Introduction to Parsing in Retrieval Augmented Generation (RAG)

Yogesh Haribhau Kulkarni

Outline

Parsing is the key

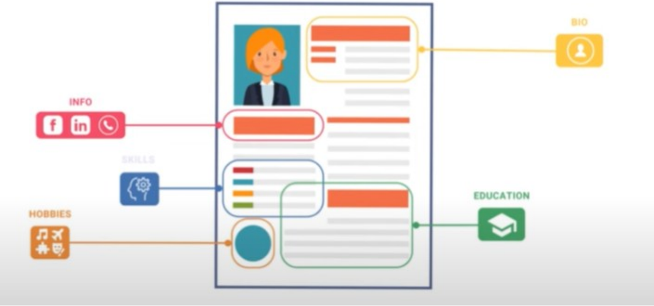
(Ref: Key to RAG Success: Document Parsing Explained - EyeLevel)

Document Parsing: The Foundation of RAG

- Parsing is the first step in any RAG pipeline.
- Bad parsing undermines even the best RAG strategies.
- Garbage in, garbage out: poor inputs = poor outputs.
- Many overlook parsing in favor of flashy AI tools.
- Without good extraction, nothing else matters.
- Most language models require clean, structured text.
- RAG applications depend on text quality from source docs.
- Advanced RAG still fails without reliable input.
- Models can't fix broken, messy data.
- Real-world systems have failed due to poor parsing.

What is Document Parsing?

- Converts formats like PDF, DOCX, HTML into usable text.
- Extracts meaningful content for language model input.
- Cleans, structures, and normalizes the data.
- Essential step before chunking or embedding.
- Involves handling many formats and edge cases.



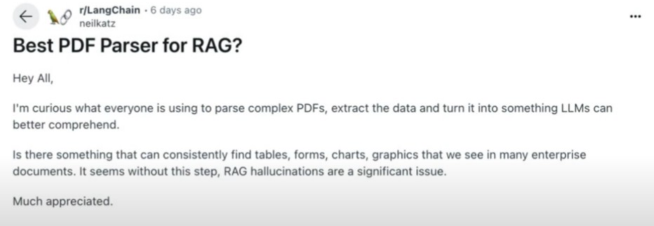
(Ref: Key to RAG Success: Document Parsing Explained - EyeLevel)

Common Misconceptions

- Engineers often ignore parsing during development.
- Focus tends to be on model tuning or retrieval logic.
- Parsing is wrongly assumed to be solved or trivial.
- Most systems lack formal evaluation of parsers.
- Homemade or ad-hoc solutions dominate practice.

The Reddit Survey Insight

- Survey on LangChain subreddit revealed no consensus.
- 57 replies yielded 30+ different parsing techniques.
- Most users hacked together informal solutions.
- Few performed proper parser evaluation or comparison.
- Highlights need for standardized testing and benchmarking.



(Ref: Key to RAG Success: Document Parsing Explained - EyeLevel)

Popular Parsing Tools Compared

- PyPDF – well-known, older, basic PDF extraction.
- Tesseract – OCR-based, handles scanned documents.
- Unstructured – handles messy formats, layout-aware.
- Tools vary widely in output quality and reliability.
- Choice depends on document type and project needs.

- Start by identifying your document types.
- Evaluate parsers with real-world examples.
- Compare outputs side by side.
- Look for structural fidelity, cleanliness, completeness.
- Test rigorously — don't rely on "it seems to work".

Real-World Example: Medical Bill

- Parsing tested on de-identified medical bill.
- Chosen for layout complexity and format irregularities.
- Shows strengths and weaknesses of each parser.
- Realistic example of what RAG apps encounter.
- Highlights need for resilient parsing strategies.

