**Introduction**

Organizations worldwide generate millions of PDF documents annually, with a substantial portion of enterprises relying on document analysis tools to extract meaningful insights from their files. Yet despite widespread adoption of PDF technology, we face a significant challenge in advancing AI-powered document processing.

The problem is fundamental: while Large Language Models have transformed how we work with text, they encounter substantial difficulties when processing PDFs directly. When AI systems require clean, structured input but receive poorly formatted documents instead, processing inefficiencies multiply throughout the workflow.

In high-stakes industries like healthcare, finance, and law, these conversion challenges aren't merely inconvenient—they pose real risks. Medical records, legal contracts, and financial reports contain precisely structured information that must remain intact during processing. Even minor formatting errors can lead to misinterpreted data, flawed analysis, or costly operational mistakes.

**The Research Mission**

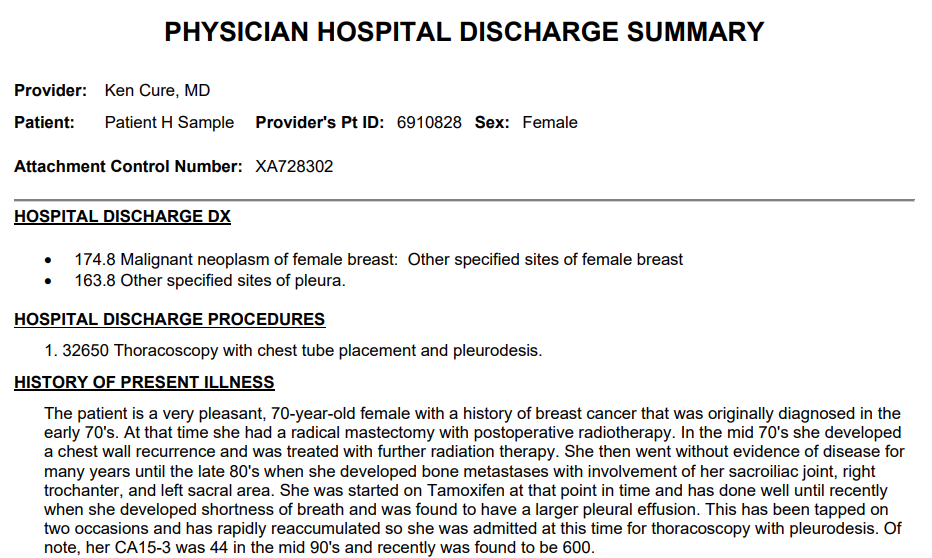
This study tackles a crucial question: which AI model performs best at converting PDFs to markdown format for downstream AI applications? Rather than simply measuring how well text gets extracted, we're evaluating what really matters for AI processing: whether the converted documents preserve semantic structure, retain factual accuracy, and maintain contextual relationships.

For organizations like Corpus Aid—which uses AI to help professionals create and review documents—choosing the right conversion technology directly impacts service quality and customer satisfaction. By identifying which models excel at preserving both content and context, this research enables more accurate document analysis, better AI-generated outputs, and significantly reduced manual correction work.

**Experimental Setup**

Evaluating PDF-to-markdown conversion requires documents that represent real-world complexity. We selected five clinical documents that span from straightforward cases to establish baseline performance to complex, multi-page records that truly test a model's limits.

**Document 1: Physician Hospital Discharge Summary (1 page)** This serves as our baseline—a well-structured medical document with standard sections like patient demographics, diagnoses, and discharge instructions.



**Document 2: Pneumonia Discharge Summary – John Smith (2 pages)** Here's where things get interesting: identical medical information presented in two different styles—narrative prose and structured format.

A medical report with a medical report

AI-generated content may be incorrect.

**Document 3: Emergency Department Observation Chart (1 page)** This document throws spatial complexity into the mix with multi-column layouts, checkbox lists, and blank form fields.

A medical application form with a red border

AI-generated content may be incorrect.

