Fivetran

A primer for data readiness for generative AI

How to build a solid data foundation and lead your organization to generative AI.

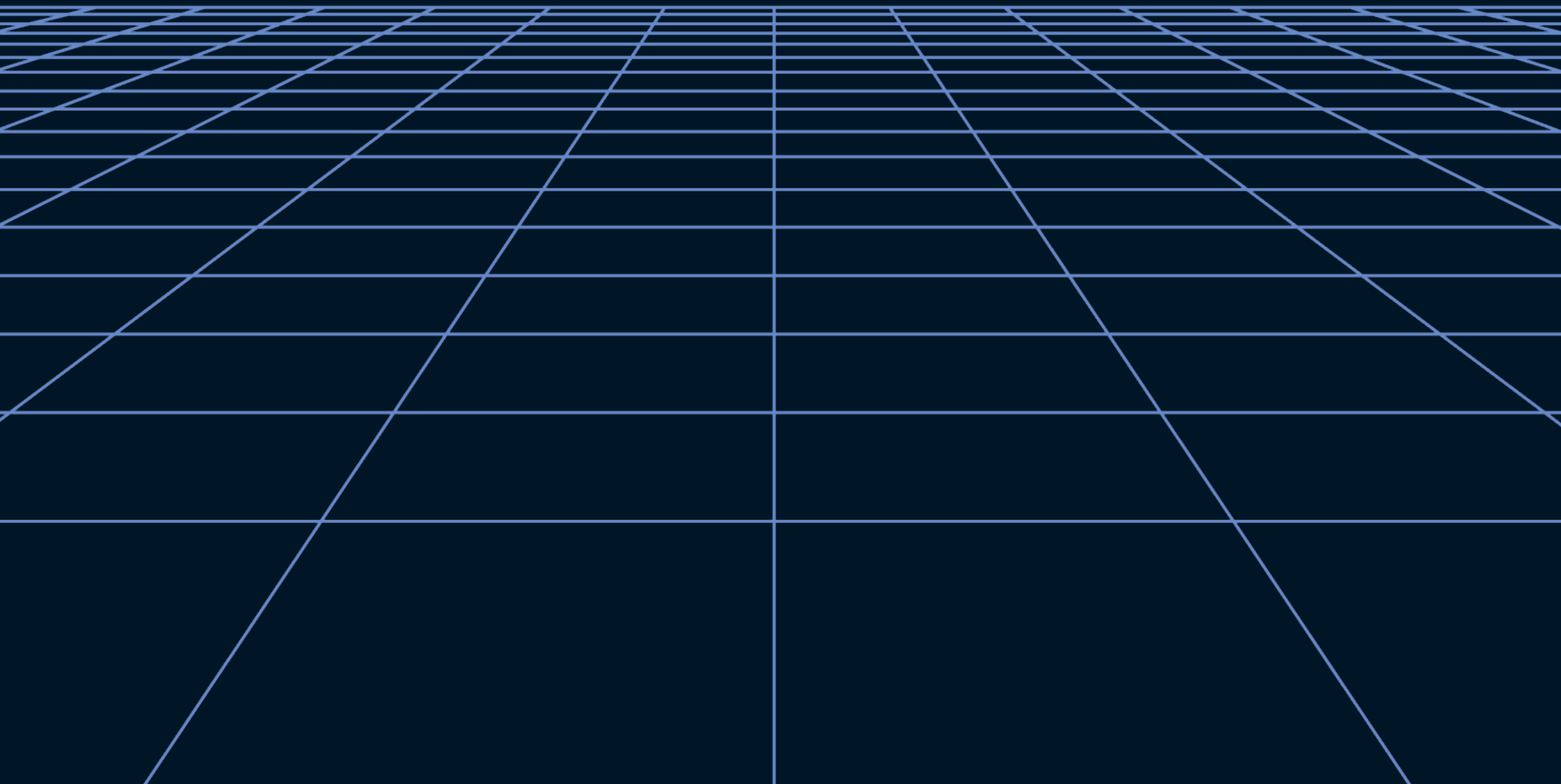


Table of contents

Executive summary

Since late 2022, generative AI has quickly demonstrated its value and potential to help businesses of all sizes innovate faster. By generating new media from prompts, generative AI stands to become a powerful productivity aid, multiplying the effect of creative and intellectual work of all kinds. Salesforce reports that 61 percent of office workers use (or plan to use) generative AI in some professional capacity.[1]

