your company's data security and privacy.




# Building a multi-year generative AI strategy

# Remember Amara's Law




> *We tend to overestimate the effect of a technology in the short run and underestimate the effect in the long run.*
> *– Roy Amara*

Advances in generative AI capabilities are likely to come at a breakneck pace over the next few years.

The generative AI tools available today aren't likely to immediately transform how professionals produce most content and otherwise help perform business operations.

Still, the tools that will become available over the next couple of years are very likely to be transformative.

Therefore, generative AI is a capability that leaders should start focusing on immediately.

They should build a multi-disciplinary v-team to closely monitor developments in the generative AI space.

# End-to-end maturity roadmap

Fractal has identified four key stages - **Crawl, Walk, Run, and Fly** - that organizations should consider following for their generative AI journey. Each stage requires investments in technical and human capabilities to maximize the potential of generative AI and ensure internal adoption while guaranteeing enterprise data privacy and information security.

## Crawl (Test)

In the “crawl” stage, organizations experiment with easy-to-deploy use cases and then quickly measure the ROI of those use cases. They also identify potential at-scale use cases that show promise for further exploration and development.

## Walk (Deploy)

In the “walk” stage, organizations deploy a few (one to three usually) use cases at scale. They again measure their ROI and use the results to expand generative AI to more business processes. This unlocks further potential for model customization, optimization, and scaling up.

## Run (Scale)

In the “run” stage, organizations analyze most or all their existing processes to build a multi-year AI transformation roadmap that comprehensively integrates generative AI across the enterprise.

## Fly (AI native)

In the "fly" stage, organizations take the next step towards becoming an AI-native company. If for digital transformation, companies went from digitizing analog processes to building new ones that were digital from the ground up, native AI companies will (re)design processes from scratch using AI as a core building block.

## Crawl

- Generative AI in 30 days PoC
- SenseForth FractalGPT customization and optimization

Long term industry experience & solutions

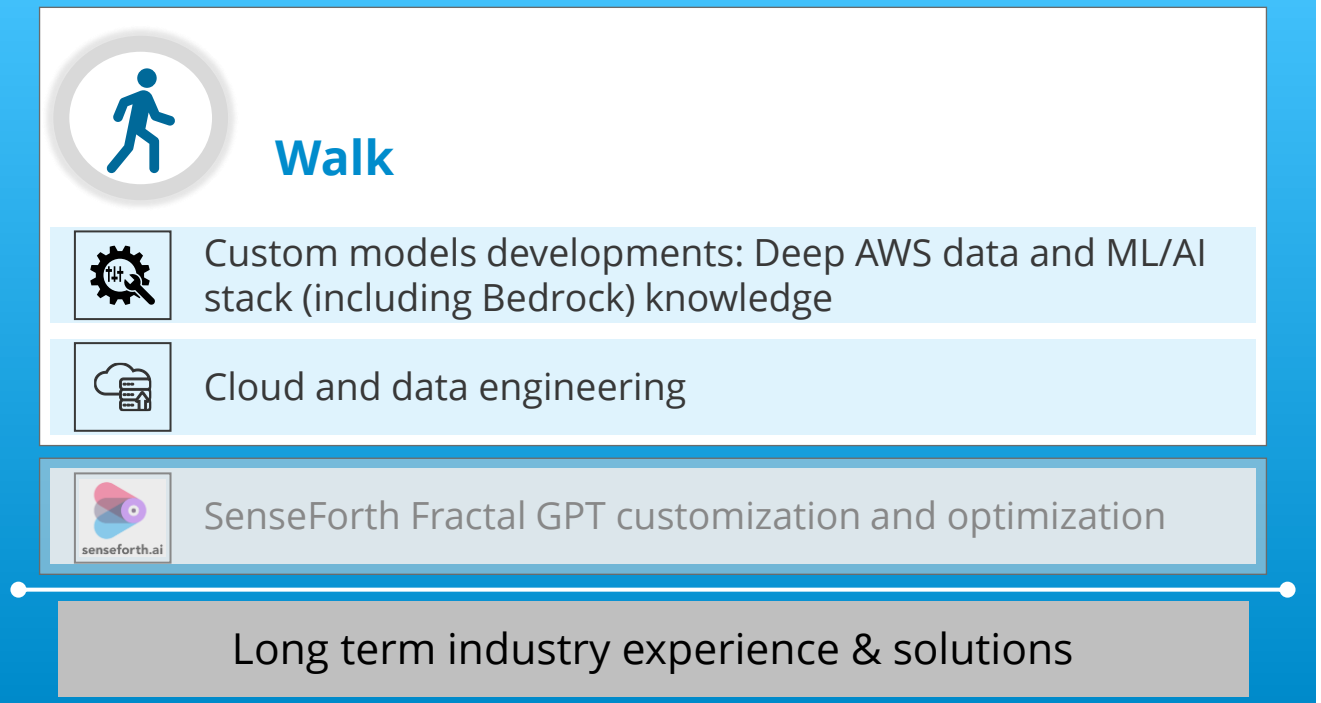
# End-to-end generative AI journey: The “crawl” stage

