

2025-08-13

Access the code and data at <https://github.com/psimm/meet-your-new-client>

Meet Your New Client: Writing Reports for AI – Benchmarking Information Loss in Market Research Deliverables*

Paul F. Simmering[†]
Q Agentur für Forschung GmbH
paul@simmering.dev

Benedikt Schulz
Q Agentur für Forschung GmbH
benedikt.ps@gmail.com

Oliver Tabino
Q Agentur für Forschung GmbH
oliver.tabino@teamq.de

Dr. Georg Wittenburg
Inspirient GmbH
georg.wittenburg@inspirient.com

ABSTRACT As organizations adopt retrieval-augmented generation (RAG) for their knowledge management systems (KMS), traditional market research deliverables face new functional demands. While PDF reports and slides have long served human readers, they are now also “read” by AI systems to answer user questions. To future-proof reports being delivered today, this study evaluates information loss during their ingestion into RAG systems. It compares how well PDF and PowerPoint (PPTX) documents converted to Markdown can be used by an LLM to answer factual questions in an end-to-end benchmark. Findings show that while text is reliably extracted, significant information is lost from complex objects like charts and diagrams. This suggests a need for specialized, AI-native deliverables to ensure research insights are not lost in translation.

KEYWORDS retrieval-augmented generation; knowledge management; market research; large language models

*We would like to thank the participants of the AI Forum of the German Society for Online Research (DGOF) for the discussion about knowledge management that inspired this study. We are also grateful for helpful comments by participants at the General Online Research 2025 conference. Paul Simmering acknowledges inference credits granted by Fireworks AI and compute credits granted by Modal.

[†]Corresponding author.

1 Introduction

As organizations implement retrieval-augmented generation (RAG) in knowledge management systems (KMS), the audience for market research deliverables is shifting. Reports in PDF and PowerPoint (PPTX) formats, once designed for human readers, are now also consumed by AI systems. For consultancies and market research agencies, this means that effective packaging of results for AI readability is becoming a core requirement. Information loss during document ingestion can undermine downstream retrieval and generation, making the choice of format and structure critical for future-proofing deliverables.

Recent surveys and case studies (e.g., M. Cheng et al. (2025), Gao et al. (2024), PricewaterhouseCoopers (2024)) highlight the growing adoption of RAG and the importance of document structure for effective AI use. However, standard delivery formats may not be optimal for machine processing, especially when information is embedded in and/or implied via complex layouts or images.

This study quantifies the information loss that occurs when documents are converted to Markdown for RAG. We systematically compare how well documents in PDF and PPTX formats can be used by large language models to answer factual questions. Our end-to-end evaluation uses four open-source Markdown conversion libraries (Docling, Marker, Markitdown, Zerox OCR) and three vision and question answering models. We introduce a new QA benchmark and corpus of 41 PPTX documents, expanding on prior work focused on PDFs. Our findings inform both practitioners and researchers on the limitations of current formats and tools and provide concrete guidance for agencies seeking to deliver AI-ready reports.

2 Related work

Research on the topic comes from two directions: (1) document layout analysis benchmarks for the accuracy of document segmentation and conversion, and (2) question-answering benchmarks that evaluate the accuracy of an answer generation model. As an end-to-end benchmark, this study bridges the gap between the two.

2.1 Document layout analysis (DLA) benchmarks

DLA benchmarks focus on the accuracy of document segmentation using image models. DocLayNet by Pfitzmann et al. (2022) is the largest benchmark for PDF layout segmentation, featuring 11 layout elements, including text, tables, and pictures, across 80863 manually annotated pages. Later, H. Cheng et al. (2023) added M⁶Doc, which includes scanned documents, 74 fine-grained layout elements, and a large Chinese corpus.

OmniDocBench by Ouyang et al. (2025) introduced PDF slides as a document type, relevant to our study. Going beyond DLA, it evaluates tools like MinerU, Marker, Mathpix, and VLMs by comparing their Markdown output to ground truth. Images are removed in pre-processing. Pipeline tools and VLMs performed comparably, with MinerU excelling on English and Qwen2-VL-72B by Alibaba

on Chinese. Qwen2-VL-72B also performed best on the JSON extraction OCR benchmark by OmniAI (2025).

Specialized benchmarks for specific layout elements include PlotQA by Methani et al. (2020) and ChartQA by Masry et al. (2022) for chart understanding, HybridQA by W. Chen et al. (2020) for table-based QA, and PubTables-1M by Smock, Pesala, and Abraham (2022) for table extraction from scientific papers.

