LitAI: Enhancing Multimodal Literature Understanding and Mining with Generative AI

Gowtham Medisetti, Zacchaeus Compson, Heng Fan, Huaxiao Yang, and Yunhe Feng
University of North Texas, Denton, TX, USA
gowthammedisetti@my.unt.edu, {zacchaeus.compson, heng.fan, huaxiao.yang, yunhe.feng}@unt.edu

***Abstract*—Information processing and retrieval in literature are critical for advancing scientific research and knowledge discovery. The inherent multimodality and diverse literature formats, including text, tables, and figures, present significant challenges in literature information retrieval. This paper introduces *LitAI*, a novel approach that employs readily available generative AI tools to enhance multimodal information retrieval from literature documents. By integrating tools such as optical character recognition (OCR) with generative AI services, *LitAI* facilitates the retrieval of text, tables, and figures from PDF documents. We have developed specific prompts that leverage in-context learning and prompt engineering within Generative AI to achieve precise information extraction. Our empirical evaluations, conducted on datasets from the ecological and biological sciences, demonstrate the superiority of our approach over several established baselines including Tesseract-OCR and GPT-4. The implementation of *LitAI* is accessible at https://github.com/ResponsibleAILab/LitAI.**

***Index Terms*—Literature Mining, OCR, Generative AI, Prompt Engineering, ChatGPT, GPT-4**

I. Introduction

Literature information retrieval and understanding play an important role in learning existing works and discovering future research directions. It is very important to retrieve the information embedded in such literature, including text, tables, and figures. However, it is challenging to parse multimodal information from literature formatted in PDF or images due to the huge diversity and lack of standards in presentation. For example, each paper may adopt different formats like single-column and double-column; table information can be organized into arbitrary rows and columns and even nested arbitrary rows and columns; figures can use diverse colors, shapes, and other visualization elements.

Many existing efforts have been made to extract accurate information from literature sources. Optical Character Recognition (OCR) serves as one of the core technologies utilized for this purpose [1]. Utilizing OCR to process textual information, such as paper titles, abstracts, and main body text, is relatively straightforward. However, the effectiveness of OCR results can be compromised by the quality of literature papers presented in images and PDF formats. To address elements more complex than plain text, such as tables and references, OCR is also employed, although it often struggles with recognizing and processing the structure and logic inherent in these components [2].

In addition to text-rich content, literature documents often include figures that convey substantial information, illustrating key concepts, primary experimental results, and more. While OCR can extract and interpret text embedded within these figures, it may encounter challenges such as low resolution and distractions. Techniques like image captioning have emerged as potential methods to parse figures. However, many of these solutions overlook the contextual information of figures, specifically the text description of the figure provided in the main body of the paper.

To address these research gaps, we introduce *LitAI*, an off-the-shelf generative AI-enhanced approach for multimodal literature understanding that incorporates existing text recognition tools. We have utilized the zero-shot capabilities of Generative Pre-trained Transformer (GPT) services to enhance text recognition and the parsing of tables and figures. Rather than relying solely on off-the-shelf generative AI as an end-to-end solution, we strategically integrate it with established literature analysis tools. ChatGPT effectively corrects typos and inaccuracies in text extracted by OCR. Additionally, *LitAI* enables users to categorize extracted text into sections such as the Abstract, Introduction, and Conclusion. It also assists in reorganizing poorly formatted reference lists through carefully crafted prompts in ChatGPT. For table parsing, *LitAI* introduces several prompt engineering techniques to extract nested structures and data formats. Lastly, we propose a context-aware prompt engineering method to query and retrieve relevant content from figures contained in PDF images, enhancing the interpretability of the visual data.

To demonstrate the effectiveness of *LitAI*, we conduct extensive qualitative and quantitative experiments on literature from the ecological and biological domains. *LitAI* consistently outperforms both AI-free tools and end-to-end generative AI solutions in analyzing the main text, references, tables, and figures. These findings underscore the generalizability and adaptability of *LitAI* across various fields.