This is the initial stage of implementing generative AI solutions.

Here, organizations start by exploring, testing, and understanding the basic concepts and capabilities of generative AI-powered applications and business processes.

This step involves:

- conducting research,
- building foundational knowledge,
- engaging teams across company functions, and
- experimenting with small-scale projects to gain familiarity with generative AI.




# End-to-end generative AI journey: “Walk” stage

In the “walk” stage, organizations progress further by moving from experimental projects to the practical at-scale deployment of generative AI-powered solutions.

They begin to develop custom models that provide more suitable answers adapted to their specific business needs.

This customization involves not only adapting prompts and embeddings but also curating the data used to create those models.

It necessitates establishing a robust data infrastructure to collect, validate, protect, and manage business data into the models to ensure their continuous updates and accuracy.

This stage focuses on building the data platform infrastructure, developing internal expertise, optimizing performance, and validating the effectiveness of the generative AI solution.

It also emphasizes the importance of tailored model development and robust data management practices.

## Run

- Center of Excellence
- Business cases prioritization
- Model productization and integration: Software engineering, application integration
- Enterprise-wide scale up (size) and scale out (number of projects): DevOps, MLOps

- Custom models developments: Deep AWS data and ML/AI stack (including Bedrock) knowledge
- Cloud and data engineering

- SenseForth Fractal GPT customization and optimization

Long term industry experience & solutions

# End-to-end generative AI journey: “Run” stage

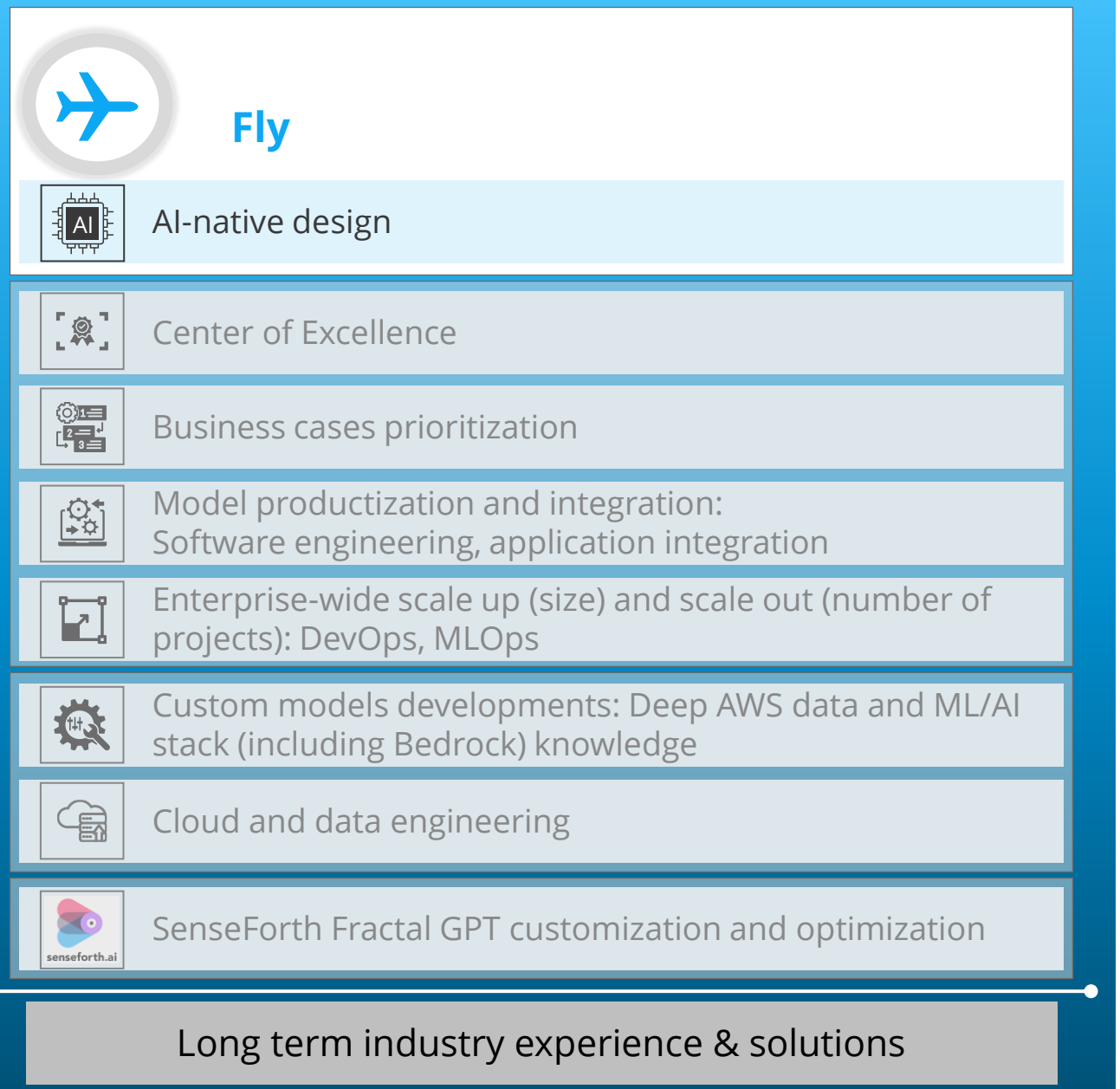
During the “run” stage of their generative AI journey, organizations focus on enterprise-wide scale-up (size) and scale-out (number of projects).

