AI & Machine Learning

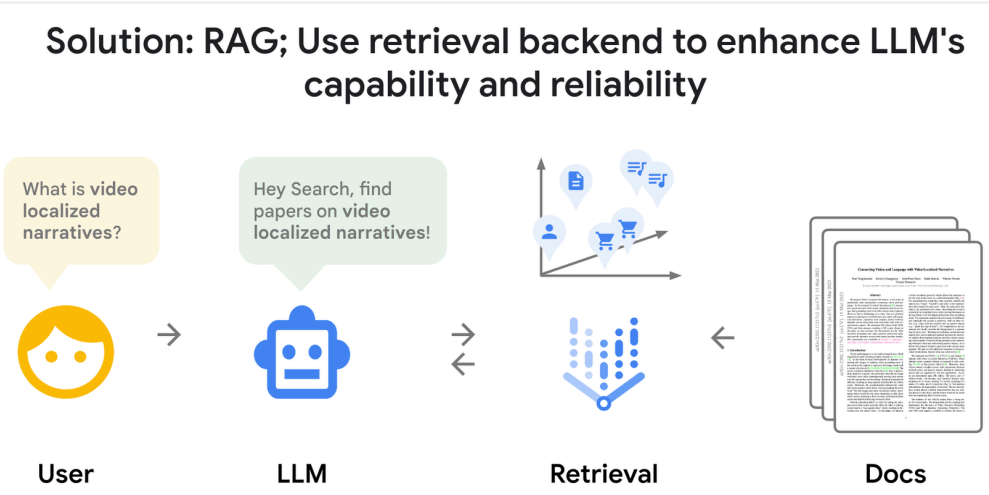
Your RAGs powered by Google Search technology, part 1

February 13, 2024

Kaz Sato
Developer Advocate, Google Cloud

Guangsha Shi
Senior Product Manager, Google Cloud

When a large language model (LLM) doesn't have enough information or has no contextual knowledge of a topic, it is more likely to hallucinate and provide inaccurate or false responses. Developers are increasingly excited about generative AI and Retrieval Augmented Generation (RAG) — an architecture pattern that combines LLMs with backend information retrieval from other information sources. This approach 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limited to the scope of training data, lack of relevant context from enterprise data, and data that is not fresh or outdated.



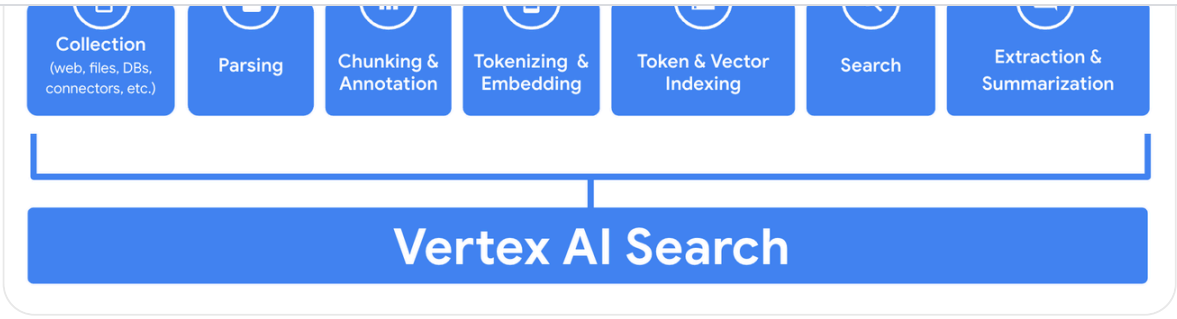
As RAG-based approaches have grown in popularity, it's become clear that a RAG system's efficacy is completely dependent on the **search**

the power if the backend cannot provide high-quality search capabilities for scanning large amounts of proprietary information.

AI researchers and developers have been exploring numerous RAG technologies, including text chunking, query expansion, hybrid search, knowledge graph, reranking, and others. In short, an LLM needs the right data so that it doesn't have to make things up, and you need the best search technology to get the best data.

