The data leader's primer for generative AI

Your guide to building a strong data foundation for generative AI

Table of contents

Executive summary

Since the public release of ChatGPT in late 2022, generative AI has shown considerable potential to augment creative and intellectual work through information retrieval and rapid ideation. According to McKinsey, generative AI may add up to $6.1 trillion to $7.9 trillion to the global economy annually in the coming decades.[1]

According to a 2023 Gartner report, **55%** of organizations plan to use generative AI and **78%** of executives believe the benefits of AI adoption outweigh the risks.[2]

The world is already being transformed by AI-assisted medicine, education, research, and more.[3] Researchers at Stanford recently created a new model, SyntheMol, to accelerate drug discovery.[4] Education provider Khan Academy personalizes learning using an AI tutor, Khanmigo.[5]

With generative AI, organizations can augment foundation models trained on a broad base of public data with their unique, proprietary data. By combining public and proprietary data, a model may become the most knowledgeable entity in an organization, creating many opportunities for innovation.

However, generative AI is only as good as its data. To responsibly and effectively leverage AI, an organization needs a solid foundation of tools, technologies, and best practices.

Data readiness for generative AI depends on two key elements:

- Refined, centralized data from disparate data sources that is continuously and automatically updated.
- Robust data governance frameworks to maintain data quality, compliance, and ethical use as AI initiatives expand.

Introduction

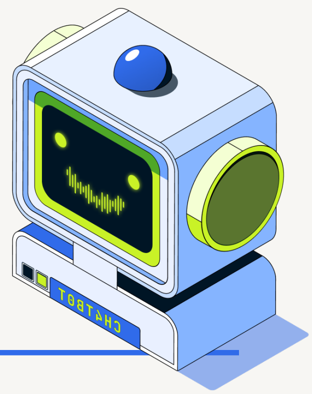
Generative AI differs from other forms of machine learning and artificial intelligence in its capacity to create – i.e. generate – new data in all forms of media, including text, code, images, audio, video, and more.

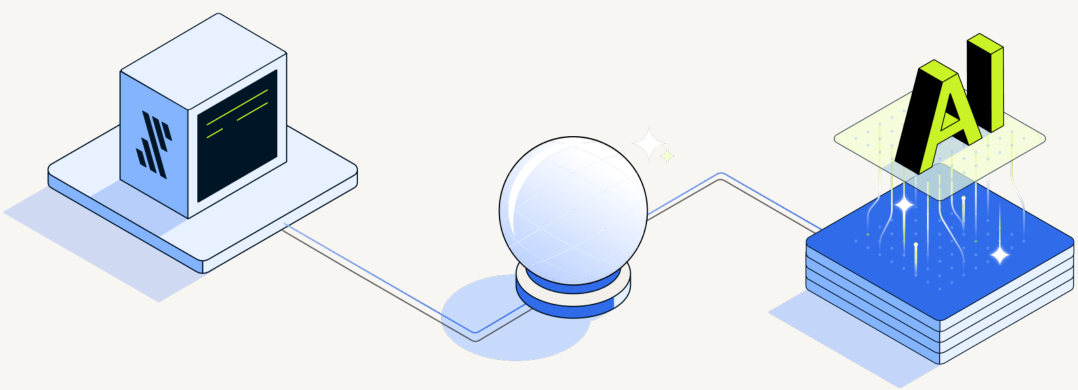
Large language models (LLMs) are a type of generative AI that is trained on enormous volumes of data from across the internet and produces text in response to prompts. This output typically takes the form of natural language, though LLMs can also be used to generate text that follows other patterns, such as code. In this ebook, we will mostly discuss LLMs, although the same principles apply to non-text media.

Off-the-shelf generative AI models trained on huge volumes of publicly available data are also known as foundation models. Commercial applications of generative AI commonly combine foundation models built by a third party with an organization's proprietary data in an architecture called **retrieval-augmented generation (RAG)**. This architecture offers a highly practical and accessible alternative to building models from scratch, a costly and difficult undertaking beyond the means of most organizations.

What generative AI is not

Despite its ability to mimic human language and other creative or intellectual output, generative AI fundamentally lacks consciousness. Under the hood, like other machine learning, generative AI is linear algebra, statistics, and computation on a gargantuan scale.