To support this growth, DevOps and MLOps practices are implemented to ensure generative AI model integration, monitoring, and management of AI models in their overall IT and Data Science infrastructure.

Model productization and integration become key priorities, requiring strong software engineering capabilities and the utilization of platforms such as Power Platform to integrate those models across existing enterprise business processes.

Business cases across functions are prioritized to identify high-impact areas where generative AI can drive significant value.

With the support of third-party partners, such as Fractal, organizations can build a Center of Excellence (CoE) that can help centralize expertise, increase knowledge sharing, and enforce governance.




# End-to-end generative AI journey: “Fly” stage

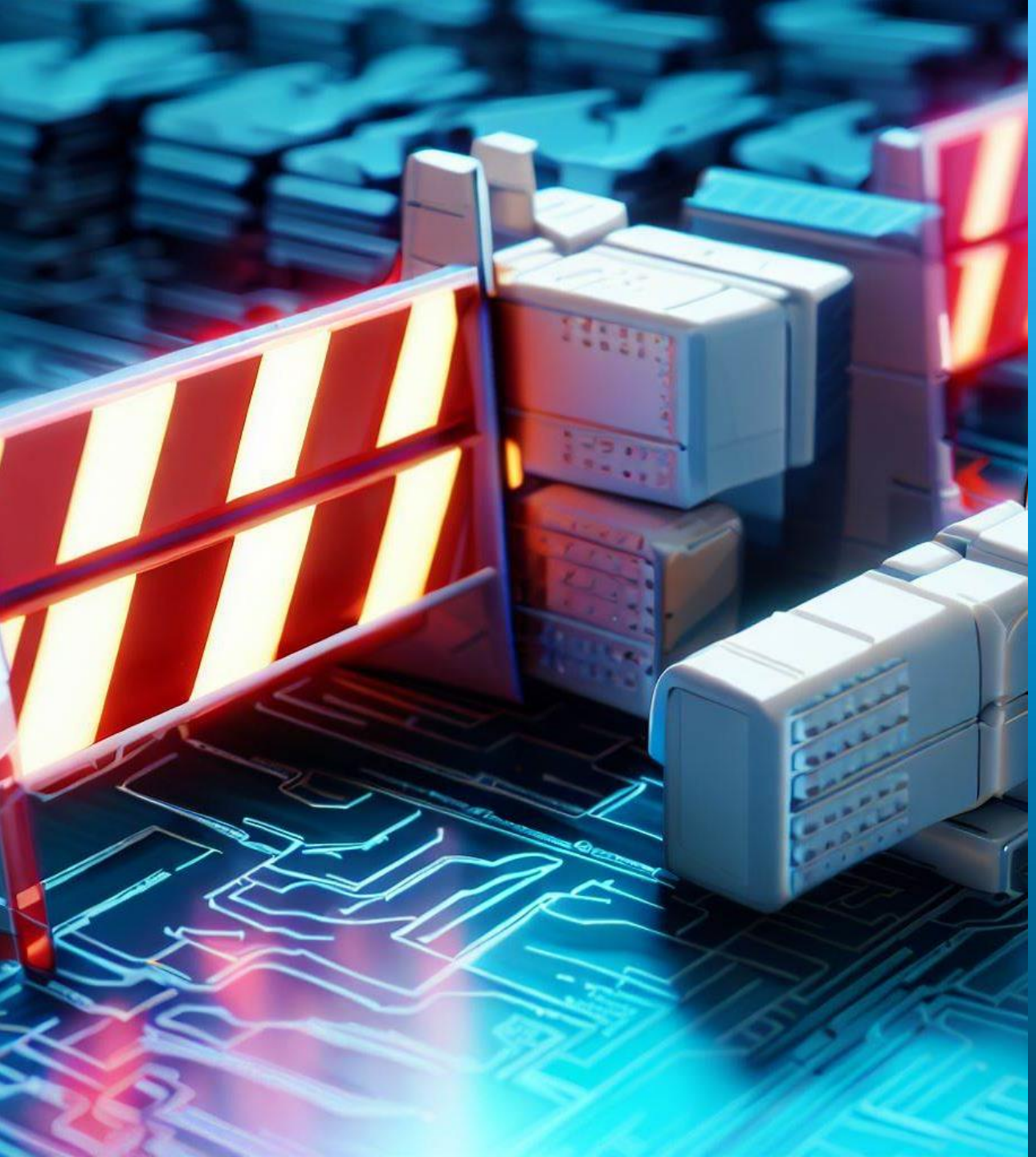
In the “fly” stage, organizations make a shift towards becoming an "AI native" organization. In this stage, organizations can design new processes or redesign existing ones using AI as a core building block.

Instead of adding AI to existing processes, they leverage AI from the ground up to maximize results.

This approach parallels the shift seen during the early days of digital transformation, where companies added digital elements to existing processes instead of designing them as digitally native ones from the ground up.

Similarly, AI-native companies build processes by incorporating AI from the ground up.

The fly stage represents a transformative phase where generative AI becomes a driving force, enabling organizations to unlock new business opportunities and create unique customer experiences.