2.2 Question-answering (QA) benchmarks

Question-answering benchmarks evaluate how reliably models can produce correct answers from document-derived evidence. Comprehensive surveys map this space and its task formulations (Cambazoglu et al. 2021; Wang 2022; Rogers, Gardner, and Augenstein 2023). In the taxonomy of Rogers, Gardner, and Augenstein (2023), the present study falls under probing questions with free-form answers over expert materials in a monolingual (English) setting. It also works directly with realistic unstructured documents, rather than starting from cleaned text-only representations in classic benchmarks such as HotpotQA (Z. Yang et al. 2018).

FinanceBench (Islam et al. 2023) is closest in spirit among end-to-end evaluations over real documents. It targets financial filings and reports. The benchmark is limited to text and tabular layout elements, starting from PDFs that are converted to text using PyMuPDF and Langchain, without image extraction. A key finding is that a long-context approach, where the full document text is provided to the model, outperforms vector-store variants. Their best configuration (long context with GPT-4 Turbo) reaches 79% accuracy. FinanceBench further reports that placing relevant context before the question improves performance. These observations motivate our own use of a long-context setting and prompt.

The UDA benchmark (Hui, Lu, and Zhang 2024) advances realism by operating on raw PDFs and HTML rather than pre-cleaned text, aggregating diverse datasets and question types. It concentrates on text and tables and compares several extraction and parsing strategies. Notably, a VLM-based table parsing approach outperforms classic extraction, and a sequential “parse first, then answer” pipeline can surpass direct image-questioning. UDA also contrasts long-context with retrieval pipelines, reporting broadly similar performance on knowledge tasks but an advantage for RAG on numerical reasoning. While UDA omits data charts, diagrams, and images as layout elements, its approach and focus on real-world usefulness motivate our study.

Building on these lines of work, we extend end-to-end QA to PPTX and to diagrams, data charts, and images. We foreground the conversion step, comparing four Markdown conversion libraries, and quantify how format (PDF vs PPTX) and layout affect QA accuracy. The emphasis is practical: guidance for market research teams on format and layout choices that preserve answerable content in KMS.

3 Benchmark

3.1 Overview

Figure 1 illustrates the experiment setup. It consists of 468 questions relating to 41 PPTX documents. Each document is manually converted into PDF format using PowerPoint. Then, the PPTX and PDF versions of the documents are each converted to Markdown via conversion libraries (see Section 3.6). This results in two Markdown files per original document. These form the knowledge base for the question answering task. A language model is prompted with each question and given the corresponding document’s full Markdown text as context. Finally, the answers are compared to the human-annotated ground truth by an LLM as a judge.

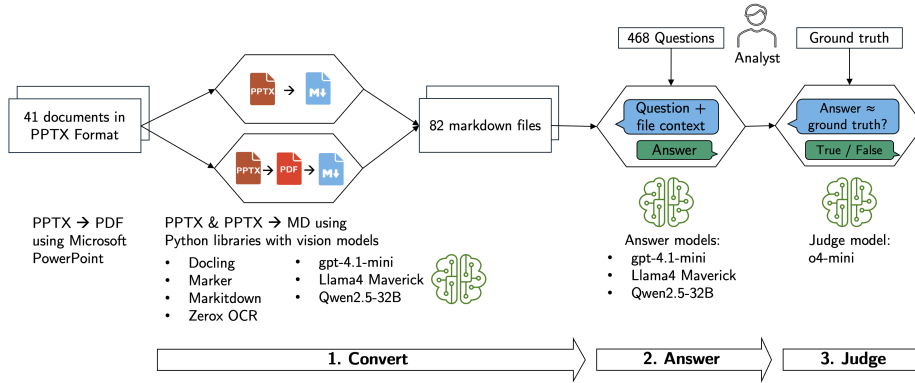


Figure 1: Experiment setup diagram.

3.2 File formats

A key difference between the file formats lies in their structure. Livathinos et al. (2025) categorize the formats into low-level formats (PDF and PNG) and markup-based formats (PPTX).

- **PPTX** is a semantic, XML-based format that stores objects like text and tables in a structured hierarchy. In theory, this structure is highly amenable to machine processing. Images are an exception and require a VLM to be captioned.
- **PDF** is a visual format describing page layout, not semantic content. Reconstructing the document’s structure from low-level drawing instructions is complex and prone to information loss.

Each step involving a VLM or other stochastic model carries a risk of information loss or hallucination. Based on the differences in structure, a starting hypothesis is that PPTX files will be converted to Markdown with less information loss than PDF files.