1

Data prerequisites for generative AI

Successful generative AI depends on a solid data foundation and high data maturity, in which an organization demonstrates mastery over both integrating – moving and transforming – data and governing its use. Without data maturity, generative AI becomes extremely difficult to deploy and can produce misleading outputs.

According to an MIT Technology Review report, **45%** of C-suite leaders cite data integration and pipelines as the top challenge for AI and **64%** of respondents consider data integration and pipelines a top investment priority.[6]

There are both technological and organizational elements to data maturity. On the technological side, the following capabilities are essential:

- A central, cloud-based data repository – data warehouse or lake – that can serve as a scalable, single source of truth
- A data integration solution that reliably and automatically centralizes structured and unstructured data from all sources at scale and features:
  – Fast, timely updates
  – Reliability and the ability to quickly recover from failures
- The capability to model and transform data in a collaborative, version-controlled manner
- Data governance capabilities such as:
  – The ability to block and hash sensitive data before it arrives in a central repository
  – Access control
  – Data cataloging

Automation is the central theme of efficient, reliable, and scalable data integration. Much like generative AI itself, automated data integration is a formidable force multiplier, freeing up engineering resources for higher-value analytical and operational activities.

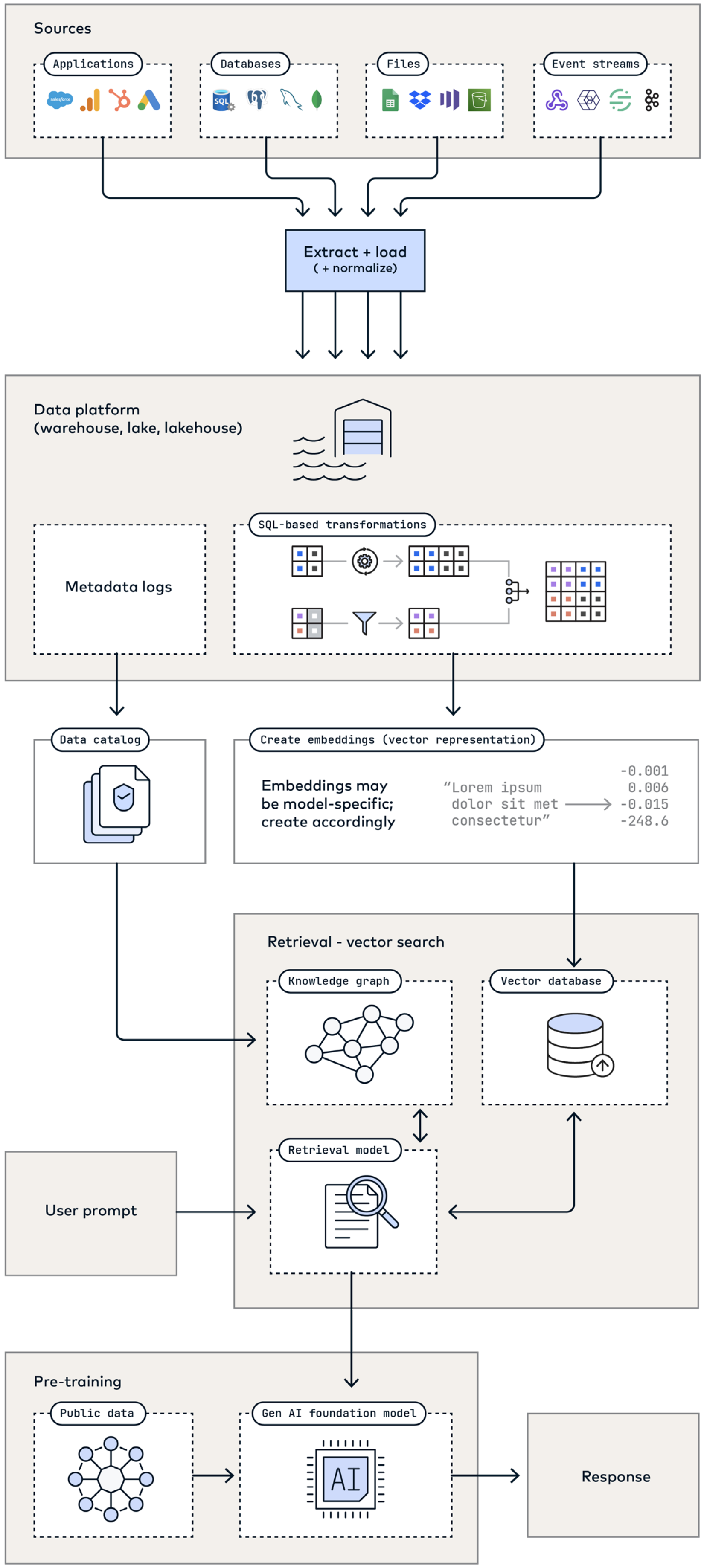
On the organizational side, your team will need the following practices and structures in place:

- Automation to manage infrastructure costs and efficiently produce data assets
- Data literacy to ensure that data assets are high quality and AI-ready
- Visibility, governance, and stewardship to ensure compliance with data-related legal, regulatory, and ethical considerations
- Product development mindset and the ability to align data to AI use cases and tailor data products to the needs of stakeholders

These practices and structures demonstrate that your team effectively and responsibly handles data and is ready to pursue generative AI.

Your data platform architecture for generative AI

Building a generative AI from scratch is a complex undertaking with the potential to cost many millions of dollars and months of development time. A more practical and less risky option is to augment a foundation model with your organization's unique, proprietary data using a RAG architecture. A RAG architecture should look something like the diagram on the following page:


Sources
Applications
Databases
Files
Event streams
Extract + load
(+ normalize)
Data platform
(warehouse, lake, lakehouse)
Metadata logs
SQL-based transformations
Data catalog
Create embeddings (vector representation)
Embeddings may
be model-specific;
create accordingly
"Lorem ipsum
dolor sit met
consectetur"
-0.001
0.006
-0.015
-248.6
Retrieval - vector search
Knowledge graph
Vector database
Retrieval model
User prompt
Pre-training
Public data
Gen AI foundation model
AI
Response

The beginning of this architecture resembles that of other analytics use cases. You need high performance pipelines to extract and load structured and unstructured data from a wide range of sources to a cloud destination, such as a data lake. Transformations should be performed at the destination rather than in-flight for maximum flexibility. Data from the same source can be transformed into two separate data models, one optimized for reporting and other for generative AI.

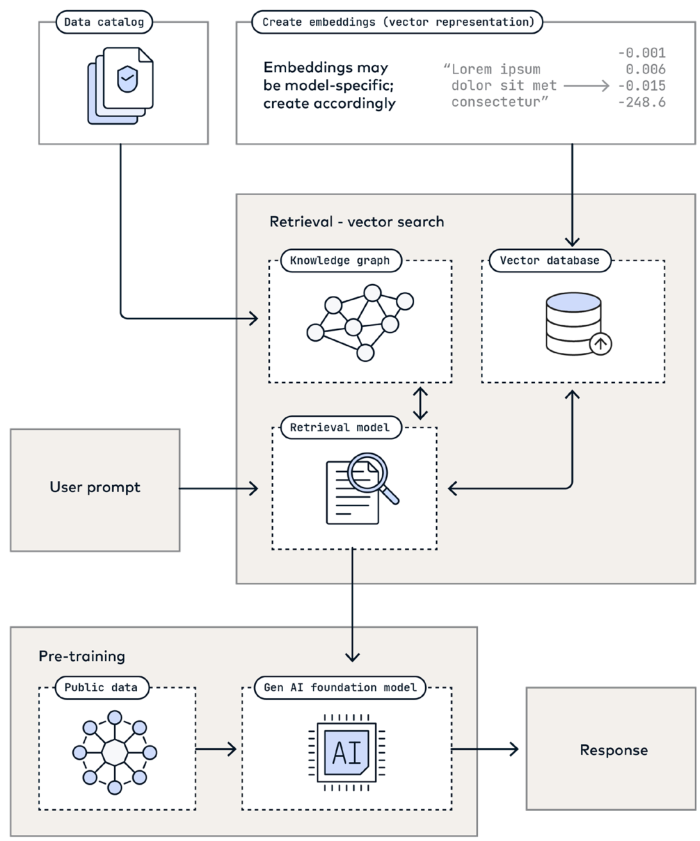
What is a vector database?

