Google Cloud

Grounding generative AI in enterprise truth

Beyond foundation models

It’s gen AI’s moment of truth.

We continue to be awed by the potential of gen AI. It can synthesize information, make coding recommendations to developers, assist analysts with market research, jumpstart docs and slides, and help marketers with creative campaigns. And that’s just getting started.

But the arrival of gen AI hasn’t been without its challenges—with leaders across industries needing to simultaneously integrate the new technology while racing to get ahead of competitors. The pace is relentless.

Foundation models trained on general training data are suited for a multitude of purposes, like content generation, summarization, and simple natural language-based classification. And while they develop a vast “world knowledge” during training, many businesses find that the current generation of foundation models alone can’t be relied upon for the trusted, accurate responses today’s enterprise use cases demand.

To adopt gen AI with confidence and unlock its full potential, businesses need to ground foundation model responses in enterprise systems and fresh data. This can help ensure that models output the accurate and complete responses that companies rely on for ROI and investment opportunities.

Google Cloud

“2023 was gen AI’s breakout year. In 2024, enterprises are asking how far they can take it to drive genuine impact. And grounding AI models in enterprise truth is how businesses are turning the opportunities of gen AI to reality.”

Oliver Parker

Vice President, Global Generative AI GTM, Google Cloud

Google Cloud

To deliver personalized customer experiences, offer assistance to employees, and improve processes, enterprises are augmenting foundation models—to ground them in real-time information and enterprise data.
At Google Cloud, we call this approach to grounding a foundation model "**enterprise truth.**"

Gain a competitive edge with insights from your own data, not someone else's.

Google Cloud

Foundation models alone have limitations, particularly for knowledge-based applications.

01 Foundation models are only as up-to-date as their training data, known as the model's "knowledge cutoff."

02 Foundation models can't verify their responses against other sources, meaning they sometimes produce incorrect responses, or "hallucinations."

03 Enterprise use cases typically require the model to have access to factual internal data, which is a complicated and novel engineering challenge.

04 The model's knowledge is unlikely to be sufficiently precise for enterprise use, either due to training data limitations or proclivities toward incorrect or hallucinated responses.

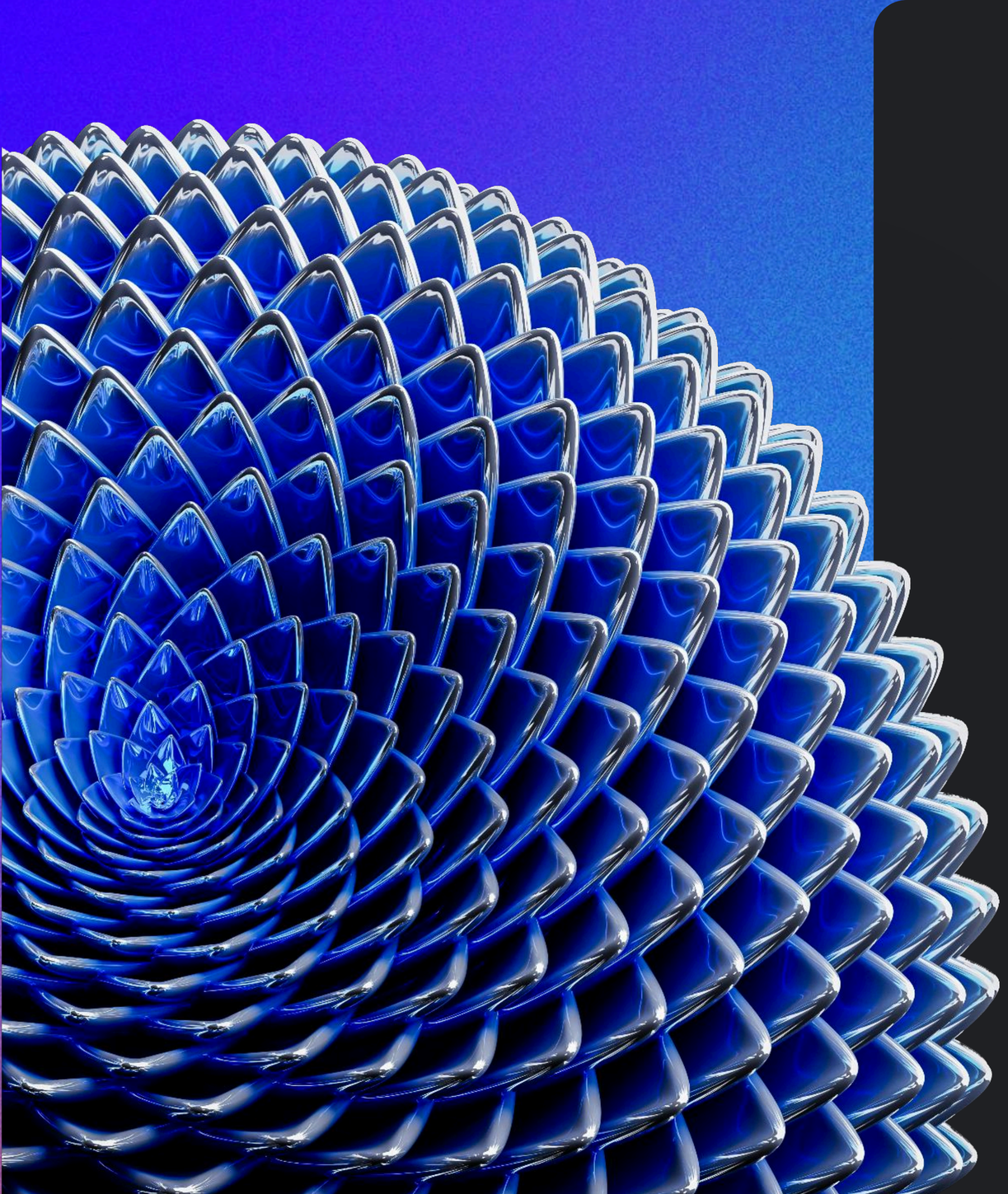
The more enterprises can ground foundation models in the enterprise truth that's specific to their products and customers, the more powerful their AI models.

