Interpretation of Agent Project Report Based on PDF Document of Large Language Model

Gonghao

July 10, 2025

Abstract

This project constructs a large language model-based PDF document interpretation intelligent agent system[1], aiming to address the intelligent understanding and analysis of academic literature. The system adopts a multimodal technical architecture, integrating PaddleOCR's PPStructureV3 model for document structure parsing, combining Qwen visual language model for image content recognition, and building a Retrieval-Augmented Generation (RAG) system based on the LangChain framework, utilizing Chroma vector database for semantic retrieval. The frontend employs Streamlit for web interface construction, while the backend integrates DeepSeek large language model for text generation and question-answering processing. Through hierarchical text chunking strategies and multimodal information fusion, the system can accurately extract text, images, and table content from PDF documents, automatically generate literature summaries, and support natural language question-answering based on original text, including cross-modal queries targeting specific images. All responses are accompanied by relevant original text segments as evidence, ensuring traceability and accuracy of answers.

1 Introduction

With the rapid development of artificial intelligence technology, Large Language Models (LLMs) have achieved breakthrough progress in the field of natural language processing, providing new technical pathways for intelligent document understanding and analysis. In the academic research domain, researchers need to process large volumes of academic literature in PDF format, which typically contain complex textual structures, mathematical formulas, charts, and images among other multimodal information. Traditional document processing methods often struggle to effectively handle such multimodal, structured content, leading to incomplete information extraction and inaccurate understanding.

1.1 Background

Intelligent interpretation of academic literature faces multiple technical challenges. In terms of PDF document parsing, traditional text extraction methods based on libraries such as PyMuPDF often encounter issues such as formula recognition errors, text order confusion, and format loss when processing AI domain English literature containing complex mathematical formulas, multi-column layouts, and special characters. To address the above problems, this project has turned to optical character recognition (OCR) technology for PDF content extraction. However, traditional OCR technology also faces challenges when processing complex layout arrangements: issues such as text order recognition errors, multi-column content mixing, and table structure destruction still persist.

Additionally, image content in academic literature (such as experimental charts, algorithm flowcharts, and result presentation diagrams) contains rich semantic information that requires specialized visual understanding technology for accurate interpretation. Intelligent question-answering systems based on literature content also need to establish efficient retrieval mechanisms that can quickly locate relevant information from large volumes of text and generate accurate, traceable answers.

In recent years, Retrieval-Augmented Generation (RAG) technology has provided effective technical solutions to address the above problems. RAG combines information retrieval and text generation

[1] During the coding phase, I primarily used Cursor to help build the Streamlit framework and implement the interfaces I needed using appropriate prompts.

technologies to generate accurate and reliable answers based on retrieved relevant document segments. Meanwhile, the development of multimodal large language models enables systems to simultaneously process textual and visual information, laying the foundation for constructing unified document understanding frameworks.

1.2 Research Objectives and Contributions

This research aims to construct a large language model-based PDF document interpretation intelligent agent system, with main objectives including:

1. Multimodal Document Parsing: Develop a PDF parsing module capable of simultaneously processing text, images, and tables, ensuring completeness and accuracy of information extraction.
2. Intelligent Content Understanding: Integrate visual language models to achieve semantic understanding and description of image content in PDFs.
3. Efficient Retrieval Mechanism: Construct a semantic retrieval system based on vector databases, supporting fast and accurate information localization.
4. Traceable Question-Answering System: Develop a RAG-based question-answering module, ensuring all answers have corresponding original text evidence.
5. User-Friendly Interface: Provide an intuitive web interaction interface supporting literature upload, content display, and intelligent question-answering functionality.

The main contributions of this research include: proposing a PDF document interpretation framework that integrates multimodal technologies, solving key technical problems in intelligent understanding of academic literature; designing hierarchical text chunking strategies to improve retrieval accuracy and efficiency; implementing cross-modal question-answering functionality supporting queries targeting specific images; constructing a complete prototype system that validates the effectiveness of the proposed methods.

