Weekly Report 1

September 6, 2025

Abstract

This report documents my process of familiarizing myself with the RAG-Anything project through empirical evaluation. I implemented an LLM-as-a-judge framework to assess the system's performance. Initial experiments with LLM-generated QA-pairs from a single PDF, which revealed a need for larger-scale testing. Consequently, I evaluated the system on 230 QA items across text, table, figure, and multimodal types from the MMDocIR benchmark[1], The results demonstrates the model's relative strength in textual tasks and potential challenges in complex multimodal reasoning, also revealing a need for new evaluation methods for multimodal RAG systems. This work serves as a foundation for understanding the project and guiding future improvements.

1 Experiments I

1.1 Evaluation Method

I first attempted to evaluate the model with LLM-generated QA-pairs on a single PDF file (the original LightRAG paper).

I modified `process_with_rag()` to extract text content and multimodal items, format context into a QA-pair generation prompt, and call the `llm_model_func()` and `vision_model_func()` from the original code to generate a QA-pair for each text-chunk or multimodal item.

The questions are scored by LLM from 1 to 5 in three dimensions, namely groundness, relevance and stand-alone. The QA-pairs with low-scoring questions are removed.

Then, RAG-Anything is queried with the questions and has the responses collected. The responses and the QA-pair are formatted into an evaluation prompt, utilizing another model to score the RAG response from 1 to 5.[2]

The prompts in this section are adapted from Hugging Face[3].

1.2 Results

Table 1 shows the performance of RAG-Anything on four types of QA-pairs. The "Threshold" in the table refers to the question quality threshold. For instance, when Threshold $= 2$, questions scored below 2 in any of the three dimensions are removed. When Threshold $= 1$, no questions are removed.

The results indicate that RAG-Anything may score higher with better QA-pairs. However, the number of multimodal questions from a single file is too small to be statistically meaningful. Even a minimal increase in the question quality threshold filters out most questions. This indicates that it is difficult to generate high-quality QA-pairs and that there is a need to evaluate the model on larger, pre-reviewed datasets.

[1] `https://huggingface.co/datasets/MMDocIR/MMDocIR_Evaluation_Dataset`

[2] Evaluation script: `https://github.com/rnluo/RAG-Anything/blob/main/examples/evaluation_on_single_file.py`

[3] `https://huggingface.co/learn/cookbook/en/rag_evaluation`

Type	Threshold = 1		Threshold = 2	
	Number of QA-pairs	Avg. Score	Number of QA-pairs	Avg. Score
Text	65	4.28	18	**4.39**
Table	6	4.00	0	N/A
Image	6	3.83	0	N/A
Equation	2	3.00	1	**5.00**

Table 1: Average LLM-judged score of RAG-Anything on four types of LLM-generated multimodal QA-pairs with two question quality thresholds.

2 Experiments II

2.1 Evaluation Method

I utilized a subset of the evaluation dataset from MMDocIR(1), a benchmark for multimodal retrieval. The dataset consists of document page screenshots annotated with questions, answers, modality type of the question, and bounding box coordinates of ground truth evidence on pages, etc.

I extracted the first 100 "Pure-text (Plain-text)" questions, 100 "Chart"/"Table"/"Figure" questions, and 30 "multimodal-t"/"multimodal-f" questions. For text questions, I save the provided OCR text into temporary files. For multimodal questions, I crop relavant screenshots with provided bounding box coordinates and save them as temporary files. The temporary files are added into the knowledge base of RAG-Anything, which is queried by corresponding questions. The RAG-responses are scored from 1 to 5 with the same evaluation prompt in section 1. [4]

Since the QA items are based on independent contexts, RAG-Anything must re-initialize and rebuild the knowledge base for every question, which is computationally expensive and time-consuming. This reveals the need for a specially curated dataset in the form of a single, long, multimodal-item-rich document with QA-pairs.

2.2 Results

Table 2.2 shows the performance of RAG-Anything on various types of QA items.

