Google Cloud

Better search, better business

Your guide to enterprise search in the gen AI era

Search well or get lost in the noise.

Generative AI has changed the standard for enterprise applications. Customers and employees now demand personalized, conversational interactions—requiring AI agents to seamlessly understand complex questions that draw on diverse data sources. And as the data grows, so does the role of search.

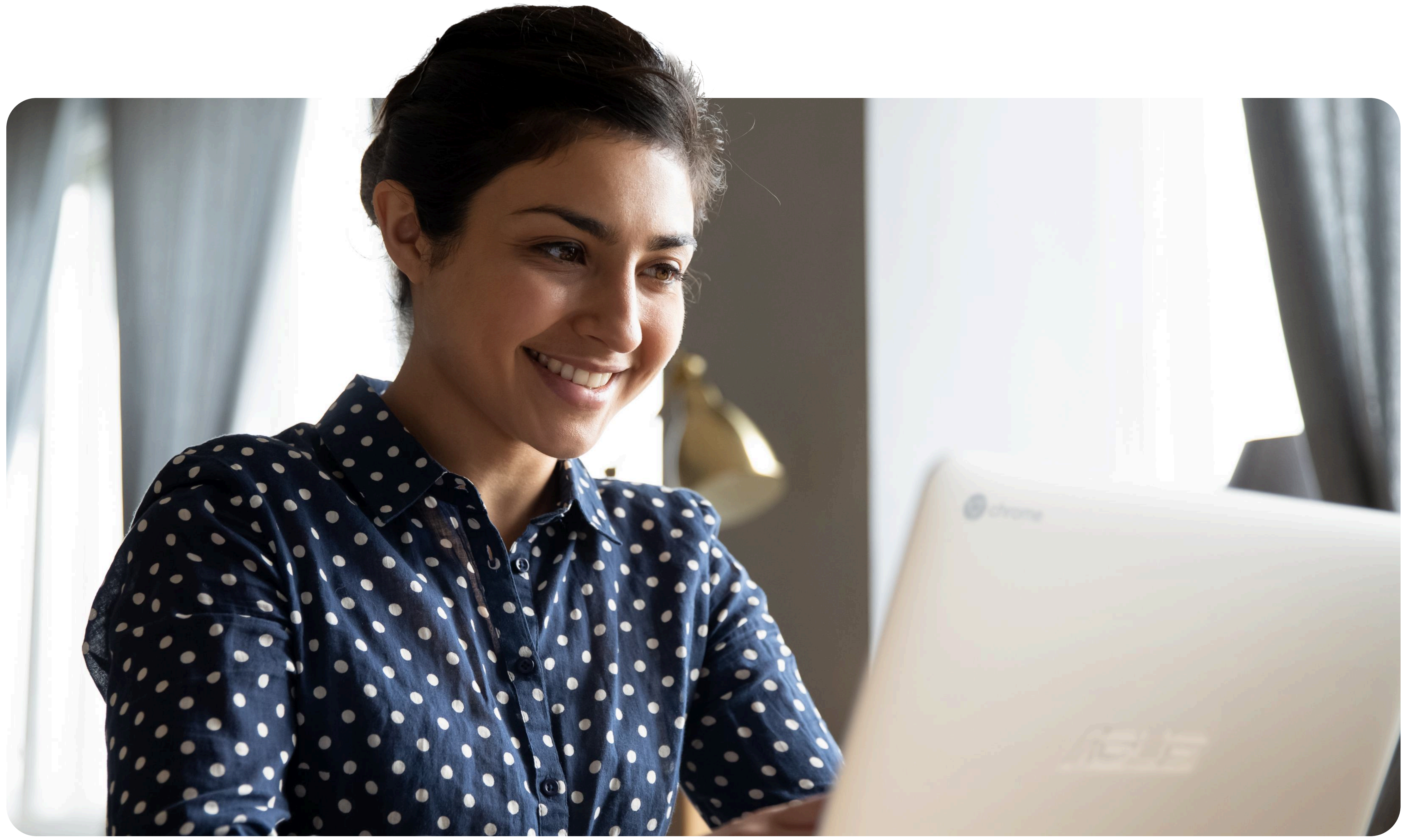
Even though search capabilities have become more accessible, enterprises still struggle to offer high-quality search across their customer-facing websites, gen AI applications, and employee experiences. Enterprise data is often siloed across various systems and spans structured and unstructured formats. Ensuring every search returns highly relevant, actionable results poses a significant technical hurdle.

How can you harness your data to get ahead of customer expectations? We'll show you how you can improve the quality of your search experiences when building gen AI applications.

We'll cover:

- Common roadblocks enterprises face in offering smarter search for their customers
- An overview of technologies powering today's search, including an in-depth look at retrieval augmented generation
- How to tackle search that's tailored for your specialized industry needs
- Getting secure, compliant search that's enterprise ready

Your consumer's search expectations are leaping ahead.



Make sure you're not left behind. Generative AI and advances in search capabilities are making search smarter for consumers.

Enterprise apps can now extract information from formats that were previously unimaginably difficult to parse, like tables embedded in PDFs. And the widespread availability of LLMs has ushered in an era where conversational experiences and robust Q&A systems are not just possible, but expected. Semantic search was once the domain of researchers or massive tech companies.

But advances in AI has made this technology accessible for enterprises to build their own basic systems.

Despite these technological advances, many enterprise websites and apps haven't kept up to consumer expectations. Customers want intuitive, relevant results that go beyond keyword matching to understand intent, context, and specific data relationships.

The gap between enterprise capability and customer desires is frustrating for end users and costing businesses.

Here are some common roadblocks enterprises face:

01 Fragmented and siloed enterprise data across different systems and formats makes incorporating this information into AI applications difficult

02 Generative AI's reliance on broad training data, without built-in ways to verify against your organization's knowledge, can lead to inaccurate responses or hallucinations

03 Grounding LLMs with manual fact-checking or trying to build your own retrieval augmented generation (RAG) system from scratch can be complex, time-consuming, and unscalable