1.3 Paper Structure

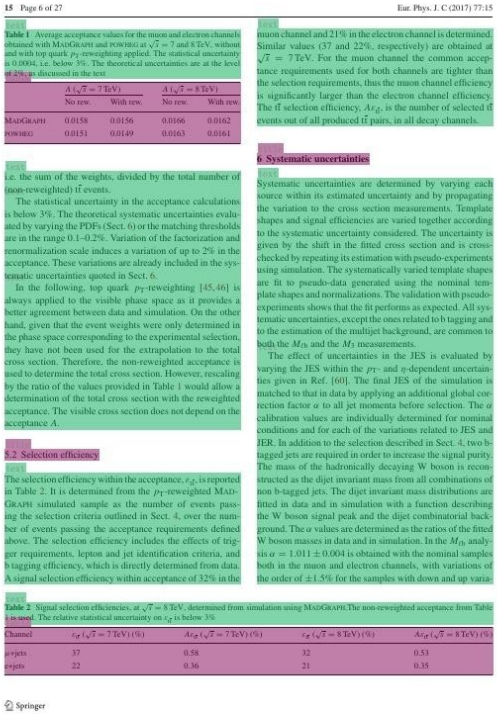
The remainder of this paper is organized as follows: Section 2 introduces the overall system architecture design; Section 3 provides detailed descriptions of the technical implementation of each core module; Section 4 demonstrates system functionality and experimental results; Section 5 summarizes the paper.

2 system architecture

The system adopts a modular layered design, with an overall architecture comprising a frontend interaction layer, PDF parsing module, image understanding module, vector retrieval and RAG module, and large language model question-answering module. These modules work collaboratively to achieve efficient parsing, content understanding, and intelligent Q&A for AI domain English PDF literature. The frontend is implemented using Streamlit for user interaction, the PDF parsing module utilizes PaddleOCR to extract multimodal content, the image understanding module integrates the Qwen visual language model, the vector retrieval and RAG module is based on LangChain and Chroma for efficient semantic retrieval, and the large language model Q&A module employs DeepSeek and other models to generate precise answers based on original text. Through multi-module collaboration, the system achieves structured parsing and intelligent Q&A for complex academic literature.

3 Technical Implementation

This chapter provides detailed descriptions of the technical implementation schemes for each core module of the system, including PDF document parsing, image content understanding, text vectorization and retrieval, and intelligent question-answering, among other key technical aspects.



15 Page 6 of 27 Eur. Phys. J. C (2017) 77:15

Table 1 Average acceptance values for the muon and electron channels obtained with MADGRAPH and POWHEG at $\sqrt{s} = 7$ and 8 TeV, without and with top quark p_T-reweighting applied. The statistical uncertainty is 0.0004, i.e. below 3%. The theoretical uncertainties are at the level of 2% as discussed in the text.

	A ($\sqrt{s} = 7$ TeV)		A ($\sqrt{s} = 8$ TeV)	
	No rew.	With rew.	No rew.	With rew.
MADGRAPH	0.0158	0.0156	0.0166	0.0162
POWHEG	0.0151	0.0149	0.0163	0.0161

i.e. the sum of the weights, divided by the total number of (non-reweighted) $t\bar{t}$ events.

The statistical uncertainty in the acceptance calculations is below 3%. The theoretical systematic uncertainties evaluated by varying the PDFs (Sect. 6) or the matching thresholds are in the range 0.1–0.2%. Variation of the factorization and renormalization scale induces a variation of up to 2% in the acceptance. These variations are already included in the systematic uncertainties quoted in Sect. 6.

In the following, top quark p_T-reweighting [45,46] is always applied to the visible phase space as it provides a better agreement between data and simulation. On the other hand, given that the event weights were only determined in the phase space corresponding to the experimental selection, they have not been used for the extrapolation to the total cross section. Therefore, the non-reweighted acceptance is used to determine the total cross section. However, rescaling by the ratio of the values provided in Table 1 would allow a determination of the total cross section with the reweighted acceptance. The visible cross section does not depend on the acceptance A.

5.2 Selection efficiency

The selection efficiency within the acceptance, $\varepsilon_{t\bar{t}}$, is reported in Table 2. It is determined from the p_T-reweighted MADGRAPH simulated sample as the number of events passing the selection criteria outlined in Sect. 4, over the number of events passing the acceptance requirements defined above. The selection efficiency includes the effects of trigger requirements, lepton and jet identification criteria, and b tagging efficiency, which is directly determined from data. A signal selection efficiency within acceptance of 32% in the muon channel and 21% in the electron channel is determined. Similar values (37 and 22%, respectively) are obtained at $\sqrt{s} = 7$ TeV. For the muon channel the common acceptance requirements used for both channels are tighter than the selection requirements, thus the muon channel efficiency is significantly larger than the electron channel efficiency. The $t\bar{t}$ selection efficiency, $A\varepsilon_{t\bar{t}}$, is the number of selected $t\bar{t}$ events out of all produced $t\bar{t}$ pairs, in all decay channels.

