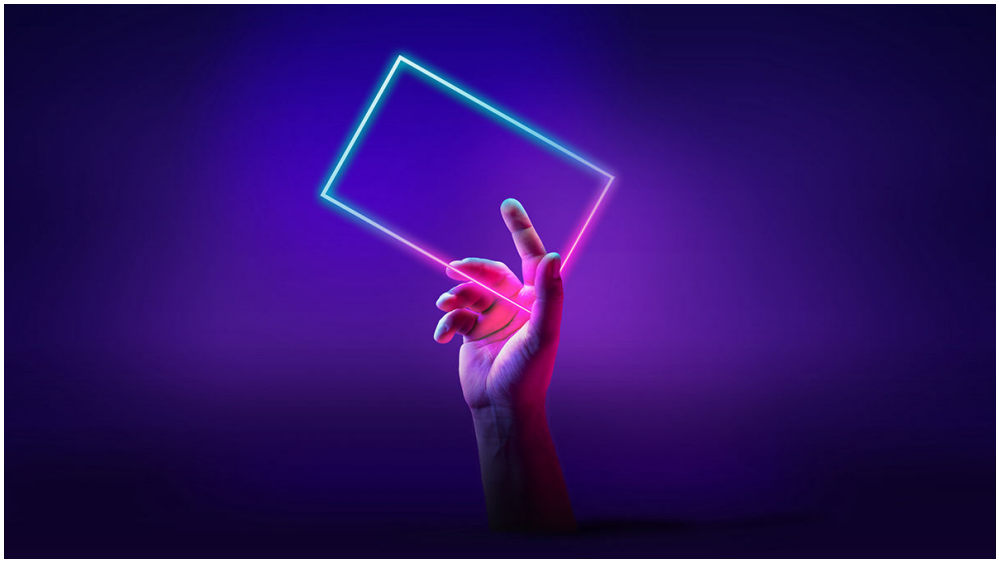
Harvard Business Review | Digital Article

Technology and Analytics



How to Train Generative AI Using Your Company's Data

Three ways companies like Bloomberg, Google, and Morgan Stanley are embedding their intellectual capital into large language models.

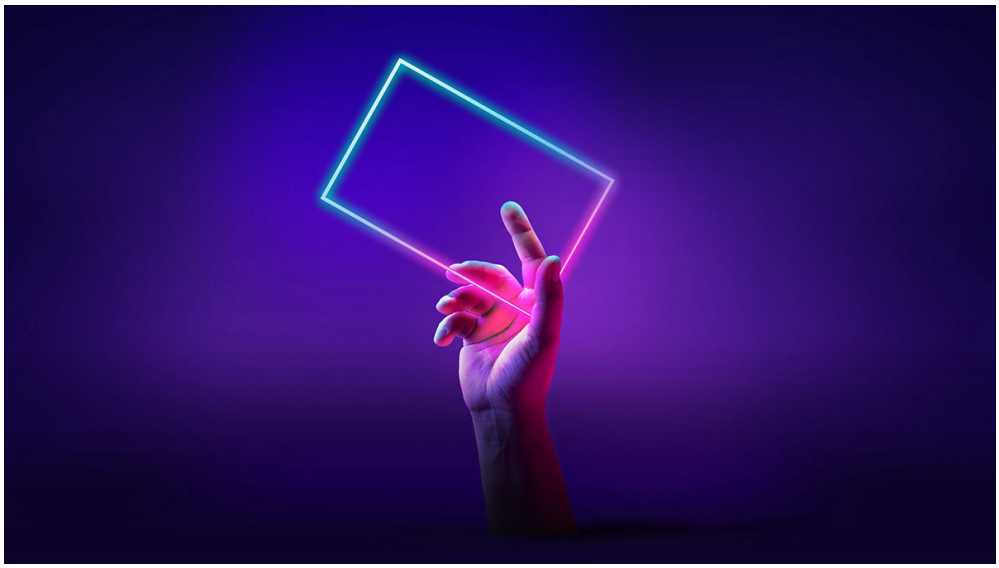
by Tom Davenport and Maryam Alavi

How to Train Generative AI Using Your Company's Data

Three ways companies like Bloomberg, Google, and Morgan Stanley are embedding their intellectual capital into large language models.

by Tom Davenport and Maryam Alavi

Published on HBR.org / July 06, 2023 / Reprint H07PNV



Anton Vierietin/Getty Images

Many companies are experimenting with ChatGPT and other large language or image models. They have generally found them to be astounding in terms of their ability to express complex ideas in articulate language. However, most users realize that these systems are primarily trained on internet-based information and can't respond to prompts or questions regarding proprietary content or knowledge.