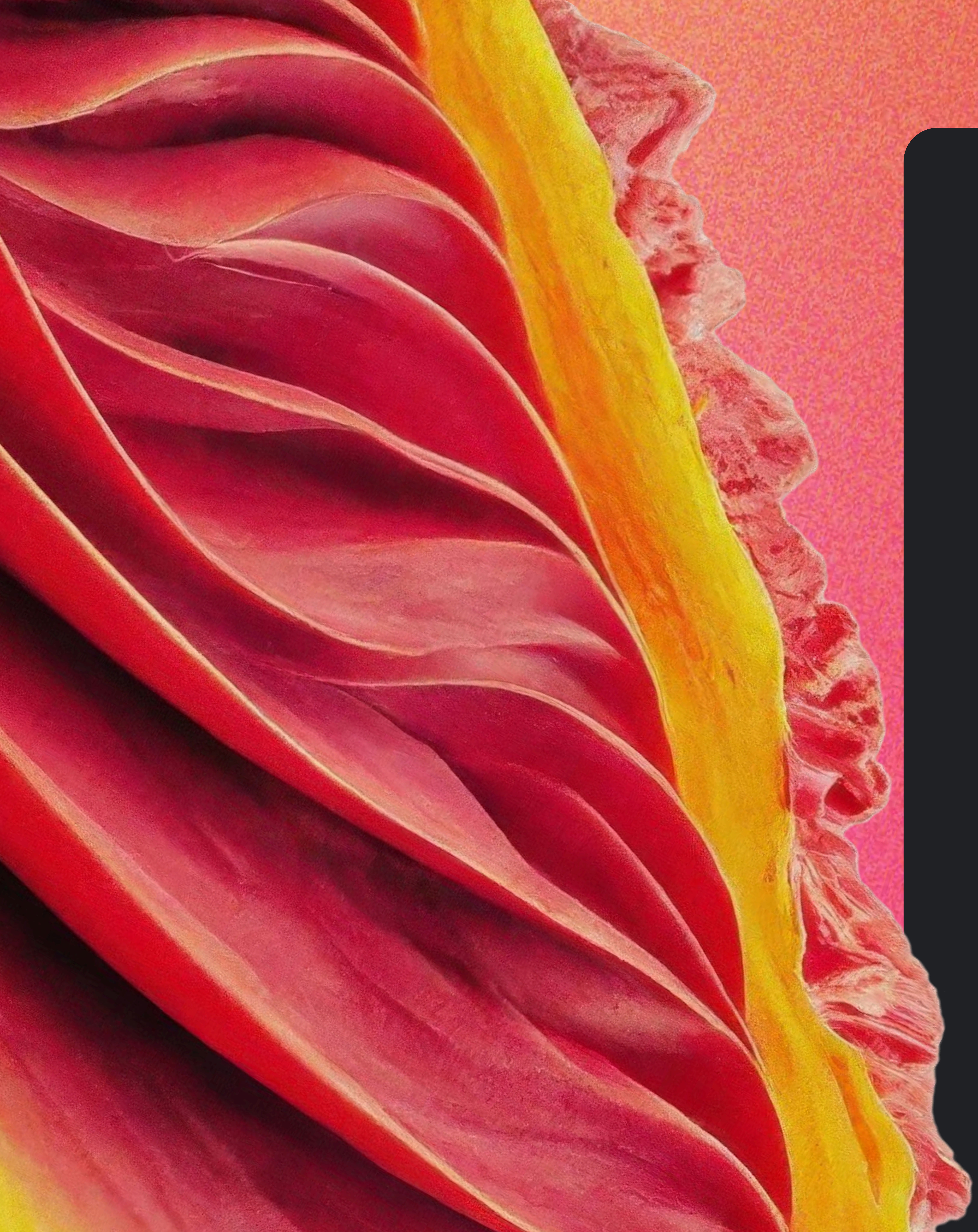
Better search improves experiences and boosts your bottom line.

By improving your search experience, you can make customers happier, increase conversions, and ultimately create a greater competitive advantage for your business.

Let’s get started.

The technologies powering today’s search.

Google Cloud



Creating the personalized, relevant experiences that customers demand begins with grounding your search applications in enterprise data.

Let's dive into the technologies that are powering today's search experiences.

Semantic search goes beyond basic keyword search by using natural language processing (NLP) to understand the intent behind the query, and returns results that match —regardless of whether or not the query and response contain identical keywords. This technology which was once reserved for the very few has now become readily available. And in fact, its prevalence has only fueled customers' high expectations in search experiences.

What if you're looking to build complex applications like product recommendations? **Vector embeddings** transform text, images, or other data into numerical representations (vectors) that capture their meaning and relationships. Going beyond the contextual information used by semantic search, vector search empowers enterprises to deliver more intelligent and helpful applications. For example, a customer support chatbot powered by vector search can understand a question like "How do I reset my password?" even if the knowledge base article is titled "Troubleshooting login issues."

Retrieval augmented generation (RAG) is a technique that improves the accuracy of large language models by combining them with your enterprise knowledge base of documents or data. RAG overcomes two key challenges with LLMs:

1. They do not have an enterprise’s data in their training data
2. They do not have new information past their training cutoff date

Before an LLM generates a response, RAG retrieves the most relevant information, ensuring answers are better informed and grounded in facts. At its heart RAG is a system to search and ground gen AI powered apps in your enterprise data.

How does your technology stack up?

To achieve the best search technology, you'll want to ensure smooth integration across your business. When implementing search, look for solutions that:

- ✓ Deliver results across different types of documents and files
- ✓ Meet your developers where they're at—and can accommodate a range of skill levels
- ✓ Are useful for both customers and employees
- ✓ Integrate with other connectors
- ✓ Handle the unique search challenges of your specific industry

Meet Vertex AI Search.

Google Cloud

Many organizations face difficulty getting enterprise-quality semantic search off the ground. Google Cloud Vertex AI Search is your shortcut to a modern gen AI-powered search experience—building on decades of Google search innovation.

Vertex AI Search enables developers of all skill levels to confidently deliver reliable and highly scalable search experiences tailored to your data—enhanced by foundation model training data or optional integration with trusted public datasets.

Semantic search with Google technology

RAG can help overcome some of the most significant limitations of LLMs, such as knowledge limited to the scope of training data, lack of relevant context from enterprise data, and data that is outdated.