The main contributions of this paper can be summarized as follows:

- We propose *LitAI*, designed to process multimodal elements such as text, tables, and figures in literature papers.
- We perform thorough evaluations of *LitAI* on literature from two distinct domains, demonstrating its superior performance over existing baselines.
- To facilitate reproducibility, we release the source code of *LitAI* at https://github.com/ResponsibleAILab/LitAI.

II. Related Work

Many works have been undertaken to facilitate the retrieval and understanding of information from literature papers. We categorize these efforts into two main areas: pure text understanding, and comprehension of tables and figures.

When processing pure text literature information, the primary reliance is on optical character recognition (OCR). For example, Esposito et al. [3] is one of the first to employ OCR to extract text information from scholarly articles. OCR++ [4] was designed to improve the robustness of OCR in scholarly article information retrieval. Günter [5] combined crowdsourcing with OCR to correct OCR errors. Research conducted by Ray Smith [6] provided a comprehensive overview of the Tesseract OCR engine, illustrating its evolution from a research project at HP Labs to its subsequent adoption as an open-source tool by Google. This historical narrative illuminates the development trajectory of OCR technologies and their profound impact on the digitization of documents. Saoji et al. [7] in their research, examined OCR technologies using Pytesseract, focusing on improving text detection accuracy with comprehensive preprocessing techniques such as noise reduction and binarization. Their research adds valuable insights into the efficacy of OCR tools in processing complex text from images. Tools such as PyMuPDF and Tesseract-OCR have emerged as pivotal solutions for bulk PDF-to-text conversions [8].

Regarding the parsing of figures and tables within the domain of literature information processing, significant progress has been achieved, particularly in multimodal data extraction techniques. For instance, Oro and Ruffolo [9] developed heuristic methods for extracting tables from PDF documents using OCR. Lopez et al. [10] introduced a novel system for the automatic extraction of figures and captions from biomedical PDFs. By harnessing the structure of PDF layouts and implementing a finite state machine, their approach markedly enhances the processing efficiency of biomedical texts, minimizing manual intervention. Works such as *PaperMage* by Lo et al. [11] and studies conducted by Huang et al. [2] have enhanced the efficiency and accuracy of extracting various elements, including text, tables, figures, and references, from complex document formats. Recent advancements have addressed challenges associated with PDF conversion, particularly in handling non-searchable documents and scanned PDFs [8], [12].

Building upon these foundations, *LitAI* aims to improve literature information processing by integrating OCR, natural language processing (NLP), and generative AI. Employing a multi-step approach, *LitAI* begins with OCR for text extraction, incorporates NLP for semantic understanding, and leverages generative AI to enhance data retrieval and synthesis. This comprehensive integration allows *LitAI* to effectively handle complex document elements like text, tables, and figures. Particularly, it improves the retrieval and categorization of visual content [13], enhances table extraction and formatting for higher accuracy [14], and streamlines reference management with advanced scripting and generative AI techniques, thereby enriching dataset construction for further language model training. In summary, *LitAI* offers novel advancements in literature information processing, facilitating more efficient knowledge discovery and synthesis.

III. Methodology

In this section, we begin by presenting an overview framework of *LitAI*. Then, we detail how *LitAI* processes text, tables, and figures in literature papers.

A. Framework Overview

As shown in Figure 1, *LitAI* is intricately designed to optimize the processing and analysis of academic papers. It starts with Optical Character Recognition (OCR) using pytesseract to convert scanned and digital texts into a machine-readable format, ensuring accurate capture of all textual content. *LitAI* utilizes specific prompts to systematically extract key sections such as the Abstract, Introduction, Methodology, Results, Discussion, Conclusion, and References. This enables section-wise literature analysis. For tables and figures, the framework processes tables by converting them to text and then to CSV format for easier manipulation, while figures are handled using GPT-4 Vision for extraction and captioning, along with generating detailed descriptions to enhance interpretability. The final output of *LitAI* is a comprehensive, section-wise compilation of all processed content, meticulously organized and formatted to facilitate ease of navigation and readiness for further application. This methodology ensures thorough and efficient management of academic documents, maximizing the accessibility and utility of the information contained within.

