Assessing Large Multimodal Models on Complex Technical and Procedural Documents*

Roberto Zanoli[1,*], Alessandro Dal Pozzo[2], Alberto Maria Massatani[2], Manuela Speranza[1], Ravi Kiran Chikkala[3] and Bernardo Magnini[1]

[1]*Fondazione Bruno Kessler, Trento, Italy*
[2]*Rete Ferroviaria Italiana S.p.A, Italy*
[3]*Saarland University, Germany*

Abstract

This paper investigates the ability of Large Multimodal Models (LMMs) to interpret complex technical and procedural documents. We present experiments conducted on documents from Rete Ferroviaria Italiana (RFI), which are critical for infrastructure planning and safety compliance and that combine textual and graphical elements, making automated interpretation challenging. We used a state-of-the-art multimodal model, i.e., GPT-4o, and designed five practical tasks that correspond to potential applications for RFI: (i) interpretation of the structure of the document as well as its graphical elements; (ii) detection of semantic inconsistencies; (iii) generation of summaries; (iv) re-writing according to targeted modifications; and (v) generation of flow diagrams from textual descriptions. Results indicate good performance in extracting and integrating information. However, the model shows limitations in generating detailed summaries for longer documents and in identifying semantic inconsistencies, particularly when the inconsistency type is unspecified or the context is extensive. Overall, the model achieves an average score of 70% across the evaluated tasks.

Keywords

large multimodal models, document analysis, technical texts

1. Introduction

Technical documents (e.g., technical guidelines for managing the development of roads, bridges and other transport networks) and procedural documents (e.g., regulations addressing actions and processes to be put in place in specific circumstances) are produced in high quantities and are of utmost utility for companies. However, due to the intrinsic complexity of such documents, their automatic analysis has, until now, required complex pipelines of specialized components, with high costs of maintenance and poor capacity of generalization and portability across different domains. Recently, several vision-language models integrating vision and text understanding (e.g., Qwen 2.5VL, LlamaOneVision, Molmo, Intern, etc.) have been made available, making it possible new progress in the interpretation of multimodal documents.

More specifically, in the paper we investigate the ability of Large Multimodal Models (LMMs) to analyze and interpret complex technical and procedural documents from Rete Ferroviaria Italiana (RFI). These documents provide guidelines for railway operations, safety, and maintenance and are therefore important for planning and building infrastructure projects. Technical documents, especially in civil engineering, are difficult to process automatically due to their mix of detailed descriptions, tables and figures. Additionally, they are usually very long, sometimes hundreds of pages, and come in different digital formats, such as text-based PDFs or scanned PDFs. Similarly, regulatory and procedural documents are challenging to interpret due to their specialized and legal language, even though they contain fewer graphical elements than technical reports.

Several studies have explored the use of Generative AI to process both technical documents in engineering and construction and procedural documents such as manuals, regulations and guidelines.

Ital-IA 2025: 5th National Conference on Artificial Intelligence, organized by CINI, June 23-24, 2025, Trieste, Italy
*Corresponding author.
zanoli@fbk.eu (R. Zanoli); a.dalpozzo@rfi.it (A. Dal Pozzo); a.massatani@rfi.it (A. M. Massatani); manspera@fbk.eu (M. Speranza); rach00004@teams.uni-saarland.de (R. K. Chikkala); magnini@fbk.eu (B. Magnini)

A comprehensive survey on text mining in the construction industry is presented in [1], while [2], [3], and [4] focus on the application of NLP and deep learning in this domain. For procedural content, [5] propose methods for extracting structured procedures, and [6] use LLMs to extract procedural knowledge from text and populate a Knowledge Graph. Despite these advances, no studies have yet tested the ability of new state-of-the-art Large LMMs to analyze technical and procedural documents.

To address this gap, we extended the work of [7] by evaluating GPT-4o on four key tasks, each corresponding to a potential application within the RFI domain: (i) *interpretation of structure, graphics, and content of documents* to automate the review of technical materials for compliance; (ii) *detection of semantic inconsistencies* to identify procedural conflicts across large document archives; (iii) *generation of summaries* to condense complex content for quicker understanding; (iv) *re-writing of document portions according with targeted modifications*, to adapt standard procedures to specific project needs. In addition, we conducted a preliminary study on a fifth task, (v) *generation of flow diagrams from textual descriptions*, aimed at enabling rapid verification of procedural workflows.

Each task was performed by asking the model a set of questions in Italian using a zero-shot approach, i.e. without providing the model with any examples. The responses of the model were manually verified to calculate the model's performance in terms of accuracy; responses were rated on a scale of 0 to 2 points based on their completeness and correctness and overall performance was then normalized into a percentage value. Additionally, to further evaluate the reliability of the model, in some cases we included "trap" questions aiming to provoke incorrect responses.