McKinsey estimates that, in the coming decades, generative AI may add up to $6.1 trillion to $7.9 trillion to the global economy annually.[2] Moreover, AI is a highly active field of continuing research and innovation, with the number of published papers on the subject doubling every two years.[3]

According to Gartner, **55%** of organizations have plans to use generative AI and **78%** of executives believe the benefits of AI adoption outweigh the risks.[4]

The world will be transformed by AI-assisted medicine, education, scientific research, law and more.[5] Researchers at the University of Toronto use generative AI to model proteins that don't exist in nature.[6] Similarly, pharmaceutical giant Bayer now uses generative AI to accelerate the process of drug discovery.[7] Education provider Khan Academy has developed an AI chatbot/tutor, Khanmigo, to personalize learning.[8] The list of examples across all industries only continues to grow.

Generative AI is not just a general-purpose productivity aid that surfaces information the way a search engine does; with gen AI, organizations can combine their unique, proprietary data with foundation models that have been pretrained on a broad base of public data. Trained on a combination of public and proprietary data, generative AI may become the most knowledgeable entity within an organization, opening up innumerable opportunities for innovation.

However, as with all analytics, generative AI is only as good as its data. To fully leverage AI, an organization needs mastery over its proprietary data. This means a solid foundation of data operations technologies and organizational norms that facilitate responsible and effective use of data.

> "The most important thing is not just collect the data, but cleanse, categorize the data and make sure it's in a usable format. Otherwise you're just paying to store meaningless data.[9]
>
> Rob Zelinka, CIO of Jack Henry, tells the Wall Street Journal"

Data readiness for generative AI depends on two key elements:

- The ability to move and integrate data from databases, applications and other sources in an automated, reliable, cost-effective and secure manner
- Knowing, protecting and accessing data through data governance

This kind of data readiness is perennially overlooked and has historically derailed many attempts to leverage the power of big data and data science. One metric suggests that as many as 87 percent of data science projects never make it to production, often because of siloed and ungoverned data as well as underdeveloped data infrastructure.[10]

This primer is an introduction to generative AI – what it is, how your organization can successfully prepare a solid data foundation for it and how to effectively use it.

Introduction

Traditional machine learning and artificial intelligence are typically concerned with numerical or categorical prediction, pattern recognition or automated decision-making. Generative AI differs from other forms of machine learning and artificial intelligence in its generative nature. Rather than merely analyzing data, generative AI produces new data in all forms of media – text, code, images, audio, video and more.

Under the hood, generative AI models are supported by machine learning models called artificial neural networks. Inspired by the architecture of brains, neural networks are designed to model complex, non-linear relationships using a graph data structure.

The key benefit of neural networks is that they learn relationships and patterns through exposure to examples instead of explicit programming. The output of a generative AI model is often refined through a process called reinforcement learning with human feedback (RLHF) – meaning that the model continuously refines itself on the basis of whether humans accept or reject its outputs.

A type of neural network model called a transformer forms the backbone of most modern generative AI models. Transformers can be massively parallelized and scaled, process inputs in a non-sequential manner (i.e. analyzing sentences, paragraphs and essays holistically instead of word-by-word) and support positional embeddings, an essential characteristic of modern large language models (LLMs).

Off-the-shelf transformer-based models trained on huge volumes of publicly available data are also known as base or foundation models. Most commercial applications of generative AI combine foundation models built by a third party with an organization's proprietary data.

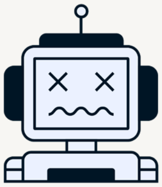
What are large language models?

Large language models (LLMs) are a type of generative AI that is trained on and produces text in response to prompts. This output typically takes the form of natural language, though LLMs can be used to generate text that follows other patterns, such as code, as well.

The "largeness" of an LLM comes from both the enormous volumes of data involved as well as the complexity of the model. Models such as ChatGPT are trained on petabytes of data collected from across the internet, consisting of trillions of tokens, or unique units of text (i.e. punctuation, characters, parts of words, words, phrases and more). These tokens are assigned unique identifiers and arranged into embeddings, or positions, in a massive, multidimensional space near positions of semantically and contextually similar tokens.

