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 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.*

Introduction

Operational workflows in the insurance industry rely heavily on structured data describing insured assets, coverage limits, premiums, and deductibles. However, this information is often embedded in heterogeneous documents such as Excel files and PDF reports with varying layouts and formatting.

In previous work [1], we explored the use of large language models (LLMs) for extracting structured data from such documents after converting them into Markdown format. While the approach demonstrated satisfactory effectiveness on well-structured Excel input data, it revealed significant limitations when applied to complex PDF documents.

The main issue lies in the loss of layout and structural information during text conversion, especially for documents containing multi-page tables, merged cells, or scanned content. As a result, extraction quality decreases noticeably and outputs become inconsistent or incomplete.

To address these limitations, this work investigates the use of multimodal large language models (MLLMs), which can process both textual and visual information simultaneously. The objective of this work is not traditional OCR or plain text recognition, which are already well established, but semantic reconstruction of structured insurance information from heterogeneous multimodal documents.

The proposed approach focuses on interpreting relationships between textual content, layout structure, grouped tables, and domain-specific financial fields in order to generate a unified machine-readable representation.

Problem Overview

Insurance companies process large volumes of policy documents daily, received from multiple insurers and brokers. These documents are highly heterogeneous, exhibiting varying formats (e.g., Excel, PDF), complex layouts with mixed content (text and tables), and inconsistent structures across documents.

This variability makes traditional rule-based or template-based extraction approaches unreliable, as they depend on stable formats and predefined patterns. The challenge is particularly pronounced in PDF documents, where visually structured information, such as multi-page tables, merged cells, and embedded content, is difficult to accurately interpret using text-only processing.

As a result, manual data extraction remains widely used in practice, despite being slow and error-prone. Therefore, robust automation requires approaches capable of handling both the semantic content and the visual structure of heterogeneous documents. Insurance policies are particularly challenging due to the presence of grouped financial tables, object-specific coverage structures, deductibles, and domain-specific terminology that must be interpreted consistently across different document layouts.

Task Definition

The objective of this work is to investigate the capability of multimodal large language models to extract structured data from heterogeneous insurance policy documents. Given a document in Excel or PDF format, the system is expected to identify relevant content, interpret its structure, and generate a unified JSON representation of insured objects and their associated attributes.

The extracted information includes hierarchical insurance entities and domain-specific financial relationships represented in a unified JSON schema. Depending on the document type, the extracted structure may include insured vessels, vehicles, or land-based equipment together with associated characteristics such as brand, model, vessel type, tonnage, IMO number, classification society, insured sums, premium values, deductible structures, policy periods, currencies, and grouped insurance risks.

The extraction process therefore requires not only text recognition, but also contextual association between visually separated document sections, tables, object-specific attributes, and financial entities distributed across heterogeneous layouts. The complexity of the task increases further when multiple insured objects share partially overlapping financial sections or grouped deductible tables that require contextual interpretation.

The performance of the models is evaluated by comparing their outputs against human-interpreted reference data in terms of precision, consistency, and completeness.

Motivation for Multimodal LLMs

The limitations of text-based processing approaches highlight the need for methods that can better capture the structural complexity of real-world documents. In particular, converting PDF documents into plain text or Markdown often results in the loss of critical layout information, including table boundaries, column relationships, and hierarchical structure. Although OCR technologies have achieved high text-recognition accuracy for many document types [7], semantic reconstruction of heterogeneous structured documents remains challenging, particularly when complex layouts and domain-specific relationships must be interpreted consistently.

Recent work in document understanding has explored multimodal approaches such as Document Visual Question Answering (DocVQA) [3] and spatially-aware architectures like LayoutLM [2], which explicitly incorporate layout information. More recently, multimodal large language models, including GPT-4V [4], Gemini [5], and Qwen-VL [6], have further advanced the field by enabling joint reasoning over text and images. However, these approaches remain limited when applied to highly heterogeneous real-world documents.

