NATURAL LANGUAGE PROCESSING

AUTOMATIC DOCUMENT FORMAT AND CONTENT RECOGNITION FOR ACADEMIC PAPERS

Pranjul Mishra
Nazira Tukeyeva
Saurabh Singh

AGENDA

- PROBLEM STATEMENT
- BACKGROUND WORK
- METHODLGY
- ARCHITECTURE
- RESULTS AND CHALLENGES
- PERFORMANCE COMPARISION
- FUTURE SCOPES

PROBLEM STATEMENT

Primary goal:

- create an automated system that processes academic documents and extracts:
    - titles and author(s) with affiliations
    - section content organized by headings (e.g., abstract, introduction, results)
    - non-textual elements such as tables and images.

Key research questions:

1. How can we design a method to automatically detect and extract essential components from academic documents?
2. Which NLP techniques are most effective for segmenting document content by headings and sub-headings and in hierarchical way?

BACKGROUND WORK

- **Document Structure Recognition:**
  - Evolution from rule-based methods (e.g., CERMINE) [1] to transformer-based model like BERT [2].
- **Content Segmentation:**
  - Shift from traditional template-based approaches to applying RNNs to segment text by learning contextual patterns within document content [3].
- **Non-Textual Element Extraction:**
  - Tools like DeepDeSRT, Camelot, and Tabula enhance detection of tables and figures [4].
- **Challenges in Document Analysis:**
  - Diversity in document layouts (formatting styles, heading structures), and limited annotated datasets for training document analysis models [5].

METHODOLOGY

Datasets:

- PubMed Central Open Access Subset (PMC-OAS)
- arXiv Dataset
- ICDAR Competition Datasets

Data Processing:

- Text Extraction: PDFMiner (extracts raw text from PDF)
- Tokenization and Segmentation
- Noise Removal

Document Segmentation and Content Extraction:

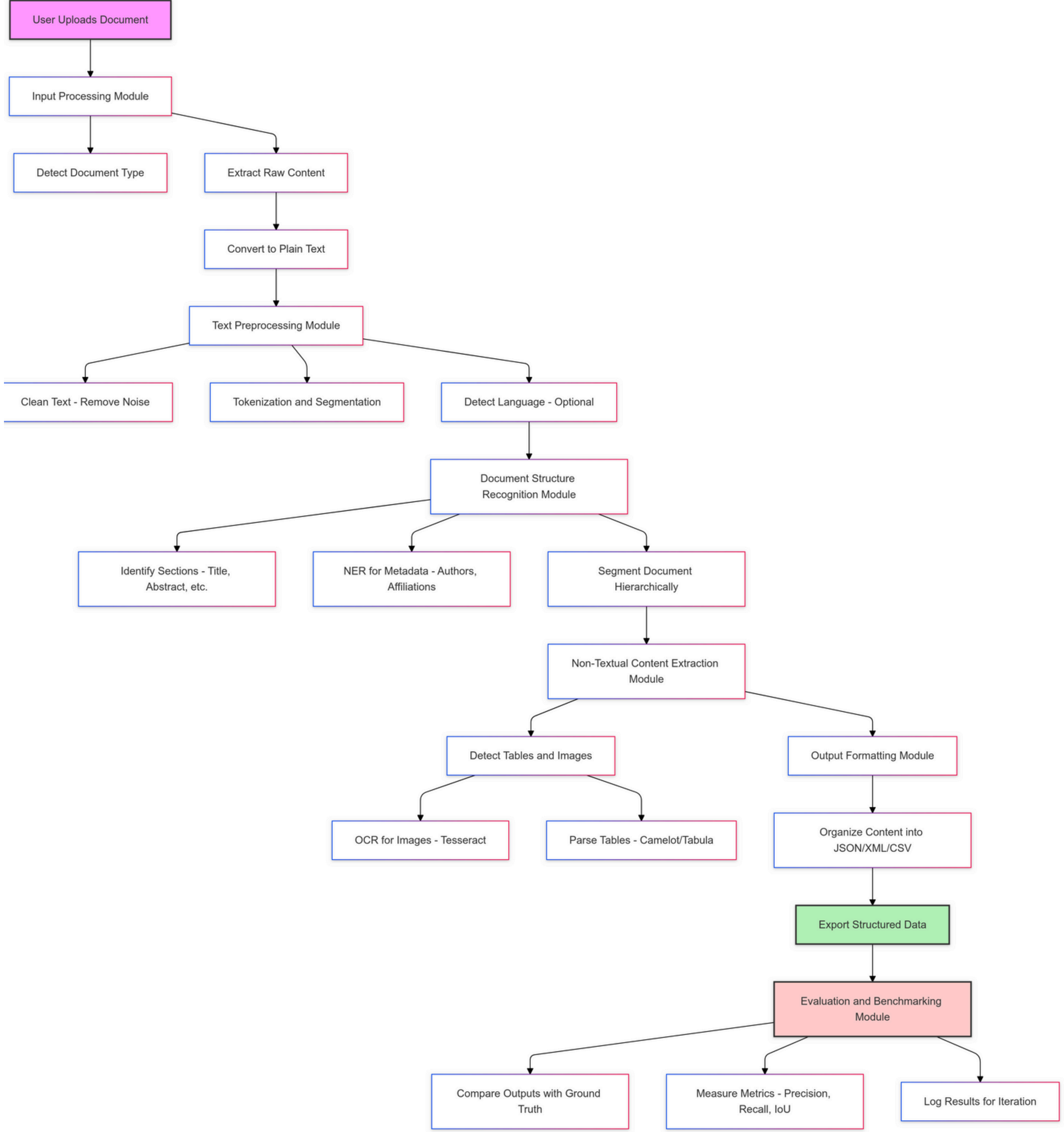
- NLP Models (eg. BERT): understand and segment the document based on contextual relationships within the text hierarchically.
- NER: identify entities
- Regular Expressions: identify commonly formatted sections, such as references or bibliography