For instance, by grounding foundation model responses in ERP systems, businesses can create AI agents that provide accurate shipping information in response to a question like "when will my order arrive," instead of just parroting the standard delivery timeframes. By grounding responses in documentation and manuals, they can deliver more helpful answers to product questions and for troubleshooting.

Let's dive in.

The technology powering the next phase of AI for enterprises.

Google Cloud



The approach that enables grounding in enterprise truth is retrieval augmented generation (RAG)—extending the capability of foundation models to incorporate external data before generating a response.

RAG is a flexible methodology to ground a model's responses. It has two stages:

- **Retrieval**: The system searches external data sources and retrieves the information that can best answer the user's query
- **Generation**: The model uses the retrieved facts to generate a grounded response

This approach enables the model to answer questions about information it didn't see during training. And when an enterprise's data is updated—a regular occurrence with operational, transactional, or analytical databases—the enterprise application using RAG generates timely, relevant responses, without retraining. All of this means the app can perform knowledge-intensive tasks in new contexts, and ensure up-to-date outputs.

The types of information enterprises can use for grounding foundation models are limitless: web information; enterprise data like databases and data warehouses; enterprise applications like ERP, CRM, and HR systems; and many other sources.

What's the difference between prompting, RAG, and fine-tuning?

All three of these techniques are ways to improve the output of foundation models, but they differ in implementation and impact.

Prompting is how you interact with an AI model, and improving your prompting strategy is the quickest and easiest way to achieve better outputs. Effective prompt engineering requires understanding the factors that influence the model, and shifting your language accordingly.

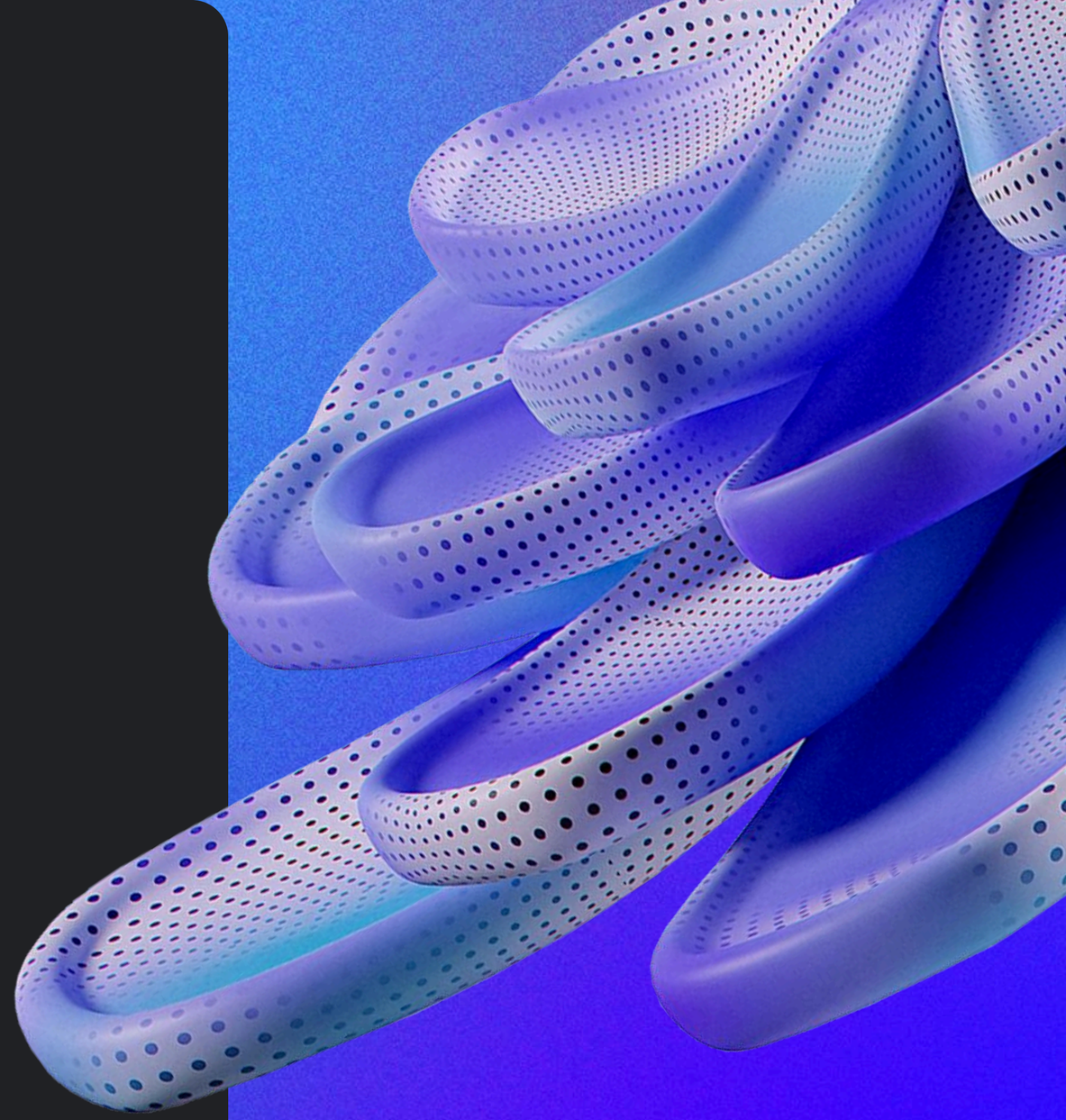
However, its potential for impact can't go beyond the knowledge of the model. And as these refinements are about what the user types in, prompting modification isn't a technique for improving the output of AI models in customer-focused scenarios, such as an AI agent.