Multimodal large language models (MLLMs) offer a promising alternative by combining textual and visual understanding within a single framework. Unlike traditional LLM-based pipelines, which rely solely on textual representations, multimodal models are capable of directly interpreting document images alongside extracted text. This enables them to preserve spatial relationships and better understand complex layouts, such as multi-page tables and irregular formatting.

Unlike plain OCR extraction, the proposed approach attempts to preserve semantically relevant structural relationships between visually grouped document elements rather than only reproducing raw text formatting.

By leveraging both text and visual cues, multimodal LLMs improve the extraction of key entities, financial values, and structural relationships within insurance documents. This makes them particularly suitable for handling heterogeneous and visually complex inputs, where traditional approaches struggle.

Methodology

The proposed approach extends previous text-based extraction pipelines by incorporating multimodal large language models capable of processing both textual and visual information in order to reconstruct structured object-level representations from heterogeneous insurance policy documents.

1. Preprocessing

The input documents are first analyzed according to their file type. For Excel documents, the system extracts structured sheet content and renders sheets as images. For PDF documents, it extracts page-level text and renders each page as an image.

Depending on the document type, this step includes:

- extraction of text or structured sheet content;
- rendering of pages or sheets as images;
- detection of table-like content.

Document pages and sheets are rendered as images using default preprocessing settings while preserving textual information and visual layout cues required for multimodal interpretation.

2. Content Selection and Prompt Construction

Relevant document content is selected and used to construct the model input. For PDF files, the full extracted text is provided together with images of pages detected as containing table-like content. These images preserve layout information. For Excel files, structured sheet text and rendered sheet images are included.

The prompt defines the extraction task and specifies the target JSON schema. It instructs the model to use textual content for exact values and visual input for interpreting layout, merged cells, grouped headers, and visually complex table structures.

All models are evaluated in a zero-shot setting using a consistent prompt structure that defines the target JSON schema and includes detailed task instructions, without task-specific fine-tuning or few-shot examples.

3. Multimodal Model Inference

The constructed prompt is submitted to one or more multimodal LLMs. The models process both textual and visual inputs in order to extract structured information about insured objects.

This enables the system to:

- interpret complex layouts and formatting;
- identify insured objects and related attributes;
- extract values from tables and mixed-content document sections.

Using multiple models allows their output to be compared in terms of robustness and extraction quality.

4. Response Processing and Normalization

The model output is first cleaned and validated as a JSON object. It is then parsed and transformed into a structured representation.

During normalization, field names are aligned, values are converted to consistent formats, and missing fields are explicitly represented as null where applicable. Additional normalization includes mapping of risk names, numeric value standardization, and deterministic ordering of extracted objects.

The resulting structured representation constitutes the model's final prediction for the given document.

5. Evaluation

The extracted results are evaluated against a human-interpreted reference representation of the documents. The evaluation focuses on correctness, consistency, and completeness of the extracted information.

Each extracted object-field pair is compared against the reference JSON representation. Correctly extracted fields, missing fields, incorrectly assigned values, and hallucinated fields are considered during evaluation. Due to the limited number of documents, the study combines qualitative assessment with approximate object-field level comparison rather than large-scale statistical benchmarking.

The qualitative labels "Low", "Medium", "High", and "Very High" correspond to increasing levels of structural and semantic alignment with the reference JSON representation, ranging from major extraction failures ("Low") to near-complete reconstruction ("Very High").

Experimental Setup

The experiments are conducted on the same set of heterogeneous insurance policy documents used in our previous study [1], consisting of two Excel files and two multi-page PDF reports. The documents vary in layout, table structure, and representation of financial fields, covering both relatively regular and highly irregular cases. The study is intended as an exploratory comparative evaluation rather than a large-scale benchmark. The selected documents were chosen to represent structurally heterogeneous real-world cases frequently encountered in insurance workflows.

Due to confidentiality restrictions associated with real insurance policies, the evaluated documents cannot be publicly released. The study therefore focuses on comparative analysis across models using the same document set.

A gold-standard JSON representation of insured objects and their attributes is used as a reference for evaluation, enabling direct comparison with the previously studied text-based approach.

