International Research Journal of Education and Technology

Peer Reviewed Journal

ISSN 2581-7795

LLM Parsing For Data Entry Using AI

Sowndarya R, Rajeshwari K

Dept of Computer Applications
Adhiyamaan College of Engineering, Hosur

---***---

Abstract - *The project examines the application of large language models (LLM) in automatic and enhancing data entry tasks. The system removes important information from unnecessary text, maps the data into a structured field, and detects errors in real time. The approach improves accuracy, scalability and flexibility in various industries, reduces human error and increases productivity. LLM for data entry project uses devices such as Pytessract for Parsing such as optical character recognition (OCR) and poppler for PDF parsing. The text extraction module removes raw text from various document formats, while the data mapping module converted the text into structured data, aligning it with predetermined areas. Document classification module classification classifies documents using the learning algorithm. The natural language interface module enables interactive input, improves access and efficiency. The project can reduce human error, increase productivity, and improve the overall efficiency of data admission processes.*

Key Words: *Large Language Model, Optical Character Recognition (OCR), Text Extraction, Data Mapping, Natural Language Interface.*

1. INTRODUCTION

LLM Parsing is an innovative solution for data entry project that takes advantage of Large Language Model (LLM) to automatically automatic data entry tasks. The objective of this project is to streamline the workflow by removing important information from unnecessary text and detecting errors in real time. The use of LLMS enables the system to learn from data and improve its accuracy over time. The project uses a five module to automatically automatically extends the extraction and processing of data from unnecessary text. These modules include lesson extraction, data mapping, document classification, natural language interfaces and integration. The **Text Extraction Module** uses Optical Character Recognition (OCR) for PDF parsing and pytessract for popplar. The **Data mapping modules** convert the extracting text into structured data, aligning it with predetermined areas. **Document classification module** classification classifies documents using the learning algorithm. The **natural language interface module** enables interactive input, improves access and efficiency. By using LLM power, the project can significantly reduce human error and increase productivity in data entry functions.

1.1 Text Extraction

The **Text Extraction Module** uses an OCR tool, an OCR tool to remove raw text from various document formats such as PDFs, pictures, and scanned files. Recognizing the characters within the pytessaract images and converts unnecessary material into a machine-elective text. This process ensures that the data is ready for further analysis and processing. By taking out the text with efficiency, the module lays the foundation for later data mapping and classification works

1.2 Data Mapping

The **Data mapping modules** convert the extracting text into structured data, aligning it with a predetermined areas in the database or system. This identifies the major pieces of information, such as the date, volume and name, and maps them in their respective fields. This ensures that the data is accurately represented and is ready for further process or integration. By automating this mapping process, the system increases efficiency and reduces the risk of human error**.**

1.3 Document Classification

The **Document classification** in this project classifies module documents into predetermined types, such as invoices, contracts and receipts. Using the machine learning algorithm, it analyzes the structure and material of each document to apply appropriate processing rules. This classification ensures that correct data extraction and mapping techniques are used for each document type. By automating this stage, the system improves accuracy and reduces the need for manual intervention.

2. Natural Language Interface

The natural language interface module in this project enables users to interact with the system through the converted language, making the process more comfortable and adapted to the user. Instead of relying on complex forms or technical inputs, users can input data, request updates, or ask for information in simple, natural language. This approach reduces obstacles for use, allowing non-technical users to easily connect with the system. By supporting flexible communication, it streamlines the data entry process and increases the overall system efficiency. The module not only improves access, but also encourages widespread adoption in diverse user groups.

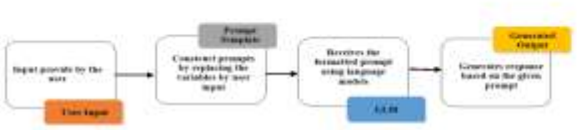
Input provide by the user
User Input
Prompt Template
Construct prompts by replacing the variables by user input
Receives the formatted prompt using language models
LLM
Generated Output
Generates response based on the given prompt

Fig -1: **LLM Parsing**

3. CONCLUSIONS

In Conclusion, **LLM Parsing for Data Entry Using AI** provides a transformative approach to automatic and enhance data extraction and processing tasks. By integrating advanced machine learning models, system streamlines workflows, reduces human error, and improves efficiency in various industries. The combination of text extraction, document classification, data mapping and natural language interfaces ensures that user can basically and accurately interact with the system. Real-time processing and predictive analytics further enhance the capabilities of the system, making it favourable and scalable for various business needs. This solution not only increases productivity, but also promotes greater access to diverse users. With its ability for future enhancement and comprehensive integration, the project has the ability to reopen data entry practices in areas such as finance, healthcare and legal. Ultimately, it determines the stage for more intelligent and efficient document management systems, providing both operational and strategic benefits.

ACKNOWLEDGEMENT

This journal paper was truly prepared by myself I agree the terms and conditions.

REFERENCES

[1] V.-H. Le and H. Zhang, "Log-based anomaly detection without log parsing", *2021 36th IEEE/ACM International Conference on Automated Software Engineering (ASE)*, pp. 492-504, 2021.

[2] H. Dai, H. Li, C. S. Chen, W. Shang and T.-H. Chen, "Logram: Efficient log parsing using n-gram dictionaries", *IEEE Transactions on Software Engineering*, 2020.

[3] "Natural Language Processing with Spark NLP: Learning to Understand Text at Scale" by Alex Thomas, published in July 2020.

[4] Foundations of Large-Scale Multimedia Information Management and Retrieval" by Edward Y. Chang, published in 2011.