What generative AI is not

Despite its ability to mimic human language and other creative or intellectual output, a generative AI model fundamentally does not have any consciousness, emotions, subjectivity or semantic, human-like understanding of the data it ingests. Under the hood, like all machine learning, generative AI is just linear algebra, statistics and computation on a gargantuan scale.



1

Data prerequisites for generative AI

Generative AI depends on a foundation of data maturity, in which an organization demonstrates mastery over both integrating data – moving and transforming it – and governing its use. Without data maturity, the prototyping, deployment and testing of generative AI – or indeed, any kind of analytics – becomes extremely difficult.

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, lake or lakehouse – that can serve as a single source of truth
- A tool that reliably and automatically ingests data from sources at scale and features:
  – Fast, timely updates
  – Reliability and the ability to quickly recover from failures
- A tool or capability that supports collaborative, version-controlled modeling and transforming of data
- Data governance capabilities such as:
  – The ability to block and hash sensitive data before it arrives in a central repository
  – Access control
  – The ability to catalog data
  – Automated user provisioning

Automation is the common theme of all of these tools and technologies. It is an essential prerequisite for efficient, reliable and scalable data movement and integration. Much like generative AI itself, automated data movement is a formidable force multiplier, freeing up your engineering resources to make far greater analytical or operational impacts.

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

1. A scaled analytics organization where, in addition to a core team of analysts, you also have domain experts assigned to specific functional units within your organization.
2. Reports issued on a regular cadence, and stakeholders in your organization who access and regularly use dashboards to support decisions
3. Product thinking in analytics, in which the reports, dashboards, models etc. that your team builds are tailored to the needs of stakeholders
4. Good visibility into your data, as exemplified by cataloging of data assets

These practices and structures demonstrate that your team effectively and responsibly handles data, using it to support decisions.

“

Right now, when digging deeper (into the details of corporate generative AI projects), you find more experiments and less production use cases.[11]

Naveen Zutshi, CIO at Databricks

”

Productionizing AI remains a common and serious challenge because organizations often haven’t laid an adequate foundation for it. Together, these technological and organizational capabilities indicate that your team is ready to hire the talent and assemble the infrastructure needed to adequately support generative AI.

Your data platform architecture for generative AI

In practice, it is very unlikely that you will build your own generative AI from scratch. Building a generative AI from scratch is a colossal undertaking, with the potential to cost hundreds of millions of dollars and the equivalent of hundreds of years of non-parallelized compute time. Both the scale of the data and the complexity of the models involved can be gigantic.

Your organization is most likely to use a base or foundation model – a commercially available model already trained on huge volumes of public data, giving it considerable multipurpose capabilities – and then further train it with your organization’s bespoke data.