Here are a few RAG technologies that can help you get the most of search:

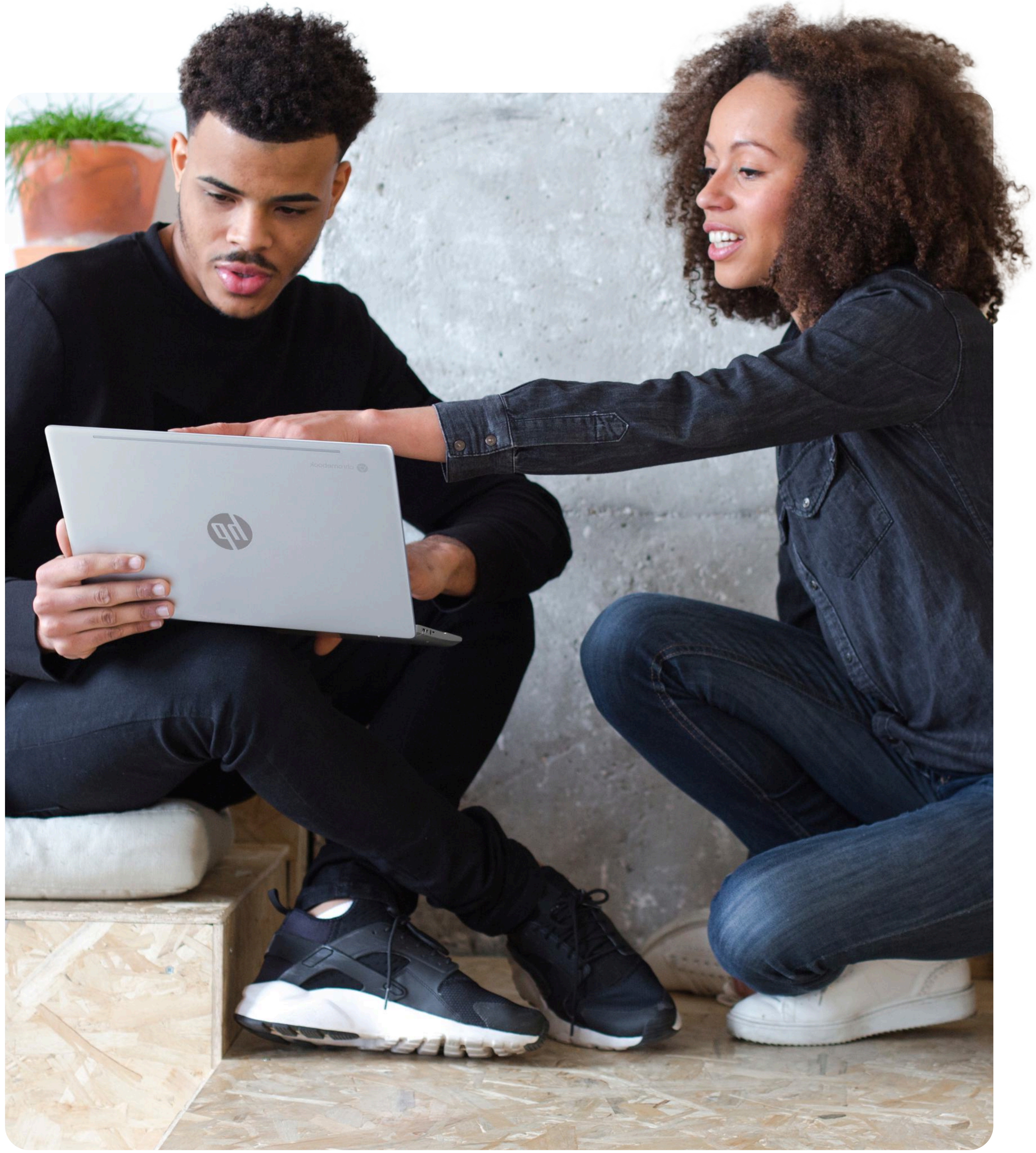
- **Knowledge Graph** helps discover additional information and context from graph relationships, over and above semantic search
- **Hybrid search** simultaneously performs both keyword and semantic searches for each query
- **Query understanding** revises and expands the user's input (such as correcting spelling errors) to return more relevant search results

For more, read the full blog post.

Vertex AI Search understands query intent and relationships between queries and documents, and the hybrid search approach (semantic and keyword) ensures highly relevant results even for nuanced or open-ended questions.

Vertex AI Search streamlines LLM integration, empowering developers to leverage its superior search capabilities for robust, reliable, and production-ready search systems.

> “This collaboration represents a multi-year commitment to bring about real change to the drug development process. Reducing the timeline for drug development and enhancing the delivery of drugs to appropriate patients would represent a major breakthrough for the application of AI in the life sciences ecosystem."
>
> — **Will Lewis, CEO, Insmed**



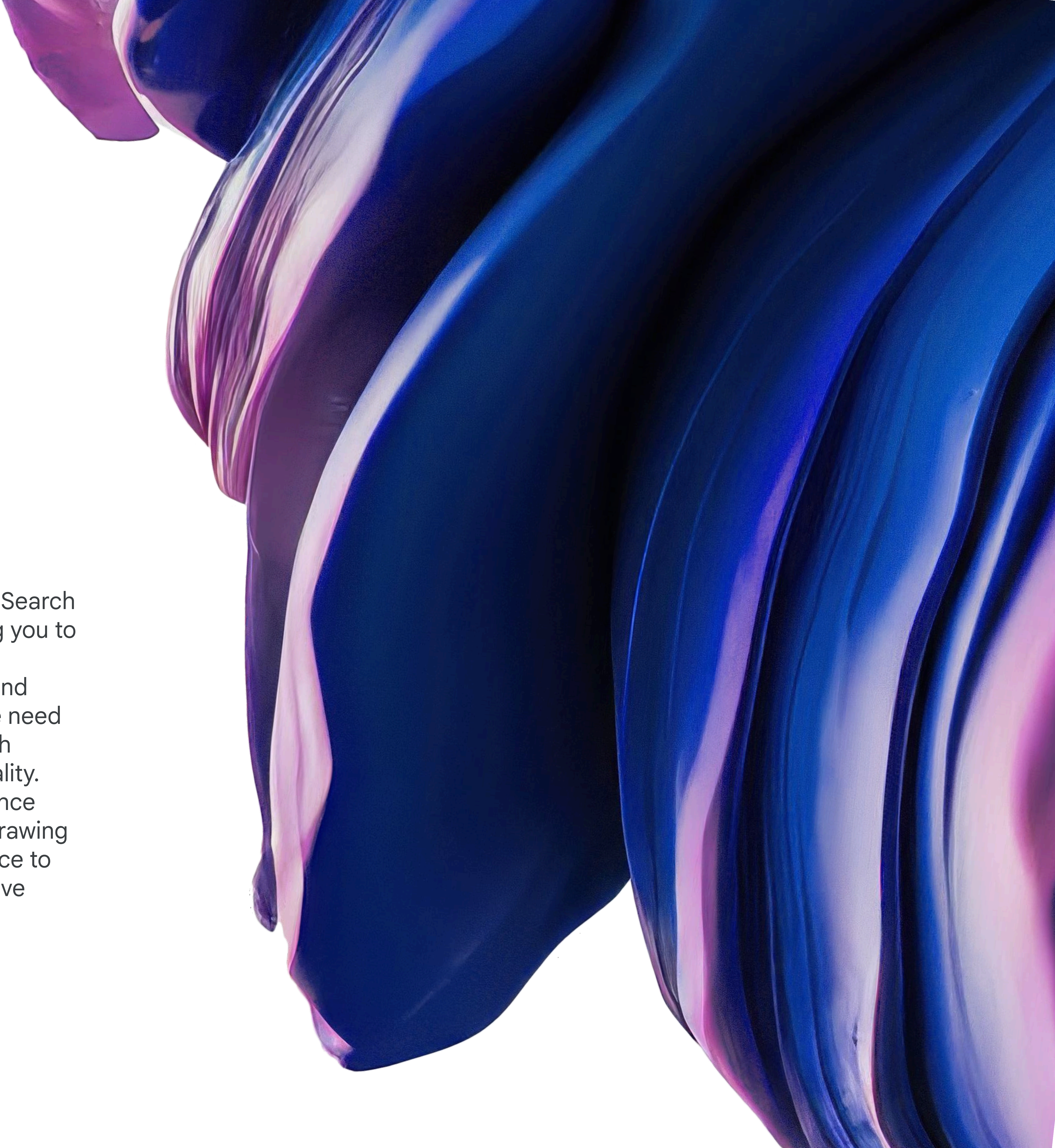
Search out of the box.

