Tab 1

[author names]

Document parsing is crucial for natural language processing (NLP) tasks, especially in retrieval-augmented generation (RAG) tasks and pipelines. While various PDF parsing tools exist, visual-language models (VLMs) have emerged as a possible option for this task typically assigned to parsers. This study investigates hallucination in multimodal document questioning by comparing traditional RAG parsing pipelines with VLMs. Using a set of 10 biology papers containing >= 5 visual elements including figures, tables, and graphs, a total of 30 questions with ground truth answers were manually created, evaluating model reasoning, inference, and overall retrieval capabilities. Three Ragas evaluation metrics were applied: Answer Accuracy, Answer Correctness, and Factual Correctness. For the vision pipelines, GPT-4o achieved the highest accuracy (0.4167), followed by Qwen 2.5 (0.3750), and Gemini 2.5 Pro (0.358). The multimodal textual pipelines had Unstructured.io at the highest accuracy (0.358333), LlamaParse (0.3833), and Docling (0.366667). Optical Character Recognition (OCR) was enabled across all three textual pipelines. Although the visual models consistently outperformed parser + LLM pipelines, factual correctness remained low across all systems (<0.35), demonstrating the persistence of hallucination even with visual grounding. There were less discrepancies between correctness and factual scores, however. Future work will expand the experimental design by including additional question sets (e.g. MMLongBench) and by testing ablations that remove or alter visual content, references, and other document components to better understand the sources of hallucination.

Blitz Talk Transcript

“So this slide shows our project roadmap, which basically describes how we’ll evaluate hallucination and faithfulness in VLMs.

Step 1 is dataset selection. We started by choosing PubMed biomedical articles since they reflect real-world scientific documents — which is exactly where faithfulness matters most.

Step 2 is where we actually make our data. We manually scrape the documents, write questions, and create ground-truth answers that we can later use to check if the model is telling the truth or just hallucinating.

Step 3 is when we start building our system. So we adapt our Retrieval-Augmented Generation, or RAG, pipeline so it works with VLMs. The idea is to smartly chunk long scientific papers so we can pass in only the most relevant parts to the model.

This is where we are right now.

Moving on, Step 4 is our first real evaluation for faithfulness. Here, we insert these context blocks — these are basically key sentences the model should rely on. If the model is truly faithful, it should root its answers in these blocks instead of inventing new information, so that helps us measure the groundedness.

Step 5 turns up the difficulty — we change the document by removing or altering evidence and see how the model responds. If it starts hallucinating all of a sudden, we know it was depending on shallow signals instead of real understanding.

Finally, Step 6 is when we write everything up. This is where we would compare how VLMs compare against standard LLMs on this biomedical interpretation task, and hopefully get a manuscript out of it.