3.3 Collection of documents

The PPTX documents were sourced from publicly accessible web sources and additional documents were supplied by Q Agentur für Forschung, a market research agency. The majority of the files were found in GitHub repositories. These were found via the GitHub API and then manually inspected. Only English documents that contain information across at least three different layout elements were included. Document length ranges from 8 to 106 slides, with an average of 38.3 slides per document.

In regard to content, the documents are market research reports, conference presentations and lecture slides. They cover a range of topics, from cosmetics trends to political science. Due to the nature of GitHub as a source, computer science is the most common topic. A complete list of the documents, including their sources and licenses, is available in the GitHub repository.

3.4 Question formulation

Each item in the benchmark is a triplet (D, q, a) , where D is a PPTX document in its original format, q is the question answerable using information in D , and a is the ground truth answer to the question. This mirrors the format of UDA benchmarks (Hui, Lu, and Zhang 2024). A total of 468 such questions were formulated across the 41 documents, amounting to an average of 11.4 questions per document.

The content that questions relate to is distributed across the slides. Some relate to the first slide, others up to slide 99. The average content position is at 14.9 slides.

The questions were manually formulated by two of the authors. When formulating the questions, care was taken to ensure that they could be answered unambiguously. To isolate the impact of different formats, each question was designed to be answerable using information from a single slide and a single layout element. However, information can be present in multiple forms in the same document. For example, a text box can repeat information that is also present in a data chart or a table. When asked a question about the report, the model could use either layout element or a combination to find the answer. It may also infer the correct answer from the surrounding text. As Rogers, Gardner, and Augenstein (2023) note “we need to reconsider the idea that whether or not a given reasoning skill is ‘required’ is a characteristic of a given question. It is rather a characteristic of the combination of that question and the entire dataset”. In some instances, text boxes serve as axis labels for a data chart, requiring the model to understand the relationship between the chart and the text.

3.5 Layout elements


Questions pertain to information presented in different layout elements. We distinguish between five layout elements found in PowerPoint files, shown in Figure 2.

- **Text:** Information within a single, contiguous text field.

- **Image:** Visual elements, including photographs and screenshots of other layout elements such as a table saved as an image.
- **Table:** Structured data arranged in rows and columns.
- **Diagram:** Combinations of text and graphical elements, such as flow charts.
- **Data Chart:** Graphical representations of numeric data, such as bar or line charts. Only native PowerPoint charts are included; charts inserted as images are classified as *Image*.

Other data in PowerPoint, such as videos and speaker notes, are outside of the scope of this study.

To avoid an outsized impact from a single unusually easy or difficult object on the result, a maximum of three questions were asked regarding one layout element, for example a diagram on a particular slide.



Text
Q: Is GameStop's income positive or negative?
A: Negative

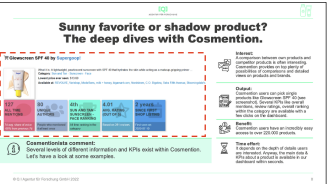
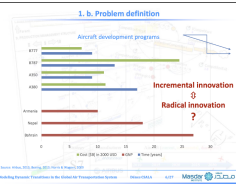


Image (red highlight not in original document)
Q: How many unique authors mentioned Glowscreen SPF 40 by SuperGoop?
A: 80



Data chart
Q: Which aircraft development program has been there longer, A380 or A350?
A: 40%

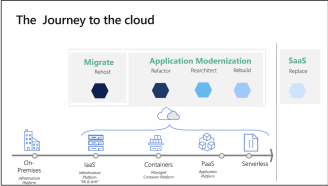


Diagram
Q: What comes after containers in the journey to the cloud?
A: PaaS

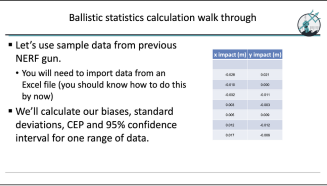


Table
Q: What is the y impact in m for the bullet with the x impact of -0.028m?
A: $4.03 * 2.23$

Figure 2: Example slides, questions and answers for the five layout elements.

3.6 Document conversion libraries

A common entry point into a text-based RAG system is to convert documents to Markdown. The plain text format is token-efficient, human-readable, and supports many formatting options. Images are supported as external references with optional captions. LLMs can also generate Markdown themselves (Z. Chen et al. 2025), making it useful for summaries and intermediate outputs. The two main conversion paradigms are (1) pipeline-based decomposition using specialized object detection models and (2) end-to-end processing using a vision language model (VLM).