Leveraging a company's propriety knowledge is critical to its ability to compete and innovate, especially in today's volatile environment. Organizational Innovation is fueled through effective and agile creation, management, application, recombination, and deployment of knowledge assets and know-how. However, knowledge within organizations is typically generated and captured across various sources and forms, including individual minds, processes, policies, reports, operational transactions, discussion boards, and online chats and meetings. As such, a company's comprehensive knowledge is often unaccounted for and difficult to organize and deploy where needed in an effective or efficient way.

Emerging technologies in the form of large language and image generative AI models offer new opportunities for knowledge management, thereby enhancing company performance, learning, and innovation capabilities. For example, in a study conducted in a Fortune 500 provider of business process software, a generative AI-based system for customer support led to increased productivity of customer support agents and improved retention, while leading to higher positive feedback on the part of customers. The system also expedited the learning and skill development of novice agents.

Like that company, a growing number of organizations are attempting to leverage the language processing skills and general reasoning abilities of large language models (LLMs) to capture and provide broad internal (or customer) access to their own intellectual capital. They are using it for such purposes as informing their customer-facing employees on company policy and product/service recommendations, solving customer service problems, or capturing employees' knowledge before they depart the organization.

These objectives were also present during the heyday of the "knowledge management" movement in the 1990s and early 2000s, but most companies found the technology of the time inadequate for the task. Today, however, generative AI is rekindling the possibility of capturing and disseminating important knowledge throughout an organization and beyond its walls. As one manager using generative AI for this purpose put it, "I feel like a jetpack just came into my life." Despite current advances, some of the same factors that made knowledge management difficult in the past are still present.

The Technology for Generative AI-Based Knowledge Management

The technology to incorporate an organization's specific domain knowledge into an LLM is evolving rapidly. At the moment there are three primary approaches to incorporating proprietary content into a generative model.

Training an LLM from Scratch

One approach is to create and train one's own domain-specific model from scratch. That's not a common approach, since it requires a massive amount of high-quality data to train a large language model, and most companies simply don't have it. It also requires access to considerable computing power and well-trained data science talent.

One company that has employed this approach is Bloomberg, which recently announced that it had created BloombergGPT for finance-specific content and a natural-language interface with its data terminal. Bloomberg has over 40 years' worth of financial data, news, and documents, which it combined with a large volume of text from financial filings and internet data. In total, Bloomberg's data scientists employed 700 billion tokens, or about 350 billion words, 50 billion parameters, and 1.3 million hours of graphics processing unit time. Few companies have those resources available.

Fine-Tuning an Existing LLM

A second approach is to "fine-tune" train an existing LLM to add specific domain content to a system that is already trained on general knowledge and language-based interaction. This approach involves adjusting some parameters of a base model, and typically requires substantially less data — usually only hundreds or thousands of documents, rather than millions or billions — and less computing time than creating a new model from scratch.

Google, for example, used fine-tune training on its Med-PaLM2 (second version) model for medical knowledge. The research project started with Google's general PaLM2 LLM and retrained it on carefully curated medical knowledge from a variety of public medical datasets. The model was able to answer 85% of U.S. medical licensing exam questions — almost 20% better than the first version of the system. Despite this rapid progress, when tested on such criteria as scientific factuality, precision, medical consensus, reasoning, bias and harm, and evaluated by human experts from multiple countries, the development team felt that the system still needed substantial improvement before being adopted for clinical practice.

The fine-tuning approach has some constraints, however. Although requiring much less computing power and time than training an LLM, it can still be expensive to train, which was not a problem for Google but would be for many other companies. It requires considerable data science expertise; the scientific paper for the Google project, for example, had 31 co-authors. Some data scientists argue that it is best suited not to adding new content, but rather to adding new content formats and styles (such as chat or writing like William Shakespeare). Additionally, some LLM vendors (for example, OpenAI) do not allow fine-tuning on their latest LLMs, such as GPT-4.

Prompt-tuning an Existing LLM

Perhaps the most common approach to customizing the content of an LLM for non-cloud vendor companies is to tune it through prompts. With this approach, the original model is kept frozen, and is modified through prompts in the context window that contain domain-specific knowledge. After prompt tuning, the model can answer questions related to that knowledge. This approach is the most computationally efficient of the three, and it does not require a vast amount of data to be trained on a new content domain.

