Coding Nexus · Follow publication

Member-only story

Google Just Made RAG Ridiculously Easy with the New File Search Tool

3 min read · 11 hours ago

Civil Learning Following

Listen Share More

If you've ever tried building your own RAG setup, you probably understand the challenge — managing embeddings, vector databases, chunking text properly, and ensuring everything interacts with your model without breaking the bank.

Well, Google just made that entire mess disappear.

They quietly *introduced a new **File Search Tool** within the Gemini API, which handles all the RAG heavy lifting for you*. You throw in your files, ask your question, and it figures out the rest.

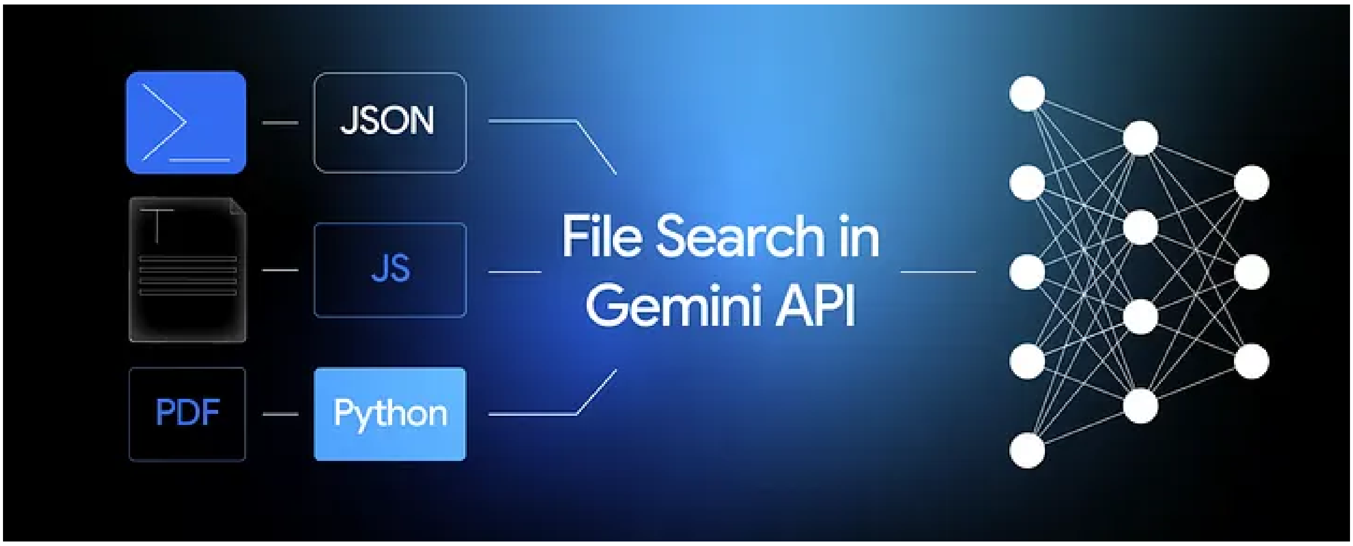

Image by — Google

...

What is this thing?

At its core, File Search enables Gemini to 'understand” your data. You can upload PDFs, DOCX files, text, JSON, or even code files. When you ask Gemini a question, it doesn’t just guess — it examines your uploaded files, finds the relevant parts, and uses them to respond.

It’s like connecting your personal brain directly to Gemini. No separate vector database, no retrieval pipeline, nothing to manage.

Just files in, answers out.

...

It’s cheap. Like, really cheap.

This is the part that surprised me. You don’t pay for queries or storage; you only pay once — when you index your files.

Embedding creation costs **$0.15 per 1 million tokens**, using the `gemini-embedding-001` model. That’s peanuts compared to building your own pipeline with tools like Pinecone or Weaviate.

After that, you can freely query those files as much as you want.

...

How it actually works

File Search simplifies RAG by automatically chunking files, generating embeddings, storing and retrieving them, and injecting context into your Gemini prompts.

It’s all handled inside the same `generateContent` API call you already use.

When you make a query, it performs a **vector search** behind the scenes using the latest Gemini Embedding model. So it understands *meaning*, not just keywords.

Even better: Gemini's answers include **citations** — it explicitly states which file and section it sourced from. You can click through and verify it. No more guessing if your model hallucinated something.

. . .

Use case: Beam's crazy fast game generation

One of the early testers, **Phaser Studio**, is using File Search on their AI-powered game platform, Beam.

They have a library of over 3,000 files — including templates, code snippets, design documents, and other internal data. File Search allows them to query all of it in less than 2 seconds. Previously, it took *hours* to manually find the same information.

Richard Davey, their CTO, summed it up perfectly:

> "Ideas that once took days to prototype now become playable in minutes."

That's pretty wild.

. . .