Fine-tuning optimizes a foundation model's performance on a specific task or pattern of data—adjusting the parameters of the model or introducing new data inputs.

It is an efficient technique for adapting the model, however, the data must be similar to what the model was already trained on. And even with fine-tuning, the model will give outputs outside the facts it was trained on, so hallucinations are still a problem. It also can not incorporate external or dynamic data past the new training period.

RAG can be significantly more impactful than fine-tuning for overall performance. Because the content of a knowledge base can differ significantly from the original training set, RAG technology can shift to contexts which fine-tuning would not be able to accomplish. RAG is the only technique of these three that can dynamically reference new, external information after the model's training.

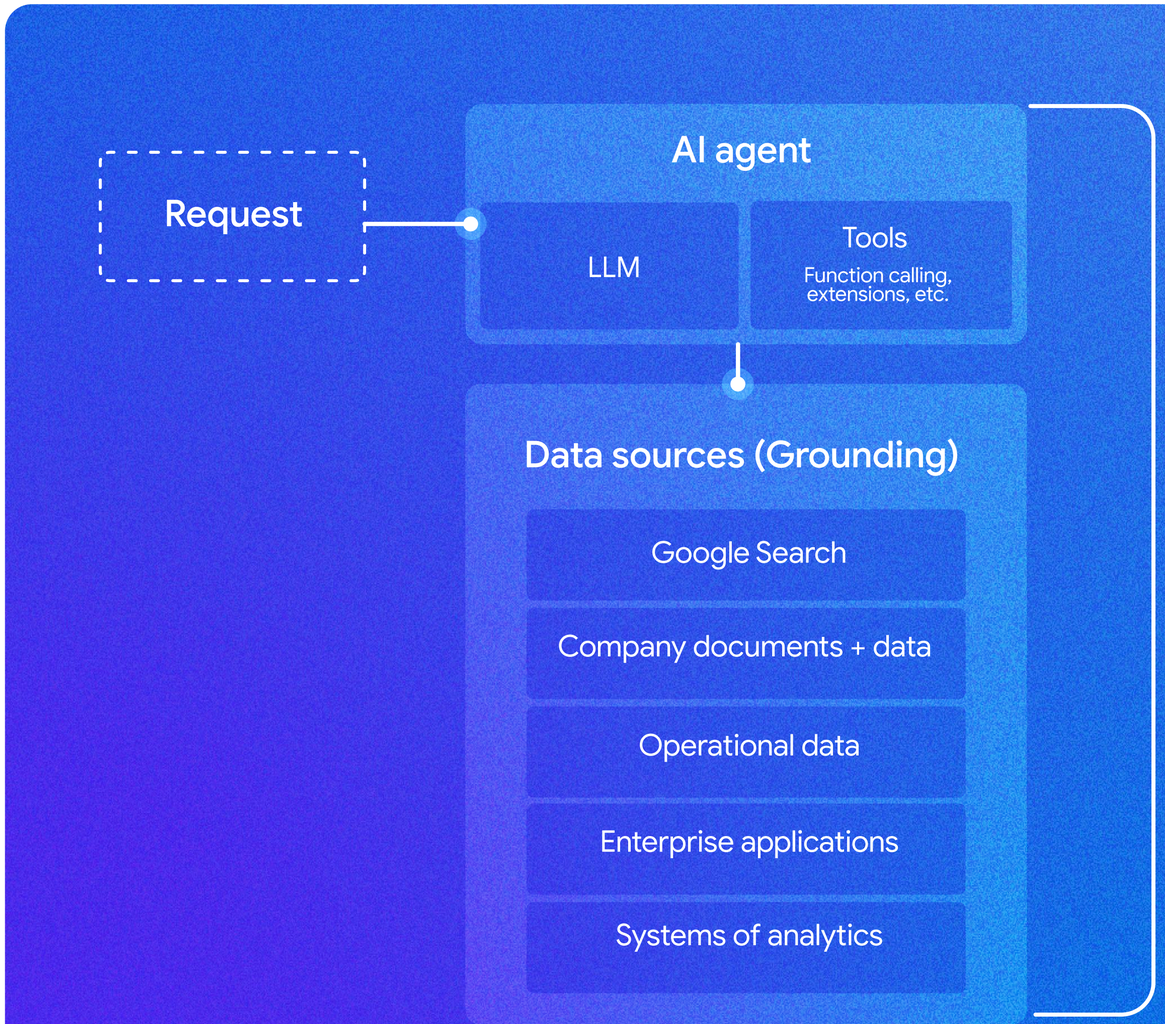
Ideally, an enterprise would use multiple techniques, using prompting and/or fine tuning to influence the behavior of the output and RAG to retrieve those relevant facts.