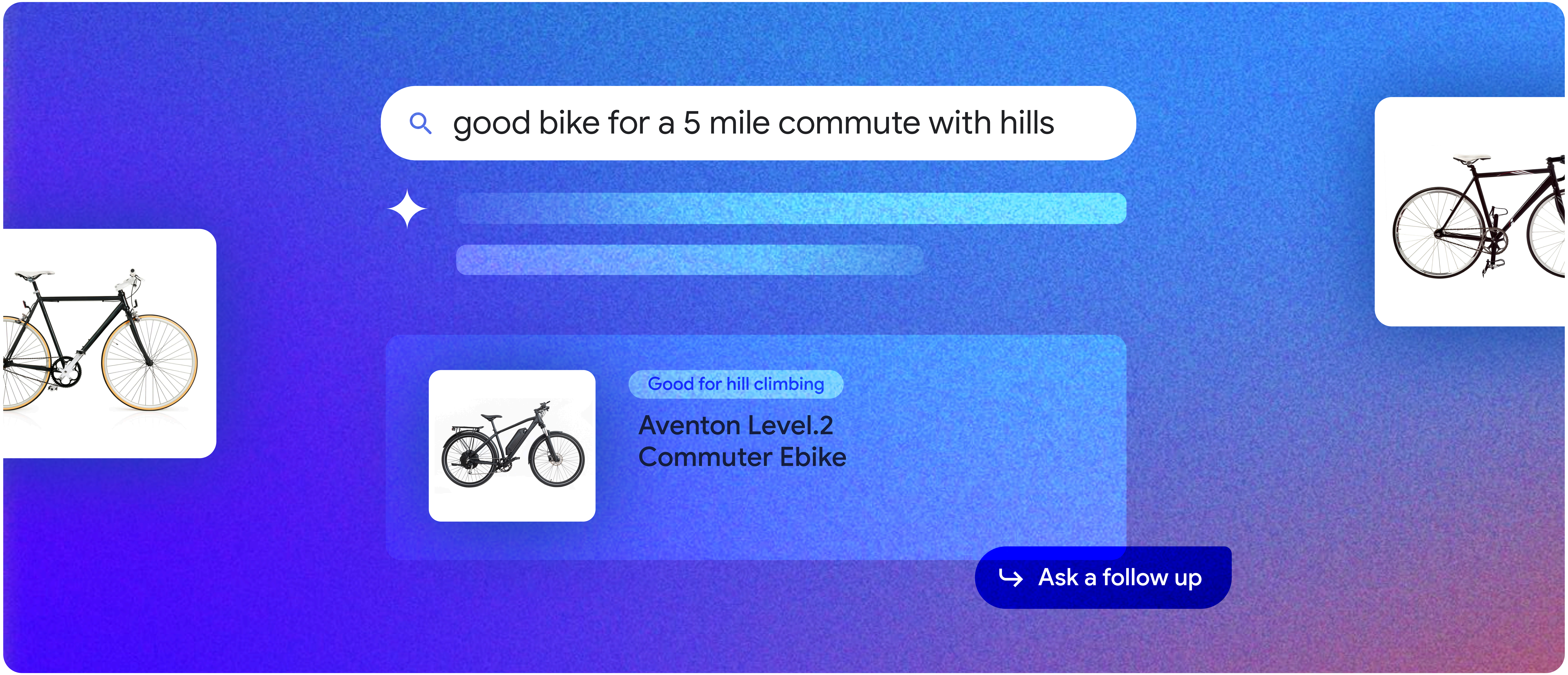
Need a quick win? Vertex AI Search is ready to go out of the box. Use it to power your website search and see the difference immediately. And you can easily augment its capability with RAG and vector search to tailor your search application to your specific needs.

Get your entire ecosystem on the same page.

Search doesn't exist in a vacuum. Vertex AI Search embraces an ecosystem approach to connecting with your apps and data sources. It offers seamless integration via both 1st and 3rd party connectors for read-only access to popular applications and content management systems like JIRA, Salesforce, Confluence, and extensions that can take actions on behalf of your user. Vertex AI Search provides built-in support whether you need to index content from websites, storage systems, databases, or industry-specific data formats. It's cross-functionality that enables you to create a unified search experience that breaks down silos and centralizes information access.

What is exciting now is that Vertex AI Search enables blended search, empowering you to run a single query seamlessly across structured data, unstructured data, and public web pages. This eliminates the need to build and integrate separate search engines for each data type and modality. The result is a unified search experience that delivers higher-quality results, drawing from the strengths of each data source to provide comprehensive and informative answers to user queries.




good bike for a 5 mile commute with hills
Good for hill climbing
Aventon Level.2 Commuter Ebike
Ask a follow up

The new consumer search experience

As a consumer, when using the Google Search Generative Experience you will notice your search results page is organized in a new way to help you get more from a single search. With generative AI in Search, you can Ask new kinds of questions that are more complex and more descriptive. Get the gist of a topic faster, with links to relevant results to explore further.

Get started on something you need to do quickly, like writing drafts or generating imagery right from where you are searching. Make progress easily, by asking conversational follow-ups or trying suggested next steps. It's a good example of how search is different with generative AI.

