DATASTAX

WHITE PAPER

Generative AI: You Can't Afford to Wait

What business leaders need to know about this technology revolution—and how they can take advantage of it to drive business results

Read this guide to gain a clearer understanding of autonomous agents, large language models, vector databases—and how the technology is available right now to build generative AI applications that impact your business.

The internet. Open source software. The cloud. Mobile. Each of these technology revolutions created a future that was radically different from the past, in very significant ways.

Generative AI (artificial intelligence) promises to be even more significant than any of these. It is already revolutionizing content creation and promises to redefine how businesses drive innovation, interact with customers, gather market intelligence, train employees, do market research, and even build software.

Consider this: A recent McKinsey report identified 63 generative AI use cases across 16 business functions that could deliver total value in the range of $2.6 trillion to $4.4 trillion in economic benefits annually. By way of comparison, the UK's entire GDP in 2021 was $3.1 trillion.

Very soon, "using AI" will be akin to "using the internet." Every function and persona in your organization—and among your customers—will be intimately involved with autonomous AI agents to make them more productive, or to find what they are looking for, or to accomplish their goals.

Most business leaders grasp the importance of infusing generative AI into their organizations and into their customer-facing applications, but many aren't sure of how to get started and turn all this AI excitement into impact and business results. There are some strategies that clearly won't work: a "wait and see" approach will result in losing ground to organizations that are diving into generative AI right now. Another anti-pattern: turning AI-driven productivity gains into headcount reductions; instead, it should be a tool to redefine how every employee works and to reshape customer interactions.

AI is becoming pervasive today – and the pace of new innovations in AI technology is accelerating at a ferocious pace. The way to make the most of these advances is to accelerate the "pace of ingenuity" among your people to match. The reward for getting this right gives you an edge at creating new experiences and increasing the scope and scale of your data, thereby expanding the horizon for what it is possible for your people to envision and deliver next.

And there's more good news: the technology is available today to start taking advantage of generative AI in your organization. From something as simple as employing ChatGPT or other ready-to-use generative AI technologies to accelerate content creation, to securely using your organization's proprietary data to build chatbots that can answer questions about employee statistics or sales leads, the opportunity is here now to view each discipline at your organization, be it HR, finance, sales, marketing, as a new use case to put generative AI to work and drive productivity and accelerate creativity.

Keep reading to learn the important terms and concepts behind what is potentially the biggest technology revolution of our lives; grasp the importance of autonomous agents and powering them with the right data; and how feasible it has become to launch and accelerate your organization's generative AI efforts to drive real, sustainable business impact.

The language of GenAI: LLMs, agents, and vectors

ChatGPT reached 100 million users faster than any other application preceding it – it did so in under three months! Back in January 2023, when this AI tool first started catching everyone's attention, it was garnering 13 million visitors per day.

In a very short amount of time, thanks in large part to the impressive results and broad accessibility of ChatGPT, generative AI has caught fire. At the center of all this excitement is the innovation made possible by large language models, or LLMs.

Large language models

An LLM is a kind of algorithm that harnesses deep learning and huge datasets to understand, summarize, and generate new content. In essence, it processes natural language inputs and predicts the next words based on what it's already seen–until it provides a complete answer. ChatGPT is the generative AI interface, or "agent," that's built atop GPT-4, which is an LLM. Interfaces like ChatGPT enable us to query LLMs with questions or "prompts" and can produce text that is very similar to what a human would write.

When some people hear "agent" in the context of AI, they think about the simple chatbot that appears as a pop-up window that asks how it can help when they visit an e-commerce site. But LLMs and the autonomous agent interfaces can do much more than respond with simple conversational prompts and answers pulled from an FAQ.

One important characteristic of LLMs: they don't have memory. In other words, it can't remember the context it was provided with previous interactions. But with access to the right data—and lots of it—applications built on LLMs can "remember" past interactions that help drive very advanced, conversational ways to interact with us that deliver expertly curated information that is more useful, specific, rich — and often uncannily prescient (see sidebar).

