Gemini API file slicing: send only the pages, cells and clips a model needs

26 November 2026 · 4 min read · blogs.iictp.org/gemini-api-file-slicing-controls-token-cost-explained

Short answer: Gemini API file slicing is a set of controls, launched by Google DeepMind on 23 November 2026, that lets a developer or agent ask the server to load only chosen pages, spreadsheet ranges or media segments from an uploaded file. Google says you are billed tokens only for the parts the model inspects, and that Office and OpenDocument files are now accepted natively. It matters because long documents and recordings are a common source of wasted context and cost in agent workflows.

TL;DR The short version

- Gemini API file slicing, launched on 23 November 2026, lets a developer or agent load only chosen pages, spreadsheet ranges or media segments from an uploaded file.
- Google says you pay tokens only for the parts the model inspects, and Office and OpenDocument files are now accepted natively.
- Long documents and recordings are a common source of wasted context, so test slicing on your heaviest agent workflows and compare cost and accuracy.

Read online, listen to it, or change the text size: https://blogs.iictp.org/gemini-api-file-slicing-controls-token-cost-explained

What is file slicing in the Gemini API?

Instead of converting a whole file to text, or sending every page as an image, you pass a Files API URI with processing options that say which range to read and in what form. Rohan Doshi and Ben Caine of Google DeepMind say it works across all Gemini 4 models and is available in Google AI Studio and the Gemini platform. Existing file size and page limits still apply to the uploaded file.

How it works

- **Documents and slides** (PDF, Office, OpenDocument): choose a start and end page, a resolution of low, medium or high, and text or page-image layers.
- **Spreadsheets** (.xlsx, .xls, .ods): choose a sheet and a cell range, and layers for data, formulas and visuals.
- **Video:** clip a time interval, set frame rate and resolution, and pick transcript, frames or audio.
- **Audio:** clip a time interval.
- **Inside agent loops:** the file can be referenced in a function response as well as a user turn, so the agent chooses what to open next.

What is new compared with whole-file ingestion

Google’s argument is that when a model API accepts only whole files, agent harnesses end up running office converters, PDF renderers and video tools on the client to cut files up. That uses memory and disk and, for shared platforms, forces per-user isolation just to slice a document. Server-side slicing moves that step into the API and lets one upload be reused across turns.

Defaults reported by the source

- If you omit processing options, documents and slides are read as layout-preserving page images.
- Spreadsheets by default include evaluated values, formulas and sheet names for every tab.
- The announcement shows a two-step pattern: skim all pages as extracted text with a smaller model to find the right section, then open that one page at high resolution with a stronger model.
- The post says partners validated token savings before launch; it gave no figures in the text we read, so measure your own.

Limitations and open questions

- Savings depend on how well the first, cheap pass finds the right range. A wrong page means a wrong answer delivered at low cost.
- Text-only skimming can miss information that exists only in charts, scans or layout.
- The controls are specific to this API, so portable agent code needs an adapter.
- Check data handling and retention terms before uploading sensitive files. Sources were read on 26 November 2026. Products, rules and standards change, so check the linked originals before you rely on any detail.

How to try it

1. Pick one workflow that currently sends whole files, and record tokens and accuracy per task.
2. Upload each file once with the Files API and keep the URI.
3. Skim with text extraction or transcripts at low resolution to locate the relevant range.
4. Re-request only that range at higher resolution and ask the question again.
5. Compare cost and error rate with your baseline, including cases where the answer sits in a chart or a formula.

What it means for developers

The design choice is a retrieval step before reasoning, applied to a single file. Log which ranges the agent opened so that a reviewer can check the evidence behind an answer. For tool-heavy agents, the related technique is to load tool definitions only when needed: see Gemini API tool retrieval.

Frequently asked questions

- **What is Gemini API file slicing?** Server-side controls that load only selected pages, cell ranges or media segments of a file into the model.

- **Which file types are supported?** The post lists PDF, Office and OpenDocument documents and slides, spreadsheets, video and audio.
- **Which models support it?** Google says all Gemini 4 models.
- **Does it lower cost?** Google says tokens are billed only for the inspected parts. The size of the saving depends on your files and prompts.

Related reading

- Gemini API tool retrieval
- TokenEconomizer: cutting coding agent token cost
- What is an LLM context window?
- AI Practitioner certification

> Related credential: AI Practitioner. Explore the exam facts and syllabus.

Sources and further reading

1. Rohan Doshi and Ben Caine, Google DeepMind. Introducing file slicing controls in the Gemini API. 23 November 2026. blog.google/innovation-and-ai/technology/developers-tools/file-slicing-controls

Published by IICTP, an independent technical credential body, on 26 November 2026. The current version of this post is at https://blogs.iictp.org/gemini-api-file-slicing-controls-token-cost-explained. More guides: blogs.iictp.org.