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]

[1]Google Inc., [2]XKDR Forum

1. Motivation

- Goverments periodically release fiscal data that is critical for research
- The data format (PDF) is difficult to analyze
- How can LLMs understand these formats?
- How can we create research-ready datasets at scale?

2. Challenge

Large Language Models (LLMs) are great at language, yet poor at multi-page tabular extraction

- 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

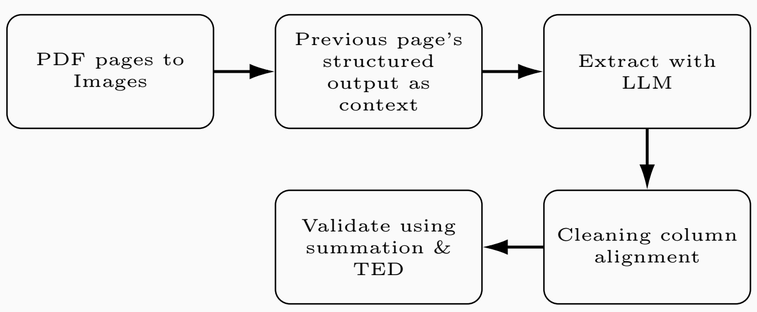
Sources of complexity

- Tables span multiple pages; tables have varying structures
- The files are large: roughly hundreds of pages (84-241 pages in the 7 volumes). Since tables span multiple pages, page-by-page extraction missed vital information
- Regional languages mix with English in inconsistent patterns
- Units and formatting vary across documents

3. Our approach

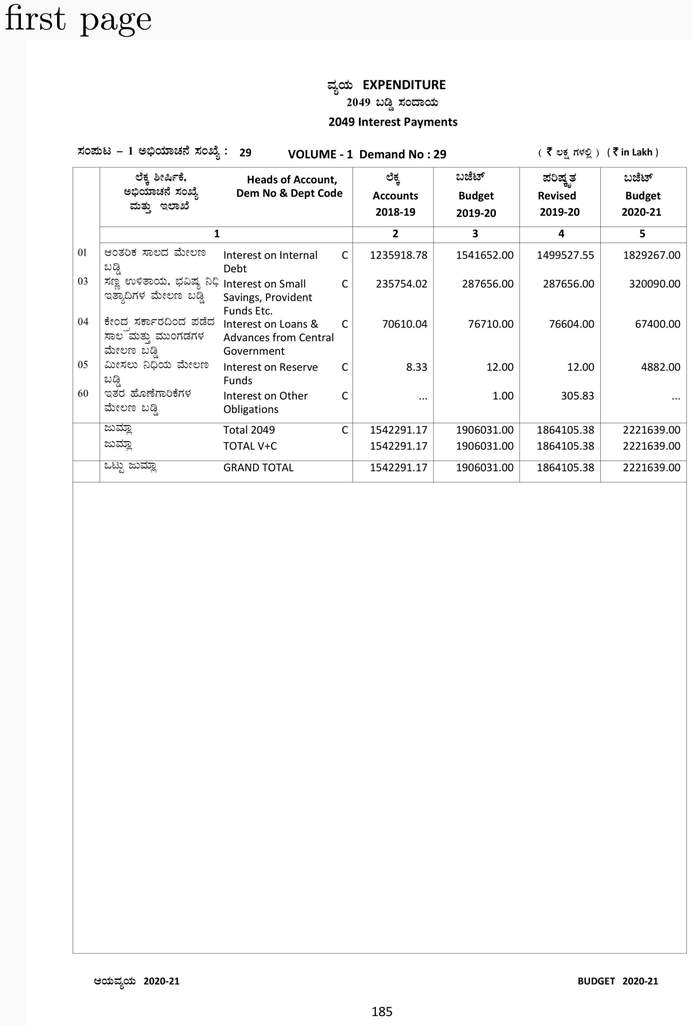
Our improvements to a naïve LLM document extraction