**Document 4: Hospital Discharge Summary – Statistical Report (1 page)** Numbers matter in healthcare. This document focuses on quantitative precision with large tables containing percentages and decimal values.

A screenshot of a medical report

AI-generated content may be incorrect.

**Document 5: Outpatient Clinic Summary – Complex Multi-page (4 pages)** Our most challenging test case: a hierarchical, multi-page document requiring preservation of nested sections and chronological progression across multiple treatment visits.

A close-up of a document

AI-generated content may be incorrect.

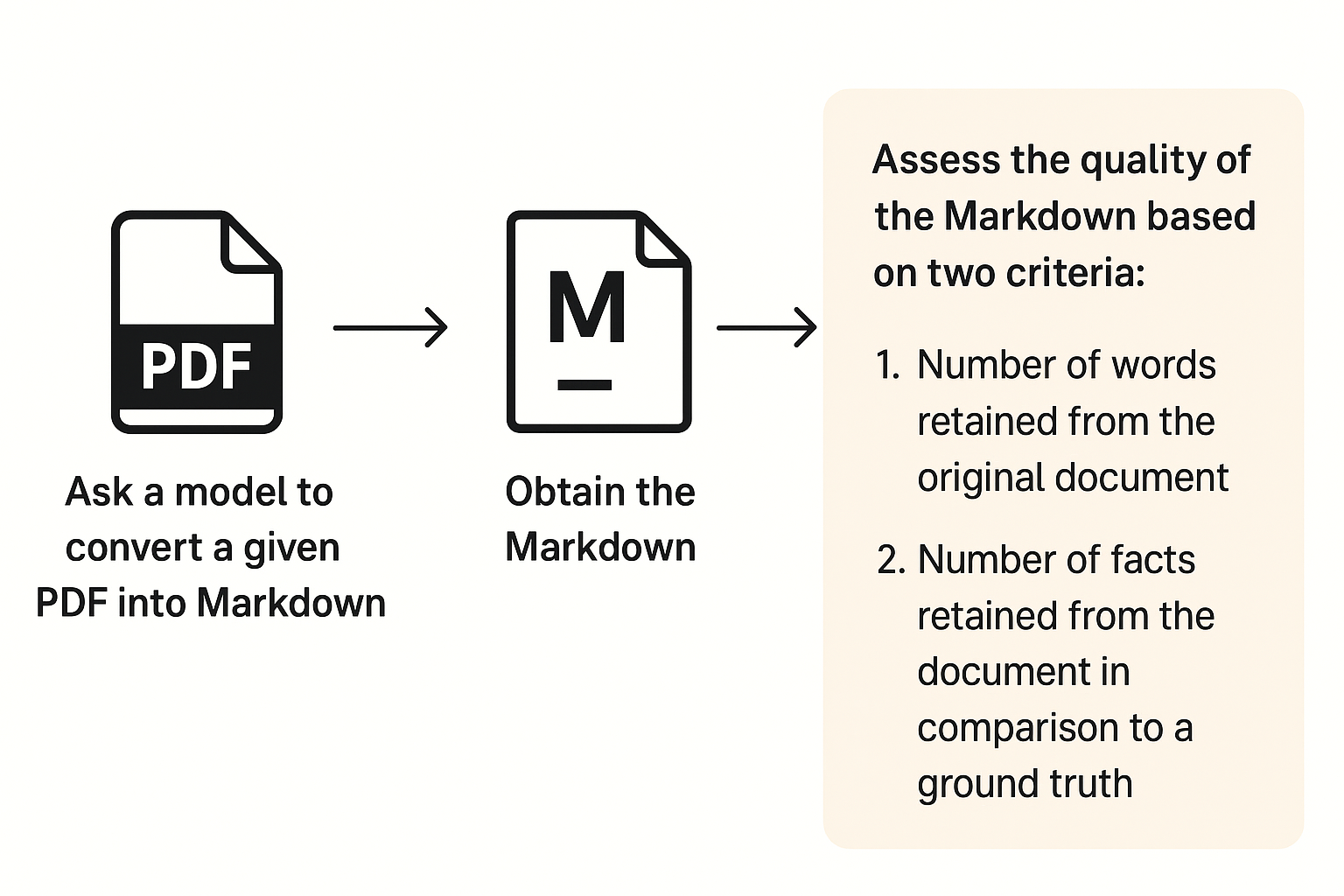
**Choosing the Right Models to Test**

We evaluated two distinct categories of AI models, each representing different approaches to the conversion challenge.