6 Systematic uncertainties

Systematic uncertainties are determined by varying each source within its estimated uncertainty and by propagating the variation to the cross section measurements. Template shapes and signal efficiencies are varied together according to the systematic uncertainty considered. The uncertainty is given by the shift in the fitted cross section and is cross-checked by repeating its estimation with pseudo-experiments using simulation. The systematically varied template shapes are fit to pseudo-data generated using the nominal template shapes and normalizations. The validation with pseudo-experiments shows that the fit performs as expected. All systematic uncertainties, except the ones related to b tagging and to the estimation of the multijet background, are common to both the M_{t} and the M_3 measurements.

The effect of uncertainties in the JES is evaluated by varying the JES within the p_T- and η-dependent uncertainties given in Ref. [60]. The final JES of the simulation is matched to that in data by applying an additional global correction factor α to all jet momenta before selection. The α calibration values are individually determined for nominal conditions and for each of the variations related to JES and JER. In addition to the selection described in Sect. 4, two b-tagged jets are required in order to increase the signal purity. The mass of the hadronically decaying W boson is reconstructed as the dijet invariant mass from all combinations of non b-tagged jets. The dijet invariant mass distributions are fitted in data and in simulation with a function describing the W boson signal peak and the dijet combinatorial background. The α values are determined as the ratios of the fitted W boson masses in data and in simulation. In the M_{t} analysis $\alpha = 1.011 \pm 0.004$ is obtained with the nominal samples both in the muon and electron channels, with variations of the order of ±1.5% for the samples with down and up varia-

Table 2 Signal selection efficiencies, at $\sqrt{s} = 8$ TeV, determined from simulation using MADGRAPH. The non-reweighted acceptance from Table 1 is used. The relative statistical uncertainty on $\varepsilon_{t\bar{t}}$ is below 3%

Channel	$\varepsilon_{t\bar{t}}$ ($\sqrt{s} = 7$ TeV) (%)	$A\varepsilon_{t\bar{t}}$ ($\sqrt{s} = 7$ TeV) (%)	$\varepsilon_{t\bar{t}}$ ($\sqrt{s} = 8$ TeV) (%)	$A\varepsilon_{t\bar{t}}$ ($\sqrt{s} = 8$ TeV) (%)
μ+jets	37	0.58	32	0.53
e+jets	22	0.36	21	0.35

Springer

Figure 1: Layout analysis using PaddleOCR

3.1 PDF Document Parsing Module Implementation

Current open-source solutions for PDF parsing in the field are Langchain's Unstructured and the recently popular project gptpdf. Unstructured's advantage lies in integrating the complete OCR and layout analysis process, which can output rich text chunks, facilitating subsequent retrieval and Q&A. However, it has limited support for multi-modal content such as images and charts in documents. gptpdf uses PyMuPDF for layout parsing, merging text areas through rules and labeling image and chart areas, then submitting all content to GPT-4o or Qwen-VL and other multimodal large models for recognition, generating structured markdown documents. This method improves structural reconstruction and multimodal parsing capabilities, but relies heavily on large models and has accuracy and efficiency bottlenecks in complex documents.

In contrast, PaddleOCR and its PPStructureV3 model represent another technical direction based on OCR combined with deep learning. This approach leverages advanced neural network architectures for both text recognition and layout analysis, enabling robust extraction and structuring of diverse content types, including text, images, and tables, even in complex document layouts. Compared to Unstructured and gptpdf, PPStructureV3 demonstrates greater stability and efficiency in multimodal content parsing and complex layout reconstruction, and offers strong scalability for practical deployment. Therefore, this project implements high-quality structural parsing of AI domain English PDF literature based on PPStructureV3, providing a solid foundation for subsequent multimodal understanding and intelligent Q&A.

3.2 Image Understanding Module Implementation

The image understanding module is built based on the Qwen visual language model, which possesses powerful multimodal understanding capabilities. During implementation, the system first preprocesses images extracted from PDFs, including image denoising, resolution adjustment, and format standardization. Then, the module utilizes Qwen's visual encoder to extract image features, converting image content into understandable text descriptions through visual-language alignment mechanisms.