- **Image-based processing**: Convert PDF pages to high-resolution JPGs (300 DPI). This 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-process to improve column alignment



Input PDF to output table

(clockwise) Two pages that capture complexity of the PDF; the output table extracted from the first page



ವ್ಯಯ EXPENDITURE
2049 ಬಡ್ಡಿ ಸಂದಾಯ
2049 Interest Payments

ಸಂಪುಟ – 1 ಅಧಿಯಾಚನೆ ಸಂಖ್ಯೆ : 29 VOLUME - 1 Demand No : 29 (₹ ಲಕ್ಷ ಗಳಲ್ಲಿ) (₹ in Lakh)

	ಲೆಕ್ಕ ಶೀರ್ಷಿಕೆ, ಅಧಿಯಾಚನೆ ಸಂಖ್ಯೆ ಮತ್ತು ಉಲಾಟೆ	Heads of Account, Dem No & Dept Code		ಲೆಕ್ಕ Accounts 2018-19	ಬಜೆಟ್ Budget 2019-20	ಪರಿಷ್ಕೃತ Revised 2019-20	ಬಜೆಟ್ Budget 2020-21
	1			2	3	4	5
01	ಆಂತರಿಕ ಸಾಲದ ಮೇಲಿನ ಬಡ್ಡಿ	Interest on Internal Debt	C	1235918.78	1541652.00	1499527.55	1829267.00
03	ಸಣ್ಣ ಉಳಿತಾಯ, ಭವಿಷ್ಯ ನಿಧಿ ಇತ್ಯಾದಿಗಳ ಮೇಲಿನ ಬಡ್ಡಿ	Interest on Small Savings, Provident Funds Etc.	C	235754.02	287656.00	287656.00	320090.00
04	ಕೇಂದ್ರ ಸರ್ಕಾರದಿಂದ ಪಡೆದ ಸಾಲ ಮತ್ತು ಮುಂಗಡಗಳ ಮೇಲಿನ ಬಡ್ಡಿ	Interest on Loans & Advances from Central Government	C	70610.04	76710.00	76604.00	67400.00
05	ಮೀಸಲು ನಿಧಿಯ ಮೇಲಿನ ಬಡ್ಡಿ	Interest on Reserve Funds	C	8.33	12.00	12.00	4882.00
60	ಇತರ ಹೊಣೆಗಾರಿಕೆಗಳ ಮೇಲಿನ ಬಡ್ಡಿ	Interest on Other Obligations	C	...	1.00	305.83	...
	ಜುಮ್ಲಾ	Total 2049	C	1542291.17	1906031.00	1864105.38	2221639.00
	ಜುಮ್ಲಾ	TOTAL V+C		1542291.17	1906031.00	1864105.38	2221639.00
	ಒಟ್ಟು ಜುಮ್ಲಾ	GRAND TOTAL		1542291.17	1906031.00	1864105.38	2221639.00

ಆಯವ್ಯಯ 2020-21 BUDGET 2020-21

185

ವ್ಯಯ EXPENDITURE
2049 ಬಡ್ಡಿ ಸಂದಾಯ
2049 Interest Payments

ಸಂಪುಟ – 1 ಅಧಿಯಾಚನೆ ಸಂಖ್ಯೆ : 29 VOLUME - 1 Demand No : 29 (₹ ಲಕ್ಷ ಗಳಲ್ಲಿ) (₹ in Lakh)