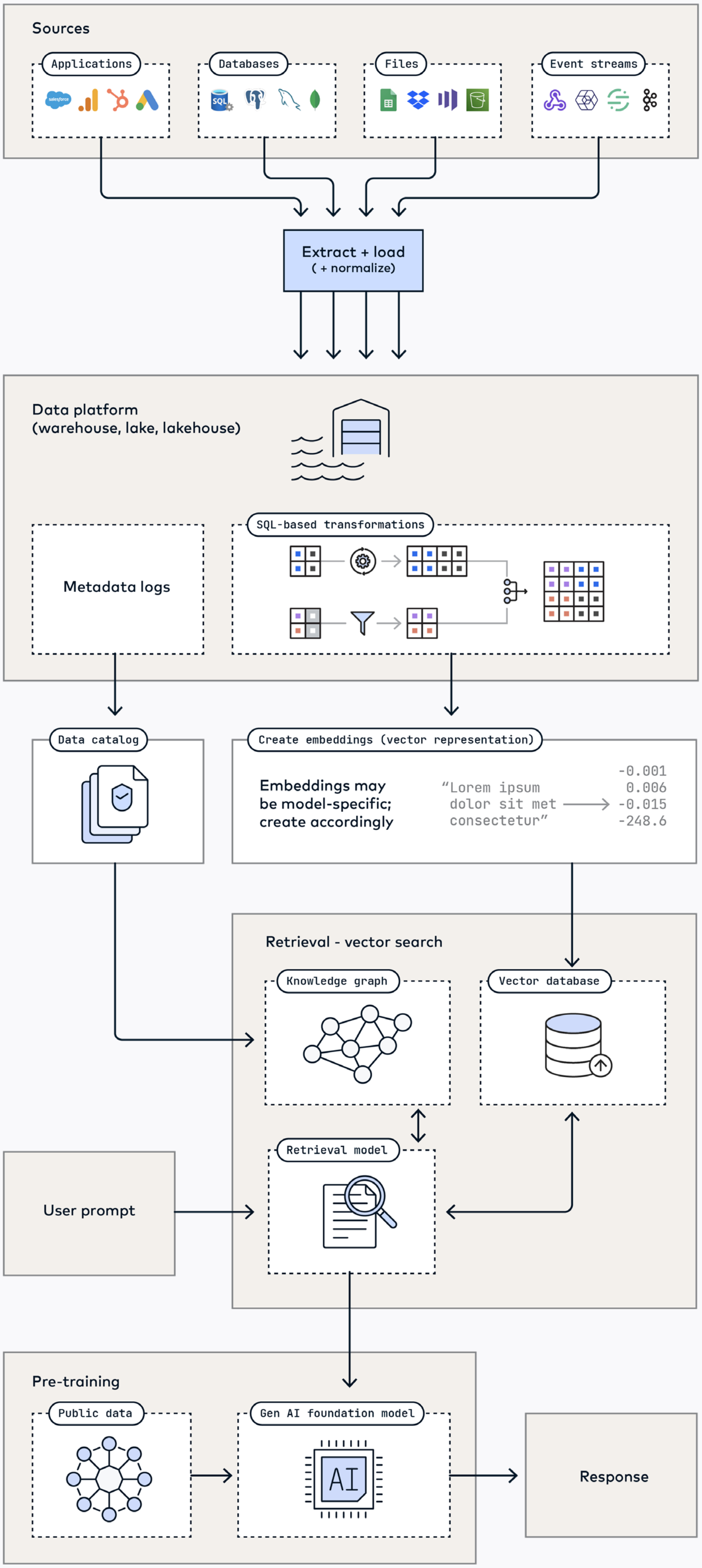
Like other forms of analytics, generative AI relies on that data being timely and of high quality. The architecture your organization will set up to support generative AI (in particular, a large language model) should look something like this:

See the full diagram on the next page →

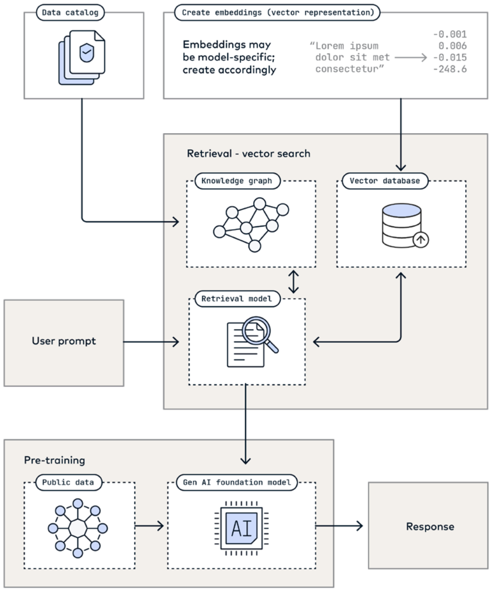
The first few stages of this architecture are shared in common with more basic analytics use cases. You need a pipeline to extract and load data from a wide range of sources to a destination, such as a data lake. Transformations should be performed at the destination to turn raw (or nearly raw) data into models that can be used to support reports, dashboards and other data assets.

What comes afterward is unique to generative AI. You can supplement an off-the-shelf generative AI model with your data in two ways:

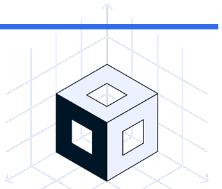
❶ The most important way is to turn long strings of text into enumerations and then store these in a vector database that your foundation model can access. This enables the generative AI to add your organization’s data to its long-term memory, enabling it to return meaningful results both from the general knowledge provided through its initial training and your organization’s unique data.


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

2. The second way piggybacks off of any data catalog you might construct for data governance. You can combine large language models with knowledge graphs, explicitly encoding semantic understanding into the model, not just statistical word associations.