What an autonomous agent can do today (this isn't science fiction)

It's finally summer, and you want to start building that deck for the backyard parties you've been dreaming of all winter. You open the mobile app provided by your favorite hardware retail chain, and ask it to build you a shopping list. It comes back with the materials you need, their prices, and their availability.

But what if it could do more? What if it asks you the dimensions of your deck and the features you want to include, and then offers up some visualizations of what your project could look like?

What if, based on your postal code, the application suggests nearby contractors who could help you with the job. What if, because of its access to construction code data, it tells you that your project requires a permit, helps you get one—and then helps you schedule a visit by a building inspector. What if the app estimated the amount of time it will take the project to be completed will take deck stain to dry (even including the seasonal climate trends for where you live) and how long it'll be until you can actually have that party on your deck that you've been planning.

This isn't science fiction. Building applications that offer the kind of customer engagement that autonomous AI agents can supply is possible *right now* (keep reading to learn how).

For an entertaining and eye opening dive into this topic, check out "The Complete Beginner's Guide to Autonomous Agents."

Vectors

There can be no AI without data—and not just any kind of data.

For LLMs to "understand" words, they need to be stored as text "vectors," which are sets of numerical data that capture the meanings and usage patterns of words. Vectors are, you might say, the lingua franca of AI.

Vectors have been around for a while, but with the popularity and accessibility of the generative AI interface ChatGPT, they've become the essential ingredient to enable building business-critical applications and services atop LLMs. The most useful applications that organizations will build with these technologies will leverage their own private data for LLMs by composing their own vectors.

Vector search

By representing data as vectors—a list of numbers that capture each item's key characteristics—organizations can efficiently search and compare very big datasets comprising multiple different data formats. Millions of customer profiles or images or articles that are represented as vectors can be combed through very quickly with "vector similarity search" (or "nearest neighbor search").

Unlike traditional keyword-based search, which matches documents based on the occurrence of specific terms, vector search focuses on the similarity of queries; for instance, are their semantic meanings similar?

Using the English language as an example, the words "happy", "cheerful", and "joyful" all have similar meanings, but in a traditional keyword-based search, documents matching "cheerful" and "joyful" would not be returned for a query of "happy". Vector search, however, understands meaning, enabling users to describe what they are searching for without the need to be exact.

In other words, this capability enables finding similar items based on their vector representations. Similarity search algorithms can measure the "distance" or similarity between vectors to determine how closely related they are.

Vector databases

To deliver the responsiveness and scale demanded by AI applications, vectors need to be stored in and retrieved from fast and scalable databases.

A critical requirement of a database that enables vector search is speed. Traditional databases have to compare a query to every item in the database. In contrast, integrated vector search enables a form of indexing and includes search algorithms that vastly speed up the process, making it possible to search massive amounts of data in a fraction of the time it would take a standard database.

In a business context, this is extremely valuable when using AI applications to recommend products that are similar to past purchases, or identify fraudulent transactions that resemble known patterns, or anomalies that look dissimilar to the norm.

There are a handful of databases today that offer vector search as a feature – and fewer still that offer the scalability and speed required to handle the extreme demands of generative AI (these demands will only grow as AI use cases spread throughout an organization and, consequently, more data needs to be stored as vectors).

One example is DataStax's Astra DB, which is built on the highly scalable, high-throughput, open source Apache Cassandra® data store. Cassandra has already been proven at scale to power AI by the likes of Netflix, Uber, and Apple.. The addition of vector search makes Astra DB a one-stop shop for high-scale database operations—enabling the storage of vectors in the same database as an organization's operational data, which leads to faster response times in real-time apps, for example.

Integrating vector search with an extremely scalable database like Astra DB enables calculations and ranking directly within the database, eliminating the need to transfer large amounts of data to external systems. This reduces latency and improves overall query performance.

This is particularly important because the LLMs that serve as the foundation for agents are "stateless" – they don't save data generated in one user session for use in the next session with that user. In other words, they have no memory. But LLMs and agents are truly powerful and useful when they can build their responses by referencing previous interactions.

