

Google's Gemini API Just Made Document Search Truly Multimodal — Here's What That Means



Google has rolled out a major upgrade to its Gemini API, extending retrieval-augmented generation (RAG) so that text and images can now live inside the same unified vector space. For developers wrestling with enterprise documents that mix paragraphs, diagrams, scanned pages, and tables, the change is a long-awaited shift. The release also introduces page-level citations and metadata-based filtering — two features that together push Gemini closer to the precision and traceability that demanding industries like healthcare, law, and engineering have been asking for. Here is a closer look at what changed, why it matters, and how the pipeline works under the hood.

From Text-Only Retrieval to Genuinely Multimodal Search

Until recently, most RAG systems treated text and visuals as two separate problems. Documents had to be flattened into plain text, with diagrams and images either ignored or pushed through a separate pipeline. Gemini's new release collapses that divide. Developers can now query both modalities simultaneously inside one shared embedding space, which means a single question can pull relevant information from a written description and the accompanying schematic in a single pass.



Picture running a query against a thousand-page product manual. With multimodal retrieval, the API can surface both the troubleshooting paragraph and the labeled wiring diagram that goes with it — without forcing the developer to glue together a custom workflow to handle the two formats separately.

Metadata Filtering for Sharper, More Relevant Results

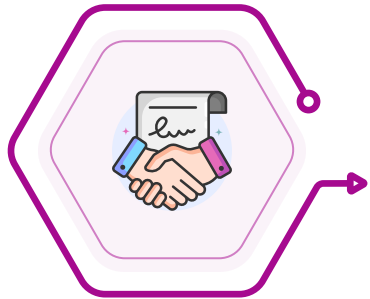
The second headline addition is metadata-based filtering. Every document uploaded to the API can now carry custom key-value tags such as "department: finance," "region: North America," or "year: 2025." Queries can then be narrowed down using those tags before the retrieval step even begins.

For more content like and follow me:



@bertblevins

INCGPT.COM



For enterprises sitting on sprawling repositories that span legal contracts, engineering specifications, HR policies, and financial reports, this is a practical fix for one of RAG's quieter weaknesses: irrelevant matches buried inside otherwise useful results. Instead of returning the top semantically similar chunks from across the entire dataset, the API returns only the ones that match the filter, cutting noise and reducing wasted compute.

Page-Level Citations Bring Real Traceability

Another standout feature is page-level citation support. When the API returns an answer, it now points to the exact page within the source document where the supporting information lives. That single change has outsized implications for any workflow where accountability matters.



A legal team reviewing a contract can jump straight to the clause that backs up an answer. A medical researcher pulling figures from a study can verify them against the original page in seconds. For compliance reviews, audits, and any AI-assisted decision that may need to be defended later, citations stop being a nice-to-have — they become the connective tissue between a model's output and the original source of truth.

How the Pipeline Actually Works

The new multimodal capability runs on a clean five-stage pipeline that is worth understanding even at a high level.



First, ingestion: documents are uploaded through the API, including PDFs, raw images, and scanned pages. Second, chunking: text is broken down into token-bound segments, while images are split into smaller tiles for processing. Third, embedding: text and image chunks are both converted into vectors using Gemini's embedding models and dropped into the same shared vector space. Fourth, storage: those vectors and their associated metadata are placed in a file search store. And fifth, querying: when a user issues a prompt, the API retrieves the top-ranked chunks using metadata filters where applicable, then returns grounded responses with page-level citations attached.

The result is a single, unified retrieval flow that treats text and visuals as equal citizens, rather than bolting one onto the other.

Where the Update Actually Lands

The applications stretch across just about every document-heavy industry.



In healthcare, a clinician can query both written patient notes and diagnostic images inside one prompt — particularly useful when a diagnosis depends on aligning a radiology scan with the corresponding clinical history. In engineering, technicians can pull schematics alongside the written troubleshooting steps that describe them. In legal work, attorneys can run searches that respect both the body of a document and the exhibits attached to it.



Financial services teams stand to gain as well. Analysts often handle reports that pair text-heavy commentary with charts, embedded tables, and screenshots from internal dashboards. Multimodal retrieval lets them pull insights from both layers in one query, rather than relying on a separate process to summarize the visual side.



Pricing Designed for Both Hobbyists and Enterprises

Google has structured the pricing to keep the door open for solo developers and small teams while still scaling for large organizations. Individual files are capped at 100 MB. The free tier includes one gigabyte of total storage — enough room to experiment without putting a credit card down. Vector storage itself and query-time embeddings are also free.



Charges apply to document ingestion and to token usage during generation. For organizations dealing with serious document volume, the model scales upward predictably, without the sudden cost spikes that often catch teams off guard with managed vector databases.

A Smooth Path Forward for Existing Users

For teams already running on the Gemini file search API, the migration story is reassuringly simple. The new multimodal features layer onto the existing infrastructure, so workflows do not need to be rebuilt from the ground up. Developers can adopt the new capabilities incrementally — adding metadata to existing documents, enabling multimodal queries on the collections that need them, and switching on page-level citations where traceability matters most. The path is gradual enough that teams can pilot the new features on a single high-value document collection before rolling them out across the broader stack.

The Bigger Picture

Gemini's multimodal RAG update is part of a broader shift in how enterprise AI handles unstructured data. By collapsing text and visuals into one searchable layer, adding filters that respect business context, and exposing citations at the page level, Google is making it easier to trust AI-generated answers in environments where being wrong has consequences. For developers building serious tools on top of large language models, that combination — precision, traceability, and multimodality — is exactly the foundation that has been missing.