Non-Textual Element Extraction:

- OCR for Embedded Text: extract text from images
- Table Extraction Tools: detect and extract tabular data from PDF files (eg. Camelot, Tabula)
- Layout Processing with OpenCV: analyze layout structure

ARCHITECTURE

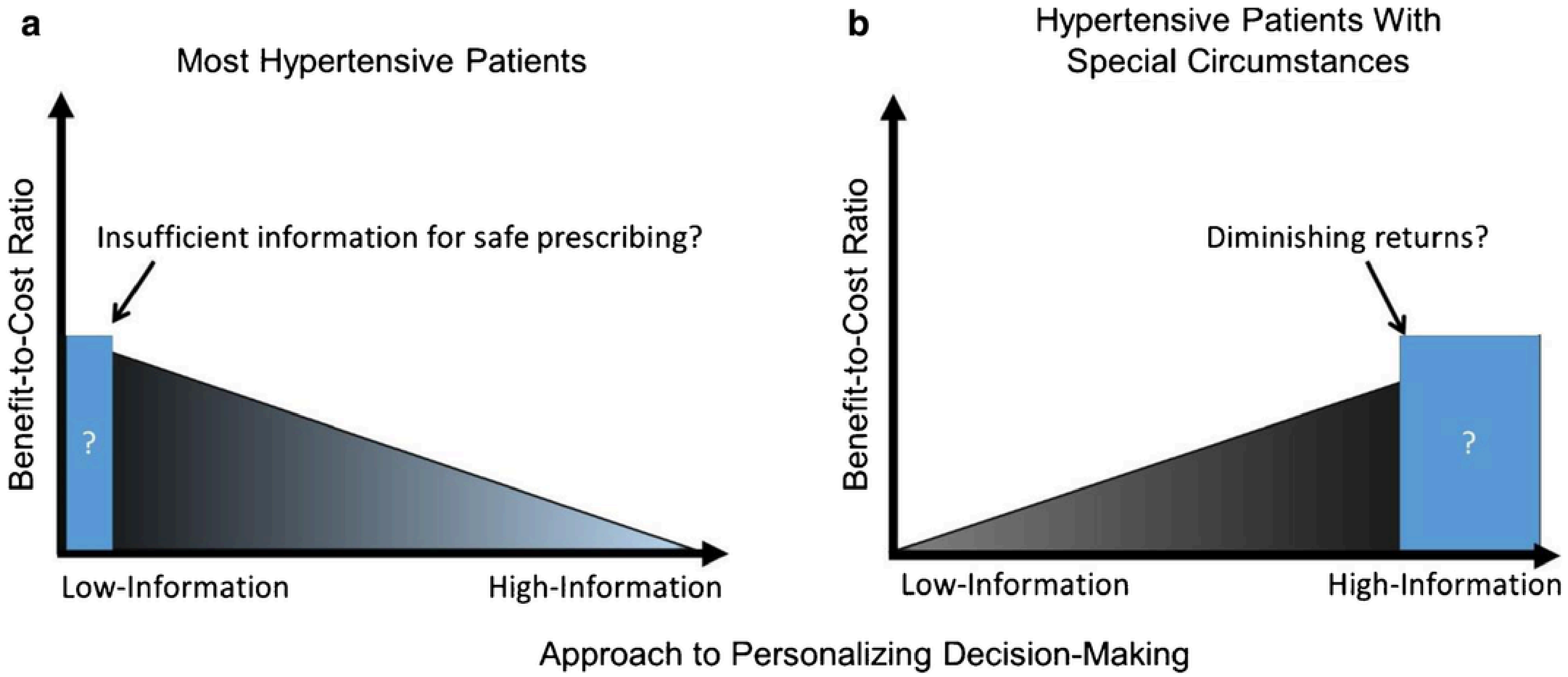
In addition to this modular approach we additionally added a small text based mode “t5-small” to be computationally efficient and as an additional help to hierarchically structure the output json file.


User Uploads Document
Input Processing Module
Detect Document Type
Extract Raw Content
Convert to Plain Text
Text Preprocessing Module
Clean Text - Remove Noise
Tokenization and Segmentation
Detect Language - Optional
Document Structure Recognition Module
Identify Sections - Title, Abstract, etc.
NER for Metadata - Authors, Affiliations
Segment Document Hierarchically
Non-Textual Content Extraction Module
Detect Tables and Images
Output Formatting Module
OCR for Images - Tesseract
Parse Tables - Camelot/Tabula
Organize Content into JSON/XML/CSV
Export Structured Data
Evaluation and Benchmarking Module
Compare Outputs with Ground Truth
Measure Metrics - Precision, Recall, IoU
Log Results for Iteration

RESULTS & CHALLENGES

- Challenges :
  a. We found a few images as nuances in the a paper along with the relevant images , like some logo, or dot or some date images
  b. Whether to ignore the footnotes or not ?
  c. Efficiency of the model since addition of a Language model’s performance usually varies depending on the machine on which it is used.

- Results Achieved:
  - We were able to extract the textual and non -textual elements accurately, with some ambiguities (because of foot notes , side bar notes and sometimes additional cite notes)
  - Additional Segmentation of Texts in chunks and parallelly processing them.


a
Most Hypertensive Patients
Benefit-to-Cost Ratio
Insufficient information for safe prescribing?
?
Low-Information
High-Information
b
Hypertensive Patients With Special Circumstances
Benefit-to-Cost Ratio
Diminishing returns?
?
Low-Information
High-Information
Approach to Personalizing Decision-Making