A vector database stores and retrieves data as "embeddings" – high-dimensional representations of text, images, audio, video, and other media. Think of an embedding as a long list of numbers that represent coordinates. This coordinate system allows you to search and retrieve data based on similarity.

What comes afterward is unique to generative AI. Modern tools such as Databricks Mosaic AI, Snowflake Cortex, Google Vertex AI, and other offerings by leading cloud providers often abstract away the need to separately provision a vector database or produce embeddings from data.

Under the hood, you can augment an off-the-shelf generative AI model with your data in two ways:

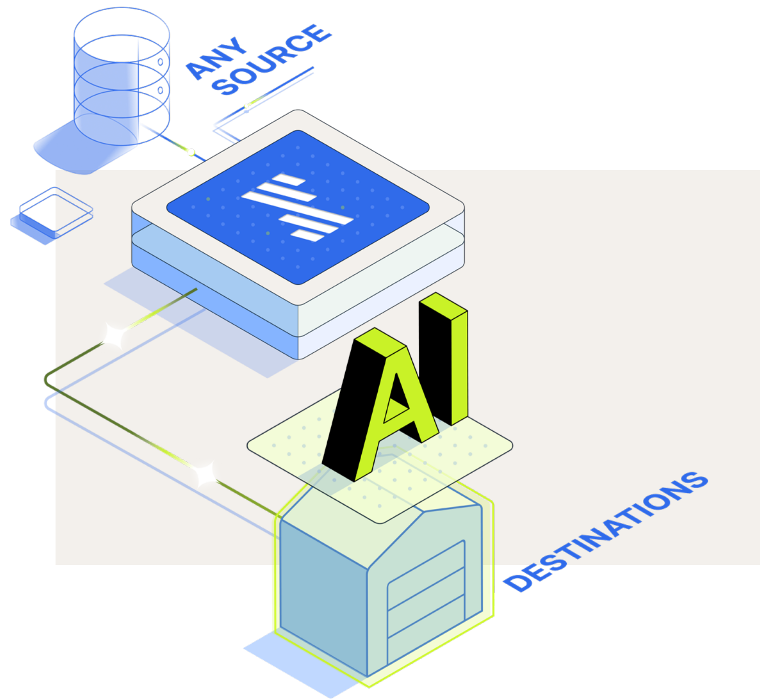
1. You can turn long strings of text into embeddings and store these in a vector database for a foundation model to access. This enables the model to produce outputs based on both the general knowledge provided through its initial training and your organization's unique data.
2. You can combine large language models with knowledge graphs, explicitly encoding semantic understanding into the model. Knowledge graphs represent data in a graph, with relationships as connections between data points, e.g. "Harry" (node) "wears" (edge) "pants" (node).



When you have an adequate technological and organizational foundation for generative AI, make sure your team has the following capabilities:

- Prompt design/engineering and verifying the output of artificial intelligence
- A solid understanding of how generative AI can solve problems in your particular business domain
- Understanding and mitigating security and ethical concerns
- Infrastructure management
- Data literacy for accurate and ethical use of AI

Most of this engineering effort will likely be spent quality-controlling and tuning your applications' outputs and managing your production infrastructure.



2

How generative AI will evolve the nature of work

All organizations produce huge volumes of media like contracts, blogs, call transcripts, chat applications, project management tools, emails, and documentation. A large language model trained on such a corpus can answer domain-specific questions, summarize text, translate between languages, adjust tone, analyze sentiments, and more. A large language model with access to an organization's accumulated data can not only act as the most knowledgeable "member" of an organization but perform a wide range of valuable tasks, such as:

- Acting as internal and external help desks, replacing the need to manually sift through documents or ask for another person's time
- Shortening turnaround time for sales and marketing content, including media of all kinds – written collateral, images, animations, and more
- Accelerating software engineering by generating boilerplate code or translating between programming languages

There are countless potential industry-specific use cases, as well:

- Automatically producing financial documentation[7]
- Enabling retail customers to virtually "try on" clothing and accessories[8]
- Accelerating drug discovery by uncovering new configurations of proteins and other molecular structures[9]
- Accelerating legal and paralegal work, such as legal research, drafting documents, and communication with clients[10]
- Enabling doctors to diagnose rare and unusual diseases and devise personalized treatment plans[11]
- Helping engineers optimize the geometry of hardware components through generative design[12]
- Generating assets for video games, films, and other digital media[13]
- Helping educators personalize lesson plans and instructional materials[14]

Generative AI can support nearly all creative or intellectual business activities, mainly by decreasing the time and effort required to summarize, ideate, iterate, and prototype.

3

How to use generative AI

Successful implementations of generative AI require good prompts and preventing hallucinations, i.e. plausible but factually incorrect results. Upstream of the user experience, they also require governance over the proprietary data used to augment foundation models.

Prompt design and prompt engineering

As with a search engine, a prompt is the text input a user supplies to an AI model to receive an output. Prompt design and prompt engineering are centered around 1) crafting prompts that elicit helpful responses and 2) systematizing and governing the use of prompts, respectively.

The key to crafting good prompts is to clearly communicate in a way that minimizes ambiguity and provides as much useful context as possible. This is typically done through the prompt interface itself, though accessing a generative AI through an API may offer direct access to parameters as well.

The following principles originate from the Udemy course The Complete Prompt Engineering for AI Bootcamp by Mike Taylor and James Phoenix.

Sample Prompt

```
The following is an example of a
naive prompt:

Can I have a list of marketing
headlines for blogs?
```

```
A more helpful prompt provides as
much context as possible:

Can I have a list of ten marketing
headlines for blogs?

Description: Explain the benefits
of automated data integration

Audience: analysts, data engineers,
data scientists

Seed words: data integration,
data engineering, analytics,
idempotence, change data
capture, automation, efficiency,
reliability, ETL, ELT

Examples: What is the architecture
of automated data integration?, the
ultimate guide to data integration
```

1. any explicit examples, guidance, or context, is also known as zero-shot prompting. This forces the model to consult very general knowledge, with a high likelihood of producing poor results.

2. **Provide examples –** The quality and consistency of responses can be vastly improved by providing both positive and negative examples for the model to anchor itself.

3. **Format the response –** Specify the format of the response you want. If you need information presented as a numbered list or bullet points, say so explicitly. This consideration is especially important for generating code; if you need data presented in JSON or tables, a foundation model should be able to produce those, as well.

4. **Divide labor –** Not everything has to be accomplished with a single prompt. You can take advantage of your session's short-term memory with a succession of prompts that add additional details, direction, formatting, etc. to continuously refine the output. You can add context at any point during the process using embeddings from a vector database. Finally, you can import the output from one model into another, more specialized one to further refine your results.

5. **Evaluate outputs and iterate –** Prompting is an iterative process. The "engineering" side of prompt engineering is a matter of systematically testing, evaluating, and iterating through different configurations of prompts.

Poor or even malicious prompting will not only return misleading or nonsensical results, but it can also expose confidential information or even endorse antisocial behavior. Microsoft's Tay chatbot became infamous within hours of deployment after interactions with malicious users resulted in bigoted responses[15]. Worse, hacks known as prompt injections can be used to misdirect a model and produce any response the hacker wants, including social engineering like phishing attempts.

The ability of generative AI to produce content on a large scale while mimicking well-known styles and patterns also poses a danger in terms of intellectual property theft. Another malicious application of generative AI is to intentionally produce content that is false and violating in some way, such as combining deepfakes with explicit content.

There are several ways to mitigate these challenges:

- Train the users of your products, whether internal or external, how to write prompts
- Provide default prompt structures for your users
- Carefully curate the data used to augment the model, excluding sensitive data
- Build safeguards directly into the model, e.g. refusals to answer abusive prompts
- Test use cases internally before rolling them out externally

Responsible AI: Minimizing hallucinations and protecting your business

A serious pitfall of generative AI is that it can produce results that aren't factual or connected to reality in any way. The stakes can be high; there has already been at least one documented instance of a lawyer citing hallucinated court cases.[16] There are several ways to minimize the frequency and impact of hallucinations.