2. Task 1: Interpretation of Structure, Graphics and Content

This test evaluates the ability of GPT-4o to interpret both textual content (including structure, and graphical elements such as tables, images) and mathematical expressions in technical and procedural documents [7].

Dataset: The dataset consists of four documents: two technical documents related to railway and civil engineering, and two procedural documents from the RFI's Safety Management System. It includes both selectable text and scanned documents. The technical documents contain a variety of elements: 24 tables, 4 photos, 50 figures and 2 graphs. The procedural documents are mainly based on text, with a limited number of visual elements: 14 tables and 2 figures.

Methodology: Table 1 presents the two main categories of questions asked to the model. Questions related to text focus on bibliographic extraction, document structure identification (e.g., page numbers, tables and figures), text interpretation and summarization. Questions related to visual elements involve interpreting tables, describing images, analyzing figures, understanding mathematical expressions and graphs. "Trap" questions, such as asking for information about a column in a table that does not exist, are used to test the robustness of the model.

Results: GPT-4o achieves an average accuracy of 83.66% on textual content and 88.00% on graphic content, with an overall accuracy of 85.83%. However, the accuracy drops significantly to 80.25% when answering "trap" questions. Table 2 reports the model's accuracy in detail for textual and graphical content, based on a 0–2 scoring scale reflecting completeness and correctness.

3. Task 2: Detection of Semantic Inconsistencies

This test assesses the GPT-4o's ability to identify inconsistencies by verifying, for example, whether descriptions of the same procedure written in different documents or reported in different parts of the same document are consistent with each other.

Table 1
Example of questions for evaluating textual and graphical content comprehension. Trap questions are highlighted in boldface.

Content	Question
Bibliographic Info.	Extract the full names of the authors of the document.
Doc. Structure	Provide the exact number of pages/tables in the document.
Text Interpretation	What norms must be followed according to the technical specifications? **What is the main cable length of the cable car?**
Tables	What is the tensile strength of the 4 mm lower membrane? **What is the highest value in the fifth column of Table 12.8.1-1?**
Photos	Describe the objects or people in Figure 12.8.4.2.6.a. **How many trees are in the figure?**
Figures	Describe the contents of Figure 12.8.4.2.5.c. **What does the red-colored object in the figure represent?**
Math Expressions	What does the mathematical expression $11 \leq n \leq 40$ in Table 12.14.3.7 refer to? **How is the product in the mathematical expression interpreted?**
Graphs	What is represented in the graph of Figure 1?

Table 2
Accuracy on regular questions (left) and "trap" questions (right) for technical and procedural documents.

	Tech. Docs.		Proc. Docs.		Tech. Docs.		Proc. Docs.	
Content	**RFI**	**Other**	**RFI**	**RFI**	**RFI**	**Other**	**RFI**	**RFI**
Bibliographic Info.	1.00	1.00	1.00	1.00	—	—	—	—
Doc. Structure	0.50	0.67	0.92	0.75	—	—	1.00	1.00
Text Interpretation	0.80	1.00	0.62	0.76	0.50	1.00	0.71	0.71
Tables	1.00	1.00	0.80	0.90	0.00	1.00	1.00	0.75
Photos	0.50	1.00	–	0.75	0.75	1.00	—	1.00
Figures	0.50	1.00	–	0.75	0.00	1.00	—	0.50
Math. Expressions	1.00	1.00	–	1.00	0.00	1.00	—	0.50
Graphs	1.00	–	–	1.00	1.00	—	—	1.00

Dataset: The evaluation involves two procedural documents from RFI. All inconsistencies have been intentionally introduced by RFI to evaluate the model's analytical capabilities.

Methodology: Inconsistencies are detected both within a single document (intra-document) and between two documents (inter-document), with the evaluation focusing on two dimensions: Search Space and Prompt Type. For Search Space, the process starts by analyzing the entire document (whole document(s)) and then focuses on smaller portions: whole sections (minimal section +1), sections without subsections (minimal section), and finally, specific text segments like paragraphs or excerpts (paragraph and excerpt). This approach tries improving the model's accuracy by reducing the search space. Three levels of detail are used in the prompts. Generic prompts, such as "Check if the information in the following text(s) is consistent" (Gen. 1) or "Find and report any contradictions between the following texts or within the following text" (Gen. 2), are applied. More precise prompts (Specific) ask, for instance, whether the descriptions of certain systems or processes are consistent across the texts. The most explicit prompts (Explicit) directly mention that a contradiction exists in the text and instruct the model to identify and report it.