# Barriers and risks associated with generative AI

# Barriers to generative AI Adoption

### Data quality & availability

Limited access to reliable and relevant internal and external data makes it difficult to effectively use generative AI algorithms for accurate results and insights.

### Lack of business alignment

Organizations use generative AI for multiple purposes, like innovating a product/service or using it to improve existing processes.

However, to harness its potential, organizations must have a clear vision, strategy, and roadmap that align with their goals, needs, and values.

### Regulatory & ethical challenges

Concerns around privacy, security, bias, and ethical implications create regulatory barriers and raise questions about the responsible use of generative AI.

Also, a lack of awareness leads to misconceptions and fears about AI taking over human jobs rather than augmenting their ability to improve their productivity and efficacy.

# Barriers to generative AI Adoption (Continued)

### Culture & change management

Resistance to change, organizational culture, and lack of awareness about the benefits of generative AI can impede adoption and hinder the necessary cultural shift for its effective use.

### Technical complexity & expertise

The complexity of implementing and managing generative AI systems can pose challenges in effectively leveraging the technology if the organization lacks the necessary new and specialized expertise.

### Lack of infrastructure & resources

Generative AI models require substantial storage capacity, the right data platform, network bandwidth, and a skilled workforce to run efficiently and reliably.

# Risks and their mitigation strategies

## Hallucination or errors, bias

Generative AI models can sometimes produce outputs that are inaccurate, nonsensical, or contradictory to the input or the training data leading to confusion or even affecting decision-making or communication.