Table 1 gives an overview of open-source libraries to convert PDF and PPTX to Markdown. In addition, there are proprietary solutions like LlamaParse and Mathpix. The pipeline library `unstructured` is omitted as it does not support

export to Markdown at the time of writing. PyMuPDF and pypdf are included because they are used in related benchmarks, however, they do not support PPTX conversion. MinerU does not support PPTX files, but the wrapper library magic-doc does. However, the version of MinerU supported by magic-doc lags behind the current version of MinerU.

Library	Developer	Paradigm	PDF	PPTX	License
Docling	IBM	Pipeline	Yes	Yes	MIT
GPTPDF	Chen Li	VLM	Yes	No	MIT
PyMuPDF	Artifex Software	Pipeline	Yes	No	AGPL-3
pypdf	Martin Thoma	Pipeline	Yes	No	BSD-3
Marker	Datalab	Pipeline	Yes	Yes*	GPL-3**
Markitdown	Microsoft	Pipeline	Yes	Yes	MIT
MinerU	OpenDataLab	Pipeline	Yes	No	AGPL-3
Zerex OCR	OmniAI	VLM	Yes	Yes*	MIT

Table 1: Comparison of document conversion libraries. *Marker and Zerex OCR convert PPTX to PDF at the start of the conversion. **model weights for Marker are licensed cc-by-nc-sa-4.0.

This study compares the following four libraries that support both PDF and PPTX in their current version:

- *Docling* by Livathinos et al. (2025) at IBM. Version 2.38.0 was used (released 2025-06-23).
- *Markitdown* by Microsoft (2024). Version 0.1.2 was used (released 2025-05-28).
- *Marker* by Datalab (2024). Version 1.7.5 was used (released 2025-06-11).
- *Zerex OCR* by OmniAI (2024). Version 0.1.06 was used (released 2024-12-18).

All libraries were run with support from a VLM to caption images. The captions were placed into the Markdown documents at the location of the images.

3.7 Conversion library limitations

Missing image captions: Docling does not support image captions for images in PPTX files. Some images in PDF files are not captioned. PDF files converted to Markdown by Markitdown exhibit duplicate text and missing image captions. Markitdown does not support image captions for images that are in PDF format when they are embedded in PPTX files.

Conversion errors: Marker and the underlying library pillow were unable to handle Windows Metafile (WMF) files, a legacy format encountered in 15 benchmarked documents, on the Debian Linux system used for this study. Markitdown had an issue with 17 PPTX files that contain shapes with **None** as top, left, height or width attributes. These combinations of libraries and documents were excluded from further analysis.

3.8 Models

The following LLMs are compared as vision models and as question answering models:

- *Llama4 Maverick* by Meta ([Meta AI 2025](#)), hosted on the Fireworks AI API
- *Qwen2.5-32B Instruct* by Alibaba ([A. Yang et al. 2025](#)), hosted on the Fireworks AI API
- *gpt-4.1-mini-2025-04-14* by OpenAI ([OpenAI 2024](#)), hosted on the OpenAI API

All of these models have a sufficiently long context length to fit the longest document in Markdown form. Prompts, temperature, and other token generation parameters are kept constant. They are documented in the appendix.

The LLM-as-judge method ([Zheng et al. 2023](#)) is employed to evaluate the correctness of the answers. In contrast to word-based evaluations like BLEU, ROUGE, and METEOR, the LLM-as-judge method is able to semantically evaluate the answers. It is not limited to lexical matches. Further, it can be instructed on specific judging criteria. A reasoning model, *o4-mini-2025-04-16*, carries out the judging in a true-false manner. We provide the prompt in the appendix. The nature of the questions leaves little room for interpretation. The sum of correct answers divided by the total number of questions is the accuracy, the primary metric of this study.

3.9 Configurations

The experiment has a factorial design, combining the four libraries, each of the three models as vision model for image captioning and as the question answering model. At the time of writing the Zerox OCR library was not compatible with Llama4 or Qwen2.5 on hosted inference APIs, therefore only gpt-4.1-mini was used with it. When Zerox OCR converts PPTX to Markdown, it starts by converting the PPTX to PDF. This internal conversion is not the subject of this study, so Zerox is only used with PDF files. In total, 30 configurations were tested.