B. Text Processing

We categorize literature papers in PDF format into two distinct groups: searchable and non-searchable. Searchable PDFs consist of text that has been digitally identified and recognized, enabling users to perform text searches within these documents. Conversely, non-searchable PDFs are primarily image-based and do not allow text recognition or searches.

1) Text Processing for Searchable PDFs: For text extraction from searchable PDFs, *LitAI* utilizes PyMuPDF[1] to directly retrieve text. To enhance the scalability of text processing from searchable PDFs, we have developed a robust PyMuPDF-based Python script that supports batch processing. This script facilitates the efficient conversion of multiple PDF files into text format, effectively managing multiple files and incorporating comprehensive error handling to ensure uninterrupted operation. This automated conversion process greatly improves workflow efficiency and expands analytical capabilities, thereby streamlining the processing and analysis of data within PDF documents.

[1] https://pymupdf.readthedocs.io/en/latest/

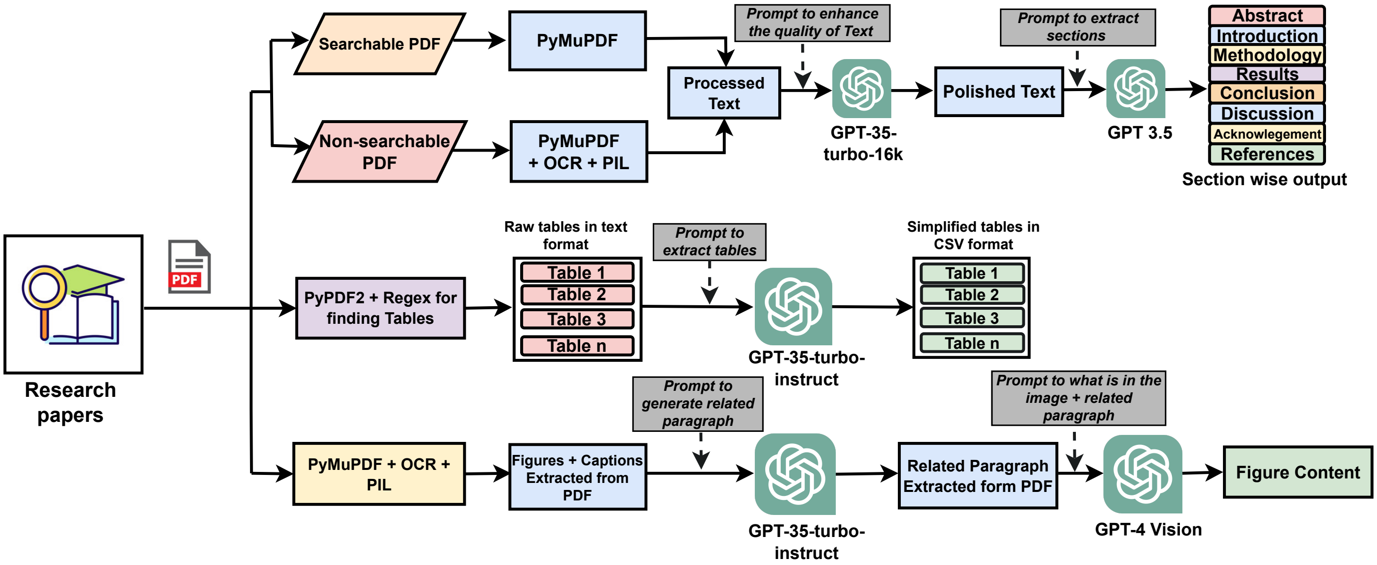


Fig. 1: *LitAI* framework overview. *LitAI* supports both searchable and non-searchable PDFs for text processing. Considering the efficiency of table and figure extraction efficiency, it only supports searchable PDFs for table and image processing. For all three types of literature processing, generative AI is employed to enhance the parsing and understanding quality.

2) Text Processing for Non-searchable PDFs: To extract text from non-searchable PDFs, we harness the combined capabilities of PyMuPDF and Tesseract-OCR for accurate text retrieval. Specifically, PyMuPDF converts each PDF page into an image and Tesseract-OCR is utilized to convert pixel-based text in these images into a searchable format. This approach facilitates text extraction from non-searchable PDFs, significantly broadening the scope of PDF data accessible for detailed analysis.