Retail

- Search & Recommendations for Retailers optimized for eCommerce catalogs and KPIs
- Signals from Google Search
- Custom-tune LLM to their unique product catalog and shopper search patterns

Media

- Search & Recommendations for Media customers optimized for media catalogs and KPIs
- Signals from Google Search

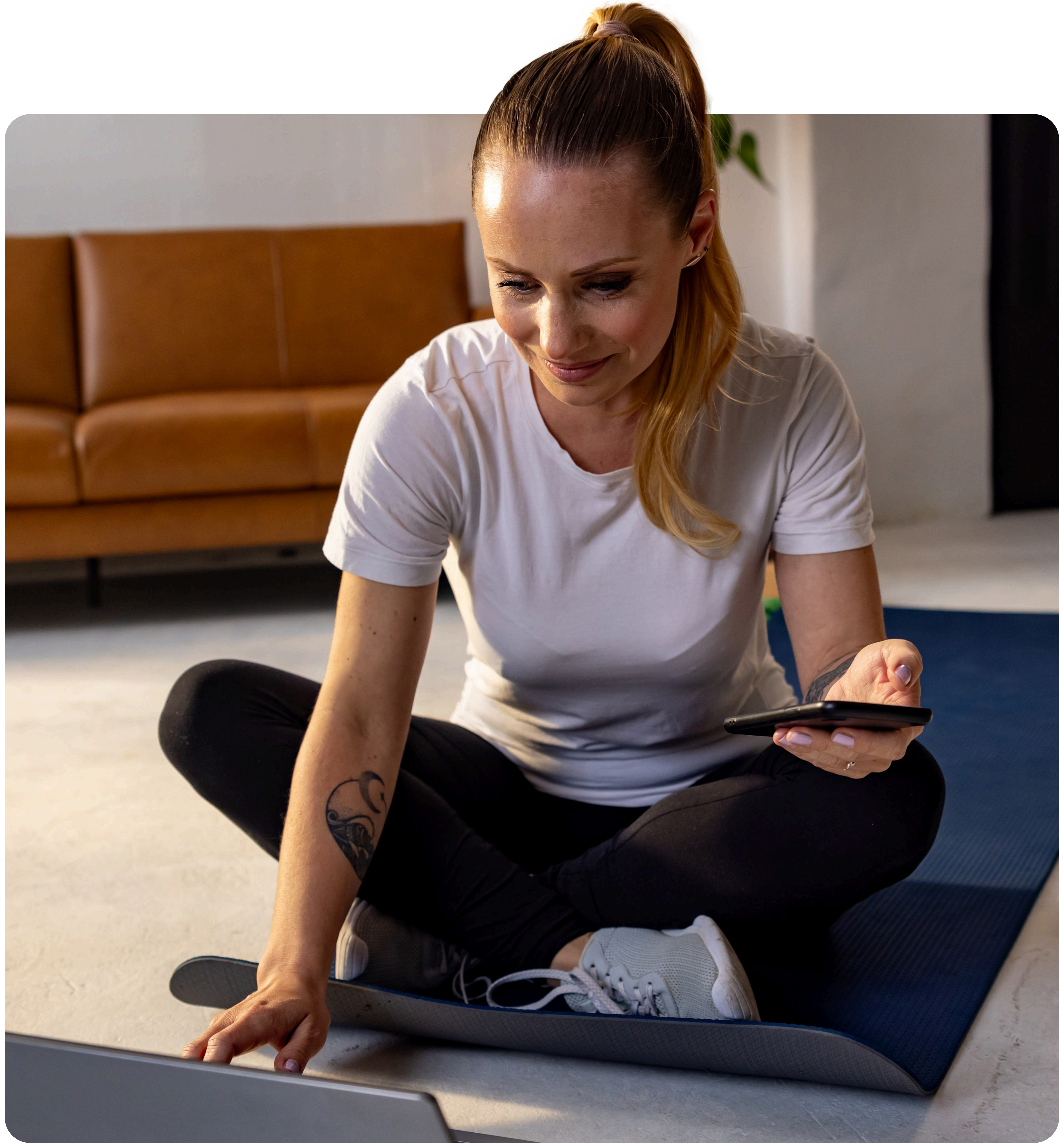
Healthcare

- Search for Healthcare providers, health plans, and ISVs optimized for medical terminology and patient data
- Significantly higher search quality than existing systems due to medical search technology from Google Research

Tailored for your specialized industry needs

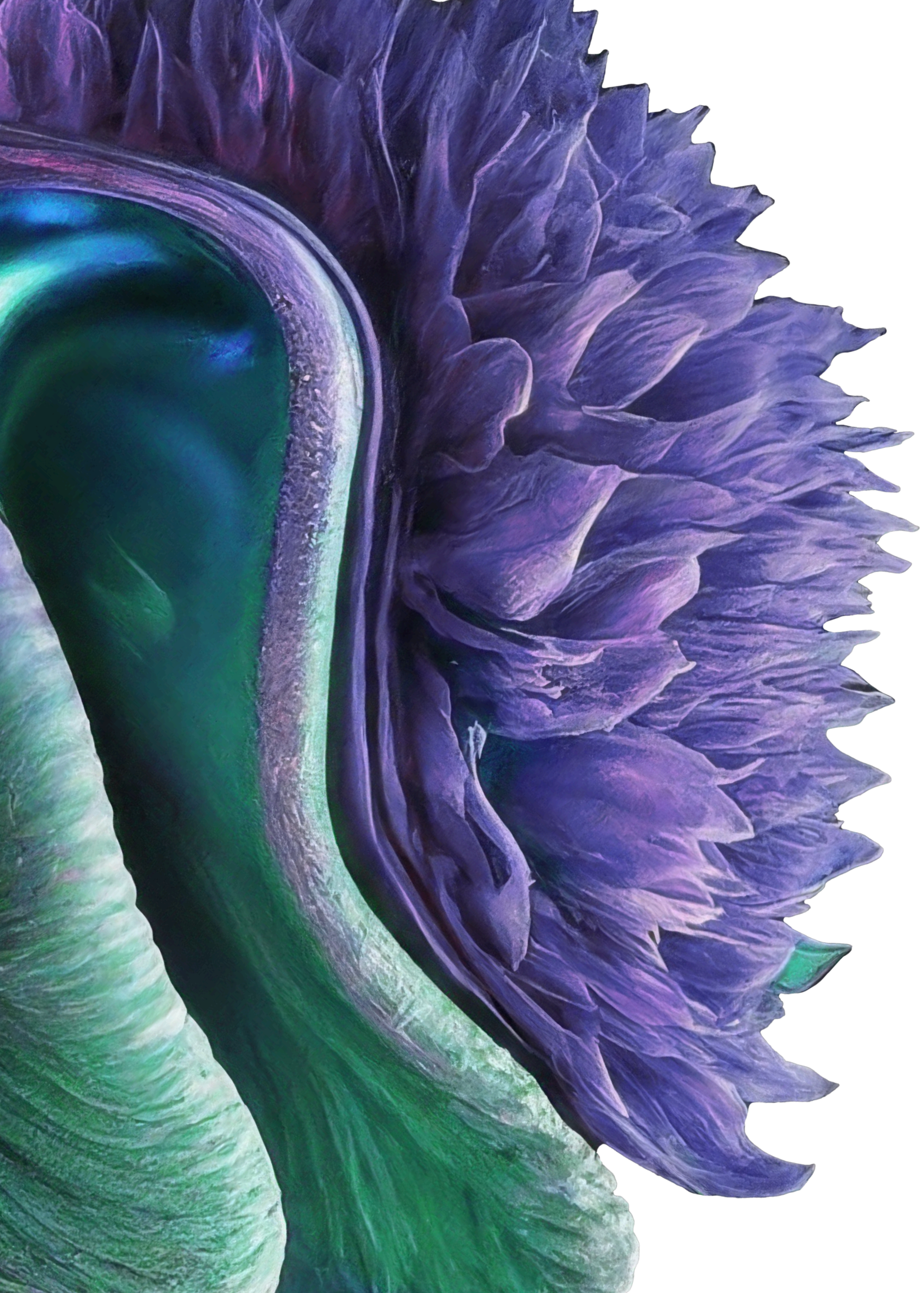
Retail, media, and healthcare are just a few industries with unique search challenges. Retail needs to handle massive product catalogs, media has complex libraries, and healthcare deals with sensitive clinical data. A one-size-fits-all search just doesn't cut it.