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

What is a vector database?



A vector databases enables you to store your own data in a manner interpretable by a foundation model. Vector databases specialize in the storage and retrieval of "embeddings" – high-dimensional representations of text, images, audio, video and other media. Think of an embedding as a long list of numbers that represents coordinates. This coordinate system allows you to search and retrieve data on the basis of vector distance, or similarity. In a vector database for a large language model, tokens, or units of text, are grouped together on the basis of associative and semantic similarity.

Note that assembling this workflow may not (yet) be as straightforward as a simple plug-and-play. Even with the help of an increasing number of off-the-shelf tools for managing data infrastructure with generative AI, it is likely that you will need to lean heavily on engineering, data science and AI expertise to make the parts function properly with each other and build usable applications on top of the architecture.

You can, however, ensure that the expert talent you hire is deployed effectively by first building a solid technological and organizational foundation for your data operations. Once you have laid an adequate technological and organizational foundation for generative AI, you may hire or train data scientists, AI researchers and machine learning engineers in earnest. Look for the following skills:

- Prototyping, productionizing and tuning machine learning models, including neural networks
- Prompt design/engineering and verifying the output of artificial intelligence
- Understanding and mitigating security and ethical concerns

How generative AI will change the nature of work

The simplest way generative AI, particularly large language models, can boost productivity is in the same role as search engines: surfacing critical information for users but with a greater ability to simulate human-like semantic understanding of its output. In this regard, generative AI is a general-purpose productivity aid that makes quality information freely available to those who want it.

A more powerful application of generative AI is to leverage the unique data and context associated with an organization from its operations. All organizations produce a huge corpora of text through contracts, blogs, call transcripts, chat applications, project management tools, emails and internal documentation of all kinds.

A large language model trained on such a corpus can meaningfully answer domain-specific questions, summarize text, translate between languages, adjust tone, extract issues, themes and sentiments and more. In effect, a large language model with access to an organization's accumulated data can act as the most knowledgeable "member" of an organization.

Practical, general examples of how this capability can support a business include:

Supporting customers, whether as a fully automated chatbot or helping human customer service representatives access obscure information or troubleshoot issues

Shortening turnaround time to create sales and marketing content, including media of all kinds – written collateral, images, animations and more

Accelerating the software engineering process by generating boilerplate code or translating between programming languages

Rapidly brainstorming and prototyping new products and concepts

There are countless potential industry-specific use cases, as well. Generative AI may be able to:

Automatically produce financial documentation

Enable retail customers to virtually “try on” clothing and accessories

Accelerate the discovery of new drugs by pharmaceutical researchers by uncovering new configurations of proteins and other molecular structures

Accelerate legal and paralegal work, such as legal research, drafting documents and communication with clients

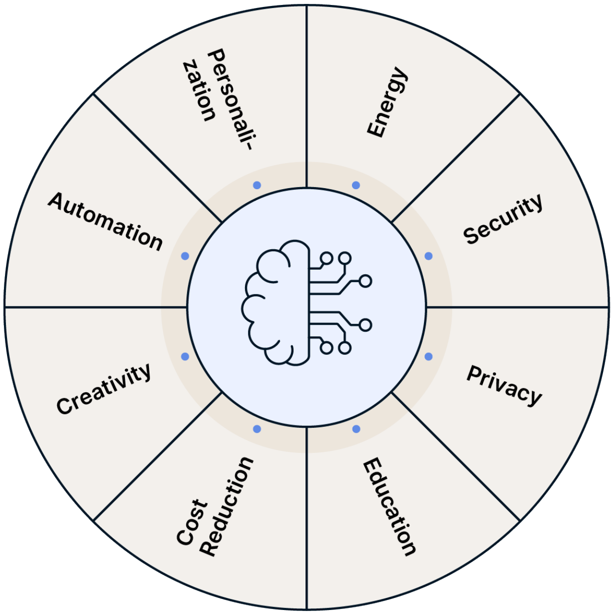
Enable doctors to more readily diagnose rare and unusual diseases and devise personalized treatment plans

Help aerospace engineers optimize the shape of new airframes

Procedurally generate CGI assets for video games, films and other digital media

Help educators create personalized lesson plans and instructional materials

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



Although generative AI will primarily affect occupations that heavily engage in intellectual or creative work, there is growing evidence that, within those occupations, relatively lower performers stand to benefit the most.[12]

In other words, generative AI stands to raise the average level of performance at a given organization.