3) Enhancing Extracted Text Quality with Generative AI: The text extracted from PDFs, particularly non-searchable ones, may include typos or errors due to the inherent limitations of PDF parsing and OCR tools, as well as the structural complexity of literature papers (refer to Figure 2a). To refine this raw output, *LitAI* integrates GPT-3.5 Turbo (gpt-35-turbo-16k) into the text processing workflow, enhancing the correction of grammatical errors and improving overall textual coherence. *LitAI* employs the following prompt to facilitate this enhancement:

> **OCR Enhancement Prompt:** You are a typo correction tool assigned to refine a research paper. Your task is to identify and correct any typographical errors while ensuring that the original words and sentences remain unchanged. Please review the provided research paper and rectify any typographical errors without altering the original content.

This combined approach results in a more refined and polished output, thereby laying a robust foundation for advanced analysis and processing, and ultimately contributing to a more accurate and reliable interpretation of the data extracted from the PDF documents. Figure 2b illustrates an example of the text refined by GPT-3.5 Turbo. In addition, *LitAI* can categorize the refined text into sections like Abstract, Introduction, and Reference automatically through prompt engineering.

C. Table Processing

Tables are frequently used in scholarly papers to organize and present structured information systematically. The proposed *LitAI* sets up a robust automated framework for extracting and formatting tables from PDFs by integrating PyPDF2 [2] with OpenAI's GPT-3.5 Turbo. Rather than processing all text converted from PDFs through GPT-3.5 Turbo, PyPDF2 initially identifies and extracts text specifically related to tables using targeted regular expressions. Each extracted table is then saved as an individual CSV file, enabling separate file management for each table to facilitate seamless processing within the token constraints of GPT models.

Due to the challenges associated with converting tables of varying organizational formats into a standardized and easy-to-use format, *LitAI* leverages the few-shot learning capabilities of the GPT-3.5 Turbo Instruct models to refine the extracted table information. The prompt used for refining tables is outlined below:

> **Table Refining Prompt:** Please format the following simple table into a structured CSV.
> ***Follow these instructions***:
> - Ensure each row corresponds to a single line in the output CSV with clear, descriptive headers and any subtotals or annotations as separate rows.
> - Ensure all data is accurately preserved and entries with commas are properly quoted.
>
> ***Example:***
> Header1, Header2, Header3
> data1, data2, data3
> subtotal1,,subtotal3
> Note: Descriptions or special instructions
> ***Content:***
> Raw table text information extracted by PyPDF2.

[2] https://pypi.org/project/PyPDF2/

11s not surprising tat the maj of ties concerning aquatic inset predate. ey elaonsipe vole the aque stages of mosquitoes asthe ey 2640), a someiastanes the predate, 18,68) Most mosquito larvae et reat nd maintain, and make excelent pe fr ¢ wide varity of agua eranisms In 'dil, the bing nuance and daca vector sipcance of many specie of Mult mosques has encouraged considerate tenon bing ives ote tral Othe than mosquito ava, sui larva and chironomid larae ae the mes feequtly slid invertebrate peey in the fesbinater habitat.

(a) Raw Tesseract-OCR results

It's not surprising that the majority of studies concerning aquatic insect predation focus on the aquatic stages of mosquitoes (e.g., 2640). In some instances, they predate mosquito larvae at rest and maintain, making excellent prey for a wide variety of aquatic organisms. In addition, the increasing nuisance and disease vector significance of many species of mosquitoes have encouraged considerable attention being given to their control. Other than mosquito larvae, similar larvae and chironomid larvae are the most frequently studied invertebrate prey in the freshwater habitat.

(b) Refined results with *LitAI*

Fig. 2: Comparison of raw Tesseract-OCR output and its enhancement via *LitAI*. Generative AI integrated in *LitAI* can correct typos and fix grammar issues existing in raw Tesseract-OCR outputs.