That's where the specialized Vertex AI Search solutions come in. Tailored offerings come packed with industry-specific features, like signals from Google Shopping for retail or specialized healthcare language models. These solutions aren't just a nice-to-have, they're game-changers for businesses where search is mission-critical.



PUMA®

Vertex AI Search for retail will bring Google-quality search and recommendations to PUMA.com's digital properties, helping shoppers better discover PUMA's products and deliver personalized shopping recommendations to consumers based on their current interests and trends. In addition, PUMA will explore Google Cloud's generative AI and visual search tools to power future bespoke offerings, such as a generative AI shopping assistant and options to "Shop the Look"—or virtually try on sportswear items—with the help of AI-generated content.



As good on the inside as it looks from the outside.

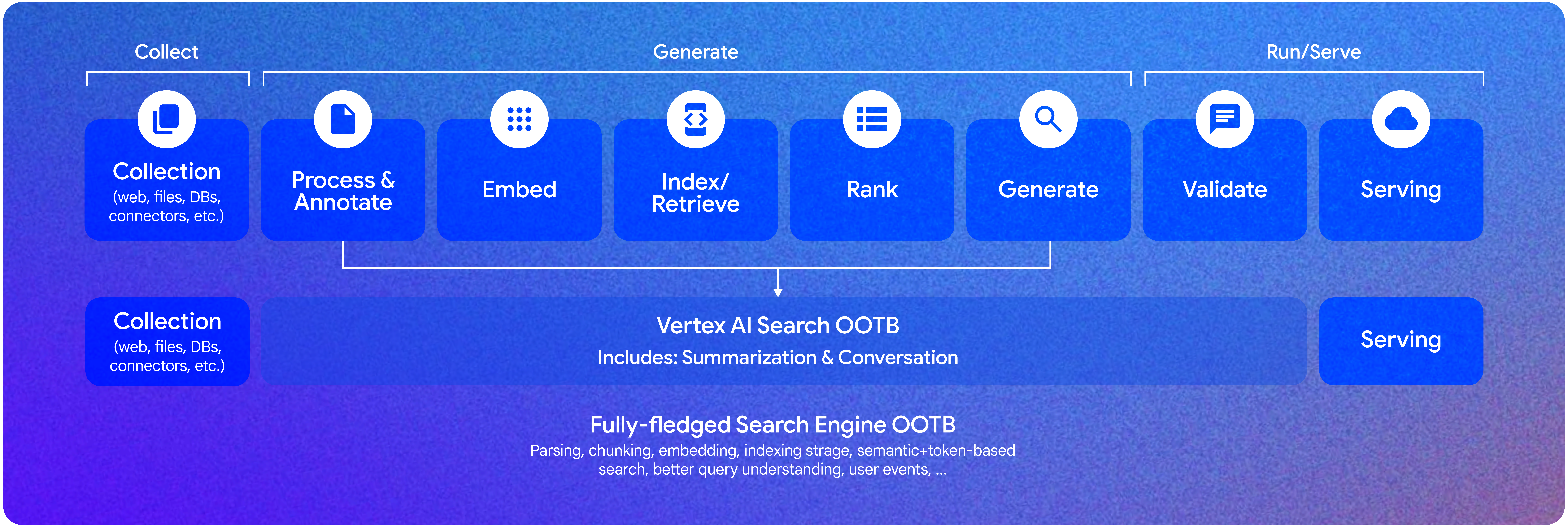
Internal search is just as important as customer-facing search. Your employees need to find documents, data, and knowledge quickly in order to do their jobs well. But creating an effective internal search system is a whole different story.

Don't let your intranet be a black hole of information. Vertex AI Search can transform it into a powerful tool for your employees. Gen AI search understands the context of queries and delivers the most relevant results—boosting productivity and saving valuable time.

Grounded models give superior results.

Leading enterprises rely on RAG techniques to ground their foundation models. If you'd like to build your own RAG-based search, we offer DIY components to do so. However, this can be a highly complex process and should be tackled by companies with experience collecting data and deciding how to perform data chunking, embedding, and indexing in order to ensure the RAG technology runs properly.

Alternatively, you could use Vertex AI Search for your RAG apps. Vertex AI Search simplifies 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.



Collect

Generate

Run/Serve

Collection
(web, files, DBs, connectors, etc.)

Process & Annotate

Embed

Index/ Retrieve

Rank

Generate

Validate

Serving

Collection
(web, files, DBs, connectors, etc.)

Vertex AI Search OOTB

Includes: Summarization & Conversation

Serving

Fully-fledged Search Engine OOTB

Parsing, chunking, embedding, indexing strage, semantic+token-based search, better query understanding, user events, ...

With Vertex AI OOTB search, you can build RAG applications in a few clicks.

For developers seeking granular control over grounding, Vertex AI Search now provides a suite of component APIs that expose the building blocks of its out-of-the-box RAG system. This grants you the flexibility to create and maintain bespoke solutions to address specific use cases or complex requirements.