Google Search technologies, backed by over two decades of experience pioneering information retrieval at scale, offer powerful information retrieval capabilities for RAG solutions. With Vertex AI Search, for instance, developers can easily combine LLMs with Google Search technology, letting them focus on building innovative AI apps instead of taking months to design and build their own advanced search engine for RAG systems.

Vertex AI Search is a fully-managed platform that lets you build AI-enabled search experiences for your public or internal websites, mobile applications, and various other enterprise search services. The product is the result of deep collaboration between the Google Search and Google Cloud teams to create a developer-friendly way to integrate LLM intelligence with Google Search technology.



In the first post of this two-part series, we'll explore the key features that power Google-quality retrieval in LLM and RAG-based applications and how Vertex AI Search has inherited these capabilities from Google Search.

Google-quality retrieval and ranking

Semantic search using deep learning has become a crucial feature for most search engines, letting developers build systems that can understand the meaning of query texts rather than simply using keyword matching.

Despite these advancements, most RAG systems still use simple similarity search in vector databases to retrieve information. This approach can often lead to the return of low-quality, irrelevant results.

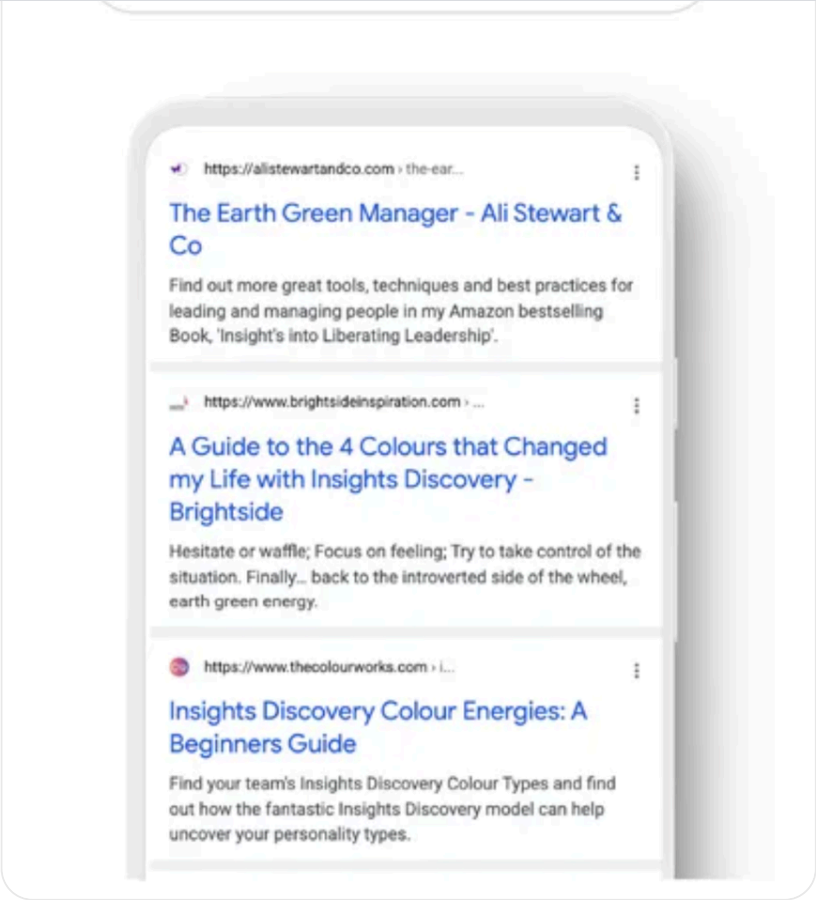
and its answer, "The scattering of sunlight causes the blue color," have distinctly different meanings. Why, then, do you use similarity search to find answers?

Semantic search is not just similarity search

In the Stack Overflow demo that we introduced in a previous post, simple similarity search was highly effective because the dataset had 8 million pairs of questions and answers. However, datasets do not usually contain pre-existing question-and-answer or query-and-candidate pairs in many real-world RAG scenarios. Therefore, it is vital for an AI model to learn and be able to predict the relationship between queries and their corresponding answers to deliver production-quality semantic search.