An example Image after performing the image extraction

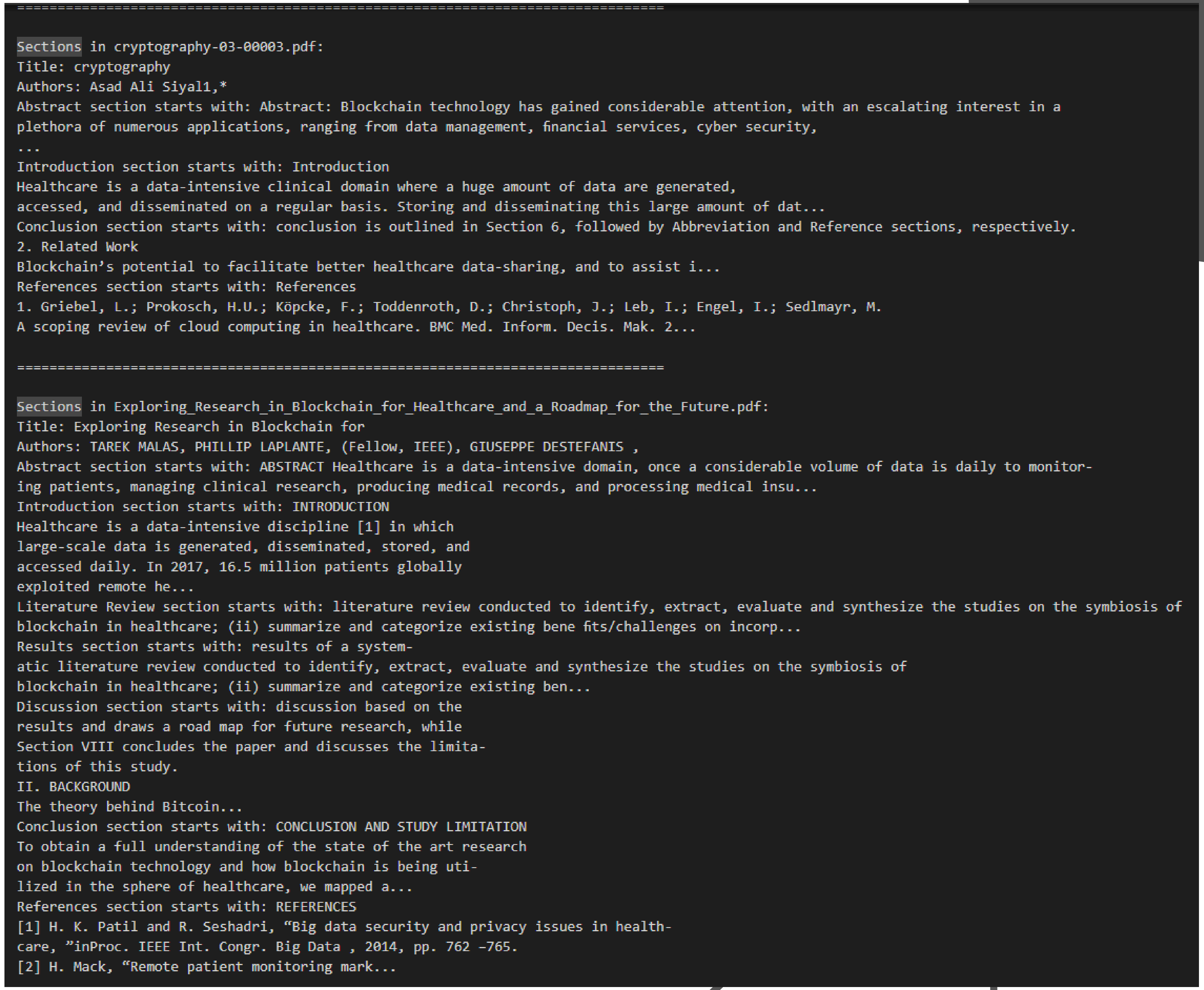
PERFORMANCE COMPARISION

Models	Performance (% in terms of extracting the correct texts from File)	Edge case Remarks
Our Model	85%	Additional images while attempting to extract images and formulas.
GROBID	95%-97%	Good in extracting formulas and images compared to our model
GPT -4o	99%	Better than our model since it is more computationally efficient and pretrained heavily
GPT 3.5 TURBO	80-90%	It fails with images most of the times and gets the tables and cite notes ambiguous sometimes

FUTURE SCOPES

1. The output format which we are focusing is in Json/XML/CSV which further can be converted into set of triples using set of rules and thus a knowledge graph can be created .
2. Upon integration with better LLM's this can be improvised especially any chance of nuances could be avoided and the data obtained could be more hierarchically organized while working with multiple pdf files at once.