The performance of multiple multimodal large language models is compared, including GPT-4.1-mini, GPT-5-mini, Gemini 3.1 Pro, and Qwen variants, applied using a consistent prompting strategy and a unified JSON schema.

Additional experiments were conducted with Mistral models, but due to their lower reliability and limited output quality, they were not included in the main comparison.

Results

Table 1. Excel Sample A

Model	Structure	Accuracy	Normalization	Notes
GPT-4.1-mini	Excellent	Very High	Clean	Best overall
GPT-5-mini	Excellent	Very High	Minor issues	Prefixes, formatting
Gemini 3.1 Pro	Excellent	Very High	Minor issues	Similar to GPT-5-mini
Qwen 2.5 (7B)	Partial	Medium	Incorrect	Duplicate names, incorrect values, and hallucinated fields
Qwen 3.5 (27B)	Good	High	Semantic errors	Incorrect mapping (e.g., IV interpreted as Inventory)

Table 2. Excel Sample B

Model	Structure	Accuracy	Normalization	Notes
GPT-4.1-mini	Excellent	Very High	Minor issues	Year truncation, over-extracted deductibles (MIN/MAX)
GPT-5-mini	Excellent	Very High	Minor issues	Formatting, over-structured deductibles (%)
Gemini 3.1 Pro	Excellent	Very High	Minor issues	Similar to GPT-4.1-mini
Qwen 2.5 (7B)	Weak	Low	Incorrect	Missing objects, no deductibles, missing rates
Qwen 3.5 (27B)	Partial	High	Partial	Missing deductible names (merged table)

Tables 1 and 2 summarize the results on the two Excel policy samples. The leading models (GPT-4.1-mini, GPT-5-mini, and Gemini 3.1 Pro) achieve very high accuracy and consistent structure, with only minor normalization issues. In contrast, Qwen models show lower reliability, including missing fields, incorrect values, and structural inconsistencies. Overall, the results indicate that extraction from structured Excel documents is largely reliable for advanced multimodal models.

Table 3. PDF Sample A

Model	Structure	Accuracy	Normalization	Notes
GPT-4.1-mini	Good	High	Clean	Combined-column errors (A/B), wrong column alignment
GPT-5-mini	Excellent	Very High	Minor issues	Best overall, prefix normalization
Gemini 3.1 Pro	Excellent	Very High	Minor issues	Similar to GPT-5-mini, missing deductible
Qwen 2.5 (7B)	Weak	Low	Incorrect	Incorrect values, field confusion, misclassification
Qwen 3.5 (27B)	Good	Medium	Partial	Missing premium values

Table 4. PDF Sample B

Model	Structure	Accuracy	Normalization	Notes
GPT-4.1-mini	Partial	High	Limited	Misclassification, duplication, incomplete values
GPT-5-mini	Good	Medium	Strong	Context mixing, incorrect value

				assignment, identifier confusion
Gemini 3.1 Pro	Good	High	Limited	Best values, weaker classification
Qwen 2.5 (7B)	Weak	Low	Incorrect	Missing objects, no deductibles
Qwen 3.5 (27B)	Good	Medium	Partial	Full coverage, missing deductibles

Tables 3 and 4 summarize the results on the two PDF samples. Performance is more variable compared to Excel due to complex layouts. GPT-5-mini and Gemini 3.1 Pro achieve the strongest results, while GPT-4.1-mini is more sensitive to structural issues. Qwen models show lower reliability, with frequent extraction errors. Overall, multimodal models improve PDF extraction, but structural challenges remain.

Table 5. Average Object-Field Extraction Performance

Model	**Excel EMA (%)**	**Excel SA (%)**	**Excel Avg. Extra Fields**	**PDF EMA (%)**	**PDF SA (%)**	**PDF Avg. Extra Fields**
GPT-4.1-mini	95.8	98.0	92	71.93	71.15	6.5
GPT-5-mini	67.5	89.0	135	72.31	78.31	6.5
Gemini 3.1 Pro	82.5	93.0	92	70.60	79.92	0.5
Qwen 2.5 (7B)	43.3	46.0	0	56.20	26.89	68.5
Qwen 3.5 (27B)	78.3	92.0	46	65.25	65.38	0

Table 5 presents an object-field extraction performance comparison across the evaluated models. The reported Excel and PDF values represent averages across the two evaluated documents of each type.