Google Search started incorporating semantic search in 2015, with the introduction of noteworthy AI search innovations like deep learning ranking system RankBrain. This innovation was quickly followed with neural matching to improve the accuracy of document retrieval in Search. Neural matching allows a retrieval engine to learn the relationships between a query's intentions and highly relevant

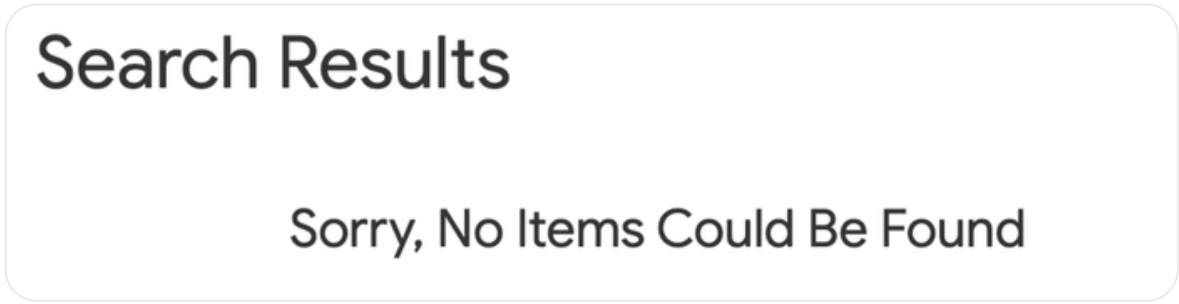
“Neural matching helps us understand fuzzier representations of concepts in queries and pages, and match them to one another. It looks at an entire query or page rather than just keywords, developing a better understanding of the underlying concepts represented in them. Take the search “insights how to manage a green,” for example. If a friend asked you this, you’d probably be stumped. But with neural matching, we’re able to make sense of it. By looking at the broader representations of concepts in the query — management, leadership, personality and more — neural matching can decipher that this searcher is looking for management tips based on a popular, color-based personality guide.”—How AI powers great search results



Neural matching learns the relationships between queries and documents

Vertex AI Search uses the same RankBrain and neural matching processes to generate query and document embeddings, which are vectors that map semantic relationships and enable Google-quality semantic search. Instead of having to "homebrew" a way to fill the gap between questions and answers in a RAG system, developers can easily

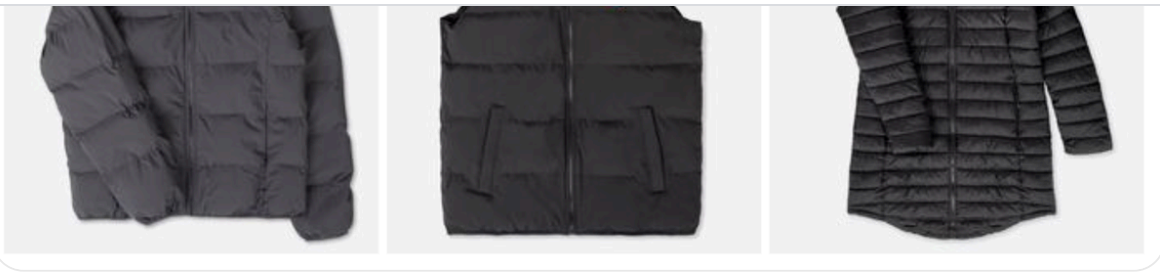
For example, with a popular keyword-based search engine loaded with all product catalogs from the Google Merchandise Store, a query "warm clothing for winter" will return the following results:



A typical keyword search result for a query "warm clothing for winter"

With Vertex AI Search loaded with the same product catalogs, the same query returns the following results:

Generative AI Summary: You can find warm clothing for winter at the Google Merchandise Store [1, 2, 5]. Some of the items include: Google Vail Unisex Grey Puffer Jacket, Google Denali Unisex Puffer Vest, Google Denali Womens Puffer Vest, Super G Tahoe Unisex Black Puffer Vest, Super G Glacier Unisex Puffer Jacket, Super G Glacier Womens Puffer Jacket [5].