4 Results

4.1 Layout elements and file formats

Figure 3 shows the accuracy by file type and layout element. Table 2 shows the same results split by conversion library. In order of overall accuracy, the layout elements are:

1. **Text** is accurately converted by all libraries and with both file types, scoring above 90% accuracy. That makes text the most reliable layout element for question answering.
2. **Tables** are accurately converted in all but one configuration, with most showing accuracy above 90%. The exception is Markdown on PDF files, which does not convert tables to Markdown tables. Instead, table cells are extracted as text snippets, losing their structural relationships.

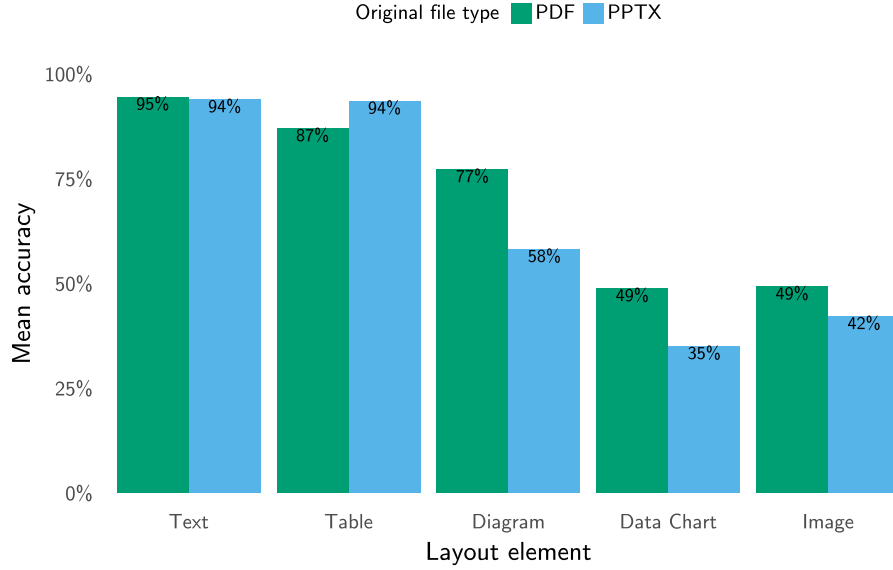


Figure 3: Answer accuracy in percent by file type and layout element. Results are averaged across conversion libraries, question answering model, and vision model. Conversions that resulted in errors are excluded.

3. **Diagrams** have lower accuracy than text and tables, with all scores below 90%. Zerox OCR achieves 88.5% on PDF files, likely due to its use of box-drawing characters. Markdown reaches 80.2% on PPTX files. Other libraries and file type combinations score below 80%.
4. **Data charts** display high variance of accuracy across libraries and file types. Markdown excels at this task in PPTX files, scoring the highest accuracy of 70.1%. Accuracy in other configurations is low.
5. **Images** show high accuracy variance. Docling cannot caption images in PPTX files, and Markdown cannot in PDF files. Docling achieves a comparatively high accuracy of 72.5% on images in PDF files, more than 10 percentage points higher than others using the same vision models. This may be due to its option for a custom image prompt, which is not available in the other libraries. The image prompt is in the appendix.

We do not find support for our hypothesis that PPTX files with their structured XML format are generally better suited for document understanding than PDF files. As an external reference to the statistics reported here, FinanceBench’s (Islam et al. 2023) best long-context configuration with GPT4-Turbo reported 79% accuracy on questions pertaining to financial PDF documents.

Library	File type	Layout element				
		Text	Table	Diagram	Data Chart	Image
Docling	PDF	92.9	90.4	74.3	57.1	72.5
Docling	PPTX	94.4	94.5	60.6	16.7	9.7
Marker	PDF	95.8	91.2	73.5	52.9	63.8
Marker	PPTX	93.4	94.2	54.0	18.2	57.7
Markdown	PDF	94.4	77.0	80.2	32.7	8.4
Markdown	PPTX	94.0	91.9	59.9	70.1	59.0
Zerox OCR	PDF	96.1	94.2	88.5	60.0	58.1

Table 2: Answer accuracy in percent by conversion library, file type and layout element. Results are averaged across question answering model and vision model.

Question answering model	Vision language model		
	Llama4 Maverick	Qwen2.5 32B	GPT 4.1 Mini
Llama4 Maverick	67.7	67.0	72.3
Qwen2.5 32B	63.9	61.7	69.6
GPT 4.1 Mini	67.0	65.9	72.4

Table 3: Answer accuracy in percent by QA model and VLM. Results are averaged across layout elements, file types, and conversion libraries. Runs with Zerox OCR were excluded, as it is incompatible with Llama4 and Qwen2.5 on hosted inference APIs.