3

How to use generative AI

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

Crafting good prompts is fundamentally about clearly communicating to the model 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 often provides 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.

Learn more →

1. **Give directions –** Add adjectives, context and other descriptors and guidance about your intended output. This includes including the user's persona in the prompt – you can tell the model what to emphasize and what kind of detail to offer. Models such as ChatGPT will often ask you to do so if you provide an especially sparse prompt. The naive approach to writing a prompt is also known as zero shot prompting, in which a request is made of a model without enumerating any explicit examples, guidance or broader context. This forces the model to generalize from preexisting knowledge, with a high likelihood of producing spurious 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 –** Show or enumerate to the model the format of the response you want. If you need information presented as a numbered list or bullet points, it doesn't hurt to 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 should be treated as an iterative process. The "engineering" side of prompt engineering is a matter of systematically testing, evaluating and iterating through different configurations of prompts.

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
```

Poor prompt design won't just return misleading or meaningless results. Poor – and malicious – prompting can lead a model to expose confidential information in its response. It can also generate responses that describe or endorse antisocial behavior. Microsoft's Tay chatbot famously became a bigoted monster within hours of deployment after interactions with malicious users.[13] Even worse, hacks known as prompt injections can be used to misdirect a model and produce any response the hacker wants, including social engineering such as 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. Yet 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.

Minimizing hallucinations

Earlier, we discussed the importance of evaluating outputs. A well-known pitfall of generative AI is that it can produce spurious results that aren't factual or connected to reality in any way (we've all seen the AI-generated pictures of people with an improbable number of fingers). The stakes can be high; there has already been at least one documented instance of a lawyer citing hallucinated court cases.[14] There are several ways to minimize the frequency and impact of hallucinations.

- The first is to simply fact-check results and reject bad ones. Recall that generative AI models learn through reinforcement; one approach may be to keep humans in the loop to extensively test the answers provided by a model until it meets some threshold for accuracy and relevance before any system built on top of a generative AI model is released to the public or used to support any serious decisions.

- The second approach is to ensure a high-quality upstream training set by carefully curating and governing the data your organization provides to its foundation model. You don't have to put all of your organization's data into a vector database to feed your foundation model. Instead, you should carefully consider the problem you are trying to solve using generative AI and select for 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.[15]

- A third approach is to supplement the foundation model with a knowledge graph that explicitly encodes real, semantic relationships between concepts. This enforces explainability.

  You should pick a foundation model that was trained on a similar base of data as your own data. There are many publicly available models that specialize in different media (text, images, audio, etc.), domains (geospatial analysis, music, etc.) and use cases (sentiment analysis, data visualization, etc.).

Protect your generative AI

There are several general principles to safeguarding your organization's use of generative AI. One principle is fundamentally an upstream data governance and security issue, and mainly concerns anonymizing, encrypting or altogether excluding sensitive data from the training set. This is a prerequisite for all lawful and ethical uses of data, per GDPR and other legal and regulatory requirements.

Knowledge graphs provide an opportunity to explicitly encode safeguards into a generative AI model. A knowledge graph can provide a model with access control policies and flag data that is too sensitive to safely expose.

Another principle is to maintain control over the prompts themselves. Rejecting or sanitizing prompts that include certain phrases or concepts before they can reach the model is one way to prevent obvious misuse.

The mirroring principle is to screen the output of a model, ensuring that end users are never exposed to outputs that contain sensitive data, contain abusive or antisocial messages, etc. In both cases, the screening can be performed through human review, a rule-based system or even another AI trained specifically to classify and evaluate content.

Finally, generative AIs are ultimately trained and improved over time with human supervision in the loop. Allowing end users of a generative AI system to report bad results will improve your model's efficacy and safety.

4

Set your team up for success