Vertex AI Search results for a query "warm clothing for winter"

As you can see in the results above, the semantics of queries and answers are vastly different in many cases. Queries often represent the user's intent (i.e., looking for warm clothing) rather than the answers (i.e., puffer jacket or vest) directly. A production-grade semantic search is not just a similarity search, but must provide smart recommendation to users.

Production-grade semantic search + LLM reasoning

In advanced RAG systems, LLM reasoning is commonly employed to overcome the limitations of simple similarity searches. By combining it with production-grade semantic search, it can greatly enhance the efficiency of an advanced RAG system.

Given that it's the beginning of winter, a customer is browsing for clothing on an e-commerce site. Winters are cold in their city. They entered "warm clothing for winter" as a search term on the site. What other search terms might they use to find related and cross-sell items?

Responses from an LLM may include the following queries:

- Type-specific: Warm winter jackets, Cozy knitwear, Thermal leggings, Waterproof snow boots
- Activity-specific: Ski clothing, Winter running gear, Work-appropriate winter outfits, Cozy homewear
- Style-specific: Cashmere sweaters, Puffer vests, Statement scarves, Athleisure-inspired winter looks

By building a RAG system that can conduct searches on Vertex AI Search with these queries, you benefit from both the power of LLM reasoning and product-grade semantic search. The result is a system that can discover a broad array of relevant products that match different requirements and attributes, including type, activity, and style.

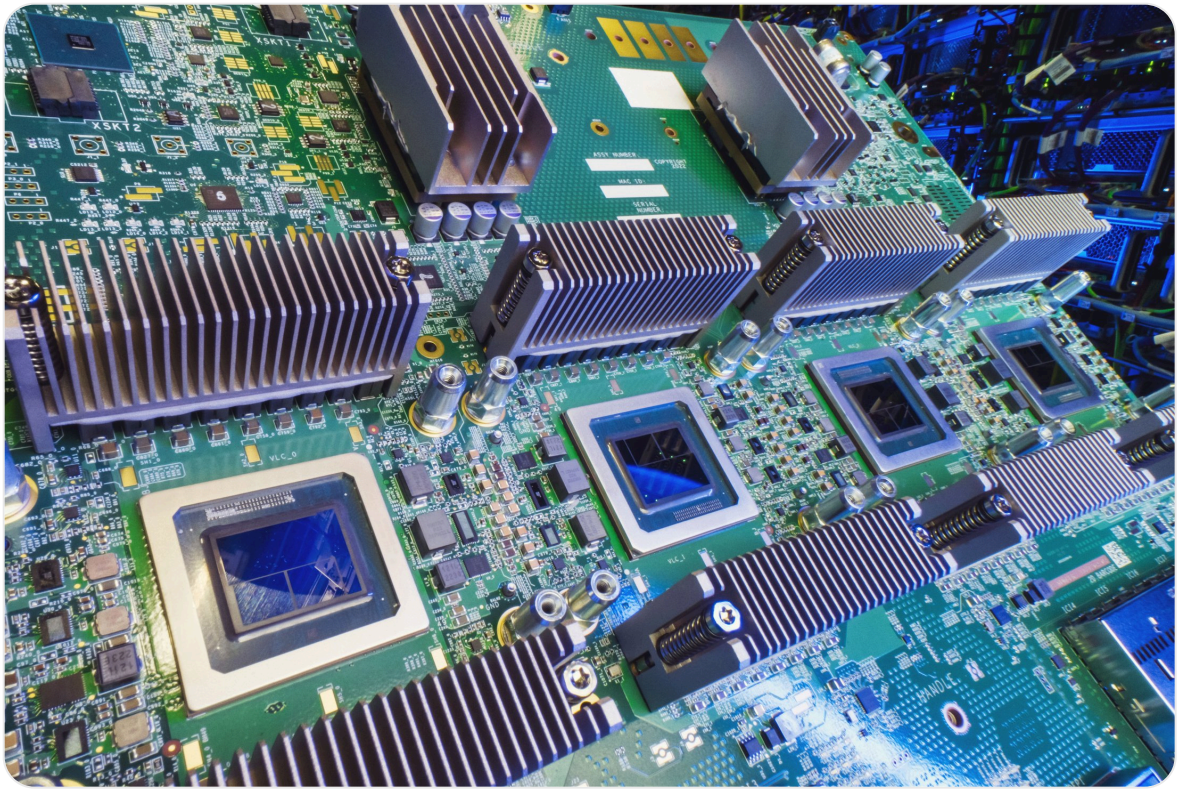
Vector search and AI processor serving billions