D. Image Processing

Understanding figures in scholarly papers is crucial as they often convey essential information. While most existing methods focus solely on the images when parsing them, *LitAI* employs generative AI to analyze images by considering both the figures and their corresponding descriptions in the PDF document. Specifically, we utilize GPT-3.5 Turbo Instruct to process the relevant text in the paper and adopt GPT-4 Vision for advanced image content interpretation. These AI models are integrated into our *LitAI* framework, providing powerful tools for extracting and enriching both textual and visual data. This multimodal approach offers deeper insights and a richer contextual understanding of images in scholarly papers. Through iterative testing and refinement, we have established a reliable workflow that effectively harnesses these AI capabilities, creating a novel approach to managing and interpreting visual data in PDF documents.

First, we utilize PymuPDF to systematically identify and extract images from PDFs on a page-by-page basis. We then employ Tesseract-OCR to associate each extracted image with its corresponding figure caption, leveraging spatial relationships and key identifiers such as *Figure* or *Fig*. To enhance OCR accuracy, we implement preprocessing techniques that improve image quality, including scaling and noise reduction.

Once the figure caption is identified, we use GPT-3.5 Turbo Instruct to match the image with the most relevant paragraph from the text, thereby providing a contextual basis for querying the image. The prompt used for this process is as follows:

Image-Text Matching Prompt: Find the paragraph containing the following figure caption and more details exact match: *figure_caption, page_text*

When the content related to the image description is ready, *LitAI* utilizes GPT-4 Vision to analyze the image within the context of the retrieved descriptions related to the image. The prompt used for this analysis is as follows:

Image Interpretation Prompt: What's in this image? Some more context of the image:{*related paragraph*}

IV. Experiments

This section presents the experimental results of *LitAI* in retrieving and analyzing text, tables, and images from scholarly articles. We evaluate the performance of our model by comparing it against two baselines: conventional methods that do not utilize AI, and end-to-end solutions that employ readily available AI technologies without further refinement.

A. Experimental Settings

To evaluate the performance of *LitAI*, we deploy it across scholarly articles within the biology and ecology fields. These articles exhibit significant diversity in writing layout, reflecting the intricate realities of processing literary works. The publication dates of these scholarly articles span from 1968 to 2023. The content of these papers includes a wealth of domain-specific terminology and features various tables and figures. We assess the performance of *LitAI* by comparing it with several established baselines in terms of information retrieval and comprehension of multimodal data, including text, tables, and figures, within these scholarly articles.

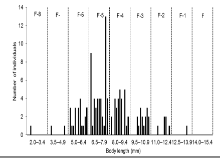
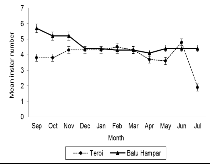
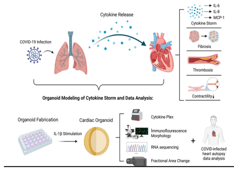
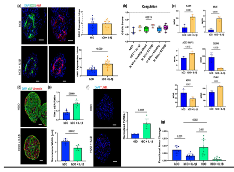
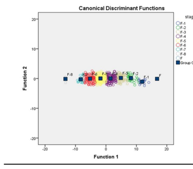
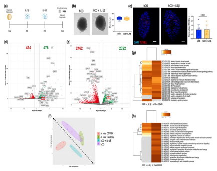
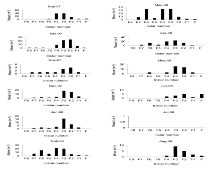
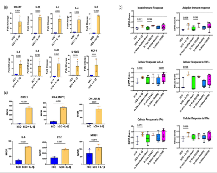
B. Text Extraction and Structuring

LitAI is adept at categorizing extracted text from scholarly papers into distinct sections such as the Abstract, Introduction, Methodology, Results, Conclusion, and References. We assess the capabilities of *LitAI* in text extraction and understanding on a section-by-section basis. Our evaluation covers 50 scholarly articles from the fields of Biology and Ecology, comparing *LitAI*'s performance against Tesseract-OCR and GPT-4 using metrics like cosine similarity of token counts and Word Mover's Distance (WMD) [15]. The comparative results are presented in Table I. These findings clearly demonstrate *LitAI*'s robust capabilities in text extraction and structuring, where it consistently outperforms Tesseract-OCR in nearly all sections based on both cosine similarity and WMD. While GPT-4 demonstrates superior performance compared to *LitAI* in processing sections such as Abstract, Discussion, and References based on WMD, we find GPT-4 is very unstable in handling PDFs, often failing to parse them. This instability significantly limits the possibility of conducting a scalable