Google Cloud

Enterprise truth visualization

As an example, let's look at a chatbot on an athletic brand's website that uses RAG to supplement its training data with the latest information from their website. When a customer asks a question, the chatbot will reference the newest information relevant to the query. The output is more relevant and accurate than a response from a generally trained foundation model, alone. And if the product information changes, subsequent queries will reference the updated information to continue to deliver accurate outputs.



Benefits + outcomes

- Reduced hallucinations
- Up-to-date information
- Increased control over response
- Enhanced decision making
- Trustworthiness
- Regulatory compliance

Grounded models produce superior results.

Let’s take a look at how grounding a foundation model in enterprise truth can improve the accuracy, contextualization, and timeliness of an app’s responses.

Google Cloud

Data authenticity

A foundation model that's grounded in enterprise truth uses information that is genuine and reliable. Using retrieval to augment a model with facts from your enterprise truth vastly improves the trustworthiness of the output. If a model frequently delivers inaccurate or unreliable responses, users won't trust its abilities. In contrast, if a generative model consistently delivers factual and grounded responses, users develop greater confidence in its abilities. This earned trust opens the door to implementing the model in sensitive areas where accuracy is critical, such as financial reporting, healthcare advice, or even legal applications.

By continually improving the completeness and accuracy of responses, enterprises can unlock unique use cases across the business and lay the groundwork for the next generation of AI agents.

Data relevance

Grounding a model in enterprise truth means that the model delivers more informative and nuanced output directly relevant to the specific context. Many applications require specificity or detail, and so it's helpful to ground a model in domain-specific data such as scientific papers, product descriptions, or customer records in databases. RAG technology powers the selection of the correct customer records or finding the right chunk of text within a scientific paper—or thousands of them—helping models achieve the data relevance enterprises require.

Minimize hallucinations

Generative models rely on statistical relationships learned during training, which sometimes causes them to generate plausible-sounding outputs that are factually incorrect or even nonsensical. By connecting the model to reliable knowledge sources, grounding enables the model to cross-reference its responses against trustworthy data or verifiable facts, significantly increasing the accuracy and relevance of outputs.

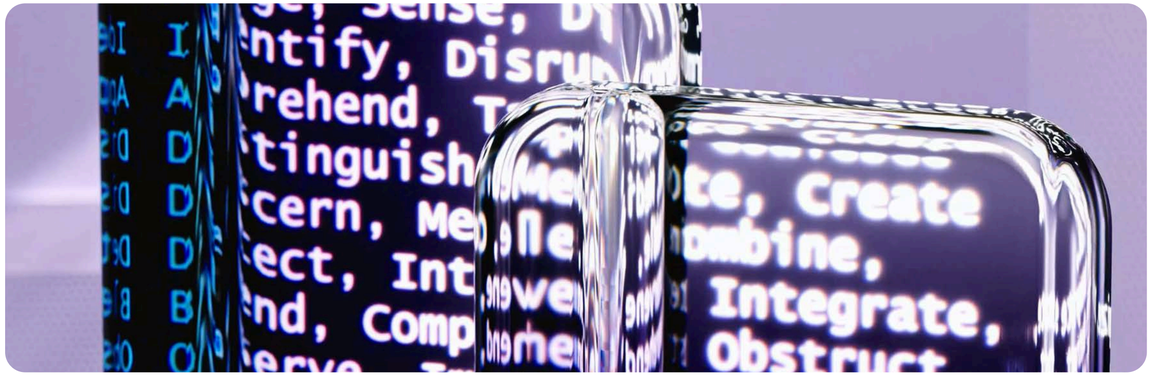
Real-world integration

Models develop significant knowledge and reasoning abilities during training—but training data is nevertheless a snapshot in time, limiting what models can achieve when use cases require precise, up-to-date information. By grounding results in current, real-world information, organizations can ensure models have access to fresh information for up-to-the-minute tasks.

It's also important to understand your AI output. If you don't know where your model got its information from, it can be hard to verify its accuracy. Models grounded in enterprise truth—which inherently reference accurate information—are an important step towards your organization's explainable AI goals. The output of RAG can also include references to the source material.

Google Cloud

Creating more helpful AI agents



Suppose an athletic brand wants to create an AI agent to help customers find and purchase shoes on their retail website.

If the company just puts an interface atop a foundation model API, the resulting app would be able to discuss shoes generally, based on its training knowledge, but it wouldn't have particular expertise in the brand's shoes or any awareness of new footwear trends that emerged after its training cutoff date.

RAG mechanisms enable the company to connect its gen AI models to real-world information and enterprise data—so the model could then access the brand's internal data such as product information, inventory levels, and manufacturing timelines.

The result? An advanced, proactive shoe-recommending agent, able to observe patterns like the customer's last several purchases all having green stripes. It would remember an earlier chat where the customer said they dislike shoes that squeak on hardwood floors, and then reference customer reviews to purge squeakiness from its recommendations. It would also generate tables on the fly so the customer can more easily compare options, and it would use up-to-date inventory and shipping information to help execute transactions.

With the right grounding in enterprise truth, the sky's the limit—and so is the value the agent can create.

Grounding gen AI with Google Search and enterprise data.

Google Cloud

Gen AI models know the most probable response, which isn't the same as being able to cite facts. This is why Google Cloud offers a variety of ways to ground foundation models in the truth relevant to its use case.

Vector embeddings: A primer

Recommendation engines, search algorithms, and many other natural language processing applications rely on vector embeddings. Vector embeddings convert text into numerical representations, encoding information about language meaning and semantic similarities between words and phrases. Vector search then enables the retrieval of the most relevant information based on the meanings of the query or question, going beyond surface-level similarity to the question.