UNIVERSITY IMAGING CENTER

SUMMARY BILL OF ALL CHARGES

TAX ID: 0893765213
PLEASE REMIT PAYMENT TO:
7835 KESTER AVE SUITE 345
VAN NUYS CA 91405

BILLING STATEMENT	
Patient ID	Statement Date
PRE 987394	11/29/2022

Grant, Victoria
5436 S. Trent street
Ontario, CA 91761

ALWAYS REFERENCE PATIENT ID
PHONE : 818 654 3876
DOB : 05/16/1992

TECHNICAL IMAGING PROCEDURES AND BILLINGS

Technical Component,(Imaging) performed by : Dr. Vince

EXAM DATE	PROC. CODE	DESCRIPTION	MOD	B.PART	DX	AMOUNT
7652 Central Street , Montclair CA 91763						
11/17/2022	72141	MRI NECK SPINE W/O DYE	TC	CSPINE	M54.2	$ 1,600.00
11/17/2022	72148	MRI LUMBAR SPINE W/O DYE	TC	LSPINE	M54.5	$ 1,600.00

PROFESSIONAL PROCEDURES AND BILLINGS

EXAM DATE	PROC. CODE	DESCRIPTION	MOD	B.PART	DX	AMOUNT
Professional Component (Radiology Report) Interpreted by: Dr. Vince						
7652 Central Street , Montclair CA 91763						
11/17/2022	72141	MRI NECK SPINE W/O DYE	26	CSPINE	M54.2	$ 600.00
11/17/2022	72148	MRI LUMBAR SPINE W/O DYE	26	LSPINE	M54.5	$ 600.00

Total Charges	Total Payments	Total Adjustments	Balance Due
$4,400.00	$0.00	$0.00	$4,400.00
Internal use Only: 9/27/2022			

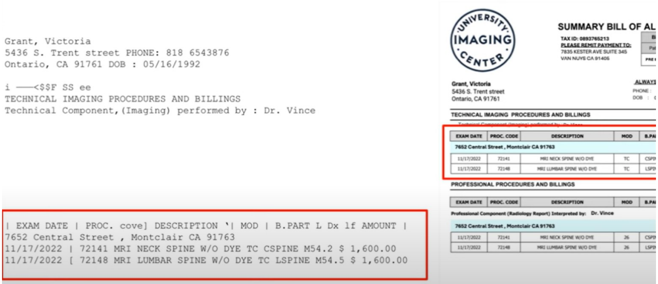
ADDITIONAL OR REVISED BILLING MAY OCCUR
PLEASE EMAIL BILLING@UIC.COM TO CONFIRM FINAL BILLING AMOUNT

Limitations of PyPDF

- PyPDF ok for texts but struggles with complex formats like tables.
- It failed to extract key data from real-world docs.
- Parsing tables often results in empty or broken content.
- Not due to PyPDF's fault—PDFs are inherently hard to parse.
- Many PDFs use inconsistent encoding and layouts.

Tesseract OCR: Strengths and Weaknesses

- Tesseract uses image-based OCR to extract text.
- Better at recognizing tables than PyPDF.
- Still introduces errors in column alignment.
- Column headers often get merged or misread.
- OCR also introduces spelling mistakes (e.g. "Cove" vs. "Code").

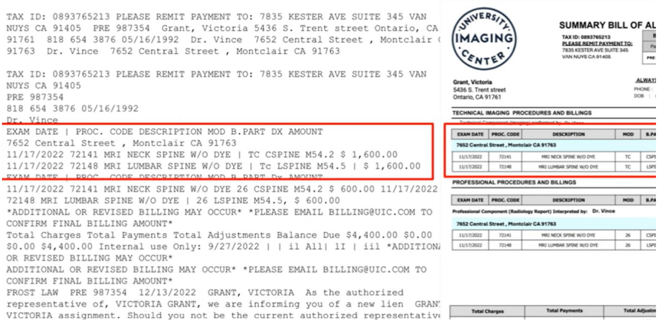


Challenges with OCR Outputs

- Language models must infer structure from broken text.
- Humans can "guess" meaning—models may not.
- Noisy extractions increase risk of incorrect answers.
- Inconsistent column separation confuses models.
- Clean layout is crucial for reliable RAG responses.

Unstructured: A Common Default

- Popular choice 'Unstructured' company; default in LangChain integrations.
- Handles layout better than OCR in some cases.
- Still suffers from column misalignments.
- Model must rely on context instead of structure.
- Reasonable quality, but far from perfect.