TABLE I: Comparison of *LitAI*, Tesseract-OCR, and GPT-4 using cosine similarity of token matrix and Word Mover's Distance

	Abstract		**Introduction**		**Materials & Methods**		**Results**		**Discussion**		**Conclusion**		**References**		**Acknowledgment**	
Model	Cosine	WMD	Cosine	WMD	Cosine	WMD	Cosine	WMD	Cosine	WMD	Cosine	WMD	Cosine	WMD	Cosine	WMD
LitAI	**0.86±0.17**	0.76±0.19	0.76±0.23	**0.64±0.15**	**0.88±0.14**	**0.73±0.14**	**0.73±0.29**	**0.68±0.19**	**0.79±0.28**	0.75±0.20	**0.84±0.22**	**0.75±0.19**	**0.65±0.23**	0.59±0.16	**0.82±0.17**	**0.74±0.20**
Tesseract-OCR	0.82±0.11	0.69±0.16	**0.79±0.12**	0.59±0.11	0.76±0.25	0.64±0.13	0.70±0.24	0.59±0.14	0.79±0.24	0.68±0.17	0.68±0.27	0.63±0.17	0.52±0.30	0.55±0.12	0.62±0.25	0.62±0.18
GPT-4 (Unstable)	0.69±0.41	**0.77±0.25**	0.62±0.22	0.53±0.04	–	–	0.51±0.42	0.64±0.24	0.75±0.41	**0.79±0.25**	0.61±0.41	0.69±0.23	0.61±0.35	**0.68±0.24**	–	–

TABLE II: Comparison of *LitAI*, Salesforce BLIP, GPT-4, and Google Gemini on figure interpretation

Sample Image	**BLIP**	**GPT-4**	**Gemini**	***LitAI* (ours)**
	a plot of the number of the different species in the species	Histogram displaying body length distribution in millimeters. Vertical bars show counts per category labeled from F-1 to F-8 and F.	Graph shows distribution of body lengths in a population, with more individuals in the mid-range of lengths.	Bar graph details monthly counts of Thalerosphyrus nymphs by body length, with categories labeled F-8 to F.
	a line graph of the average and average time of a person	Monthly trend in mean instar number of "Terai" and "Batu Hampar", showing stability and fluctuation patterns.	Line graph displays monthly visitor numbers to Teroi and Batu Hampar, highlighting seasonal popularity.	Graph depicts growth rates in Thalerosphyrus populations in two rivers, with instar numbers decreasing over time.
	a diagram of the process of the human body's organ	Diagram shows COVID-19 impact on the body, especially cytokine storm effects on the heart, using cardiac organoids.	Investigates SARS-CoV-2 impact on the heart using organoid models to study shape, gene activity, and protein levels.	Schematic illustrates COVID-19 research on cytokine impacts on heart health, with detailed research methodology.
	a diagram of the different cells in the human cell are shown	Image presents experimental data on cellular responses, highlighting differences under various conditions.	Study utilizes human heart organoids to explore COVID-19-induced damage, analyzing organoid changes and molecular responses.	Panels depict IL-1β effects on cardiac organoids, comparing control and treated conditions across several experiments.
	a plot of a plot of a number of different functions	The image displays a scatter plot titled "Canonical Discriminant Functions," color-coding data points from stages F-1 to F-8 and group centroids, differentiating them based on the first two canonical discriminant functions.	The image displays a scatter plot of canonical discriminant analysis with colored squares representing observations and axes labeled as discriminant functions.	The image shows a canonical discriminant function plot distinguishing instar stages F-1 to F-8 of Thalerosphyrus by body length, using centroids to represent stages from Batu Hampar and Teroi river nymphs.
	a diagram of the different types of cell lines and their corresponding functions	This image illustrates the comprehensive effects of IL-1β on human colon organoids, detailing timelines, size, cell viability, gene expression, PCA, and gene ontology heatmaps to investigate inflammatory responses and implications for diseases like COVID-19.	The figure depicts experiments on human cardiac organoids under cytokine storms, including cell counts, fluorescence microscopy with DAPI and TUNEL, and gene ontology graphs, emphasizing their relevance to COVID-19.	The image shows diverse data visualizations such as stained organoids, cell death graphs, GSVA, volcano and PCA plots, and heatmaps, analyzing IL-1β effects on human cardiac organoids related to cardiac fibrosis.
	a diagram of the different cells in the human cell are shown	Bar graphs depict monthly size-frequency distributions of Thalerosphyrus from September 2007 to August 2008 in Batu Hampar River, showing fluctuations in instar stages F1 to F8 per square meter (m^2), emphasizing survival challenges in later stages.	The image is a scientific chart showing "Instar number" measurements from September 2007 to August 2008, ranging from 0 to 40.	The bar graphs show the monthly size-frequency of Thalerosphyrus in Batu Hampar River (Sep 2007-Aug 2008), detailing counts per m^2 from instars 1 to F8, with a decrease at instars 3 and 4, and successful maturation at instar 5.
	a diagram of the different cells in the human cell are shown	The figure displays immune responses in human colon organoids with bar graphs of cytokine changes and box plots comparing conditions in IL-1β treated and untreated hCOs.	The image displays an experiment on the effects of medications on immune responses, with graphs for cytokines such as IL-4, MCP-1, and IL-12p70, emphasizing cardiac health.	The image displays graphical analyses of IL-1β treatment effects on human cardiac organoids, highlighting cytokine and immune gene expression changes and providing insights into cardiac responses.