The most important element is to ensure high-quality upstream data by carefully curating and governing the data your organization provides to its foundation model. Anonymize, encrypt, or altogether exclude sensitive data.

Don't put all of your organization's data into a vector database to feed your foundation model. Instead, carefully consider the problem you are trying to solve and select content that is unbiased, accurate, relevant, and internally consistent. During the COVID pandemic, diagnostic AIs failed to catch COVID cases in large part due to poor-quality data.[17]

The second is to simply fact-check results and reject bad ones. Generative AI models learn through reinforcement; human supervision in the loop may be necessary to test the answers provided by a model until it meets an acceptable threshold for accuracy and relevance before an application is released to the public or used to support any serious decisions. Allow end users of a generative AI system to report bad results – it will improve your model's efficacy and safety and ensure that end users are never exposed to sensitive data, abusive or antisocial messages, etc. Screening can also be performed through rule-based systems or even another AI trained specifically to classify and evaluate content. You can also maintain control over the prompts themselves by rejecting or sanitizing prompts that include certain phrases or concepts before they can reach the model.

A third approach is to supplement the foundation model with a knowledge graph that explicitly builds factual relationships between concepts. This also enforces explainability. Knowledge graphs also allow for the explicit encoding of safeguards into a generative AI model. For example, a knowledge graph can provide a model with access control policies and flag data that may breach those policies.

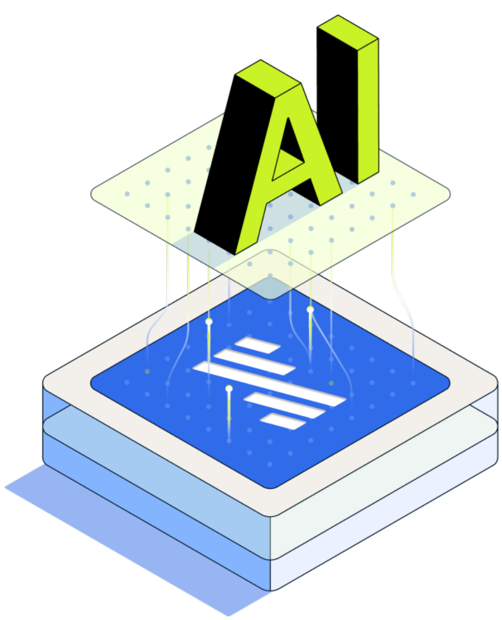
Some foundation models specialize in different media (text, images, audio, etc.), domains (geospatial analysis, music, etc.), and use cases (sentiment analysis, data visualization, etc.).

Examples of popular, general-purpose foundation models include the GPT, Llama, Gemini, Mistral AI, and Claude series. Your choice will depend on cost, performance, model size, customizability, etc.

Limitations of generative AI

As a costly, complicated tool, there are many problems for which generative AI is overkill. Common business problems can often be solved using rule-based heuristics or much simpler predictive modeling approaches, like linear regression. For instance, financial forecasting depends on methods such as moving averages, regression analysis, or the Delphi method.[18] A good rule of thumb is to reserve generative AI for problems only it can solve.

As illustrated by generative AI mishaps like bigoted chatbots, hallucinated court cases, and misdiagnoses, we are still a long way from simply entrusting machines without humans in the loop for high-stakes applications. A RAG approach with quality data minimizes many of the risks, but the growth of generative AI, like all previous waves of industrialization, won't make humans obsolete. Instead, human judgment will remain essential to channeling and guiding the power of technology.



Your team's generative AI success depends on a reliable data foundation

Generative AI promises a rich, innovative future in every industry but depends on a data foundation built on centralized, fresh, and governed data.

To prepare your organization for generative AI, you will need an automated data integration solution that works with your governance strategy as well as processes to effectively use data to support decisions. Every generative AI project starts with data, and Fivetran is trusted by Enterprises to build the reliable data foundation you need to build reliable AI.