Morgan Stanley, for example, used prompt tuning to train OpenAI's GPT-4 model using a carefully curated set of 100,000 documents with important investing, general business, and investment process knowledge. The goal was to provide the company's financial advisors with accurate and easily accessible knowledge on key issues they encounter in their roles advising clients. The prompt-trained system is operated in a private cloud that is only accessible to Morgan Stanley employees.

While this is perhaps the easiest of the three approaches for an organization to adopt, it is not without technical challenges. When using unstructured data like text as input to an LLM, the data is likely to be too large with too many important attributes to enter it directly in the context window for the LLM. The alternative is to create vector embeddings — arrays of numeric values produced from the text by another pre-trained machine learning model (Morgan Stanley uses one from OpenAI called Ada). The vector embeddings are a more compact representation of this data which preserves contextual relationships in the text. When a user enters a prompt into the system, a similarity algorithm determines which vectors should be submitted to the GPT-4 model. Although several vendors are offering tools to make this process of prompt tuning easier, it is still complex enough that most companies

adopting the approach would need to have substantial data science talent.

However, this approach does not need to be very time-consuming or expensive if the needed content is already present. The investment research company Morningstar, for example, used prompt tuning and vector embeddings for its Mo research tool built on generative AI. It incorporates more than 10,000 pieces of Morningstar research. After only a month or so of work on its system, Morningstar opened Mo usage to their financial advisors and independent investor customers. It even attached Mo to a digital avatar that could speak out its answers. This technical approach is not expensive; in its first month in use, Mo answered 25,000 questions at an average cost of $.002 per question for a total cost of $3,000.

Content Curation and Governance

As with traditional knowledge management in which documents were loaded into discussion databases like Microsoft Sharepoint, with generative AI, content needs to be high-quality before customizing LLMs in any fashion. In some cases, as with the Google Med-PaLM2 system, there are widely available databases of medical knowledge that have already been curated. Otherwise, a company needs to rely on human curation to ensure that knowledge content is accurate, timely, and not duplicated. Morgan Stanley, for example, has a group of 20 or so knowledge managers in the Philippines who are constantly scoring documents along multiple criteria; these determine the suitability for incorporation into the GPT-4 system. Most companies that do not have well-curated content will find it challenging to do so for just this purpose.

Morgan Stanley has also found that it is much easier to maintain high quality knowledge if content authors are aware of how to create

effective documents. They are required to take two courses, one on the document management tool, and a second on how to write and tag these documents. This is a component of the company's approach to content governance approach — a systematic method for capturing and managing important digital content.

At Morningstar, content creators are being taught what type of content works well with the Mo system and what does not. They submit their content into a content management system and it goes directly into the vector database that supplies the OpenAI model.

Quality Assurance and Evaluation

An important aspect of managing generative AI content is ensuring quality. Generative AI is widely known to "hallucinate" on occasion, confidently stating facts that are incorrect or nonexistent. Errors of this type can be problematic for businesses but could be deadly in healthcare applications. The good news is that companies who have tuned their LLMs on domain-specific information have found that hallucinations are less of a problem than out-of-the-box LLMs, at least if there are no extended dialogues or non-business prompts.

Companies adopting these approaches to generative AI knowledge management should develop an evaluation strategy. For example, for BloombergGPT, which is intended for answering financial and investing questions, the system was evaluated on public dataset financial tasks, named entity recognition, sentiment analysis ability, and a set of reasoning and general natural language processing tasks. The Google Med-PaLM2 system, eventually oriented to answering patient and physician medical questions, had a much more extensive evaluation strategy, reflecting the criticality of accuracy and safety in the medical domain.

Life or death isn't an issue at Morgan Stanley, but producing highly accurate responses to financial and investing questions is important to the firm, its clients, and its regulators. The answers provided by the system were carefully evaluated by human reviewers before it was released to any users. Then it was piloted for several months by 300 financial advisors. As its primary approach to ongoing evaluation, Morgan Stanley has a set of 400 "golden questions" to which the correct answers are known. Every time any change is made to the system, employees test it with the golden questions to see if there has been any "regression," or less accurate answers.