RAG workflows can use these vector embeddings to retrieve relevant data into foundation model prompts to refine them. This minimizes hallucinations, gives more context to the foundation model's answers, and provides more reliable information.

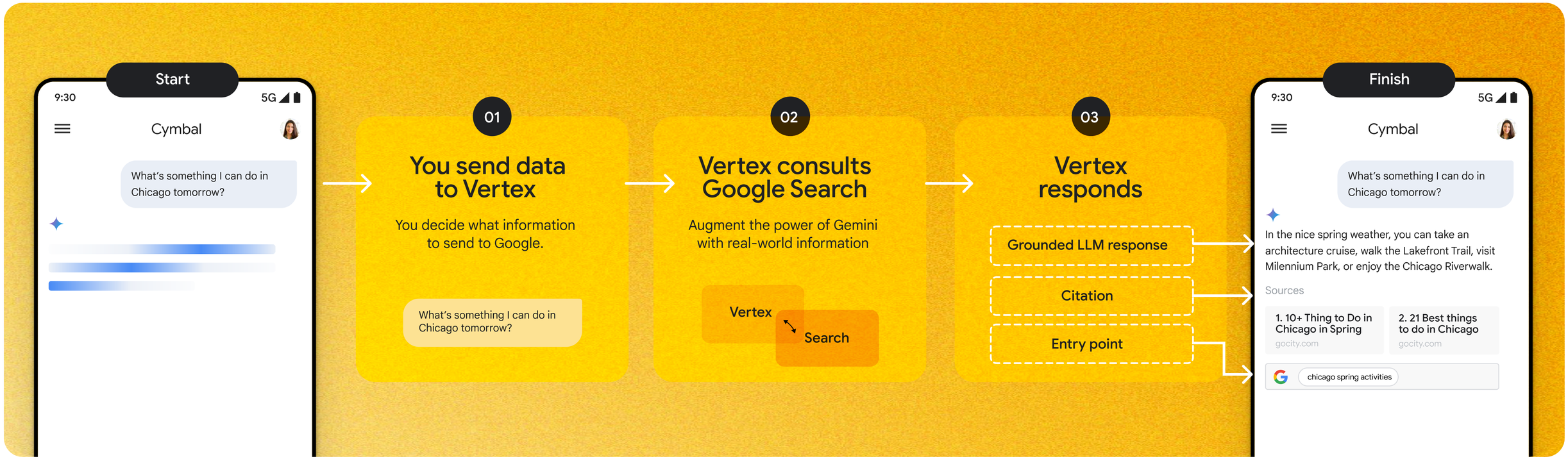
Google Cloud databases include support for vectors, meaning you don't need a specialized database. Instead, you can streamline your embedding creation and access processes using your regular database. Vertex AI Vector Search achieves low latency in vector search—ensuring that users receive instantaneous, relevant responses, no matter how large the dataset or how many parallel requests hit the system. This feature enhances the user experience across a variety of applications, including search engines, conversational AI, and personalized recommendations.

“Our aim with our research platform, Factiva, was to make the information from our dataset of over 2 billion articles more accessible to our users. As a result, we needed to create a search experience that was optimized for relevance and reliability. By applying Google Cloud's text-embeddings model, Gecko, and Vector Search, Factiva is now enabled with semantic search—allowing it to generate responses to queries with greater quality and accuracy, leading to a better customer experience that we believe will drive further efficiencies and ultimately result in greater product adoption.”

Clarence Kwei

SVP of Consumer Technology, Dow Jones

Google Cloud



Google Search

Google Search is one of the world's most trusted sources of factual and up-to-date information. With grounding in Google Search Gemini can search the web for relevant data, retrieve it, and use that content to improve the factuality and freshness of its responses. Because Gemini's training knowledge is very capable, grounding may not be needed for every query. Grounding with Google Search will soon offer **dynamic retrieval**, a novel capability that enables Gemini to dynamically choose whether to ground user inquiries in Google Search or use the more cost-efficient, intrinsic knowledge of the models.

“At Poe, we’re committed to delivering cutting-edge generative AI experiences to millions of global users. Our partnership with Google enables us to leverage grounding to help connect AI responses directly to real-time information and Google Search. For our users, this translates into more accurate, up-to-date, and trustworthy answers. We’ve been delighted with the positive feedback so far, as users are now able to interact with Gemini models with even greater confidence.”