semantic search may be among the current hot topics, it's actually the result of years of research and development. Google has been at the forefront of semantic search development for nearly a decade, starting with a strategic decision to invest in developing its family of custom, in-house AI processors back in 2013 — the Tensor Processing Unit (TPU).

TPUs are specifically tailored to provide the underlying power needed to support machine learning and AI workloads, but their genesis is rooted in the goal of supporting the deep learning needed to deliver a production-grade semantic search experience. The first TPU was deployed to Google Search production serving infrastructure in 2015. This substantial investment has helped to reduce costs and latency, enabling us to bring production-grade semantic search experience to billions of users.

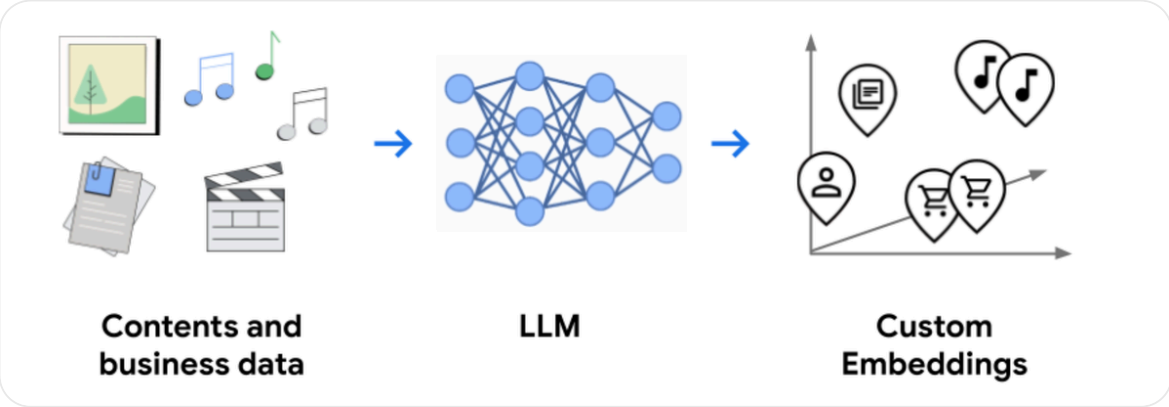
Google has spent years investing in and developing a powerful set of search technologies. For instance, Google Search processes semantic search with query and document embeddings with ScaNN, one of the world's largest and fastest vector search infrastructures. ScaNN powers Google Search and many other Google services, quickly finding highly relevant documents and content to help users get the information they need in seconds. According to ANN benchmarking graphs, ScaNN is one of the industry's core, state-of-the-art algorithms for recalling queries.

some of the most valuable technology assets built over the last decade. Inherited by Vertex AI Search, these same technologies enable the delivery of Google-quality semantic search capabilities with millisecond-level latency and at a reasonable cost, all while letting developers allocate more TPU resources to access the power of AI and large language models as a commercial service.



Vertex AI Search uses Tensor Processing Unit (TPU) to power its large-scale semantic search

In Vertex AI Search, documents are automatically converted into embeddings. Advanced developers can also expand the query-and-answer relationships with custom embeddings tailored to their business requirements.