A quick example in Python

You don't need much code to get started. Here's a simple example:

```
from google import genai
from google.genai import types
import time

client = genai.Client()
store = client.file_search_stores.create()
upload_op = client.file_search_stores.upload_to_file_search_store(
    file_search_store_name=store.name,
    file='path/to/your/document.pdf'
)
while not upload_op.done:
    time.sleep(5)
    upload_op = client.operations.get(upload_op)
response = client.models.generate_content(
    model='gemini-2.5-flash',
    contents='Summarize the research on sustainable AI.',
    config=types.GenerateContentConfig(
        tools=[types.Tool(
            file_search=types.FileSearch(
                file_search_store_names=[store.name]
            )
        )]
    )
)
print(response.text)
grounding = response.candidates[0].grounding_metadata
sources = {c.retrieved_context.title for c in grounding.grounding_chunks}
print('Sources:', *sources)
```

That’s it. Upload files, ask a question, and receive a sourced answer. Done.

. . .

Why this matters

Every AI developer encounters the same issue — models sound impressive but lack access to your company’s internal data.

File Search changes that by enabling Gemini to analyze *your* content without a complex retrieval setup.

If you’re creating anything that requires current or domain-specific information — support bots, internal tools, document Q&A — this is a game changer.

. . .

Try it yourself

You can try File Search now in **Google AI Studio**. There’s a demo called *“Ask the Manual”* — upload some files, ask questions, and see how well it grounds the answers.

Once you get the hang of it, you can remix the demo or integrate it directly into your app.

Google Cloud Platform   Rags   AI   Ai Agent   Ai Tools

Follow

Published in Coding Nexus

8.3K followers · Last published just now

Coding Nexus is a community of developers, tech enthusiasts, and aspiring coders. Whether you’re exploring the depths of Python, diving into data science, mastering web development, or staying updated on the latest trends in AI, Coding Nexus has something for you.

Following

Written by Civil Learning

3.3K followers · 6 following

We share what you need to know. Shared only for information.

Responses (1)

Bgerby

What are your thoughts?

Daniel Twum
4 hours ago

Merci beaucoup for the write-up. Is this NotebookLM repackaged or revisioned?

1 reply Reply

More from Civil Learning and Coding Nexus

In Coding Nexus by Civil Learning

Google's New LLM Runs on Just 0.5 GB RAM — Here's How to Fine-Tune It Locally"

A few days ago, Google quietly released a little AI model called Gemma 3 270M.

✦ Aug 15 👏 2.1K 💬 30

In Coding Nexus by Code Coup

Claude Desktop Might Be the Most Useful Free Tool You’ll Install This Year

I didn’t expect much when I first saw the announcement for Claude Desktop. Another AI wrapper, I thought. Maybe with a shiny UI.

✦ Oct 23 👏 953 💬 37

In Coding Nexus by Civil Learning

The Guy Who Let ChatGPT Trade for Him—and Somehow It Worked

You know how everyone says, “Don’t let AI touch your Money”? Well, someone on Reddit decided to ignore that.

✦ Oct 8 👏 274 💬 8

In Coding Nexus by Civil Learning

AGENTS.md: The File That Saves You From Dumb AI Code

If you've ever thought, "This AI code is smart but also dumb," you'll get this.

Sep 18 595 8

See all from Civil Learning

See all from Coding Nexus

Recommended from Medium

In Level Up Coding by Fareed Khan

Building a Training Architecture for Self-Improving AI Agents

RL Algorithms, Policy Modeling, Distributed Training and more.

4d ago 723 14

In Coding Nexus by Code Coup

The Top 5 Local LLMs (GLM4.5 GPT-OSS, Qwen3)

AI isn’t just for data centres anymore — it’s arriving at your desk.

✦ 6d ago 👏 97 💬 1

In Towards AI by Teja Kusireddy

We Spent $47,000 Running AI Agents in Production. Here’s What Nobody Tells You About A2A and MCP.

Multi-agent systems are the future. Agent-to-Agent (A2A) communication and Anthropic’s Model Context Protocol (MCP) are revolutionary. But...

Oct 16 3.3K 111

In Generative AI by Gao Dalie (高達烈)

DeepSeek-OCR + LLama4 + RAG Just Revolutionized Agent OCR Forever

During the weekend, I scrolled through Twitter to see what was happening in the AI community. Once again, DeepSeek has drawn worldwide...

✦ Oct 27 399 2

Joe Njenga

6 Google AI Coding Tools Making Developers Faster(And Cutting Budgets by 50%)

Gemini 3.0 is about to shake the AI coding space, and if you are not using one of these Google AI coding tools, you are behind or losing...

✦ 4d ago 👏 145 💬 2

The Atomic Architect

You Haven’t Seen AI Until You Try Claude Sonnet 4.5’s New Feature—It Redefines Insane

I built a working expense tracker in eight minutes while my coffee was still hot, and it remembered every receipt when I closed my laptop...

✦ Oct 26 👏 325 💬 19

See more recommendations