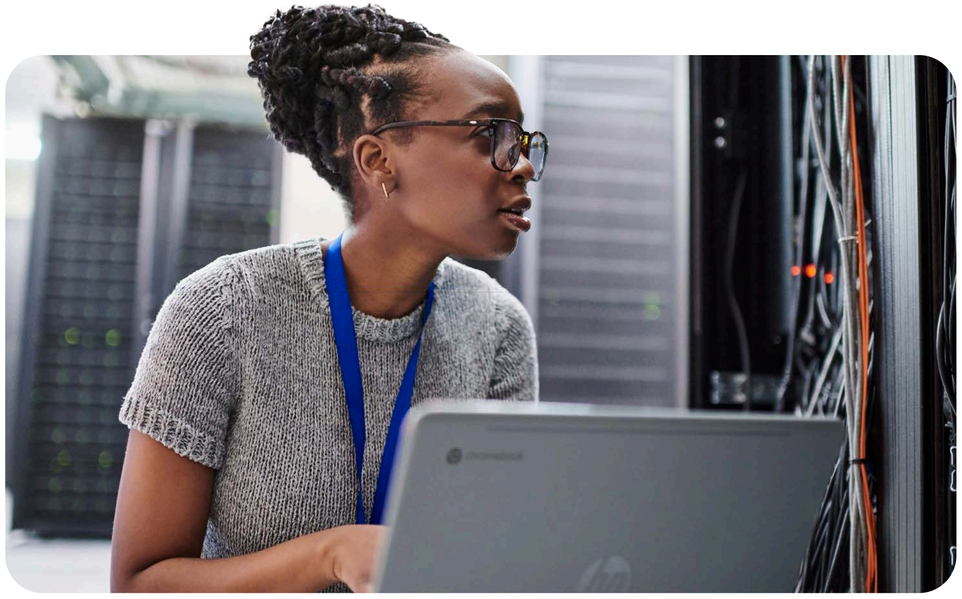
Quora

Spencer Chan

Product Lead, Quora

Google Cloud

Grounding in company documents and data



When it comes to **enterprise data**, Google Cloud offers multiple ways for businesses to ground model responses in enterprise data sources by leveraging RAG to improve the accuracy of model outputs.

Vertex AI includes a Google-quality, out-of-box RAG solution in Vertex AI Search as part of Vertex AI Agent Builder. This system can search across structured and unstructured documents and data from dozens of enterprise data connectors. Under the hood with Vertex AI Search we've simplified the end-to-end search and discovery process of managing ETL, OCR, chunking, embedding, indexing, storing, input cleaning, schema adjustments, information retrieval and summarization to just a few clicks. This makes it super easy for you to build RAG-powered apps using Vertex AI Search as your retrieval engine. Vertex AI Search also offers an easy way for enterprises to upgrade their existing applications that search across enterprise data like intranet or website search by integrating generative AI powered information retrieval and answer generation capabilities.

Coming soon, Vertex AI will enable you to ground your models and gen AI agents with specialized third-party data—unlocking unique use cases and driving greater enterprise truth across your AI experiences.

We are working with premier providers such as **Moody's**, **MSCI**, **Thomson Reuters**, and **Zoominfo** to bring their data in, too.

To meet the specialized industry requirements like searching product catalogs, media libraries, and clinical data repositories Vertex AI Search offers targeted products for retail, healthcare and life sciences, and media and entertainment. Each of these is a state-of-the-art search or recommendation engine—backed by the same technology in Google search—usable on its own, as a RAG provider, or agent tool.

If you are looking to build your own semantic search and RAG system you can use a variety of search component APIs. The wide range of available APIs extend search functionality for many specialized industries, including document processing, ranking, grounded generation, and checking grounding. The answers generated with RAG-based agents and apps typically merge the provided context from enterprise data with the model's internal training. While this may be helpful for many use cases, like a travel assistant, industries like financial services, healthcare, and insurance often require the generated response to be sourced from only the provided context.

Grounding with high-fidelity mode is a new feature of the Grounded Generation API that is purpose-built to support such grounding use cases.

The feature uses a Gemini 1.5 Flash model that has been fine-tuned to focus on customer-provided context to generate answers. The service supports key enterprise use cases such as summarization across multiple documents or data extraction against a corpus of financial data. This results in higher levels of factuality, and a reduction in hallucinations. When high-fidelity mode is enabled, sentences in the answer have sources attached to them, providing support for the stated claims. Grounding confidence scores are also provided.

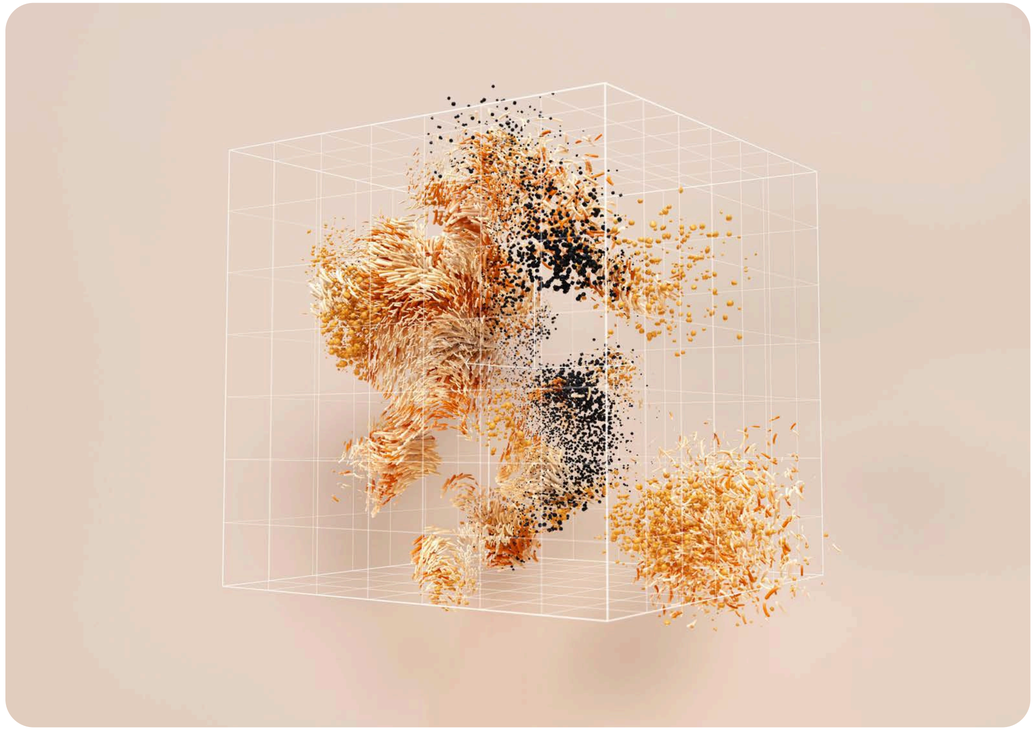
Vertex AI also offers a fully managed LlamaIndex on Vertex runtime for RAG, so you can get a fully operational search running with minimum developer investment.