4.2 Image captioning and question answering models

The experiment also involved variations in the used for image captioning (or alternatively the whole conversion in the case of Zerox OCR) and for answering questions based on the converted document. Table 3 displays accuracy by combination of models. Out of the three models, GPT 4.1 Mini posts the highest accuracy scores as a vision model. As question answering models, Llama4 Maverick and GPT 4.1 Mini reach similar accuracy. The differences illustrate that every single piece of the pipeline (file type, layout element, vision model, question answering model, conversion library) influences the accuracy.

5 Discussion

Findings indicate that the classic formats PPTX and PDF, while valuable for human interpretation, are not optimal for RAG systems. In some configurations of libraries, files types and layout elements, answer accuracy was below 30%. This is a warning sign for buyers and sellers of reports about market research and other topics. It is in the interest of both parties to ensure that information is passed correctly into a KMS. There are two main paths for this:

1. Coordinate on a file format and layout elements. Ensure that the buyer is using a conversion library that is well equipped for the file format and the layout elements used in the seller’s reports. Adjust the reports as necessary to ensure proper comprehension and work around edge cases.
2. Negotiate a special-purpose deliverable to be sent alongside the main report. More on this in Section 5.1.

A majority of the tested libraries (Docling, Marker, and Zerox OCR) prioritize PDF conversion; only Markdown emphasizes PPTX and other Microsoft Office formats. This mirrors the relative abundance of PDF files available on the web, relative to PPTX files. With current tools, PPTX is less practical than PDF for conversion, despite its foundation in structured XML.

The tested libraries were not all able to convert each document and layout element therein. We encountered errors as well as silent failures. This reflects the libraries’ recency and rapid development. Zerox OCR with its pure vision approach proved to be the most robust. This success mirrors the results of the table extraction experiment by Hui, Lu, and Zhang (2024).

5.1 Special-purpose deliverables

Instead of working around the limitations of formats made for human consumption, a secondary deliverable optimized for AI could be the solution. Text-heavy reports offer greater compatibility than styled PPTX and PDF files, yet are not optimal either, for example when methodology is presented separately from results. Our study suggests that transitioning to complementary special-purpose deliverables, designed explicitly for AI, enhances the retrieval accuracy of research insights within KMS, and thus for the client. Detailed specification of these special-purpose deliverables is left to future work, but we can already outline relevant design principles:

1. **Text-first representation:** Graphical elements should include comprehensive textual descriptions capturing both explicit content and implicit relationships, including key numbers and trend interpretations.
2. **Structured architecture:** Implement consistent hierarchical organization with explicit section boundaries. Each section should be self-contained for effective chunking.
3. **Speaker notes inclusion:** Embed speaker notes and supplementary explanatory content directly within the document to provide context that aids interpretation.

Documents that follow the Web Content Accessibility Guidelines (World Wide Web Consortium 2024) would be well-suited not just for users with disabilities, but also for KMS.

A Markdown version of a document is a candidate for a special-purpose deliverable, given the format’s ubiquitous use with LLMs and support from multiple conversion libraries. Markdown is a simple, token-efficient plain-text format that supports formatting options like image captions. The Markdown

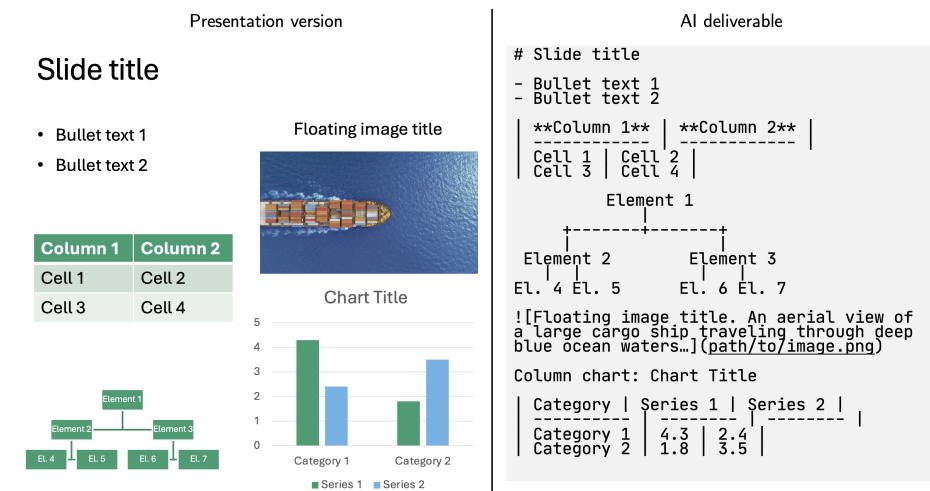


Figure 4: An idealized Markdown representation inspired by Zerox OCR. On the left is a PowerPoint slide, on the right the Markdown representation as an AI-readable deliverable. The slide features five layout elements: text, table, diagram, data chart, and image. Each element is also represented in Markdown. The data chart is converted to a table of numbers. The diagram is represented using ASCII characters. The content of the image is described in a caption.