**Multipurpose Large Language Models** like Claude Sonnet 4, Gemini 2.5 Flash, and ChatGPT-4o Mini represent the "go-to" choice for most healthcare organizations. These models are widely available through established APIs and have proven track records in structured text generation. By testing these general-purpose systems, we're answering a practical question: can the AI tools most organizations already have access to handle clinical document conversion effectively?

**Specialized Document Conversion Systems** such as Mistral OCR and Llamaparse were designed specifically for this type of work. They feature advanced layout recognition, precision OCR capabilities, and sophisticated hierarchical structure preservation. These models help us establish a performance ceiling—showing us what's possible when AI is purpose-built for document conversion.

**Experimental Workflow**



Our testing approach was deliberately straightforward to mirror real-world usage. Each model received the same simple instruction: "Convert this into markdown." No special prompting, no detailed formatting guidelines.

**Word Retention Analysis**

The first question we needed to answer was fundamental: how much of the original document actually makes it through the conversion process? This isn't just about counting words—it's about understanding two critical behaviors that can make or break a conversion's usefulness.

* **Information preservation**: Does the model retain the original content faithfully?
* **Content addition patterns**: When the converted document contains more words than the original, it suggests the model is adding interpretive or explanatory content—which could be either helpful or problematic depending on accuracy.

We employ a "bag of words" approach to make these comparisons, treating each document as a collection of individual words rather than focusing on their exact sequence.

**Fact Accuracy and Placement**

Word counts only tell half the story. Our second metric tackles a more sophisticated challenge: does critical information end up where it belongs? Imagine a model that perfectly extracts a patient's home address from a medical form—but then places that address in the "Patient Name" field instead of "Contact Information." Technically, no information was lost, but the conversion has failed catastrophically for any downstream system trying to use that data.

To measure this precisely, we created ground truth markdown templates for each PDF, manually identifying every factual element and its correct location. We then evaluate each model's output to see if it has been correctly placed, misplaced or missing.

**Results**

Word Retention: The Quantity vs Quality Trade-off

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Chatgpt | Claude | Gemini | Mistral | Llama |
| Doc 1 | 1 | 1 | 1 | 1 | 1 |
| Doc 2 | 0.8 | 1.1 | 1.1 | 1.06 | 0.98 |
| Doc 3 | 1.8 | 1.7 | 1.57 | 0 | 1 |
| Doc 4 | 1 | 1 | 1 | 1 | 1 |
| Doc 5 | 0.358 | 0.99 | 1.05 | 0.99 | 0.99 |

Fact Accuracy: Where Precision Meets Practicality

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Chatgpt | Claude | Gemini | Mistral | Llama |
| Doc 1 | 100% | 100% | 100% | 100% | 100% |
| Doc 2 | 100% | 100% | 100% | 100% | 100% |
| Doc 3 | 100% | 95% | 93% | 0% | 90% |
| Doc 4 | 98% | 100% | 100% | 100% | 100% |
| Doc 5 | 100% | 100% | 100% | 98% | 100% |

**Where models performed well**

Every model achieved perfect word and fact retention (1.0) on documents 1 and 4, apart from Chatgpt where it performed really well at 98% fact retention — suggesting that well-structured, straightforward medical documents pose minimal challenges for current AI conversion technology.

What's particularly striking is the unanimous success on Document 4, our statistical report dense with tabular data, percentages, and decimal values across multiple rows. Initially, we anticipated this numerical complexity would challenge the models significantly. Instead, every model handled this numerical complexity flawlessly, suggesting that current AI conversion technology has matured significantly in processing structured data formats.