1. Chui, M., Hazan, E., Roberts, R., Singla, A., Smaje, K., Sukharevsky, A., Yee, L., & Zemmel, R. (2023, June 14). The economic potential of generative AI: The next productivity frontier. McKinsey Digital. https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier

2. Delisi, M., Howley, C. (2023, Cotober 3). Gartner Poll Finds 55% of Organizations are in Piloting or Production Mode with Generative AI. Gartner. https://www.gartner.com/en/newsroom/press-releases/2023-10-03-gartner-poll-finds-55-percent-of-organizations-are-in-piloting-or-production-mode-with-generative-ai

3. What's the future of generative AI? An early view in 15 charts. (2023). McKinsey & Company. https://www.mckinsey.com/featured-insights/mckinsey-explainers/whats-the-future-of-generative-ai-an-early-view-in-15-charts

4. Tompa, R. (2024, March 28). Generative AI develops potential new drugs for antibiotic-resistant bacteria. *Stanford Medicine*. https://med.stanford.edu/news/all-news/2024/03/ai-drug-development.html

5. KhanMigo Education AI Guide | Khan Academy. (n.d.). Khan Academy. https://www.khanacademy.org/khan-labs

6. MIT Technology Review (2024, May 29). MIT report: AI readiness for C-suite leaders. *Fivetran*. https://www.fivetran.com/resources/reports/mit-report-ai-readiness-for-c-suite-leaders

7. Sergiienko, B. (2025, January 9). How Generative AI in Finance Addresses 10 Key Operational and Strategic Industry Challenges. *master.of.code*. https://masterofcode.com/blog/generative-ai-in-finance#Generative_AI_Use_Cases_in_Financial_Services

8. Ewen, L. (2024, September 6). Google Shopping launches virtual dress try-ons in the US. *Fashion Dive*. https://www.fashiondive.com/news/google-shopping-ai-virtual-tryon-dresses-simkhai/726326/

9. Marr, B. (2024, June 19). How Generative AI Is Accelerating Drug Discovery. *Forbes*. https://www.forbes.com/sites/bernardmarr/2024/06/19/how-generative-ai-is-accelerating-drug-discovery/

10. Gritsenko, D. (2025, January 9). Generative AI in Law: The Future of Legal Practice Through AI Lens. *master.of.code*. https://masterofcode.com/blog/generative-ai-in-law

11. Ghebrehiwet, I., Zaki, N., Damseh, R., & Mohamad, M.S. (2024, April 25). Revolutionizing personalized medicine with generative AI: a systematic review. *Artificial Intelligence Review*. https://link.springer.com/article/10.1007/s10462-024-10768-5

12. Jaffee M., Morykin, Stacey. (2024). A Case Study for Generative Design in Horizontal Infrastructure. *Autodesk University*. https://www.autodesk.com/autodesk-university/class/case-study-Generative-Design-Horizontal-Infrastructure-2022

13. echo3D. (2023, April 17). 7 Generative AI Tools for 3D Asset Creation. *echo3D*. https://medium.com/echo3d/7-generative-ai-tools-for-3d-asset-creation-97dd88153b7

14. [x]cube LABS. (2024, November 27). Personalized Learning Systems with Generative AI: Revolutionizing EdTech. *[x] cube LABS*. https://www.xcubelabs.com/blog/personalized-learning-systems-with-generative-ai-revolutionizing-edtech/

15. Kraft, A. (2016, March 25). Microsoft shuts down AI chatbot,Tay, after it turned into a Nazi. CBS News. https://www.cbsnews.com/news/microsoft-shuts-down-ai-chatbotafterit-turned-into-racist-nazi/

16. Wesier, B. (2023, May 27). Here's What Happens When Your Lawyer Uses ChatGPT. *The New York Times*. https://www.nytimes.com/2023/05/27/nyregion/avianca-airline-lawsuit-chatgpt.html

17. Heaven, W. D. (2022, August 2). Hundreds of AI tools have been built to catch covid. None of them helped. *MIT Technology Review*. https://www.technologyreview.com/2021/07/30/1030329/machine-learning-ai-failed-covid-hospital-diagnosis-pandemic/

18. 7 Financial forecasting methods to predict business performance. (2022, June 21). Business Insights Blog. https://online.hbs.edu/blog/post/financial-forecasting-methods