Generative AI is here to stay and promises a rich and innovative future. This ebook has described how you can help your organization lay a solid data foundation for generative AI. Automation and governance are essential not only to success in generative AI but in analytics more broadly.

But don’t get ahead of yourself – previous hype cycles involving data science and big data led many organizations to set unrealistic expectations and hire talent without laying the proper groundwork.

This groundwork depends heavily on technological capabilities that emphasize automation and governance as well as organizational capabilities that reflect strong norms around the effective use of data. With tools that automate data movement and integration end to end and enable governance as well as a data team well-versed in effectively using data to support decisions, your organization will be prepared to hire the specialists needed to turn artificial intelligence concepts into reality.

Besides the importance of putting together technological and human resources in the right sequence, it is also important to manage expectations. For instance, the costs of not only setting up generative AI and building out the infrastructure but running it on an ongoing basis are projected to be immense.[16] Beyond governance and security concerns, restricting access for your future models may simply be a financial necessity.

There are also many cases where generative AI may not be the appropriate response to a particular problem. Common business problems can often be solved using rule-based heuristics or much cheaper and simpler predictive modeling approaches, like linear regression. For instance, you can perform financial forecasting using methods such as moving averages, regression analysis or the Delphi method.[17] In fact, given the costly and complex nature of generative AI, a good rule of thumb is to exhaust every other option before resorting to it.

As illustrated by generative AI mishaps like bigoted chatbots, hallucinated court cases and missed COVID diagnoses, we are still a long way from simply entrusting machines without humans in the loop for high-stakes applications. 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.

1. Salesforce. (2023, September). Top Generative AI Statistics for 2023. Retrieved from https://www.salesforce.com/news/stories/generative-ai-statistics/

2. 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

3. Krenn, M. (2022, September 23). Predicting the Future of AI with AI: High-quality link prediction in an exponentially growing knowledge network. arXiv.org. https://arxiv.org/abs/2210.00881

4. Odom, E. (2023). 78% of CEOs and boards believe AI benefits outweigh associated risks. The National CIO Review. https://nationalcioreview.com/articles-insights/technology/artificial-intelligence/78-of-ceos-and-boards-believe-ai-benefits-outweigh-associated-risks/

5. Chui, M., Hazan, E., Roberts, R., Singla, A., Smaje, K., Sukharevsky, A., Yee, L., & Zemmel, R. (2023). 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

6. Oldfield, J. (2023, May 4). Researchers use generative AI to design novel proteins. Phys. https://phys.org/news/2023-05-generative-ai-proteins.html

7. Gupta, A. (2023, August 29). How 3 healthcare organizations are using generative AI. Google. https://blog.google/technology/health/cloud-next-generative-ai-health/

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

9. Lin, B. (2023, June 8). Rush to use generative AI pushes companies to get data in order. WSJ. https://www.wsj.com/articles/rush-to-use-generative-ai-pushes-companies-to-get-data-in-order-c34a7e13?mod=djemCIO

10. Staff. (2019, July 22). Why do 87% of data science projects never make it into production? VentureBeat. https://venturebeat.com/ai/why-do-87-of-data-science-projects-never-make-it-into-production/

11. Ashare, M. (2023, October 16). CIOs prioritize data upgrades as AI adoption intensifies. CIO Dive. https://www.ciodive.com/news/Databricks-MIT-Tech-Review-enterprise-AI-data-strategy/696608/

12. Smith, N. (2023, September 4). Is it time for the Revenge of the Normies? Noahpinion. https://www.noahpinion.blog/p/is-it-time-for-the-revenge-of-the

13. 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-chatbot-after-it-turned-into-racist-nazi/

14. 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

15. 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/

16. Vanian, J., & Leswing, K. (2023, April 17). ChatGPT and generative AI are booming, but the costs can be extraordinary. CNBC. https://www.cnbc.com/2023/03/13/chatgpt-and-generative-ai-are-booming-but-at-a-very-expensive-price.html

17. 7 Financial forecasting methods to predict business performance. (2022, June 21). Business Insights Blog. https://online.hbs.edu/blog/post/financial-forecasting-methods

Fivetran

Fivetran automates data movement out of, into and across cloud data platforms. We automate the most time-consuming parts of the ELT process from extract to schema drift handling to transformations, so data engineers can focus on higher-impact projects with total pipeline peace of mind.

With 99.9% uptime and self-healing pipelines, Fivetran enables hundreds of leading brands across the globe, including Autodesk, Conagra Brands, JetBlue, Lionsgate, Morgan Stanley, and Ziff Davis, to accelerate data-driven decisions and drive business growth.

For more info, visit Fivetran.com.