Implementing rigorous validation and verification processes, leveraging human oversight, and continuously monitoring and fine-tuning generative AI models can help mitigate the risks of hallucination and biases in generated content.

## Cost overrun

Sometimes, a generative AI project may require more data, infrastructure, and human expertise than expected or encounter unexpected technical or operational challenges.

So, carefully planning and monitoring resources, conducting cost-benefit analyses, and starting with smaller-scale implementations can help measure and mitigate the risk of cost overrun.

## Customization needs

Generative AI models can be difficult to customize or adapt to different domains, contexts, or user preferences.

This can limit their applicability and usefulness for various scenarios or tasks.

So, it becomes necessary to understand specific business requirements, involve end-users in the development process, and build flexible and adaptable generative AI systems.

# Risks and their mitigation strategies (Continued)

## Privacy (when used directly with OpenAI)

Generative AI models may require access to sensitive or personal data, which can be compromised or misused by third parties.

Implementing strong data protection measures, complying with privacy regulations, and considering alternative approaches like federated learning or differential privacy can help mitigate those privacy risks.


## IP protection and creator rights

Generative AI models can generate content that may infringe on the intellectual property or creator rights.

Defining ownership and usage rights through legal agreements, employing authentication mechanisms, monitoring systems, and using digital rights management (DRM) techniques can help protect both IP and creator rights.

## AI detectors

Generative AI carries the risk of being detected by AI systems or experts, leading to the potential identification of fake or harmful content.

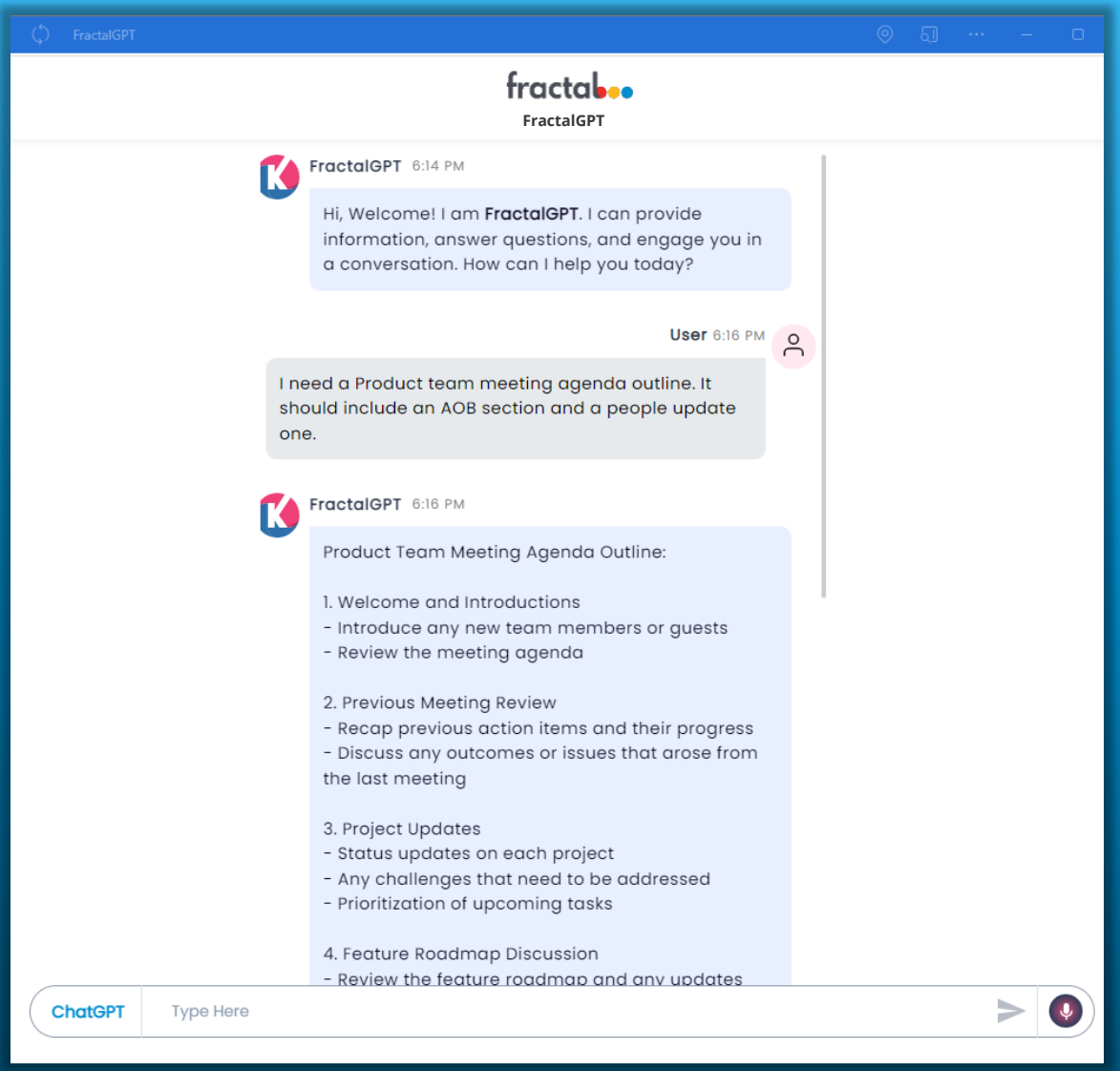
To mitigate this, organizations must create systems and implement processes to ensure generative AI tools are creating ethically appropriate content including deceptive or illegal content. The must also comply with applicable legal and regulatory frameworks governing the use of generative AI.