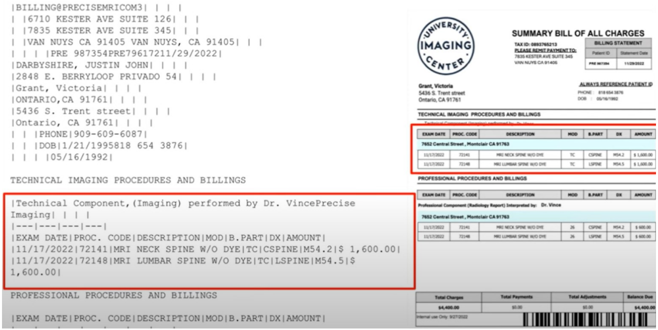


Parsing Tradeoffs: Model vs. Parser

- Many teams focus on improving the model first.
- Upgrading the parser might yield better gains.
- Better input can reduce model burden.
- Smaller or older models benefit more from clean text.
- Strong parsing reduces reliance on inference tricks.

LlamaParse: Cleaner Table Extraction

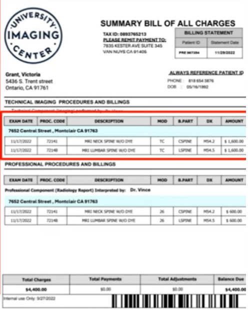
- Developed by LlamaIndex, supports markdown output.
- Clearly separates rows and columns with pipes.
- Markdown format improves model interpretability.
- Some formatting quirks but largely usable.
- Outperforms other parsers in structural clarity.



X-ray Parser: Multimodal Approach

- Combines vision models with grounding strategies.
- Detects tables and layout visually before parsing.
- Converts visual structure into usable text format.
- Produces reliable and model-friendly outputs.
- Especially effective on visually complex documents.


```
[
  {
    "summary": "The following table contains details of technical imaging procedures and billings performed by Dr. Vince at the University Imaging Center. It includes exam date, procedure code, description, modifier, body part, diagnosis, and amount."
  },
  {
    "AMOUNT": "$1,600.00",
    "B.PART": "CSPINE",
    "DESCRIPTION": "MRI NECK SPINE W/O DYE",
    "DX": "M54.2",
    "EXAM DATE": "11/17/2022",
    "MOD": "TC",
    "PROC. CODE": "72141"
  },
  {
    "AMOUNT": "$1,600.00",
    "B.PART": "LSPINE",
    "DESCRIPTION": "MRI LUMBAR SPINE W/O DYE",
    "DX": "M54.5",
    "EXAM DATE": "11/17/2022",
    "MOD": "TC",
    "PROC. CODE": "72148"
  }
]
```

UNIVERSITY IMAGING CENTER

SUMMARY BILL OF ALL CHARGES

TAX ID: 0893765213
PLEASE REMIT PAYMENT TO:
7835 KESTER AVE SUITE 345
VAN NUYS CA 91406

BILLING STATEMENT	
Patient ID	Statement Date
[illegible]	11/29/2022

Grant, Victoria
5436 S. Trent street
Ontario, CA 91761

ALWAYS REFERENCE PATIENT ID
PHONE : 818 654 3876
DOB : 05/16/1992

TECHNICAL IMAGING PROCEDURES AND BILLINGS

EXAM DATE	PROC. CODE	DESCRIPTION	MOD	B.PART	DX	AMOUNT
7652 Central Street , Montclair CA 91763						
11/17/2022	72141	MRI NECK SPINE W/O DYE	TC	CSPINE	M54.2	$ 1,600.00
11/17/2022	72148	MRI LUMBAR SPINE W/O DYE	TC	LSPINE	M54.5	$ 1,600.00

PROFESSIONAL PROCEDURES AND BILLINGS

EXAM DATE	PROC. CODE	DESCRIPTION	MOD	B.PART	DX	AMOUNT
Professional Component (Radiology Report) Interpreted by: Dr. Vince						
7652 Central Street , Montclair CA 91763						
11/17/2022	72141	MRI NECK SPINE W/O DYE	26	CSPINE	M54.2	$ 600.00
11/17/2022	72148	MRI LUMBAR SPINE W/O DYE	26	LSPINE	M54.5	$ 600.00

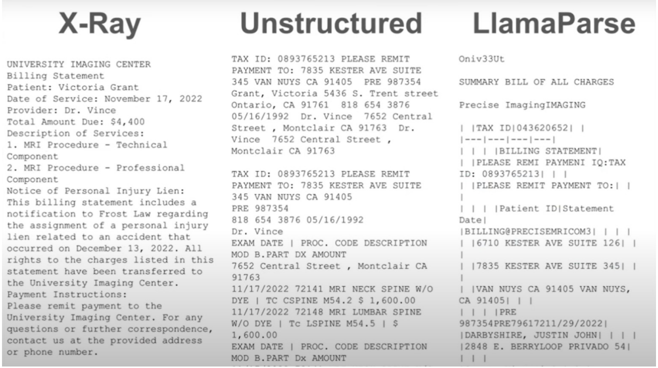
Total Charges	Total Payments	Total Adjustments	Balance Due
$4,400.00	$0.00	$0.00	$4,400.00