experiment using GPT-4, restricting our tests to five papers. Consequently, we exclude GPT-4's results from Table I when it fails to process more than one PDF out of the five papers.

C. Table Extraction and Understanding

The aim of this experiment is to assess the effectiveness of *LitAI* in detecting and formatting tables extracted from various scientific PDF documents. The evaluation focuses on the accuracy of table detection and the effectiveness of subsequent formatting to conform to structured data standards. We randomly selected 10 papers from the biology and ecology fields, each containing at least one table. The results highlight a high success rate in table detection, with a majority of the tables being accurately identified and extracted (see Table III). The overall accuracy rate for table detection stood at 93.3%, underscoring the effectiveness of the detection process. Additionally, 78.3% of the detected tables were successfully formatted into valid CSV structures, demonstrating the proficiency of the formatting algorithms.

TABLE III: Table detection and parsing accuracy by *LitAI*

Paper ID	**Year**	**# of Tables**	**Detection Acc.**	**Parsing Acc.**
1	1982	3	100%	100%
2	1978	3	33.3%	100%
3	2008	1	100%	0%
4	1968	6	100%	100%
5	2001	6	100%	83.3%
6	2018	3	100%	33.3%
7	1976	1	100%	100%
8	1981	2	100%	100%
9	2000	3	100%	66.7%
10	1974	10	100%	100%
Average	–	–	93.3%	78.3%

D. Image Interpretation

We evaluate the image interpretation capabilities of *LitAI* by comparing it to Salesforce BLIP [16], GPT-4, and Google Gemini. Table II presents the interpretation outcomes for eight images extracted from biology and ecology papers. Salesforce BLIP tends to generate basic and brief image captions. While GPT-4 and Google Gemini can produce more detailed descriptions, they often fail to contextualize the figures within the originating paper. Conversely, *LitAI* not only describes the figures but also contextualizes them based on the figure-related text description in the originating paper, significantly enhancing interpretation performance. For instance, in the second sample figure of Table II, *LitAI* successfully highlights features such as rivers, which the others overlook. Due to space constraints in the paper, Table II only illustrates the summarized figure descriptions generated by *LitAI* and baselines. Detailed descriptions are omitted but can be found on our GitHub repository.