For any vector similarity-based search engine, recommendation system, or other embeddings-based information retrieval use cases, Vertex AI offers powerful vector search capabilities. Our hybrid search feature integrates vector-based and keyword-based search techniques to ensure relevant and accurate responses for users. This system is engineered for performance, scale, and high-quality results.

Additionally, our new text embeddings models (text-embedding-004 and text-multilingual-embedding-002) are among the top-performing models on the MTEB leaderboard. They're enabling AI models to better understand meaning, context, and similarity across diverse data types and improving the performance of embeddings and vector search-based applications.

Besides these, enterprises can connect models to Google databases like AlloyDB and BigQuery for the contextual retrieval of operational data and analytics like purchase preferences, rewards, basket analysis, interaction history, and more. Google Cloud provides a host of data connectors to help businesses link their models to enterprise apps like Workday, Salesforce, ServiceNow, Hadoop, Confluence, and JIRA—enabling access to the latest data on customer interactions, issue tracking, program management, and employee records.

With a comprehensive approach to grounding that covers web search, enterprise data, and third-party enterprise applications, businesses can ensure that their models will deliver enterprise truth—wherever it is hosted.



Grounding in operational data

Grounding across a document is powerful, but when applied across operational data —systems of record and engagement— it is potentially transformative, helping to unlock real-time and immersive experiences.

Suppose the gen AI agent for the athletic brand can pull data from inventory databases to understand what shoes are in stock, the brands, and the quantities. Because it can now go beyond product documents with real-time information, the agent is now much more helpful. With these extra layers of information, it can form recommendations based not only on general product features or search history, but also richer or more actionable information, like real-time stock availability, purchase history, past interactions with the customer, or user reviews. And by offering customized promotions or recommending credit card point opportunities, the brand can grow conversion opportunities, too.

To facilitate this transformation, we are making it easier to build enterprise-ready gen AI apps using data in the databases enterprises already know and love. We support vector capabilities across our operational database portfolio: AlloyDB, Spanner, Memorystore for Redis, Cloud SQL for MySQL, Cloud SQL for PostgreSQL, and now, Firestore.

In addition, we've made it easy to connect to AI inferencing with Vertex AI, and integrated with the open-source developer ecosystem like LangChain.

Specifically for AlloyDB, our PostgreSQL-compatible database, we've further enhanced our vector capabilities with a new index type to enable faster queries with less memory overhead in AlloyDB AI. Also, **postgres_ann** is a pgvector-compatible extension that exhibits excellent query performance in real-world use cases and applications. It uses a new index type called 'scann,' based on Google's state-of-the-art approximate nearest neighbor algorithms.

When even higher performance is required, **Vertex AI's vector search** can scale to billions of vectors and find the nearest neighbors in a few milliseconds.

At Google Cloud, we're dedicated to helping organizations connect their data to AI, ensuring its outputs are relevant and grounded in operational data, for enterprise truth.

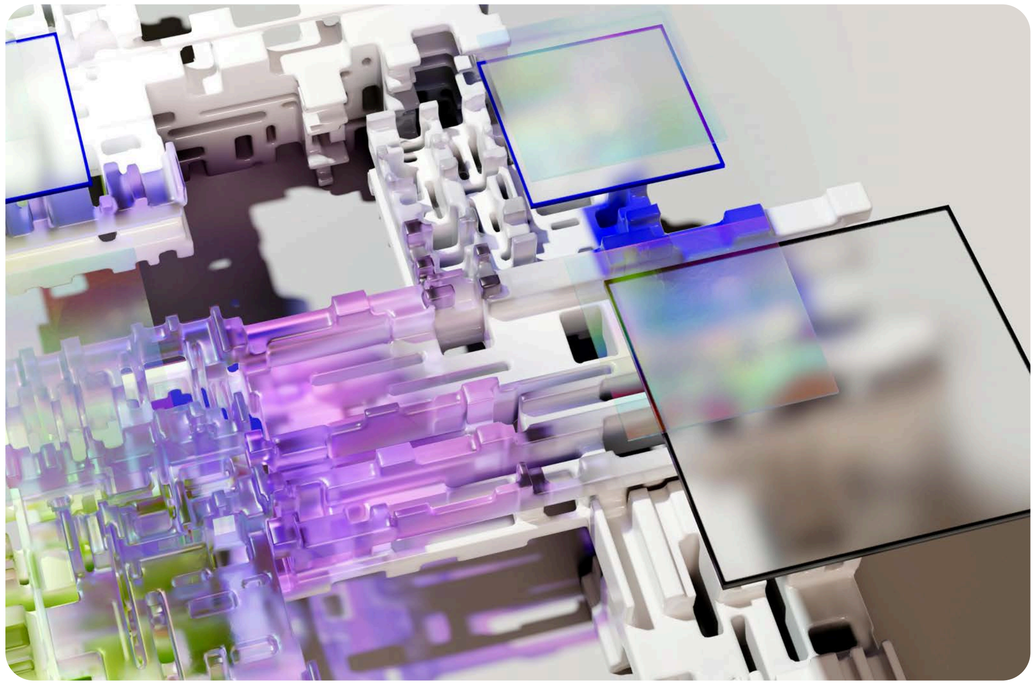
Grounding in systems of analytics

Feeding an AI model with analytics data can dramatically enhance and expand the customer experience, including anticipating the users' needs.