Internal use Only: 9/27/2022

YHK

Table Extraction Comparison

- PyPDF fails with complex tables.
- Tesseract detects tables but mangles headers.
- Unstructured does OK, but not perfectly.
- LlamaParse gives clean markdown tables.
- X-ray produces structured, grounded output.



Narrative + JSON: A Guided Format

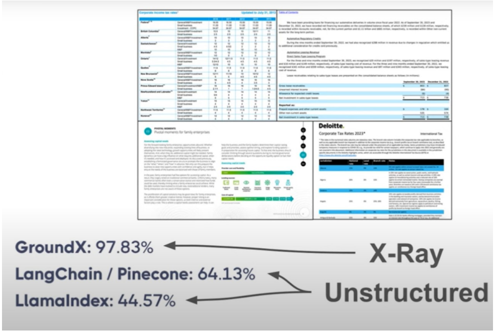
- Output begins with a narrative summary for context.
- Clearly explains the purpose and structure of the table.
- Follows with a clean JSON representation of table data.
- Format: cell-by-cell structured, easy to interpret.
- "Tell-then-show" approach improves model comprehension.

Why Output Format Matters

- Parsers may extract the same data, but format it differently.
- Output style can strongly affect model performance.
- JSON and markdown help structure information clearly.
- Human-readable structure supports better inference.
- Cleaner format = better grounding for language models.

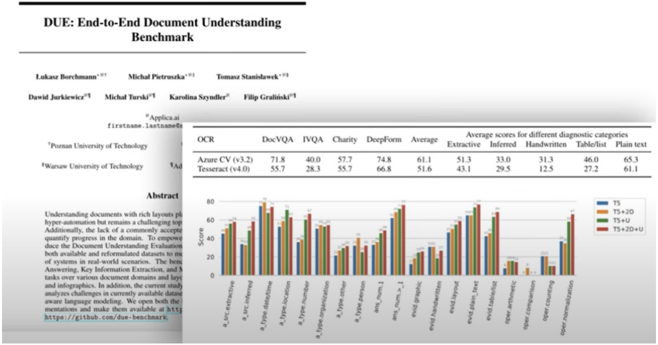
Impact of Parsing Quality

- We ran the same RAG pipeline over identical documents.
- Different parsers resulted in drastically different performance.
- Main variable: parsing quality—not model or retriever.
- Shows how foundational parsing is to good RAG results.
- A poor parser can undermine even advanced models.



Benchmark Setup (Clarified)

- **LlamaIndex**: Used PyPDF (not LlamaParse).
- **LangChain + Pinecone**: Used Unstructured.
- **GroundX**: Used X-ray with vision + grounding.
- All pipelines ran the same questions on same documents.
- Parser choice significantly influenced accuracy.



DUE: End-to-End Document Understanding Benchmark

Łukasz Borchmann, Michał Pietruszka, Tomasz Stanisławek, Dawid Jurkiewicz, Michał Turski, Karolina Szyndler, Filip Graliński

Applica.ai

Poznan University of Technology

Warsaw University of Technology

Abstract

OCR	DocVQA	IVQA	Charity	DeepForm	Average	Average scores for different diagnostic categories: Extractive	Inferred	Handwritten	Table/list	Plain text
Azure CV (v3.2)	71.8	40.0	57.7	74.8	61.1	51.3	33.0	31.3	46.0	65.3
Tesseract (v4.0)	55.7	28.3	55.7	66.8	51.6	43.1	29.5	12.5	27.2	61.1

Future Test Considerations

- Results likely to improve if LlamaParse replaces PyPDF.
- Parsing upgrades often outperform model upgrades.
- Models can only reason with what they're given.
- Better structured data = better answers, less guessing.
- Parsing is the cheapest way to level up your RAG stack.

Parsing Alone Can Move the Needle

- Academia confirms: changing only the parser impacts performance.
- Benchmark: same RAG system, different parsers → up to 20-point difference.
- Parser quality matters more than fancy downstream techniques.
- Quick wins: swap out low-quality parsers before tweaking your RAG logic.