Exact Match Accuracy (EMA) evaluates strictly identical object-field extraction relative to the reference JSON representation, including exact values, structure, and object associations.

Semantic Accuracy (SA) additionally accepts semantically equivalent normalized values and domain-specific mappings as correct. Missing fields, hallucinated fields, incorrect values, duplicate entities, and incorrect object associations are treated as errors and reduce the final semantic accuracy score.

Avg. Extra Fields represents the average number of additionally generated fields or objects that are not present in the reference JSON representation. These values primarily reflect hallucinated or redundant fields generated during extraction from complex tables and grouped financial sections. The evaluated

Excel documents contain 400 and 552 reference object-fields respectively, while the evaluated PDF documents contain 59 and 770 reference object-fields respectively.

Qualitative Error Analysis

A qualitative analysis of model outputs reveals several recurring error patterns across document types. The most common issues are related to structural interpretation, including incorrect column alignment, misclassification of fields, and incomplete extraction of table content.

For Excel documents, errors are primarily related to normalization, such as over-structured deductibles or minor formatting inconsistencies. In contrast, PDF documents exhibit more complex failure modes, including incorrect mapping of values across columns, missing fields, and confusion between object categories.

Additionally, some models produce hallucinated fields or duplicate entities, particularly in cases with ambiguous or irregular table structures. These observations highlight that, while multimodal models improve extraction performance, accurate structural interpretation remains a key challenge.

Computational Analysis

Larger multimodal models generally demonstrate stronger structural reasoning capabilities but require significantly higher computational resources and inference time compared to smaller local models. Smaller local models provide faster execution but exhibit reduced robustness on irregular layouts.

Conclusion

The results demonstrate that multimodal large language models improve the robustness of data extraction from complex PDF documents by combining textual and visual information. While extraction from structured Excel inputs is largely reliable, PDF processing remains challenging due to structural and contextual complexity.

Across the evaluated models, GPT-5-mini achieves the best overall performance, followed by Gemini 3.1 Pro and GPT-4.1-mini, which also demonstrate strong but slightly less consistent results.

No single model performs consistently across all aspects, highlighting the importance of post-processing and normalization. The study remains exploratory in nature due to the limited number of evaluated documents and the confidentiality constraints associated with real insurance data. These findings confirm that combining multimodal understanding with structured post-processing is key to achieving reliable extraction in real-world scenarios. The proposed approach represents a domain-specific application of the broader multimodal document

understanding problem. Future work will focus on improving structural understanding and exploring more advanced multimodal approaches for complex document processing.

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*

[2] Y. Xu, M. Li, L. Cui, S. Huang, F. Wei, M. Zhou, LayoutLM: Pre-training of Text and Layout for Document Image Understanding, *Proc. of KDD*, 2020, DOI: https://doi.org/10.1145/3394486.3403172.

[3] M. Mathew, D. Karatzas, C. Jawahar, DocVQA: A Dataset for VQA on Document Images, *Proc. of WACV*, 2021, DOI: https://doi.org/10.48550/arXiv.2007.00398.

[4] J. Achiam et al., GPT-4 Technical Report, 2023, DOI: https://doi.org/10.48550/arXiv.2303.08774

[5] R. Anil et al., Gemini: A Family of Highly Capable Multimodal Models, 2023, DOI: https://doi.org/10.48550/arXiv.2312.11805.

[6] J. Bai et al., Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond, 2023, DOI: https://doi.org/10.48550/arXiv.2308.12966

[7] AIMultiple Research, OCR Accuracy Benchmark, AIMultiple, [Online], available: https://aimultiple.com/ocr-accuracy, (Accessed: May 9, 2026)

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@uni-plovdiv.bg