Legal and Governance Issues

Legal and governance issues associated with LLM deployments are complex and evolving, leading to risk factors involving intellectual property, data privacy and security, bias and ethics, and false/inaccurate output. Currently, the legal status of LLM outputs is still unclear. Since LLMs don't produce exact replicas of any of the text used to train the model, many legal observers feel that "fair use" provisions of copyright law will apply to them, although this hasn't been tested in the courts (and not all countries have such provisions in their copyright laws). In any case, it is a good idea for any company making extensive use of generative AI for managing knowledge (or most other purposes for that matter) to have legal representatives involved in the creation and governance process for tuned LLMs. At Morningstar, for example, the company's attorneys helped create a series of "pre-prompts" that tell the generative AI system what types of questions it should answer and those it should politely avoid.

User prompts into publicly-available LLMs are used to train future versions of the system, so some companies (Samsung, for example) have feared propagation of confidential and private information and banned LLM use by employees. However, most companies' efforts to tune LLMs

with domain-specific content are performed on private instances of the models that are not accessible to public users, so this should not be a problem. In addition, some generative AI systems such as ChatGPT allow users to turn off the collection of chat histories, which can address confidentiality issues even on public systems.

In order to address confidentiality and privacy concerns, some vendors are providing advanced and improved safety and security features for LLMs including erasing user prompts, restricting certain topics, and preventing source code and propriety data inputs into publicly accessible LLMs. Furthermore, vendors of enterprise software systems are incorporating a "Trust Layer" in their products and services. Salesforce, for example, incorporated its Einstein GPT feature into its AI Cloud suite to address the "AI Trust Gap" between companies who desire to quickly deploy LLM capabilities and the aforementioned risks that these systems pose in business environments.

Shaping User Behavior

Ease of use, broad public availability, and useful answers that span various knowledge domains have led to rapid and somewhat unguided and organic adoption of generative AI-based knowledge management by employees. For example, a recent survey indicated that more than a third of surveyed employees used generative AI in their jobs, but 68% of respondents didn't inform their supervisors that they were using the tool. To realize opportunities and manage potential risks of generative AI applications to knowledge management, companies need to develop a culture of transparency and accountability that would make generative AI-based knowledge management systems successful.

In addition to implementation of policies and guidelines, users need to understand how to safely and effectively incorporate generative AI capabilities into their tasks to enhance performance and productivity.

Generative AI capabilities, including awareness of context and history, generating new content by aggregating or combining knowledge from various sources, and data-driven predictions, can provide powerful support for knowledge work. Generative AI-based knowledge management systems can automate information-intensive search processes (legal case research, for example) as well as high-volume and low-complexity cognitive tasks such as answering routine customer emails. This approach increases efficiency of employees, freeing them to put more effort into the complex decision-making and problem-solving aspects of their jobs.

Some specific behaviors that might be desirable to inculcate — either though training or policies — include:

- Knowledge of what types of content are available through the system;
- How to create effective prompts;
- What types of prompts and dialogues are allowed, and which ones are not;
- How to request additional knowledge content to be added to the system;
- How to use the system's responses in dealing with customers and partners;
- How to create new content in a useful and effective manner.

Both Morgan Stanley and Morningstar trained content creators in particular on how best to create and tag content, and what types of content are well-suited to generative AI usage.

"Everything Is Moving Very Fast"

One of the executives we interviewed said, "I can tell you what things are like today. But everything is moving very fast in this area." New LLMs and new approaches to tuning their content are announced daily,

as are new products from vendors with specific content or task foci. Any company that commits to embedding its own knowledge into a generative AI system should be prepared to revise its approach to the issue frequently over the next several years.

While there are many challenging issues involved in building and using generative AI systems trained on a company's own knowledge content, we're confident that the overall benefit to the company is worth the effort to address these challenges. The long-term vision of enabling any employee — and customers as well — to easily access important knowledge within and outside of a company to enhance productivity and innovation is a powerful draw. Generative AI appears to be the technology that is finally making it possible.

This article was originally published online on July 06, 2023.

Thomas H. Davenport is the President's Distinguished Professor of IT and Management at Babson College, a research fellow at the MIT Center for Digital Business, co-founder of the International Institute for Analytics, and a Senior Advisor to Deloitte Analytics. He is author of the new book *Big Data at Work* and the best-selling *Competing on Analytics*.

Maryam Alavi *is the Elizabeth D. & Thomas M. Holder Chair & Professor of IT Management, Scheller College of Business, Georgia Institute of Technology.*