- **Document AI Layout Parser** API transforms documents into structured data, making content easily accessible for information retrieval
- **Ranking API** ranks search results based on semantic similarity using LLMs, ensuring the most relevant answers are prioritized
- **Grounded Generation API** enables the creation of responses grounded in reliable data sources, including your own data or the internet

Check Grounding API validates whether gen AI statements can be supported by provided facts, offering a further safeguard for accuracy. This API can be used for both online use cases like flagging ungrounded responses to the end user or offline use cases like evaluations.

Find the nearest neighbor, quickly.

Vertex AI Vector Search finds the most relevant embeddings at scale, fast. Based on the same technology that powers core Google services, Vector Search can scale to billions of vectors and find the nearest neighbors in a few milliseconds. Developers don't need to worry about scaling the service up and down, as the service auto-scales based on the load.

Vector Search also enables customization and tunability. For example, developers can easily tune between recall rate and latency, adjusting to match their use case. And Vector search now offers hybrid search—an integration of vector-based and keyword-based search techniques to ensure the most relevant and accurate responses for your users.

New text embeddings models (text-embedding-004, text-multilingual-embedding-002) are among the top-performing models on the MTEB leaderboard, 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.

Stay up-to-date with grounding in Google Search.

With grounding in Google Search, you can integrate information from one of the world's most trusted and up-to-date data sources. With the Gemini API, you easily integrate the enhanced Gemini model into agents and apps you are building with Vertex AI.

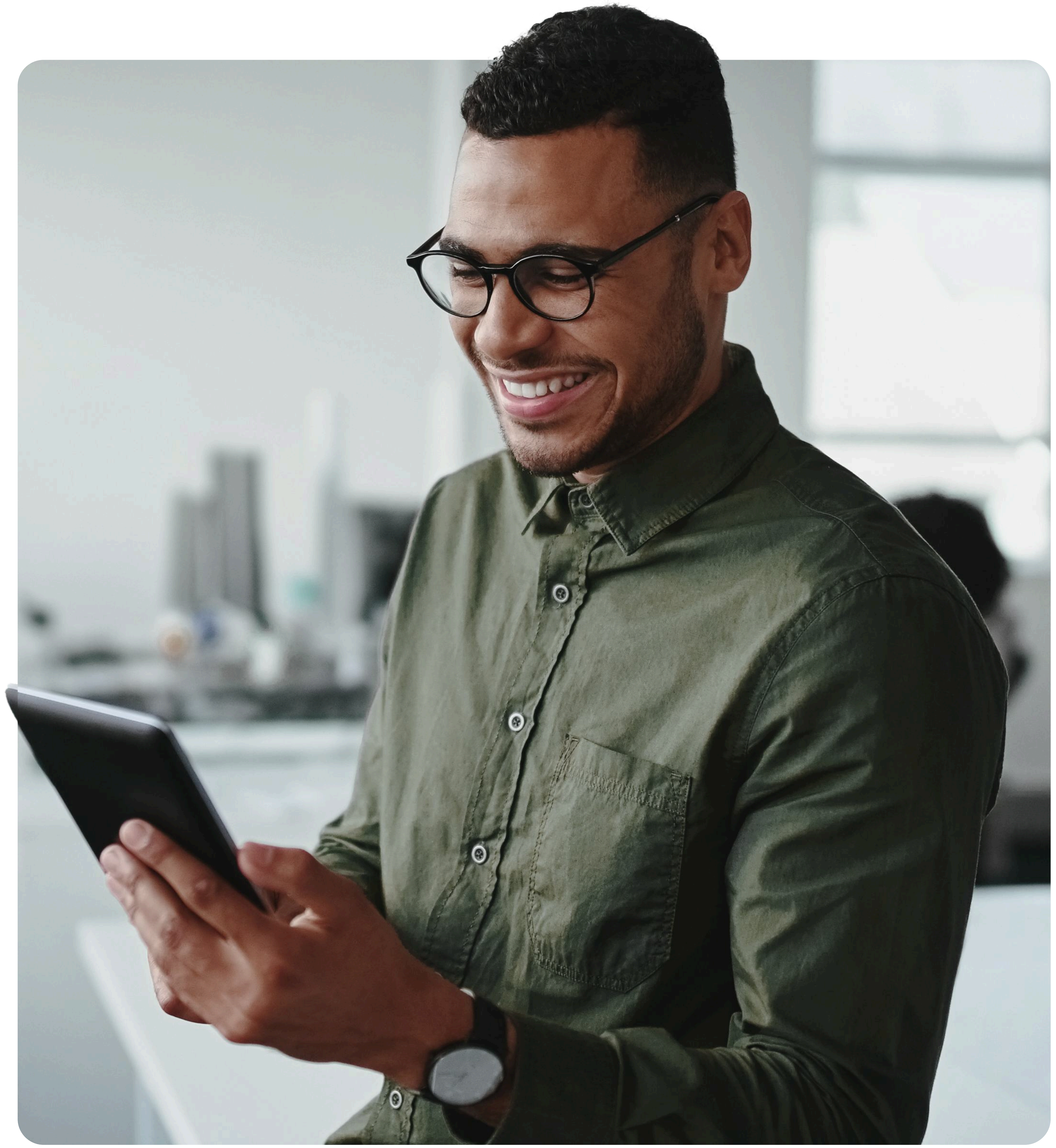
Let's say you have a new chatbot on your retail website that's helping a customer shopping for a dress and the customer mentions an outfit from a new movie. Your chatbot's training data likely doesn't have information on this but Google Search may know.

At the backend, Vertex AI translates your users' prompts into search queries and passes the results along with relevant URLs back to Gemini for grounding. This means users get results that are rooted in one of the most trusted sources of information, built on decades of experience ranking and understanding information quality.

Grounding with high-fidelity mode.

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 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 unifies gen AI development.

With Vertex AI, you can do all your gen AI development on a unified platform. Vertex AI **Model Garden** and **Model Builder** offer tools for model selection and customization.

Vertex AI **Agent Builder** is the central hub for building, deploying, and managing intelligent search and conversational AI applications. Agent Builder underscores Google Cloud's developer-centric approach, emphasizing seamless integration and a comprehensive toolkit for building with gen.