Results: Explicit prompts are the most effective in detecting inconsistencies, outperforming generic and specific prompts, especially when applied to small text segments. Both intra- and inter-document evaluations show that explicit prompts, along with Gen. 2, are more accurate in detecting inconsistencies in smaller sections such as Min. Sec., Paragraph, and Excerpt, where they achieve higher accuracy.

Table 3
Results of intra- and inter-document evaluations, showing how effectively inconsistencies are detected. The ratings range from 0 (no inconsistency detected) to 2 (perfect detection).

	Intra-Document				Inter-Document			
Content	**Gen. 1**	**Gen. 2**	**Specific**	**Explicit**	**Gen. 1**	**Gen. 2**	**Specific**	**Explicit**
Whole Document	0	0	0	0	0	0	0	2
Min. Sec. +1	0	0	0	1	0	0	0	0
Min. Sec.	0	0	0	2	0	2	0	2
Paragraph	0	0	0	2	0	2	2	2
Excerpt	0	2	0	2	1	2	2	2

4. Task 3: Generation of Summaries

GPT-4o is evaluated based on its ability to generate good summaries, i.e., complete and accurate (subtask A) and meet predefined length requirements (subtask B).

Dataset: For this task, we use five procedural documents from RFI's Safety Management System, ranging from 2,000 to 10,000 words.

Methodology: The quality of the summaries (subtask A) is evaluated by asking the model to generate summaries from single or multiple documents and evaluating the output on a scale from 0 to 2 points. To assess its ability to meet predefined length requirements (subtask B), the model was asked to generate summaries of sizes between 200 and 1,000 words.

Results: GPT-4o achieves an overall accuracy of 79% in subtask A. Regarding summary length (subtask B), the model is able to follow the instruction properly for summaries up to 400 words, whereas generating longer summaries is quite challenging. In particular, summaries requested of 600 words are sensibly shorter and gaps between requested and actual length increases even more when asking for 800- and 1000-word summaries.

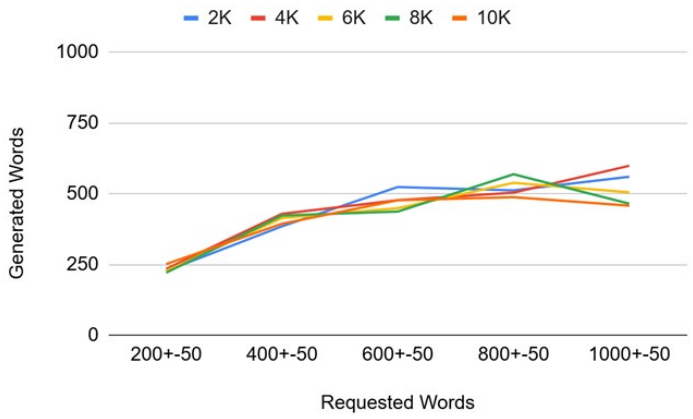

Figure 1: Requested vs. generated word count for documents with increasing lengths: 2,000 (2K), 4,000 (4K), 6,000 (6K), 8,000 (8K), and 10,000 (10K) words.

5. Task 4: Text re-writing according to Specific Modifications

This task evaluates the ability of the model to rewrite a given text by including a series of modifications specified in the prompt.

Dataset: The dataset consists of 3 RFI procedural documents of varying lengths (ranging from 8 to 23 pages).

Methodology: For the evaluation, the model is asked to produce paraphrases of specific sections of the documents (referring to them either by their number or content) while incorporating specific changes to the original text. These changes include, for example, adjustments to scheduled times, changes in responsibilities, addition of prescriptions and other indications. The paraphrases produced and the modifications incorporated are then evaluated on a scale from 0 to 2 points, based on their completeness and accuracy.

Results: GPT-4o achieves an overall accuracy of 66.67%. It successfully completes the task both with just one and with two input documents. The model's performance remains consistent whether the input section is referred to by its number (e.g., "Section II.3") or by its content (e.g., "the part about the Auditor Registry").

6. Task 5: Generation of Flow Diagrams from Textual Descriptions

The aim of this preliminary study is to investigate the multimodal capabilities of GPT-4o and also GPT-4o-mini (a more efficient version of GPT-4o) in creating flowcharts from procedures described textually in RFI procedural documents.

Dataset: The dataset includes five RFI procedural documents, split into development (two documents) and test (three documents) sets. Each document, ranging from 13 to 42 pages, contains a flowchart alongside textual descriptions clarifying specific procedures.

Methodology: Two experiments are conducted under both zero-shot and few-shot learning settings. In the first one, called Direct Flowchart Generation, the models are asked to directly produce a UML activity diagram based on the procedural text found in the RFI documents. In the second, named Flowchart Generation in PlantUML Language, the models' task is to generate the same diagram in the form of PlantUML code, which could be rendered using the PlantUML tool. This approach aims at leveraging the models' ability to convert natural language instructions into formal and domain-specific representations.

