DATA EXTRACTION FROM HETEROGENEOUS INSURANCE POLICIES USING MULTIMODAL LLMS

Simeon Monov, Nikolay Pavlov, Miglena Dodleva

Abstract. *In previous work [1] we observed limitations when large language models (LLMs) were applied to extract data from textual representations of policy documents converted to Markdown format. While this approach demonstrated strong performance for structured sources such as Excel files, the processing of PDF documents remained a bottleneck due to complex layouts, embedded tables, and scanned content. This paper presents an experimental approach for extracting information from such heterogeneous documents using multimodal LLMs capable of jointly processing text and images. The proposed method applies multiple multimodal LLMs to identify and extract key information fields and represent them in a unified JSON structure. The outputs produced by different multimodal models are analyzed and validated against human interpretation of the document content in order to evaluate precision, consistency, and completeness. The results aim to assess the potential of multimodal LLMs to improve the robustness of automated information extraction from heterogeneous insurance documentation and other visually complex financial or legal documents.*

Key words: Data Extraction, MLLMs, Multimodal Large Language Models, Heterogeneous Data, Insurance Policies, Automation, JSON, Financial Documentation, Legal Documentation

Acknowledgments

The research is supported by the project FP25-FMI-010 "Innovative interdisciplinary research in Informatics, Mathematics, and Pedagogy of Education" of the Scientific Fund of the Paisii Hilendarski University of Plovdiv, Bulgaria.

References

[1] S. Monov, N. Pavlov, M. Dodleva, Data Extraction from Heterogeneous Insurance Policies Using LLMs, *Proc. of IMEA'2025*

Simeon Monov[1,*], Nikolay Pavlov[1], Miglena Dodleva[1]
[1] Paisii Hilendarski University of Plovdiv,

Faculty of Mathematics and Informatics,
236 Bulgaria Blvd., 4027 Plovdiv, Bulgaria
Corresponding author: smonov@gmail.com