Gartner®

Quick Answer: Options for Using Your Data With Generative AI Models

Published 31 August 2023 - ID G00796708 - 7 min read

Svetlana Sicular

Initiatives: Data and Analytics Programs and Practices; Evolve Technology and Process Capabilities to Support D&A

Data and analytics leaders are asking how to use their own data to take advantage of generative pretrained AI models. This research discusses three options — all with different levels of resource, timing, price points and trade-offs — for D&A leaders to use their own data with these models.

Quick Answer

What are the options for using your data with generative AI models?

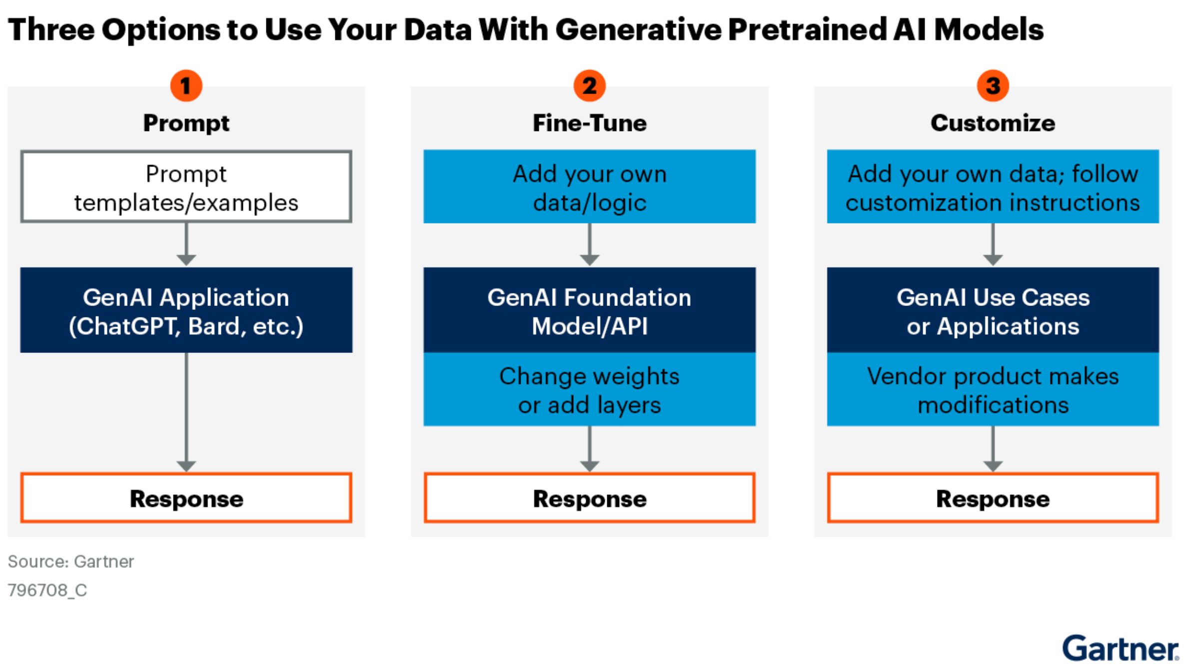
- Prompt generative pretrained AI models via applications like ChatGPT and Bard. Generative AI (GenAI) models produce outputs based on inputs — or "prompts" — that are natural language instructions, which could contain small amounts of your data.
- Fine-tune generative pretrained models, like GPT-3 or Pathways Language Model (PaLM), that are available via APIs, typically by following particular API instructions for preparing, uploading and manipulating your data. Fine-tuning involves changing weights of a model or adding extra layers to a model to accommodate your data specifics.
- Customize prepackaged off-the-shelf applications and use cases that involve generative pretrained models by adding your data and logic. Typically, providers of such solutions already performed prompt engineering, fine-tuning or even full training of a model. Customers follow vendor guidelines to use their data and logic in the solutions, such as search, coding assistance, chatbots, finance, communications or life science use cases.

More Detail

Data and analytics leaders ask how they can take advantage of powerful, generative pretrained models by using them with their own data. These models are also referred to as “foundation models”; their most popular type is large language models (see Board Briefing: Understanding ChatGPT, Other Large Language Models and Their Risks). There are three basic options to use generative pretrained models with your own data: prompting, fine-tuning or customizing specific applications provided by vendors (see Figure 1). The options vary in pros and cons for complexity, cost, time and data/model fit.

The GenAI landscape is rapidly changing, and more options to use your own data will be available in the future. Currently, the alternative of training large GenAI models from scratch is too expensive and complex for most organizations.

Figure 1: Three Options for Using Your Data With Generative Pretrained AI Models



Meanwhile, data and analytics leaders should focus on their own data readiness for AI given that more and more models come off the shelf, and given that organizational data is always a differentiator for meaningful AI results.

Data readiness will benefit not just generative pretrained AI models, but any AI implementation.

Most organizations express security, privacy and intellectual property concerns when using their own data. Although commercial and some open-source providers of generative pretrained models offer data confidentiality, data privacy and data security approaches, make sure your requirements are met by engaging your security, privacy and legal experts.

According to the 2022 Gartner AI Use-Case ROI Survey, [1] engaging legal experts at the stage of ideating AI use cases is a common practice in mature AI organizations. Currently, higher security and privacy options further increase costs of GenAI. For more information, see 4 Ways Generative AI Will Impact CISOs and Their Teams.