files produced by Zerox OCR push this format’s capabilities by transcribing diagrams into ASCII characters, an approach that achieved the highest accuracy for diagram elements in our benchmark. However, this method has its limitations: the transcription is not standardized and may be inaccurate or infeasible for complex diagrams. Markdown also generally lacks a structured way to represent the spatial layout of elements, speaker notes, or color-coding from tables and charts. Figure 4 illustrates an idealized output, showing how layout elements from a slide can be represented in Markdown.

5.2 Further research

Continued development of VLMs, layout models, and pipelines will advance both the pipeline and pure vision approach. We expect that specific feature gaps in pipeline libraries (like Docling’s PPTX image captioning) will be closed and bugs fixed.

Further end-to-end benchmarks and case studies with a focus on specific layout elements and file formats are useful in this applied field. They complement the existing QA and DLA benchmarks and provide a comprehensive practitioner’s view. Besides the technical side of document understanding, there are organizational, cultural and skill-related aspects to consider in the context of report preparation.

References

- Cambazoglu, B Barla, Mark Sanderson, Falk Scholer, and Bruce Croft. 2021. “A Review of Public Datasets in Question Answering Research.” In *ACM SIGIR Forum*, 54:1–23. 2. ACM New York, NY, USA.
- Chen, Wenhui, Hanwen Zha, Zhiyu Chen, Wenhan Xiong, Hong Wang, and William Yang Wang. 2020. “HybridQA: A Dataset of Multi-Hop Question Answering over Tabular and Textual Data.” In *Findings of the Association for Computational Linguistics: EMNLP 2020*, edited by Trevor Cohn, Yulan He, and Yang Liu, 1026–36. Online: Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.findings-emnlp.91>.
- Chen, Zhongpu, Yinfeng Liu, Long Shi, Zhi-Jie Wang, Xingyan Chen, Yu Zhao, and Fuji Ren. 2025. “MDEval: Evaluating and Enhancing Markdown Awareness in Large Language Models.” arXiv. <https://doi.org/10.48550/arXiv.2501.15000>.
- Cheng, Hiuyi, Peirong Zhang, Sihang Wu, Jiaxin Zhang, Qiyuan Zhu, Zecheng Xie, Jing Li, Kai Ding, and Lianwen Jin. 2023. “M6Doc: A Large-Scale Multi-Format, Multi-Type, Multi-Layout, Multi-Language, Multi-Annotation Category Dataset for Modern Document Layout Analysis.” In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15138–47.
- Cheng, Mingyue, Yucong Luo, Jie Ouyang, Qi Liu, Huijie Liu, Li Li, Shuo Yu, et al. 2025. “A Survey on Knowledge-Oriented Retrieval-Augmented Generation.” <https://arxiv.org/abs/2503.10677>.
- Datalab. 2024. “Marker.” 2024. <https://github.com/datalab-to/marker>.
- Gao, Yunfan, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. “Retrieval-Augmented Generation for Large Language Models: A Survey.” <https://arxiv.org/abs/2312.10997>.
- Hui, Yulong, Yao Lu, and Huanchen Zhang. 2024. “UDA: A Benchmark Suite for Retrieval Augmented Generation in Real-World Document Analysis.” <https://arxiv.org/abs/2406.15187>.
- Islam, Pranab, Anand Kannappan, Douwe Kiela, Rebecca Qian, Nino Scherrer, and Bertie Vidgen. 2023. “FinanceBench: A New Benchmark for Financial Question Answering.” <https://arxiv.org/abs/2311.11944>.
- Livathinos, Nikolaos, Christoph Auer, Maksym Lysak, Ahmed Nassar, Michele Dolfi, Panos Vagenas, Cesar Berrospi Ramis, et al. 2025. “Docling: An Efficient Open-Source Toolkit for AI-Driven Document Conversion.” <https://arxiv.org/abs/2501.17887>.
- Masry, Ahmed, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. “ChartQA: A Benchmark for Question Answering about Charts with Visual and Logical Reasoning.” <https://arxiv.org/abs/2203.10244>.
- Meta AI. 2025. “Llama 4: Advancing Multimodal Intelligence.” 2025. <https://ai.meta.com/blog/llama-4-multimodal-intelligence/>.