	ಲೆಕ್ಕ ಶೀರ್ಷಿಕೆ, ಅಧಿಯಾಚನೆ ಸಂಖ್ಯೆ ಮತ್ತು ಉಲಾಟೆ	Heads of Account, Dem No & Dept Code		ಲೆಕ್ಕ Accounts 2018-19	ಬಜೆಟ್ Budget 2019-20	ಪರಿಷ್ಕೃತ Revised 2019-20	ಬಜೆಟ್ Budget 2020-21
	1			2	3	4	5
01	ಆಂತರಿಕ ಸಾಲದ ಮೇಲಿನ ಬಡ್ಡಿ	Interest on Internal Debt					
101	ಮಾರುಕಟ್ಟೆ ಸಾಲಗಳ ಮೇಲಿನ ಬಡ್ಡಿ	Interest on Market Loans	C	1024916.40	1343614.00	1303940.00	1644445.00
102	ಸಾಲಗಳ ಮೇಲಿನ ಸೋಡಿ	Discount on Loans	C	...	...	...	5000.00
115	ಭಾರತೀಯ ರಿಸರ್ವ್ ಬ್ಯಾಂಕ್‌ನಿಂದ ಆಥೋ೯ಪಾಯ ಅಗಿಮ ಧನದ ಮೇಲಿನ ಬಡ್ಡಿ	Interest on Ways & Means Advances from Reserve Bank of India	C	...	...	4.55	10.00
123	ರಾಜ್ಯ ಸರ್ಕಾರದ ಮೂಲಕ ಕೇಂದ್ರ ಸರ್ಕಾರದ ರಾಷ್ಟ್ರೀಯ ಸ್ಣ ಉಳಿತಾಯ ಯೋಜನೆ ನಿಧಿಗೆ ವಿಶೇಷ ಭದತಾ ಠೇವಣಿ ಹಂಚಿಕೆಗಳ ಮೇಲಿನ ಬಡ್ಡಿ	Interest on Special Securities issued to NSSF of the Central Government by State Government	C	178961.95	163527.00	163527.00	147784.00
200	ಇತರ ಆಂತರಿಕ ಸಾಲಗಳ ಮೇಲಿನ ಬಡ್ಡಿ	Interest on Other Internal Debts	C	28891.49	31050.00	28595.00	27694.00
305	ಸಾಲದ ನಿರ್ವಹಣೆ	Management of Debt	C	3148.94	3461.00	3461.00	4334.00
	ಒಟ್ಟು ಜುಮ್ಲಾ	Total 01	C	1235918.78	1541652.00	1499527.55	1829267.00
03	ಸಣ್ಣ ಉಳಿತಾಯ, ಭವಿಷ್ಯ ನಿಧಿ ಇತ್ಯಾದಿಗಳ ಮೇಲಿನ ಬಡ್ಡಿ	Interest on Small Savings, Provident Funds Etc.					
104	ರಾಜ್ಯ ಭವಿಷ್ಯ ನಿಧಿಯ ಮೇಲಿನ ಬಡ್ಡಿ	Interest on State Provident Funds	C	116439.48	139473.00	139473.00	163084.00
107	ನ್ಯಾಸ ಮತ್ತು ಉಂಬಳಿಗಳ ಮೇಲಿನ ಬಡ್ಡಿ	Interest on Trusts and Endowment	C	...	1.00	1.00	1.00
108	ವಿಮೆ ಮತ್ತು ವಿಶ್ರಾಂತಿ ವೇತನ ನಿಧಿಯ ಮೇಲಿನ ಬಡ್ಡಿ	Interest on Insurance and Pension Fund	C	119314.54	146881.00	146881.00	155704.00
115	ಇತರ ಉಳಿತಾಯ ಠೇವಣಿಗಳ ಮೇಲೆ ಬಡ್ಡಿ	Interest on Other Savings Deposits	C	...	1.00	1.00	1.00
117	ನಿರ್ದಿಷ್ಟ ಕೊಡುಗೆ ಪಿಂಚಣಿ ಯೋಜನೆ ಮೇಲಿನ ಬಡ್ಡಿ	Interest on Defined Contribution Pension Scheme	C	...	1300.00	1300.00	1300.00
	ಒಟ್ಟು ಜುಮ್ಲಾ	Total 03	C	235754.02	287656.00	287656.00	320090.00
04	ಕೇಂದ್ರ ಸರ್ಕಾರದಿಂದ ಪಡೆದ ಸಾಲ ಮತ್ತು ಮುಂಗಡಗಳ ಮೇಲಿನ ಬಡ್ಡಿ	Interest on Loans & Advances from Central Government					
101	ರಾಜ್ಯ / ಕೇಂದ್ರ ಆಡಳಿತ ಪ್ರದೇಶದ ಯೋಜನಾ ಕಾರ್ಯಕ್ರಮಗಳಿಗಾಗಿ ಸಾಲದ ಮೇಲೆ ಬಡ್ಡಿ	Interest on Loans for State / Union Territory Plan Schemes	C	53690.16	62520.00	62511.00	56046.00
104	ಯೋಜನೇತರ ಕಾರ್ಯಕ್ರಮಗಳ ಸಾಲದ ಮೇಲಿನ ಬಡ್ಡಿ	Interest on Loans for Non-Plan Schemes	C	527.51	483.00	474.00	412.00
109	12ನೇ ಹಣಕಾಸು ಆಯೋಗದ ಶಿಫಾರಸ್ಸಿನ ನಿಬಂಧನೆಗಳಿಗೊಳಪಟ್ಟು ರಾಜ್ಯ ಯೋಜನಾ ಸಾಲಗಳ	Interest on State Plan Loans consolidated in terms of recommendations of	C	16364.26	13677.00	13588.00	10900.00