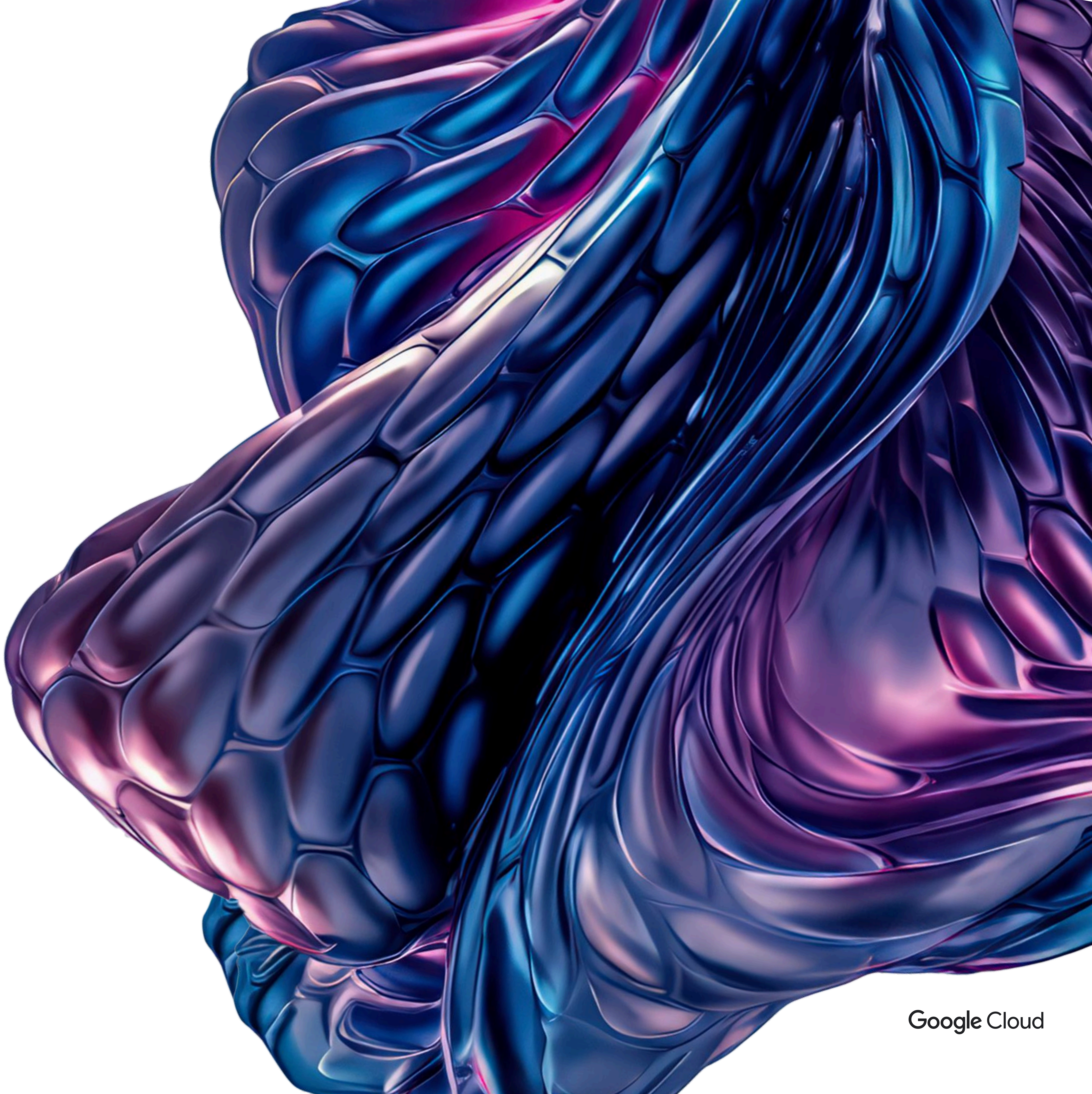
Agent Builder includes the component APIs and vector search capabilities of Vertex AI Search, along with the tools and surfaces you need to build your AI agents and applications and connect them to your data.

This simplifies the developer experience, providing a unified platform for building powerful AI experiences—even for your production workloads.

> US News saw a double-digit impact in key metrics like click-through rate, time spent on page, and traffic volume to its pages after implementing Vertex AI Search.

Get secure, compliant search that's enterprise ready.

When creating applications that use your enterprise data, security and privacy should always be top of mind. Google Cloud is committed to helping businesses leverage the full potential of gen AI with best-in-class privacy, security, and compliance capabilities. We do this by protecting systems, enabling transparency, and offering flexible, always-available infrastructure —all while grounding efforts in our AI principles.



Discover uses Vertex AI to empower its nearly 10,000 contact center agents with gen AI-driven tools with capabilities such as:

✓ **Intelligent document summarization:** Vertex AI will analyze and summarize complex policies and procedures, providing agents with information at their fingertips and insight to answer customer needs.

✓ **Real-time search assistance:** Using natural language, agents can access vast knowledge bases to suggest relevant information during live interactions, so they spend less time searching and more time helping customers.

The result? A 70% reduction in search time to reduce silent time, hold time, and transfers. (Source: Talk at Next ’24)

> “Today more than ever, customers expect exceptional service.
> By using Google Cloud's generative AI tools, we will raise the bar for customer support interactions, ensuring fast, personalized, and effective service every time."
>
> **Szabolcs Paldy,**
> Senior Vice President of Operations, Discover

Read the case study →

With Vertex AI, your data is your data.

Google Cloud does not use your input prompts, model outputs, or training data to train our own models or access this data without your explicit permission.

We also support a range of compliance and security standards, including **HIPAA, ISO 27000-series,** and **SOC-1/2/3**. These standards help to ensure the transparency, accountability, confidentiality, and integrity of your data. Vertex AI Search supports access transparency to provide customers with awareness of Googler administrative access to their data. Virtual Private Cloud Service Controls (VPC-SC) prevent your employees from infiltrating or exfiltrating data. Vertex AI Search also supports data residency in the US and EU multi-regions. We also offer Customer-managed encryption keys (CMEK) allowing enterprises to encrypt their core content with their own encryption keys.

They're the tools you need to confidently build better search experiences.

Gen AI isn't just transforming consumer search—it's revolutionizing enterprise search too. While embeddings have made semantic search accessible, achieving truly exceptional results has been a challenge. Until now.

Search that seamlessly operates across your entire product catalog, internal documents on Google Drive, and even customer data in your Confluence database is a game-changer. And the technology driving this holistic, unified approach is here.

To take advantage of the full potential of gen AI search, enterprises like yours can now build your own Google-quality search apps (BYOS). Giving you the flexibility to build the search experiences that align with your vision and the needs of your users. Vertex AI Search empowers you with best-in-class tools to craft search experiences tailored to your needs. So you can decompose the search process, leverage document understanding, ranking and retrieval, and use vector search capabilities independently.

Vertex AI Search offers you the tools to tackle your toughest search challenges, improve user experiences, and drive real business value.

Get started making search work for you.

Contact us