Garbage In, Garbage Out

Ingestion sets the ceiling on RAG quality — on-prem OCR and page-faithful capture across PDF and Microsoft Office

Retrieval-augmented generation lives or dies on what enters the index. Yet most RAG stacks pour their engineering into retrieval and generation and treat ingestion as a solved problem — strip the text, drop the layout, discard the page, move on. The hardest documents and the most decision-relevant context are lost before a single query runs. **GoodMem** is built the other way around — it does the hard work at ingestion, where the quality of every later answer is decided. Enterprise deployments add **on-prem document OCR** that keeps every page inside your network while parsing within reach of the best frontier models, and **page-faithful capture** — a rendered image of every page, across PDF and all three Microsoft Office formats, Excel included — so answers can be checked against the source and read directly by multimodal models.

4 formats
Page images at ingestion
PDF, Word, PowerPoint and Excel — most stacks keep text only

On-prem
Document OCR, zero cloud egress
runs on your hardware; nothing leaves your network

≈ frontier
Document-parsing accuracy
ranks #2 on public benchmarks, behind only the leading frontier model

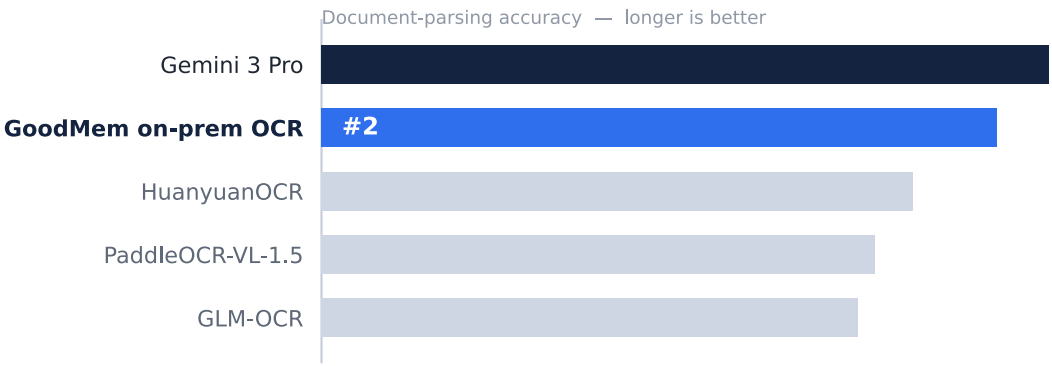

Figure 1 · Document-parsing accuracy across several standard public benchmarks. The model GoodMem deploys on-prem ranks second — behind only a leading frontier system, ahead of every other model tested — while running entirely on customer hardware. Source: published benchmark results, May 2026.

Retrieval can't recover what ingestion discards

A RAG system is judged on retrieval and generation, but both inherit a hard ceiling from ingestion: a model can only retrieve and reason over what was captured in the first place. The default pipeline — extract a stream of text, split into chunks, embed — quietly throws away three things that enterprise documents depend on. **Layout**: which row belongs to which header, which number to which column, the reading order of a multi-column page. **Visual context**: the rendered page a human uses to trust an answer. And **the hardest documents entirely**: scanned contracts, image-only PDFs, dense financial tables, slide decks that carry their meaning in their composition. What is dropped at ingestion cannot be retrieved at query time, no matter how good the embeddings.

On-prem OCR: frontier-class parsing, nothing leaves the building

For visual and scanned documents, GoodMem's enterprise deployments run document OCR locally, on a self-hosted model that stays on your own hardware. Parsing is layout-aware: titles, tables, formulas, and reading order are

detected in a single pass, with tables emitted as HTML and formulas as LaTeX rather than flattened into a text blob. Two properties make it enterprise-grade. First, it is **on-prem**: the model runs behind your firewall, images are passed to a local endpoint, and no document or page is sent to a third-party cloud OCR service. Second, it is **accurate**: on recent public document-parsing benchmarks it ranks just behind the leading frontier system and ahead of every other model tested (Figure 1) — frontier-class quality without frontier-cloud data exposure.

Pages, not just text — and Excel, not just PDF

Beyond text, GoodMem captures a rendered image of every page at ingestion: a faithful raster of the page exactly as a person would see it. This is rare — most ingestion keeps only extracted text, or at best crops of embedded figures — and it pays off twice. **Verification**: a retrieved answer can be shown against its actual source page, so a user sees where it came from and trusts it. **Multimodal reading**: the preserved page image can be handed directly to a multimodal model, which often pulls more from the rendered page — a chart, a stamped signature, a merged-cell table — than any text chunk carries. The differentiator is breadth and method: GoodMem renders page images not only for PDF but for **all three Microsoft Office formats — Word, PowerPoint, and Excel** — and does it in **pure Java, in process**, with no LibreOffice or external office engine in the deployment. The few ingestion products that render Office pages at all shell out to a headless LibreOffice install and commonly stop short of spreadsheets — one prominent multimodal RAG engine renders Word and PowerPoint but deliberately excludes Excel. We are not aware of another ingestion product that returns page images for **Excel** specifically: the densest, most table-heavy, and least-served format in the enterprise.

Ingestion capability	Typical RAG stack	**GoodMem** · enterprise
Layout-aware text extraction	often flattened	**tables + reading order**
On-prem document OCR (no cloud)	cloud API, or none	**self-hosted**
Page image — PDF	sometimes	**Yes**
Page image — Word / PowerPoint	rare (LibreOffice)	**Yes · pure Java**
Page image — Excel	≈ none	**Yes · pure Java**
Source page linked to retrieved chunk	—	**Yes · on demand**

What it changes downstream

Better inputs propagate. Layout-aware extraction yields chunks that mean something — a table row stays a row, not a scramble of numbers. On-prem OCR brings the scanned and image-only documents that plain extraction silently drops into the index at all. And a page image per result turns an opaque answer into a grounded one: the user sees the source, and a multimodal model reads the page directly.

Built for the enterprise

Every capability here is built to run inside the customer's boundary. OCR is self-hosted — no document leaves the network for a cloud parser. Page rendering is pure Java, in process: no LibreOffice install, no headless-office sidecar to operate or secure. Rendered pages are stored inline with the data. Self-hostable, built on open standards, and every answer can be checked against its source page.

Why PAIR Systems

PAIR Systems builds the autonomous information-retrieval layer for enterprise AI. GoodMem brings capabilities once confined to the world's best search teams into any product as self-hostable software — including the ingestion quality that decides everything downstream. PAIR's founder helped pioneer practical zero-shot neural retrieval at Google and productized enterprise retrieval at Vectara.

Methodology. Capability comparison reflects a June 2026 review of public documentation and source for widely used RAG ingestion systems (Unstructured, LlamaParse, LangChain, Reducto, Morphik, Vectara, Azure AI Document Intelligence, AWS Bedrock Knowledge Bases). "Page image" = a rendered full-page raster, not extracted text or cropped figures. Several render PDF page images; a minority render Word / PowerPoint via headless LibreOffice, and we found none returning Excel page images. The parsing-accuracy ranking (Figure 1) reflects published results across several standard document-parsing benchmarks (May 2026); independent standings vary. GoodMem capabilities are verified against the GoodMem server source.