A database that can handle storing huge amounts of vectors and enable queries against them essentially provides agents with that all-important memory.

Generative AI and vector search in action: SkyPoint Cloud

Generative AI can make a caregiver's job easier and more impactful—less time compiling policies and procedures means more time spent catering to patient needs. What used to be a task requiring hours of document searching can now be accomplished in seconds. This level of operational efficiency can result in significant cost savings and reduce manual and error-prone human labor.

SkyPoint Cloud Inc. uses Astra DB as a vector database to help transform senior living healthcare, an industry that is burdened with nearly 70% operational costs. With generative AI, the company ensures seamless access to resident healthcare data and insights for administrators." It's essentially a ChatGPT equivalent for senior living enterprise data that's fully HIPAA-compliant," says SkyPoint CEO Tisson Mathew. A high-speed vector database made a significant difference, he adds, stating that other vector stores were too slow to meet SkyPoint's requirements.

Astra DB's vector search capabilities ensure efficient retrieval of voluminous healthcare data, enable semantic understanding for contextual interpretation, allow personalized

recommendations based on vector similarities, and enhance AI capabilities supporting high-dimensional data models.

Finding specific policies and operating procedures is time-consuming and challenging due to scattered storage across various platforms. SkyPoint's generative AI technology streamlines this process by swiftly generating tailored care policies for each patient with citations, saving caregivers hours of work.

Working with generative AI at your organization

Now you have a basic understanding of many of the important components required for generative AI. But what's next?

Let's start with this strategic imperative to apply inside your organization: Look for AI opportunities in every business unit and functional organization at an enterprise.

Teams and functions might need support from an AI center of excellence or your central IT organization, but thinking of creative ways to use AI–and implementing them–should be everyone's job. Every employee should have the ability to test and learn how AI tools can be applied in their teams. (Check out this McKinsey podcast about using generative AI in HR for some intriguing examples.)

Cross-functional teams with AI expertise should exist within business units and functional organizations (This MIT Sloan Management Review article details this progression from experiments to organization-wide capabilities).

Building generative AI applications

As you hopefully gathered from the deck-building example earlier, building autonomous agents with public and proprietary data has become much simpler to do today, thanks to LLMs. Before generative AI became widespread, building AI applications was a laborious, complex process that entailed:

1. Finding and encoding your most relevant data as vectors
2. Training a machine learning model using those vectors
3. Deploying the model and exposing it as an API
4. Running the model in production.

Now, however, the technology has advanced to the point where you can build applications powered by generative AI without any of the steps above, and without the bespoke solutions that were previously required. It's well within the reach of most teams now—essentially, all you need to do is:

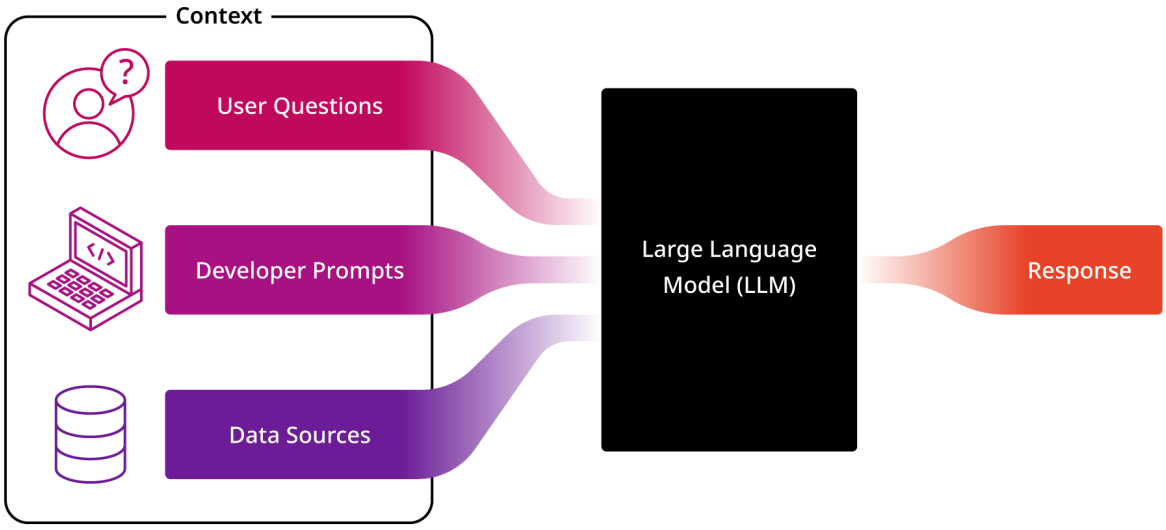
1. Get your data into an LLM
2. Describe the task you want the AI model to perform in English (this is your prompt)
3. Make an API call to an LLM API