Analytics data powers the pattern recognition and analysis of chat history critical to providing the truly personalized responses in our chatbot example.

To fuel these deeply personalized use cases, enterprises can leverage analytics data and generate **embeddings within BigQuery**—making it easy to ground foundation model outputs in BigQuery data and insights.

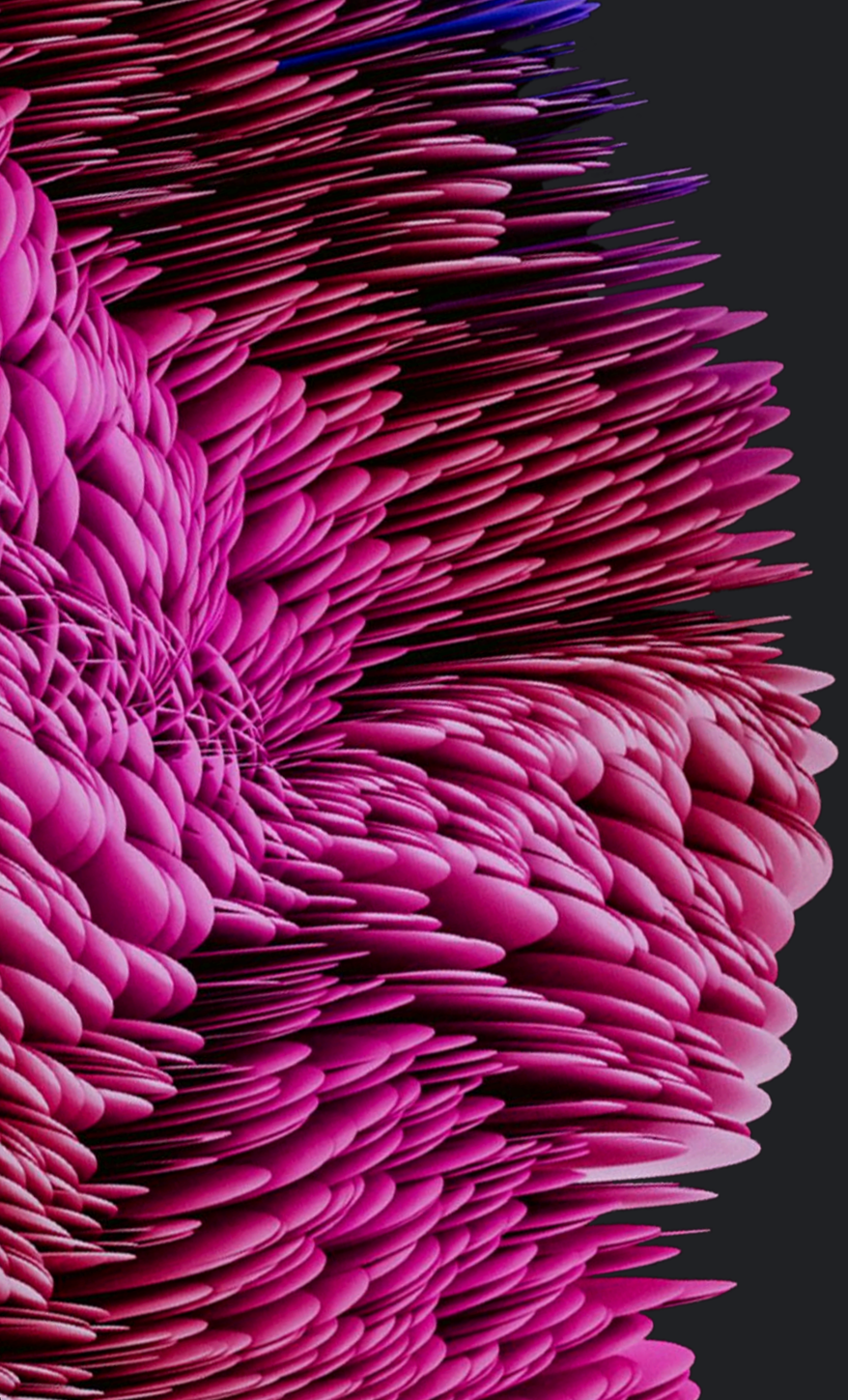
Organizations can also use **BigQuery and Vertex AI integrations** to derive insights from unstructured data formats including images, audio, and documents—and then combine them with structured data and generate embeddings using our foundation models or pre-trained models.



Predicting what's next

Many industries rely on time series forecasting to predict future events by analyzing past trends. For example, in retail, improving demand forecasting accuracy can meaningfully reduce inventory costs and increase revenue.

TimesFM is a forecasting model that's pre-trained on a large time-series corpus of 100 billion real-world data points—displaying impressive zero-shot performance on a variety of public benchmarks from different domains and granularities. And it's now part of Model Garden on Vertex AI.



Enterprise truth: Fueling gen AI innovation across businesses.

Generative AI adoption isn't just about access to capable models. It's also about grounding foundation models in first-party data and high-quality external sources—and using these connections to steer model behavior, creating more accurate, relevant, and factual gen AI experiences for businesses to offer their customers, partners, and employees.

With access to high-quality and relevant data, models can power experiences that move beyond traditional passive applications, giving rise to the next generation of AI agents grounded in enterprise truth.

That's the future we are rapidly moving towards.

Google Cloud

Get started with generative AI today.

Contact us