Option 1: Use Your Data via Prompts

Prompts allow users some control over the generated output by providing specific instructions or context. By providing relevant information, users can set the context, specify the desired format or guide the model's behavior. The new field of prompt engineering is dedicated to optimizing these instructions for effectiveness. In other words, users can give context or examples that include their own data, formats and templates as a prompt to get the best possible response.

For instance, to translate an organization-specific text, a user can provide a translation vocabulary for your own terms and abbreviations. Crafting prompts that yield accurate results requires iteration and experimentation. When the enterprise has the accurate data it wants an LLM to use in constructing its output, the popular way of prompt engineering is retrieval-augmented generation (RAG). For detailed information, refer to Prompt Engineering With Enterprise Information for LLMs and GenAI.

- **Pros:** Low in cost; medium skills requirements; low resource consumption; high speed of getting outputs.
- **Cons:** Limited size of prompts, black-box nature of the generated response; static data that might not reflect the most current information; low security if using free or inexpensive GenAI versions.

- **Recommendations:**
  - Give examples with your own context and data.
  - Experiment with examples — give examples of everything.
  - Use very detailed prompts.
  - Provide predefined prompts for beginners to accelerate their productivity with GenAI.

Option 2: Fine-Tune With Your Own Data

Fine-tuning is a process where a pretrained model is further trained on a new dataset to create a new copy of the model with a few modified layers or some adjusted underlying model weights. Typically, fine-tuning is done with the generative pretrained models that are available via APIs, such as GPT-3 and PaLM. Major commercial vendors are rapidly introducing new fine-tuning approaches, so it is important to refer to the vendor documentation for the latest information. Open-source providers such as Hugging Face, Stable Diffusion and MosaicML (Databricks) also offer the ability to fine-tune a model on your own data. The open-source LoRA (Low-Rank Adaptation of Large Language Models) is gaining popularity for fine-tuning LLMs, but it has limitations on what models can be tuned.

Fine-tuning allows organizations to accurately tailor a pretrained model to their own contexts and tasks. Fine-tuning via API typically relies on large amounts of labeled instances, which could be expensive and time-consuming to obtain. Depending on the fine-tuning approach, which is typically provided by a particular vendor, organizations can choose to use responses that are generated using only their own data or a mix of their own data and the data used in the pretrained model.

- **Pros (vs. prompting):** More examples in prompts (hundreds to thousands in fine-tuning versus tens in prompting); more accurate results for your data; could be a longer prompt size.
- **Cons (vs. prompting):** Higher cost; longer time to deliver; higher skills requirements.

- **Recommendations:**
  - To achieve accurate and productive results, combine business and technical skills.
  - Envolve data scientists (preferably with deep learning experience), DataOps engineers for data preprocessing and subject matter experts for content and output validation.
  - Ensure thorough testing.
  - Verify model licensing for fine-tuning — not all models are available for commercial use.

Option 3: Customize Prepackaged Generative AI Applications

The customization of GenAI use case assumes that vendors offer a prepackaged use case or a product that incorporates generative pretrained models. In essence, vendors use either prompt engineering or fine-tuning techniques to deliver a capability to the enterprise. In this case, customers follow vendor instructions on uploading data and customizing the solution. This is a convenient entry point for those organizations that want to reap the value of a specific use case, but that lack experience with GenAI.

Examples of prepackaged solutions include:

- Coding assistants (GitHub Copilot, Google Codey and Amazon CodeWhisperer)
- Text communications (Grammarly, Google Docs and Microsoft 365 Copilot)
- Life sciences (Huma.AI)
- Creative imaging use cases (Adobe Firefly).

GenAI models for particular use cases are available on major model hubs: Hugging Face Hub, Microsoft Azure OpenAI, Google Vertex AI Extensions and Amazon Bedrock. The cost of such solutions varies from very low to very high.

- **Pros:** Customization is a good starting point for GenAI adoption; smooth learning curve; fast results on your own data; taking advantage of vendor's experience and knowledge.

- **Cons:** Limited number of GenAI use cases; vendor variability involves a variety of approaches to using your own data, so these approaches and knowledge might not be transferable from one vendor to another; cost could be on the high side.
- **Recommendations:**
  - Start with the vendors with which you have ongoing relationships. Many of them have already incorporated some GenAI solutions to improve or simplify their products. For example:
    - Products may use simple human-language questions instead of SQL queries.
    - Products may have swapped older conversational models to LLMs, which made their communication richer and more engaging.
  - Select a coding assistant on your preferred development platform, rather than the "coolest" one — it will better suit your skillset.
  - Verify vendors' claims for quality, security and accuracy of their solutions.

Recommended by the Authors

Quick Answer: How Will Prompt Engineering Impact the Work of Data Scientists?

Glossary of Terms for Generative AI and Large Language Models

Innovation Insight for Artificial Intelligence Foundation Models

How to Choose an Approach for Deploying Generative AI

How to Pilot Generative AI

Evidence

[1] **2022 Gartner AI Use Case ROI Survey.** This survey sought to understand where organizations have been most successful in deploying AI use cases and figure out the most efficient indicators that they have established to measure those successes. The research was conducted online from 31 October through 19 December 2022 among 622 respondents from organizations in the U.S. (n = 304), France (n = 113), the U.K. (n = 106) and Germany (n = 99). Quotas were established for company sizes and for industries to ensure a good representation across the sample. Organizations were required to have developed AI to participate. Respondents were required to be in a manager role or above and have a high level of involvement with the measuring stage and at least one stage of the life cycle from ideating to testing AI use cases. Disclaimer: The results of this survey do not represent global findings or the market as a whole, but reflect the sentiments of the respondents and companies surveyed.