- Methani, Nitesh, Pritha Ganguly, Mitesh M Khapra, and Pratyush Kumar. 2020. “Plotqa: Reasoning over Scientific Plots.” In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1527–36.
- Microsoft. 2024. “Markitdown.” <https://github.com/microsoft/markitdown>.
- OmniAI. 2024. “Zerex.” 2024. <https://github.com/getomni-ai/zerex>.
- OmniAI. 2025. “Benchmarking Open-Source Models for OCR.” <https://getomni.ai/blog/benchmarking-open-source-models-for-ocr>.
- OpenAI. 2024. “GPT-4o System Card.” <https://arxiv.org/abs/2410.21276>.
- Ouyang, Linke, Yuan Qu, Hongbin Zhou, Jiawei Zhu, Rui Zhang, Qunshu Lin, Bin Wang, et al. 2025. “OmniDocBench: Benchmarking Diverse PDF Document Parsing with Comprehensive Annotations.” In, 24838–48. https://openaccess.thecvf.com/content/CVPR2025/html/Ouyang_OmniDocBench_Benchmarking_Diverse_PDF_Document_Parsing_with_Comprehensive_Annotations_CVPR_2025_paper.html.
- Pfritzmam, Birgit, Christoph Auer, Michele Dolfi, Ahmed S Nassar, and Peter Staar. 2022. “Doclaynet: A Large Human-Annotated Dataset for Document-Layout Segmentation.” In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 3743–51.
- PricewaterhouseCoopers. 2024. “Case Study: PwC Entwickelt KI-Anwendung Für Wissensmanagement Beim ITZBund.” <https://www.pwc.de/de/branchen-und-markte/oeffentlicher-sektor/case-study-ki-anwendung-fuer-wissensmanagement-beim-itzbund.html>.
- Rogers, Anna, Matt Gardner, and Isabelle Augenstein. 2023. “QA Dataset Explosion: A Taxonomy of NLP Resources for Question Answering and Reading Comprehension.” *ACM Computing Surveys* 55 (10): 1–45. <https://doi.org/10.1145/3560260>.
- Smock, Brandon, Rohith Pesala, and Robin Abraham. 2022. “PubTables-1M: Towards Comprehensive Table Extraction from Unstructured Documents.” In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4634–42.
- Wang, Zhen. 2022. “Modern Question Answering Datasets and Benchmarks: A Survey.” <https://arxiv.org/abs/2206.15030>.
- World Wide Web Consortium. 2024. “Web Content Accessibility Guidelines (WCAG) 2.2.” W3C Recommendation. <https://www.w3.org/TR/WCAG22/>.
- Yang, An, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, et al. 2025. “Qwen2.5 Technical Report.” arXiv. <https://doi.org/10.48550/arXiv.2412.15115>.
- Yang, Zhilin, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. “HotpotQA: A Dataset for Diverse, Explainable Multi-Hop Question Answering.” *CoRR* abs/1809.09600. <http://arxiv.org/abs/1809.09600>.
- Zheng, Lianmin, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, et al. 2023. “Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena.” <https://arxiv.org/abs/2306.05685>.

Appendix

Settings

The LLM calls for conversion and answering use a temperature of 0.0. The judge model does not support a temperature parameter. All other settings were left at their default values.

Prompts

Image captioning

Describe the image in detail. Extract numbers, text and everything else required to answer questions about it. Do not use line breaks or other formatting.

This image captioning prompt is only used by Docling.

Question answering

You are a helpful assistant that answers questions based on provided context.

CONTEXT:
{report_content}

QUESTION:
{question}

Answer the question based solely on the provided context.
If the information isn't available in the context, say
"I don't have enough information to answer this question."

Judging

You are the judge for an AI system evaluation.

The AI was asked a question and provided an answer.
Your task is to check whether the answer matches the ground truth.

Correct answers:

- Contain all information from the ground truth
- May contain additional information, as long as it's not contradictory

Incorrect answers:

- Contradict the ground truth, even in parts
- Lack key information of the ground truth
- State that the question can't be answered

QUESTION:
{question}

GROUND TRUTH:
{ground_truth}

ANSWER:
{answer}