Type in original dataset	Number of QA items	Avg. Score
"Pure-text (Plain-text)"	100	3.33
"Chart"	33	2.42
"Table"	26	3.08
"Figure"	41	2.17
"multimodal-t"	25	2.92
"multimodal-f"	5	2.20

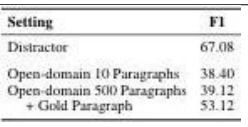
Table 2: Average LLM-judged score of RAG-Anything on various types of QA items from the MMDocIR_Evaluation_Dataset.

Pure-text questions The score of 3.33 is lower than the previous results with LLM-generated QA-pairs. This may result from a discrepancy between the model and the dataset, since RAG-Anything returns long and comprehensive responses and the factual answers in the dataset are much more concise. It reveals a need for datasets specially constructed for RAG systems.

Multimodal questions The model scores 3.08 in "Table" questions, significantly higher than in "Chart" and "Figure" questions. This may indicate that the model benefits from the provided table processor.

[4]Evaluation script: `https://github.com/rnluo/RAG-Anything/blob/main/examples/evaluation_on_dataset.py`

Table 3: Case Study: One successful case with text and one failed case with multimodal context.

Type	"Pure-text (Plain-text)"
Context	*[OCR text...]*
Question	What was the value of absolute percentage shortfall in India's GDP growth from 2002 to 2003 compared to the forecast?
Answer	2.4%
RAG-Anything	The absolute percentage shortfall in India's GDP growth from 2002 to 2003 compared to the forecast can be calculated as follows: - Forecast GDP growth: 6.7% - Actual GDP growth: 4.3% The absolute percentage shortfall is calculated as: $$\text{Shortfall} = \text{Forecast} - \text{Actual} = 6.7\% - 4.3\% = 2.4\%$$ Thus, the absolute percentage shortfall in India's GDP growth was **2.4%** from 2002 to 2003. **References** [DC] tmpji3zt16n.txt
Type	"multimodal-t"
Context	 **Setting** / **F1** Distractor 67.08 Open-domain 10 Paragraphs 38.40 Open-domain 500 Paragraphs 39.12 + Gold Paragraph 53.12
Question	By how much does the F1 score improve when the gold paragraph is added to the open-domain setting with 500 paragraphs?
Answer	The F1 score improves by 14.00 when the gold paragraph is added.
RAG-Anything	**F1 Score Improvement Overview** Based on the information available, there are various numerical values provided in the context. The relevant values that seem to relate to performance measurements, possibly including the F1 score, are 38.40, 39.12, 53.12, and 67.08. These numbers might represent metrics used in evaluating an open-domain text composition task, such as the ones mentioned regarding different paragraph settings [KG] Open-domain 10 Paragraphs - Open-domain 500 Paragraphs + Gold Paragraph. Although the exact nature of the improvement in F1 score is not explicitly stated, we can infer from the context provided that an enhancement might be associated with the addition of the 'Gold Paragraph' in the open-domain settings involving 500 paragraphs. The exact degree of improvement, however, would require detailed data that is not present in the provided context. **Conclusion** It is not possible to definitively determine by how much the F1 score improves when the gold paragraph is added based on the information available as it lacks specific comparative details directly related to the F1 score. **References** - [KG] Open-domain 10 Paragraphs - Open-domain 500 Paragraphs + Gold Paragraph - [KG] 38.40 - 53.12 - [KG] 39.12 - 67.08 - [DC] tmpx9v9v4oi.png

There is another reason why the model performs worse on multimodal questions than on pure-text questions, illustrated in Table 3. The two cases both involve subtracting two figures to obtain the answer. While the model succeeds with clear reasoning in the pure-text case, it fails to comprehend the task in the multimodal case, even though the figures are correctly extracted. This may indicate that the model can be improved in complex multimodal reasoning, especially in the effective comprehension of short texts in multimodal items.

References

[1] Kuicai Dong, Yujing Chang, Xin Deik Goh, Dexun Li, Ruiming Tang, and Yong Liu. Mmdocir: Benchmarking multi-modal retrieval for long documents, 2025.