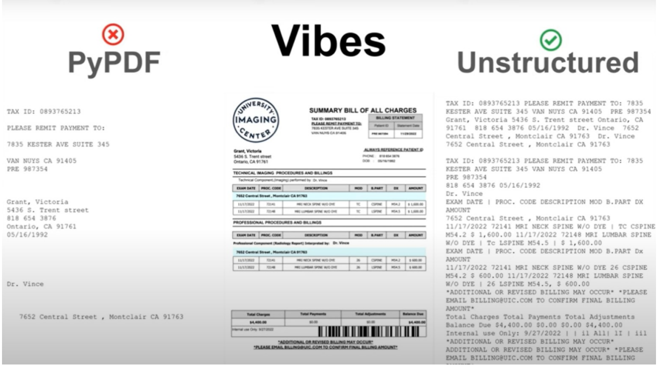
Document Context is Crucial

- Not all documents are created equal.
- Scientific papers, 10-Ks, clinical notes – all behave differently.
- Choose a parser suited to your specific domain.
- No single parser wins for all use cases.

Picking a Parser: A Two-Pronged Approach

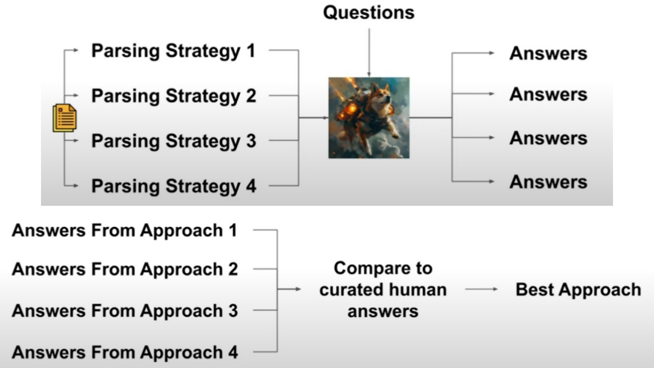
1. **Vibe check**: Run your data through multiple parsers. Look at the outputs.
2. **End-to-end eval**: Keep the RAG system constant, vary only the parser, then compare results.

"Your brain is the best model—start with your eyes."



How to Run an End-to-End Evaluation

- Change one component: the parser.
- Keep the rest of the RAG pipeline fixed.
- Feed questions through each parser's output.
- Compare generated answers to ground truth.
- Labor-intensive, but the most reliable evaluation method.



Evaluation

Auto-Eval: A Helping Hand

- Start with human-generated QA pairs (ground truth).
- Use LLMs to compare parser outputs to ground truth answers.
- Helps scale eval, but still requires initial human input.
- Avoids the trap of “models grading their own homework.”

Alternative Eval: ELO Ranking

- Useful when answers are subjective or non-falsifiable.
- Compare outputs pairwise: “Which one is better?”
- Rank parsers using ELO-style systems (used in chess).
- Great for stylistic or qualitative tasks.

Final Takeaways

- **Parsing is foundational.** Bad parsing = bad RAG, no matter the model.
- **There is no one-size-fits-all parser.**
- **Evaluate in context.** Use real documents and real questions.
- **Combine human intuition with structured evals.**
- **Opportunities exist.** Big gap in parser testing and tooling.
- Parsing is hard, but absolutely critical.
- Tools like LlamaParse, Unstructured, and X-ray are changing the game.
- Try multiple parsers and test thoroughly on your data.
- Don't trust models to validate their own output.
- Huge room for innovation in parser evaluation and automation.

Thanks . . .

- Search **"Yogesh Haribhau Kulkarni"** on Google and follow me on LinkedIn and Medium
- Office Hours: Saturdays, 2 to 5pm (IST); Free-Open to all; email for appointment.
- Email: yogeshkulkarni at yahoo dot com

(https://www.linkedin.com/in/yogeshkulkarni/, QR by Hugging Face QR-code-AI-art-generator, with prompt as "Follow me")

Temporary page!

LaTeX was unable to guess the total number of pages correctly. As there was some unprocessed data that should have been added to the final page this extra page has been added to receive it.

If you rerun the document (without altering it) this surplus page will go awa because LaTeX now knows how many pages to expect for this document.