Custom embeddings can extend the query-and-answer relationships

Custom embeddings can be used to improve search quality by allowing search engines to use additional signals to understand the relationship between a query and documents. For example, the Vertex AI Multimodal Embeddings can generate custom video embeddings, which can be used to conduct semantic search of video content with a text-based query.

model (dual encoder model) that captures business-specific relationships between queries and answers. Some common examples include finding music with similar style to a specific song, upselling and cross-selling products, or suggesting fashion items to complete a coordinated look.

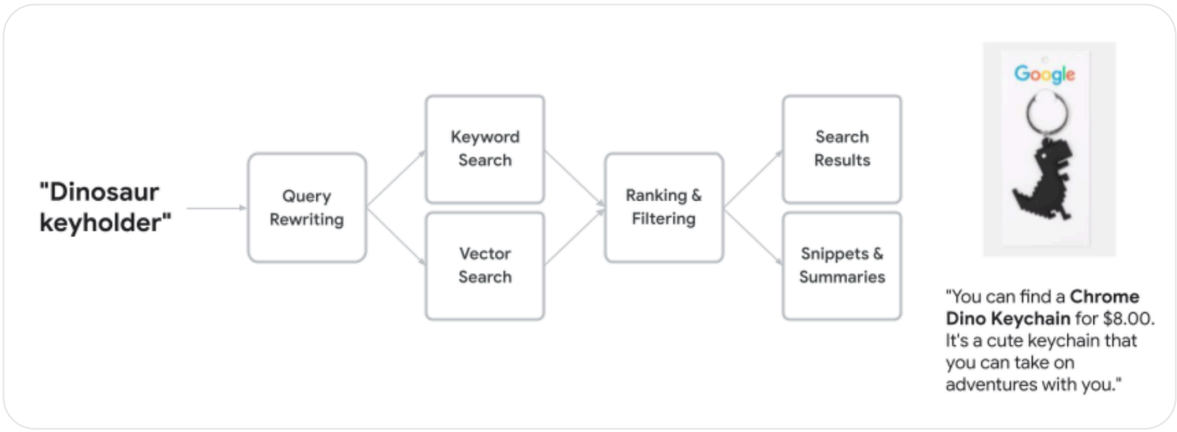
You can then generate custom embeddings from these relationships to improve the quality and personalization of recommendations. These techniques can greatly improve Vertex AI Search's ability to map queries to relevant answers based on bespoke requirements.

Keyword search with 25+ years of experience

Many RAG implementations depend only on semantic search, which entails finding documents based on the meanings ascribed to them by an embedding model. However, semantic search is not a cure-all. In some cases, the embedding model may not grasp the meaning of an item, rendering the search useless. The models may not be familiar with newly added product names or trained to memorize millions of product model numbers in its embedding space.

partial keyword matches and keyword-based filtering mechanisms. This functionality is especially crucial in sectors like healthcare and finance, where keyword searches are fundamental for accessing precise and critical patient information, financial data, or compliance-related documents.

Beyond our expertise in semantic search, we have also spent over 25 years advancing keyword-based (or token-based) search technology. Vertex AI Search builds on this by providing a hybrid search engine that simultaneously performs both keyword and semantic searches for each query. The results are then merged and re-ranked based on their respective scores, combining the best aspects of both search approaches to fill in the gaps left by each.



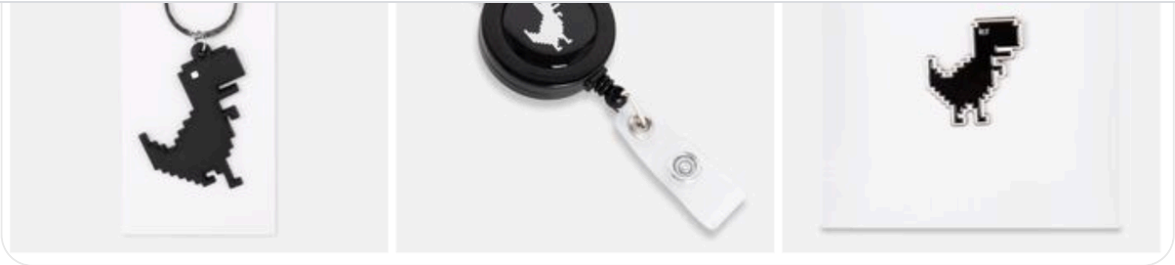
Hybrid search in Vertex AI Search

Often, users type queries incorrectly or misremember the exact names of items. In such cases, query rewriting and query expansion play crucial roles in interpreting and rectifying inputs. By default, Vertex AI Search offers automatic context-aware query rewriting and expansion in its supported languages, including:

- Word stemming and spell correction
- Adding related words and synonyms
- Removing unimportant words
- Annotating important entities with Knowledge Graph (we'll cover this in our second post)

For example, if you search for **"dinosaur keyholder"** in Vertex AI Search with the Google Merchandise Store dataset, you'll get the following results:

Generative AI Summary: You can find a **Chrome Dino Keychain** for $8.00 [1]. It's a cute keychain that you can take on adventures with you [1].

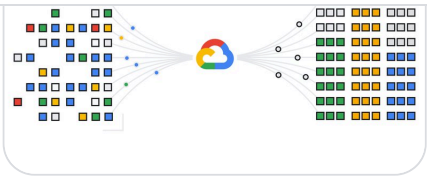


Vertex AI Search results (keychain, lanyard and pin) for "dinosaur keyholder"

With the query words "dinosaur" and "keyholder", the keyword search automatically expands the query with synonyms "dino" and "keychain." Similarly, semantic search considers the word "keyholder" as an intent-based query (i.e., you are looking for a dino accessory) closer to "lanyard" and "pin". The results above show merged keyword and semantic search results — without integrating any additional search engines or performing manual maintenance of a synonym dictionary. When needed, you can even define your own synonyms and query terms on Vertex AI Search through its serving controls.

In the second part, we will continue our discussion of the critical search technologies for RAG systems, such as re-ranking, document extraction and processing, knowledge graph, and data collection.

Vertex AI Search adds new generative AI capabilities and enterprise-ready features



Vertex AI search offers customizable answers, search tuning, vector search, grounding and compliance updates for enterprises.

By Lisa O'Malley • 6-minute read

Posted in AI & Machine Learning

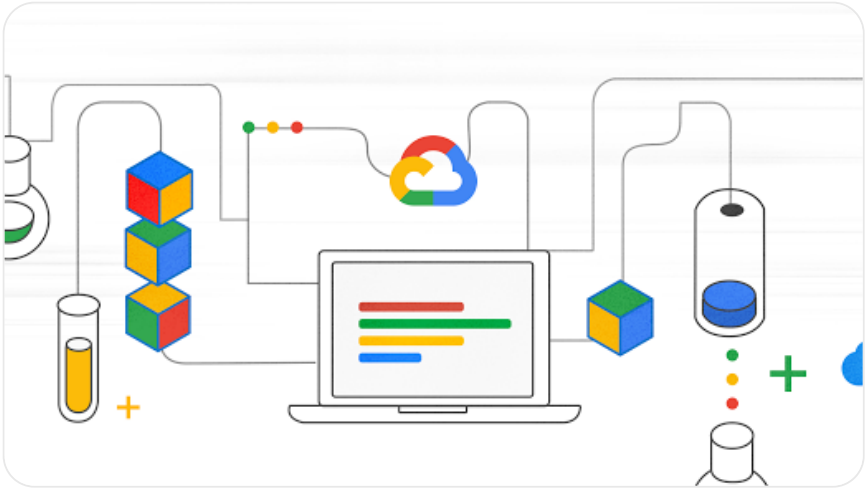
Related articles

Google Cloud

By Ankur Jain • 7-minute read

service games

By Patrick Smith • 8-minute read



Application Development

Orchestrate Vertex AI's PaLM and Gemini APIs with Workflows

By Mete Atamel • 5-minute read

AI & Machine Learning

Gemma is now available on Google Cloud

By Burak Gokturk • 3-minute read