

Why You Should Stop Uploading PDFs to Your AI Tools

A practical guide to Markdown and why it matters



The problem nobody talks about

If you use AI tools regularly, whether that's ChatGPT, Claude, Gemini, or any other model, you've almost certainly uploaded a PDF at some point. A report, a contract, a set of meeting notes, a research paper.

It feels like the obvious thing to do. PDFs are how documents get shared. They look clean, they preserve formatting, and every AI tool accepts them. So, you drag the file in, ask your question, and wait for the answer.

What you probably don't know is that uploading a PDF is one of the least efficient things you can do when working with AI. Not because the AI can't read it, but because of what the AI has to wade through to get to the content. And on paid plans, you're paying for every bit of that overhead.

What happens when you upload a PDF

AI models don't read documents the way humans do. They don't open a file and skim to the relevant section. They process every token of the input, and the format of that input determines how many tokens they consume and how cleanly the content comes through.

PDF was designed for printing and display. It's a format optimised for human eyes on a screen or page. Underneath the surface, a PDF is a complex file containing font tables, layout coordinates, repeated headers and footers on every page, column-break markers, and encoding instructions. None of that information helps an AI understand your document. But when the AI extracts text from a PDF, all of that overhead comes with it. Every redundant character gets counted as a token.

Tokens are the fundamental unit of AI processing. They roughly correspond to about three-quarters of a word in English. Every AI model has a context window, a limit on how many tokens it can hold in memory during a single conversation. When you upload a large PDF, you're burning a significant portion of that window on formatting noise before the AI even reaches your actual content. On API or business plans, you pay for every token consumed.

The numbers are striking. Research comparing the same document in PDF and Markdown format consistently shows PDFs consuming around **three times more tokens**. In one documented case study of a 307-page business book, the PDF version consumed 6.3 times more tokens than the Markdown equivalent. The PDF used 46% of a 200,000-token context window. The Markdown version used 7%.

What Markdown is and why it works better

Markdown is a plain text format with minimal, human-readable structure. It was created in 2004 as a way to write for the web without dealing with HTML. A pound sign creates a heading. A hyphen at the start of a line creates a bullet point. Double asterisks make text bold. That's most of what you need to know.

There's no hidden metadata. No font information. No layout instructions. No print formatting. Just the content, structured in a way that both humans and AI models can read cleanly and efficiently.

When you feed a Markdown file to an AI tool, the model receives the actual substance of the document with minimal overhead. It can process more content within the same context window, maintain better comprehension across long documents, and return more accurate answers because it's working with clean, well-structured input rather than a stream of text cluttered with formatting artefacts.

Every major AI model, including ChatGPT, Claude, Gemini, and the open-source models, treats Markdown as a first-class format. It doesn't need configuration or special handling. You simply paste or upload the file, and the AI reads it as intended.

When PDFs are still the right choice

This isn't an argument against PDFs as a format. PDFs are excellent for printing, for legal documents that require guaranteed formatting, for visual-heavy content with diagrams and photographs, and for archival purposes where consistent rendering matters across decades.

The point is narrower: for feeding content into AI tools, PDFs are the wrong format. They were designed for a different purpose, and that purpose introduces overhead that actively reduces the quality of your AI interactions.

If you're working with a scanned PDF, the situation is even worse. A scanned document is essentially an image, and the AI has to perform something close to optical character recognition to extract the text. The token cost is higher still, and the accuracy of the extraction can be unreliable depending on the quality of the scan.

How to convert documents to Markdown

The good news is that converting documents to Markdown is straightforward, and in most cases takes less than a minute.

From a Word document or Google Doc

Copy the content and paste it into a free Markdown editor such as Dillinger, StackEdit, or Typora. The basic structure, headings, bullets, bold text, will usually convert cleanly. Save the file with a .md extension.

From a PDF

Several free online tools convert PDF to Markdown reliably. Pandoc is a powerful command-line tool that handles complex documents well. For simpler documents, pasting the text from a PDF into a Markdown editor and adding basic structure manually takes only a few minutes and produces a cleaner result than most automated conversions.

From a webpage

Browser extensions such as MarkDownload convert any webpage to Markdown in a single click, which is useful for feeding articles, reports or documentation into AI tools.

For regular workflows

If you're frequently working with the same types of documents, it's worth setting up a simple conversion step as part of your process. A team processing dozens of contracts or reports through AI tools each week will notice a meaningful difference in both the quality of outputs and the efficiency of the workflow.

The practical difference it makes

Beyond the token efficiency, using Markdown rather than PDFs tends to produce better AI outputs for a simple reason: the AI is working with cleaner, better-structured input.

When a document is well-structured in Markdown, with clear headings, logical sections, and clean paragraph breaks, the AI can navigate it more accurately. It's less likely to confuse a repeated page header for content, less likely to lose the

thread across a page boundary, and more likely to locate the specific information you're asking about.

For long documents in particular, this matters significantly. A 50-page PDF that burns 40% of a context window on formatting overhead leaves the AI little room to also hold your question, your conversation history, and its own response. The same document as Markdown might use 15% of the context window, leaving ample space for a thorough, well-reasoned answer.

A simple habit worth building

None of this requires technical expertise. Converting documents to Markdown is a small habit change with a disproportionate return. You don't need to understand tokenisation in detail or know how context windows work under the hood. You just need to know that the format you choose changes what the AI can do with your content.

If you regularly upload PDFs to AI tools, try converting one to Markdown and asking the same question. The difference in response quality is often noticeable immediately.

The AI hasn't got smarter. You've just given it better input to work with.

Rob May is a keynote speaker, author and Executive Chairman of ramsac, one of the UK's leading managed security service providers. Through his keynotes and masterclasses, he helps leaders harness AI, strengthen cybersecurity and drive business transformation.

<https://thoughtprovoked.co.uk>