ಆಯವ್ಯಯ 2020-21 BUDGET 2020-21

186

Source Page Num	Vol Num	Demand Number	Major Head Code	Major Head Name	Sub Major Head Code	Sub Major Head Name	Full Account Code	Description	V/C Marker	Row Type	Row Level	Accounts 2018_19	Budget 2019_20	Revised 2019_20	Budget 2020_21
185	1	29	2049	Interest Payments	01	Interest on Internal Debt		Interest on Intern	C	Data	Sub-Major-Head	1235918.78	1541652	1499527.55	1829267
185	1	29	2049	Interest Payments	03	Interest on Small Savings, Provident		Interest on Smal	C	Data	Sub-Major-Head	235754.02	287656	287656	320090
185	1	29	2049	Interest Payments	04	Interest on Loans & Advances from C		Interest on Loan	C	Data	Sub-Major-Head	70610.04	76710	76604	67400
185	1	29	2049	Interest Payments	05	Interest on Reserve Funds		Interest on Rese	C	Data	Sub-Major-Head	8.33	12	12	4882
185	1	29	2049	Interest Payments	60	Interest on Other Obligations		Interest on Othe	C	Data	Sub-Major-Head		1	305.83	
185	1	29	2049	Interest Payments				Total 2049	C	Total	Major-Head	1542291.17	1906031	1864105.38	2221639
185	1	29	2049	Interest Payments				TOTAL V+C	V+C	Total	Major-Head	1542291.17	1906031	1864105.38	2221639
185	1	29	2049	Interest Payments	60			GRAND TOTAL		Total	Sub-Major-Head	1542291.17	1906031	1864105.38	2221639

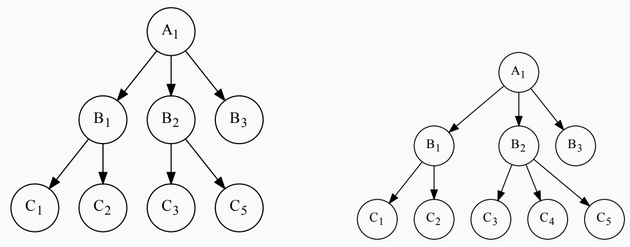
4. Validation results

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%



5. Advantages of our method

- Can handle arbitrarily long PDF files. Extract multipage tables from 200+ page PDFs
- Resilient against inconsistent Indic character encoding
- Works in the absence of ground-truth data. Information extraction is at 74%-95% accuracy
- Identifies extraction failure source to page & table location
- Create research-ready datasets for states finances
- Can also create parallel PDF & structured (CSV/JSON) corpus for LLM training