# Getting started with generative AI

Fractal offers different ways to help you quickly get started on your generative AI journey.

In addition to custom solutions, we can help you identify various use cases and prepare you to deploy generative AI in your organizations with the below offerings:

- FractalGPT: Securely deploy AI-powered chat for your enterprise
- One-hour exploratory workshop to identify suitable projects for an initial Proof of Concept (PoC)
- Generative AI in 30 days




# Start creating with FractalGPT

Securely empower your enterprise with ChatGPT capabilities in just 7 days using FractalGPT: a fast and customizable solution that safeguards your data and IP.

It is designed to be secure, scalable, and easy to use, making it ideal for businesses of all sizes and across industries.

With FractalGPT, your teams can improve productivity while ensuring your data remains secure.

Get FractalGPT on AWS marketplace:

https://go.fractal.ai/AWS/FractalGPT

# Generative AI in 30 days

The generative AI in 30 days offer is designed to help organizations analyze generative AI use cases and rapidly prepare for the deployment of generative AIs. The offer achieves this by leveraging predefined resource deployment procedures and templates from AWS and Fractal.

## Phase 1
Design workshop

- Identify generative AI scenarios for proof of concept
- Gain an understanding of the business use case
- Review the current state of data and advise on data preparation methodology
- Create initial architecture

## Phase 2
Landing zone deployment

- Establish AWS landing zone and generative AI capabilities
- Working session to customize landing zone
- Population of landing zone, showcasing implementation best practices
- Use migration accelerators
- Establish initial architecture to get data flowing

## Phase 3
Generative AI proof of concept

- Transfer knowledge on data platform, generative AI capabilities & best practices
- Develop approach to tailor LLM models
- Provide customer with a go-forward plan aligned with resourcing and architectural needs for the next phase

## Offer outcomes

- Identify generative AI use cases in discovery workshop
- Deployed POC of customer-defined generative AI use case via customer's AWS subscription
- Documentation and knowledge transfer
- Experienced AI experts to guide your team along the scenario refinement & feature definition path
- Proposal for implementation of production solution

# Why Fractal?

As a Data and Analytics AWS Partner with decades of experience developing and deploying AI solutions at scale, Fractal can help you analyze, design, and deploy your generative AI-based solution quickly and effectively.

We provide enterprise-ready solutions that adapt to each customer's own processes and data, scale up rapidly, and are cost-effective.

Fractal can support clients throughout their generative AI journey with:

- Strategic use cases selections
- Data engineering and migration
- AI Model customization, deployments, and management (MLOps)
- Cost optimization
- Best practices deployment

aws partner network

Data & Analytics Competency

## Resources

- AWS Bedrock
- Anthropic Claude
- Azure Open AI
- Google Vertex AI
- Supercharge your business with Generative AI

fractal.ai

One World Trade Center Suite 76J, New York, NY 10007 | +1 (646) 547 1600

info@fractal.ai @fractalai linkedin.com/showcase/fractal-aws-partnership

aws partner network
Advanced Consulting Partner

aws partner network
Data & Analytics Competency