The generated image descriptions are not only used to answer image-related questions, but are also re-inserted into the document text with a special token to be identified as pics as part of the text

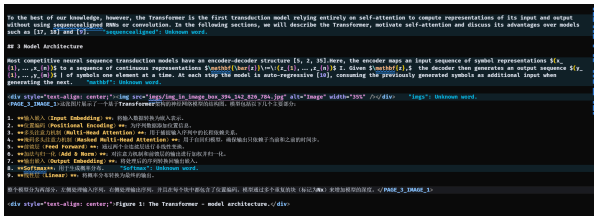


Figure 2: A description of the image with special token in markdown

chunks, thereby enhancing the contextual information of the document. This strategy helps improve the accuracy of subsequent semantic retrieval and intelligent Q&A, and enables tighter multimodal information integration.

During image description generation, the system adopts a Chinese output strategy, ensuring that generated descriptions align with Chinese users' reading habits. Additionally, the module implements associative storage of images with text content, providing support for subsequent cross-modal Q&A. Through this approach, the system can answer users' questions about specific image content, achieving true multimodal interaction.

3.3 Text Vectorization and Retrieval Module Implementation

The text vectorization and retrieval module is built based on the LangChain framework and Chroma vector database, implementing efficient semantic retrieval functionality. In text chunking, the system adopts a hierarchical chunking strategy, first performing high-level segmentation according to the document's title structure, then conducting fine-grained text chunking within each segmented block. This strategy maintains the logical structure of the document while ensuring reasonable chunk sizes.

In the vectorization process, the system uses a multilingual text vectorization model (text2vec-base-multilingual), which can effectively handle mixed Chinese-English text content. For each text chunk, the model generates high-dimensional vector representations that capture the semantic information of the text. The system stores these vectors in the Chroma vector database and establishes corresponding index structures to support fast similarity retrieval.

During retrieval, the system first converts user questions into vector representations, then finds the most relevant text segments through cosine similarity calculations. To improve retrieval accuracy, the system adopts a multi-round retrieval strategy, performing coarse-grained retrieval first, then fine-grained filtering of retrieval results. Additionally, the system implements a hybrid retrieval mechanism based on keywords, combining semantic retrieval with keyword matching to further improve retrieval recall and accuracy rates.

3.4 Intelligent Q&A Module Implementation

The intelligent Q&A module is built based on advanced large language models such as DeepSeek, implementing high-quality answer generation functionality. In question understanding, the system first analyzes the type of user question, determining whether it is an image-related question, text-related question, or mixed-type question. For image-related questions, the system retrieves corresponding image descriptions and contextual information; for text-related questions, the system retrieves relevant text segments.

In the answer generation process, the system adopts the Retrieval-Augmented Generation (RAG) technical approach. First, the system combines user questions with retrieved relevant text segments into prompts, then inputs them into the large language model for answer generation. To improve answer quality and traceability, the system explicitly requires the model to generate answers based on retrieved content in the prompts and cite relevant original text segments in the answers.

For cross-modal Q&A, the system implements special processing mechanisms. When users ask about specific image content, the system first displays the corresponding image, then generates answers based on image descriptions and related text context. This design ensures that users can intuitively see the image content while obtaining accurate interpretations based on the image.

4 System function presentation

Press the button on the side box shown in Figure 3 to upload the PDF file.

Figure 3: The index of system

Here I upload the paper "Attention is all you need " as a sample and ask some questions. After the system parses the PDF file (it might take some time), you can ask any question you want in Figure 4.

Figure 4: The system parses file successfully

Moreover, the system provides evidence from the source document along with its exact location.

During the text chunking phase, I employ a Map-Reduce approach to process the chunked text through the LLM for comprehension to get a better conclusion.