Results: In both zero-shot and few-shot learning settings, the evaluated models have difficulty in creating clear UML diagrams, often producing unclear or meaningless results. However, the models are able to generate correct PlantUML[1] code from the given descriptions.

7. Discussion and Conclusions

This study investigates GPT-4o's performance in analyzing RFI's technical and procedural documents, focusing on five tasks. **Interpretation of structure, graphics and content:** GPT-4o obtains strong performance (85.83%) on the task, as it successfully extracts bibliographic information and achieves consistent accuracy in understanding tables, photos, and graphs. However, the model fails to index pages, figures and tables. Additionally, the model is robust to hallucinations. **Detection of semantic**

[1] https://plantuml.com/

inconsistencies: The model's accuracy ranges from 0 to 100 depending on the prompt and context, with an overall average of 48.50%. The model obtains good results in detecting inconsistencies when the prompt explicitly highlights the contradiction or describes its nature. This is especially effective on short sections such as excerpts, paragraphs and minimal sections. However, performance decreases significantly when the inconsistency requires understanding larger contexts. **Generation of summaries:** GPT-4o achieves an accuracy of 79% in producing summaries. The model follows the instructions accurately when asked to produce summaries up to 400 words, especially when the prompt clearly requests a detailed summary. However, it shows limitations with longer summaries, often producing shorter texts than requested. This limitation probably originates from the model's training on shorter examples. **Re-writing with targeted modifications:** The model achieves an overall accuracy of 66.67% on the task. It effectively applies the requested changes. However, when asked to fully paraphrase text segments, the model tends to produce overly concise versions, often omitting important contextual details. **Flow diagram generation from text:** The models show difficulty in generating clear flow diagrams directly from text, but obtain better performance when using a formal language like PlantUML. This highlights a strength in translating structured text into domain-specific languages. On average, across the first four key tasks described above, the model achieves an accuracy of 70%.

Future work aims to improve these results through two main directions. First, the use of Retrieval Augmented Generation (RAG) should reduce hallucinations and improve response accuracy and completeness, especially if applied to large document collections such as those potentially available from RFI. Second, the use of few-shot learning could further improve the model's ability to handle diverse tasks.

Acknowledgments

This work has been partially supported by the PNRR project FAIR - Future AI Research (PE00000013), under the NRRP MUR program funded by the NextGenerationEU.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] H. Yan, M. Ma, Y. Wu, H. Fan, C. Dong, Overview and analysis of the text mining applications in the construction industry, Heliyon 8 (2022) e12088. URL: https://www.sciencedirect.com/science/article/pii/S240584402203376X. doi:https://doi.org/10.1016/j.heliyon.2022.e12088.

[2] Y. Ding, J. Ma, X. Luo, Applications of natural language processing in construction, Automation in Construction 136 (2022) 104169. URL: https://www.sciencedirect.com/science/article/pii/S0926580522000425. doi:https://doi.org/10.1016/j.autcon.2022.104169.

[3] A. Shamshiri, K. R. Ryu, J. Y. Park, Text mining and natural language processing in construction, Automation in Construction 158 (2024) 105200. URL: https://www.sciencedirect.com/science/article/pii/S0926580523004600. doi:https://doi.org/10.1016/j.autcon.2023.105200.

[4] A. Erfani, Q. Cui, Natural language processing application in construction domain: An integrative review and algorithms comparison, 2022, pp. 26–33. doi:10.1061/9780784483893.004.

[5] S. Agarwal, S. Atreja, V. Agarwal, Extracting procedural knowledge from technical documents, ArXiv abs/2010.10156 (2020). URL: https://api.semanticscholar.org/CorpusID:224803122.

[6] V. A. Carriero, A. Azzini, I. Baroni, M. Scrocca, I. Celino, Human evaluation of procedural knowledge graph extraction from text with large language models, in: Knowledge Engineering and Knowledge Management: 24th International Conference, EKAW 2024, Amsterdam, The Netherlands, November 26–28, 2024, Proceedings, Springer-Verlag, Berlin, Heidelberg, 2024, p. 434–452. URL: https://doi.org/10.1007/978-3-031-77792-9_26. doi:10.1007/978-3-031-77792-9_26.

[7] B. Magnini, A. Dal Pozzo, R. Zanoli, Understanding high-complexity technical documents with state-of-art models, in: F. Dell'Orletta, A. Lenci, S. Montemagni, R. Sprugnoli (Eds.), Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024), CEUR Workshop Proceedings, Pisa, Italy, 2024, pp. 540–547. URL: https://aclanthology.org/2024.clicit-1.64/.