This unanimous success across both narrative medical documentation and complex statistical tables establishes that all tested models can handle standard clinical documentation reliably, regardless of whether the content is primarily textual or numerical.

**Where Models Show Their True Colors**

The real differences emerged with documents that contained more complex layouts. Document 2 (the two-page pneumonia discharge summary) revealed the first signs of divergent strategies. While Claude, Gemini, and Mistral slightly expanded content (1.1, 1.1, and 1.06 respectively), ChatGPT compressed it to 80% of the original length, and Llamaparse stayed nearly faithful at 98%. This pattern foreshadowed more dramatic differences in complex scenarios.

Despite the differences in word retention, all the models managed to retain all of the facts from their relevant sections.

**The Spatial Complexity Challenge**

Document 3, our emergency department observation chart with multi-column layouts and form fields—proved to be the ultimate stress test. Here, we see the most dramatic variation in approaches:

* **ChatGPT, Claude and Gemini** expanded content (1.8x, 1.7x and 1.57 respectively), suggesting these models interpret complex layouts by adding explanatory structure.
* **Mistral completely failed** (0% retention), unable to process the spatial complexity at all
* **Llamaparse** maintained perfect fidelity (1.0x), living up to its reputation as a specialized document parser

However, despite adding additional fields to interpret complex layouts Claude and Gemini were not able to achieve 100% accuracy in fact retainment, as it misplaced fields into the wrong section. Chatgpt however managed to place all fields to the correct section.

A medical survey form with a number of boxes

AI-generated content may be incorrect.A black and white screen with yellow text

AI-generated content may be incorrect.

**The Multi-Page Reality Check**

Perhaps most telling was Document 5, our four-page outpatient clinic summary that mirrors real-world complexity. For word retention, ChatGPT's dramatic compression to just 36% of the original content stands in stark contrast to the other models' near-perfect retention (0.99-1.05). This suggests ChatGPT prioritizes aggressive summarization over comprehensive preservation—a strategy that could be either beneficial or problematic depending on use case.

An interesting observation was that despite the compression from ChatGPT’s markdown, it is able to retain all of the facts from the document. Something that the Mistral OCR was not able to achieve despite the additional word retention from the original document as it missed out patient address from its output document.

**Model summary**

**ChatGPT**

ChatGPT emerges as the "smart summarizer," aggressively condensing lengthy patient condition descriptions while maintaining factual accuracy. For organizations dealing with verbose documentation where concise summaries aid downstream AI processing, this could be advantageous. However, in contexts where complete information preservation is legally or clinically required, this approach poses risks.

**Claude and Gemini**

Both models demonstrate sophisticated document understanding by adding logical headers and organizational structure that weren't explicitly present in source PDFs. For example, when encountering patient information fields in a form, they create "Patient Information" headers in the markdown. This interpretive enhancement could significantly improve downstream AI comprehension, though it raises questions about fidelity to source documents.

**Mistral**

Mistral's complete failure on spatially complex documents (Doc 3) while performing adequately on others raises questions about consistency across document types. However, this dramatic performance drop may reflect our experimental methodology rather than fundamental model limitations. Further testing with varied conditions is needed to determine whether this represents a true limitation or a testing artifact.

**Llamaparse**

Llamaparse consistently delivered the most faithful reproductions, adding minimal interpretive content while preserving structure accurately. This makes it ideal for organizations prioritizing exact document fidelity, though the slightly lower fact retention scores suggest that pure fidelity doesn't automatically translate to perfect information preservation.

**Conclusion**

The choice between models ultimately depends on three key organizational priorities: fidelity requirements, processing efficiency needs, and document portfolio diversity.

For organizations in highly regulated environments where every word matters legally, Llamaparse's commitment to faithful reproduction without adding new information makes it the safest choice, despite minor accuracy trade-offs. Healthcare systems focused on improving **AI-powered clinical decision** support should consider Claude or Gemini, whose structural enhancements could dramatically improve downstream processing without compromising essential information.