V. Conclusion

In this paper, we introduce *LitAI*, a tool that leverages generative AI to redefine the retrieval of text, tables, and images from scientific documents. Our comparative analyses reveal that *LitAI* outperforms traditional methods and modern AI technologies like Tesseract-OCR and GPT-4, particularly in processing complex data formats. The integration of OCR with our generative AI framework not only improves the accuracy of data extraction but also enhances its efficiency, making *LitAI* a useful tool for researchers working with multifaceted document structures.

Future work will focus on expanding the capabilities of *LitAI* to include more languages and document formats, as well as enhancing and evaluating its application to other specialized fields of study. Through ongoing development and refinement, *LitAI* aims to remain at the forefront of literature information processing, supporting the evolving demands of scientific research and discovery.

References

[1] A. Singh, K. Bacchuwar, and A. Bhasin, "A survey of ocr applications," *International Journal of Machine Learning and Computing*, vol. 2, no. 3, p. 314, 2012.

[2] J. Huang, H. Chen, F. Yu, and W. Lu, "From detection to application: Recent advances in understanding scientific tables and figures," *ACM Computing Surveys*, 2024.

[3] F. Esposito, D. Malerba, and G. Semeraro, "A knowledge-based approach to the layout analysis," in *Proceedings of 3rd International Conference on Document Analysis and Recognition*, vol. 1, pp. 466–471 vol.1, 1995.

[4] M. Singh, B. Barua, P. Palod, M. Garg, S. Satapathy, S. Bushi, K. Ayush, K. S. Rohith, T. Gamidi, P. Goyal, *et al.*, "Ocr++: a robust framework for information extraction from scholarly articles," *arXiv preprint arXiv:1609.06423*, 2016.

[5] G. Mühlberger, J. Zelger, and D. Sagmeister, "User-driven correction of ocr errors: combining crowdsourcing and information retrieval technology," in *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*, pp. 53–56, 2014.

[6] R. Smith, "An overview of the tesseract ocr engine," in *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, vol. 2, pp. 629–633, 2007.

[7] S. Saoji, R. Singh, A. Eqbal, and B. Vidyapeeth, "Text recognition and detection from images using pytesseract," *J Interdiscip Cycle Res*, vol. 13, pp. 1674–1679, 2021.

[8] P. Sisodia and S. W. A. Rizvi, "Optical character recognition development using python," *Journal of Informatics Electrical and Electronics Engineering (JIEEE)*, vol. 4, no. 3, pp. 1–13, 2023.

[9] E. Oro and M. Ruffolo, "Trex: An approach for recognizing and extracting tables from pdf documents," in *2009 10th international conference on document analysis and recognition*, pp. 906–910, IEEE, 2009.

[10] L. D. Lopez, J. Yu, C. N. Arighi, H. Huang, H. Shatkay, and C. Wu, "An automatic system for extracting figures and captions in biomedical pdf documents," in *2011 IEEE International Conference on Bioinformatics and Biomedicine*, pp. 578–581, 2011.

[11] K. Lo, Z. Shen, B. Newman, J. Z. Chang, R. Authur, E. Bransom, S. Candra, Y. Chandrasekhar, R. Huff, B. Kuehl, *et al.*, "Papermage: A unified toolkit for processing, representing, and manipulating visually-rich scientific documents," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 495–507, 2023.

[12] B. Koning, "Extracting sections from pdf-formatted cti reports," B.S. thesis, University of Twente, 2022.

[13] Y. Zhou, H. Ong, P. Kennedy, C. Wu, J. Kazam, K. Hentel, A. Flanders, G. Shih, and Y. Peng, "Evaluating gpt-4 with vision on detection of radiological findings on chest radiographs," 2024.

[14] T. Hassan and R. Baumgartner, "Table recognition and understanding from pdf files," in *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, vol. 2, pp. 1143–1147, 2007.

[15] M. Kusner, Y. Sun, N. Kolkin, and K. Weinberger, "From word embeddings to document distances," in *International conference on machine learning*, pp. 957–966, PMLR, 2015.

[16] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," 2022.