Figure 5: A question about img

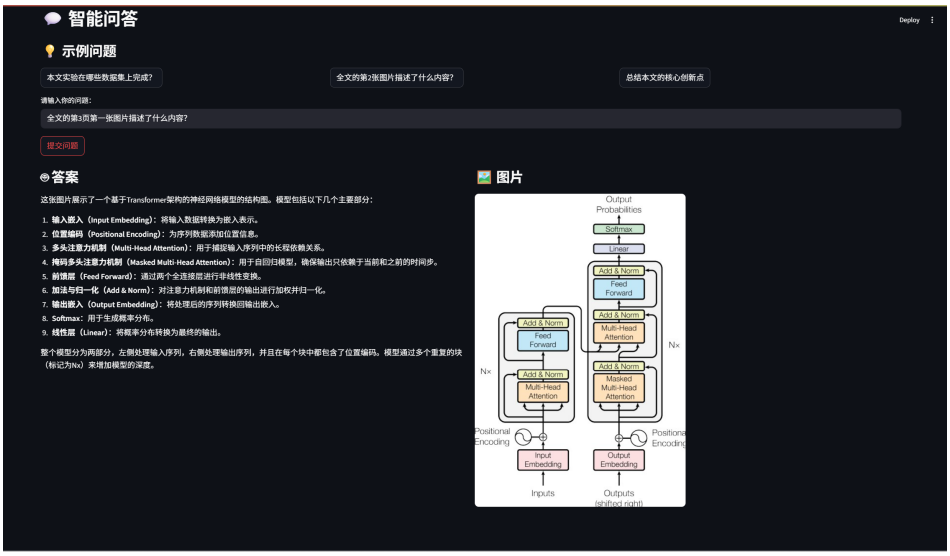


Figure 6: A question with page numbers about imgs

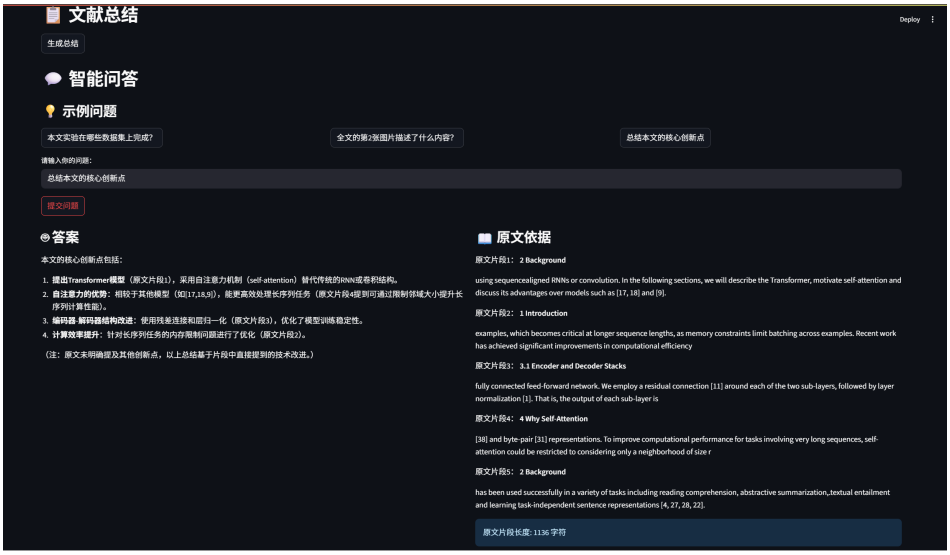


Figure 7: Sample question

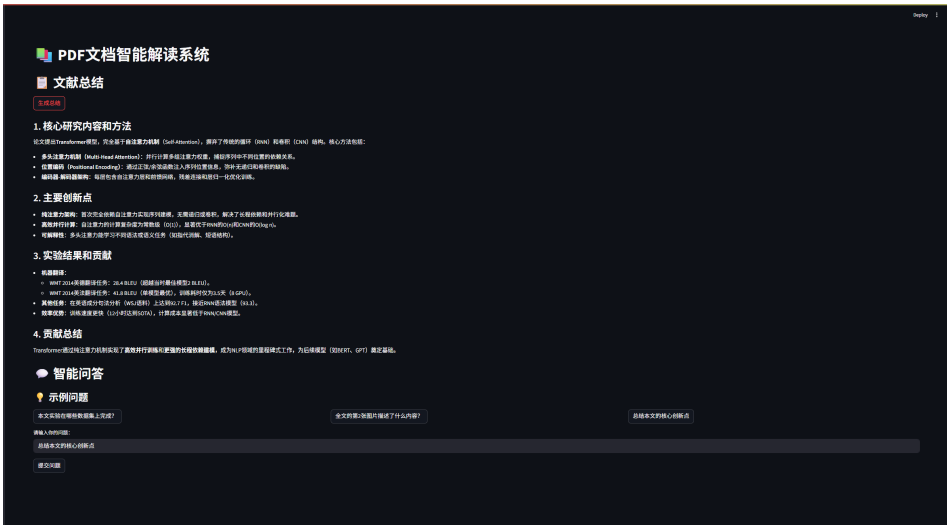


Figure 8: Generating conclusion of the paper