Information Extraction from Fiscal Documents using LLMs

Vikram Aggarwal[1], Jay Kulkarni[2], Aditi Mascarenhas[2], Aakriti Narang[2], Siddarth Raman[2], Ajay Shah[2], Susan Thomas[2]

November 16, 2025

- Goverments produce fiscal data that is critical for machine analysis
- The data format (PDF) is difficult to analyze
- How can LLMs understand these formats?
- How can we create research-ready datasets at scale?
- Whether our approach towards validation is reasonable
- Can anything improve in the extraction that will help in the validation?

Why is this difficult?

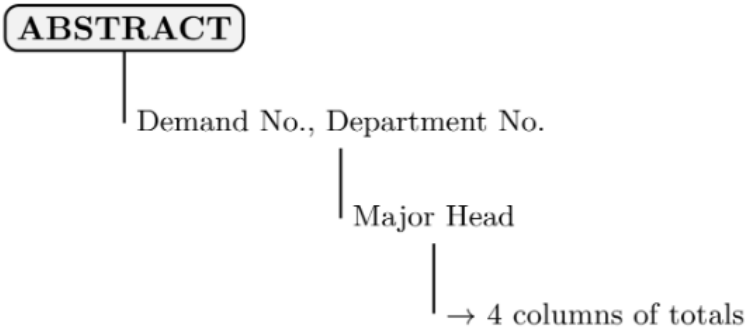
- Table structure has multi-level, complex hierarchy
- Documents are too big, larger than context window
- Documents are in Indic languages, where LLMs do worse
- Tables are inconsistently rendered, cells are split or merged
- Tables span multiple pages

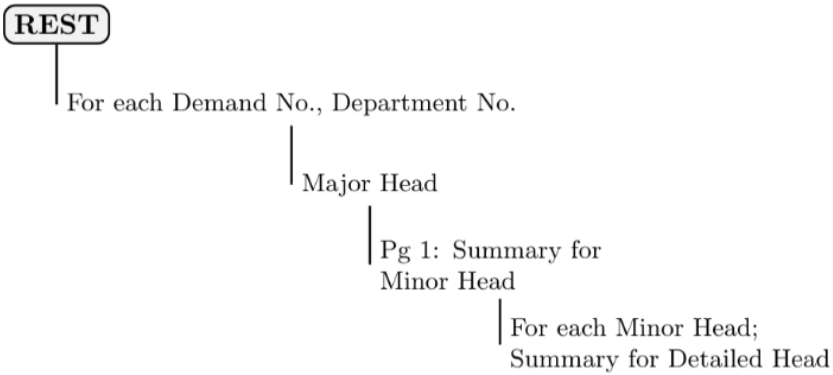
A sample of the Hierarchy

- Demand Number (2 Digit)
- Department Number (2 Digit) – Sometimes we only have the Department Name
- Major Hear (4 Digit)
- Sub Major Head (2 Digit)
- Minor Head (3 Digit)
- Sub Head (1 Digit)
- Detailed Head (2 Digit)
- Object Head (3 Digit) – smallest unit in the budget (need to distinguish from 3-digit Minor Head)

The document has two parts:

- A Summary at the “Demand No., Department No. Level” (Abstract)
- A Detailed summary that covers “Minor Heads”, “Detailed Heads”, and each “Object Head”. (The rest)


ABSTRACT
Demand No., Department No.
Major Head
→ 4 columns of totals

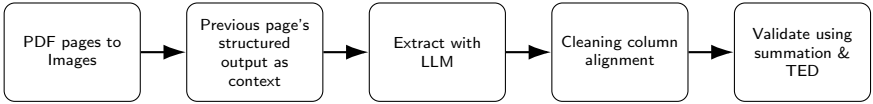


* Object Heads make the summary for "Detailed Head"

Our improvements to a naïve LLM document extraction

- **Image-based processing**: Converting PDF pages to high-resolution JPGs (300 DPI) improves LLM comprehension
- **Sequential context**: Provide previous page's structured output as an aide to understanding the current page.
- **Multi-level validation**: Use column-sums to ensure numerical consistency. Use hierarchy to validate structural consistency
- **Meta-prompting**: Provide domain context, get LLM to write the extraction prompt
- **Intelligent cleaning**: Post processing to improve column alignment

Steps

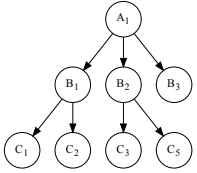

PDF pages to Images
Previous page's structured output as context
Extract with LLM
Cleaning column alignment
Validate using summation & TED

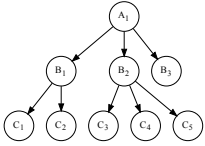
Numerical Consistency: Verify budget head sums within and across schema

Volume	Checks	Passed	Pass Rate %
Volume 1	528	463	88%
Volume 2	463	402	87%
Volume 3	289	233	81%
Volume 4	249	206	83%
Volume 5	390	316	81%
Volume 6	155	134	86%
Volume 7	237	198	84%
All	2311	1952	84%

Structural Consistency: Use Tree Edit Distance Similarity to verify tabular structure

Volume	Pages	Accuracy
Volume 1	227	95.24%
Volume 2	179	73.68%
Volume 3	139	88.00%
Volume 4	142	83.33%
Volume 5	181	96.77%
Volume 6	74	91.30%
Volume 7	112	79.17%





Advantages of our Method

- Can handle arbitrarily long PDF files
- Resilient against inconsistent Indic character encoding
- Works in the absence of ground-truth data
- Identifies extraction failure source to page & table location

Conclusion

- Extract multipage tables from 200+ page PDFs
- Information extraction is at 74%-95% accuracy
- Create research datasets for states finances
- Can also create parallel PDF & structured (CSV/JSON) corpus for LLM training