3. An efficient mechanism could be developed to deal with Complex Formulas



```
================================================================================

Sections in cryptography-03-00003.pdf:
Title: cryptography
Authors: Asad Ali Siyal1,*
Abstract section starts with: Abstract: Blockchain technology has gained considerable attention, with an escalating interest in a
plethora of numerous applications, ranging from data management, financial services, cyber security,
...
Introduction section starts with: Introduction
Healthcare is a data-intensive clinical domain where a huge amount of data are generated,
accessed, and disseminated on a regular basis. Storing and disseminating this large amount of dat...
Conclusion section starts with: conclusion is outlined in Section 6, followed by Abbreviation and Reference sections, respectively.
2. Related Work
Blockchain's potential to facilitate better healthcare data-sharing, and to assist i...
References section starts with: References
1. Griebel, L.; Prokosch, H.U.; Köpcke, F.; Toddenroth, D.; Christoph, J.; Leb, I.; Engel, I.; Sedlmayr, M.
A scoping review of cloud computing in healthcare. BMC Med. Inform. Decis. Mak. 2...

================================================================================

Sections in Exploring_Research_in_Blockchain_for_Healthcare_and_a_Roadmap_for_the_Future.pdf:
Title: Exploring Research in Blockchain for
Authors: TAREK MALAS, PHILLIP LAPLANTE, (Fellow, IEEE), GIUSEPPE DESTEFANIS ,
Abstract section starts with: ABSTRACT Healthcare is a data-intensive domain, once a considerable volume of data is daily to monitor-
ing patients, managing clinical research, producing medical records, and processing medical insu...
Introduction section starts with: INTRODUCTION
Healthcare is a data-intensive discipline [1] in which
large-scale data is generated, disseminated, stored, and
accessed daily. In 2017, 16.5 million patients globally
exploited remote he...
Literature Review section starts with: literature review conducted to identify, extract, evaluate and synthesize the studies on the symbiosis of
blockchain in healthcare; (ii) summarize and categorize existing bene fits/challenges on incorp...
Results section starts with: results of a system-
atic literature review conducted to identify, extract, evaluate and synthesize the studies on the symbiosis of
blockchain in healthcare; (ii) summarize and categorize existing ben...
Discussion section starts with: discussion based on the
results and draws a road map for future research, while
Section VIII concludes the paper and discusses the limita-
tions of this study.
II. BACKGROUND
The theory behind Bitcoin...
Conclusion section starts with: CONCLUSION AND STUDY LIMITATION
To obtain a full understanding of the state of the art research
on blockchain technology and how blockchain is being uti-
lized in the sphere of healthcare, we mapped a...
References section starts with: REFERENCES
[1] H. K. Patil and R. Seshadri, "Big data security and privacy issues in health-
care, "inProc. IEEE Int. Congr. Big Data , 2014, pp. 762 -765.
[2] H. Mack, "Remote patient monitoring mark...
```

REFERENCES

[1] Tkaczyk, Dominika, et al. ”CERMINE: automatic extraction of structured metadata from scientific literature.” International Journal on Document Analysis and Recognition (IJDAR) 18.4 (2015).

[2] Devlin, Jacob, et al. ”BERT: Pre-training of deep bidirectional transformers for language understanding.” arXiv:1810.04805 (2019).

[3] Yang, Zhilin, et al. ”Neural machine translation with recurrent attention modeling.” arXiv:1703.04675 (2017).

[4] Schreiber, Sebastian, et al. ”Deepdesrt: Deep learning for detection and structure recognition of tables in document images.” 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). Vol. 1. IEEE, 2017.

[5] Gao, Liangcai, et al. ”ICDAR 2019 competition on table detection and recognition (cTDaR).” International Conference on Document Analysis and Recognition (ICDAR). IEEE, 2019.

THANK YOU!