(For a deeper, more detailed dive into building generative AI applications, read the guide, "Vector Search for Generative AI Apps.")

Generative AI projects (including the likes of enterprise search, video and audio search, and a host of other applications) no longer require vast knowledge about machine learning or data science or ML model training. In fact, building LLM-based applications requires little more than a developer who can make a database call and an API call. Building applications that can provide levels of personalized context that were unheard of until recently is a reality that can be realized with anyone who has the right database, a few lines of code, and an LLM like GPT-4.

LLMs are very simple to use. They take context (often referred to as a "prompt") and produce a response. So, building an autonomous agent starts with thinking about how to provide the right context to the LLM to get the desired response.

Broadly speaking, this context comes from three places: the user's question, the predefined prompts created by the agent's developer, and data sourced from a database or other sources (see the diagram below).



The context provided by the user is typically just the question they input into the application. The second piece could be provided by a product manager who worked with a developer to describe the role the agent should play (for example, “you’re a helpful sales agent who is trying to help customers as they plan their projects; please include a list of relevant products in your responses”).

Finally, the third bucket of provided context includes external data pulled from your databases and other data sources that the LLM should use to construct the response. Some agent applications may make several calls to the LLM before outputting the response to the user in order to construct more detailed responses. This is what technologies including ChatGPT plug-ins and LangChain facilitate.

Build a foundation for the future

How many of your employees don’t use a tool called “the Internet” as a window into knowledge that helps them do their jobs, everyday? How about a tool called “a smartphone” as a force multiplier that enables them to accomplish more, from pretty much anywhere? Now think about both of these tools in your customers’ daily lives.

One year from now, the number of customers who don’t also use a tool called “AI” will be zero. All of them will—and it should be the same for all your employees.

If leaders haven’t done so already, they should start asking each of their direct reports—their HR leaders, their finance leaders, their sales leaders—how AI will change their functions. They should be asking their business units similar questions: How will AI change software development, or trucking, or clothing production?

The easiest problem you can solve, right away, is ensuring your technology architecture never stands in the way of delivering what your creative humans come up with for “what’s next” for AI.

DataStax Astra DB is ready for production deployment as a vector database at massive scale. It has global reach and availability and supports the most stringent enterprise-level requirements for managing sensitive data including PHI (HIPAA), PCI, and PII. Astra DB is built on Cassandra, which is well-known for its unlimited scale and high performance and is already proven as an AI engine by AI leaders like Netflix and Uber.

Additional resources

Vector Search for Generative Apps: A Guide for Developers and Architects for Using Vector Search with AI Applications

A Vector Primer: Understanding the Lingua Franca of Generative AI

How LLMs Are Transforming Enterprise Applications

About DataStax

DataStax is the real-time AI company. With DataStax, any enterprise can mobilize real-time data and quickly build smart, high-growth AI applications at unlimited scale, on any cloud. The company's offerings include the Astra DB cloud database, built on Apache Cassandra,® and the Astra Streaming event streaming technology. Hundreds of the world's leading enterprises, including Audi, Bud Financial, ESL Gaming, and SkyPoint Cloud and many more rely on DataStax to unleash the power of real-time data to deliver real-time AI. Learn more at DataStax.com.