Achieving 2x Accuracy in Knowledge Retrieval from Charts and Tables without Intensive Prompt Engineering

CambioML & Epsilla

August 2, 2024

Contents

Abstract

Every day, AI applications extract data from millions of documents, and then mostly consumed by large language models (LLMs). However, traditional Optical Character Recognition (OCR) based models often struggle with retrieval precision and recall, especially when pulling structured data such as tables and charts, and they frequently miss crucial details in complex document formats. This whitepaper introduces a state-of-the-art knowledge retrieval system on both unstructured data and structured data, which can reach up to 2x precision and 2.5x recall compared with the normal Retrieval-Augmented Generation (RAG) system. The system has two main components: Any-Parser, powered by multi-modality models designed to accurately extract text, tables, and chart information from PDF, WORD, PPT, and images; Epsilla, a no-code RAG-as-a-Service platform for building production ready LLM applications grounded by private or public knowledge base.

1 Introduction

In today's data-driven world, accurate and efficient information extraction is crucial, particularly in industries like financial services where data is predominantly stored in both unstructured data (e.g., text) and structured data (e.g., tables, charts, etc.) formats. The demand for advanced data extraction and inference solutions has never been higher. Traditional OCR-based models can detect the text itself, but often struggle with pulling layout information or extracting tables and charts, and they frequently miss crucial details in diverse document formats. These limitations lead to suboptimal results when applying cutting-edge AI applications such as RAG.

This whitepaper introduces an innovative knowledge retrieval approach on both unstructured data and structured data: integrating state-of-the-art table extraction models with RAG techniques for document question answering, which achieves 50% to 150% higher evaluation scores including content precision, recall, faithfulness, and correctness than existing commonly used RAG systems.

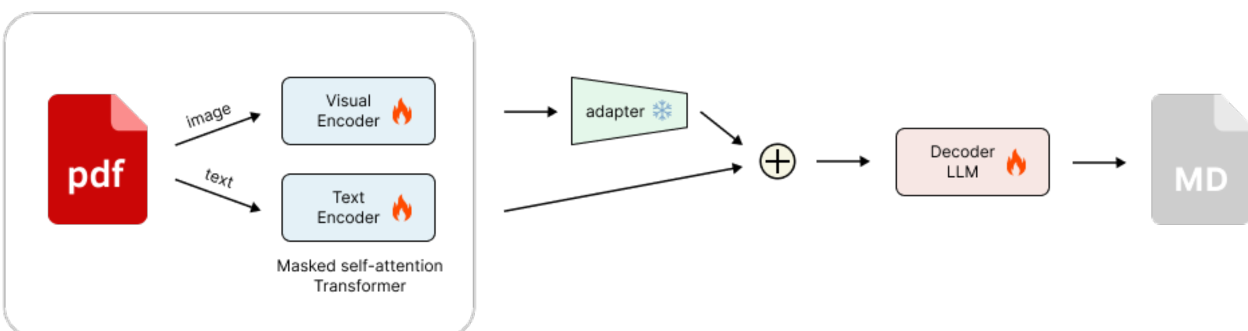


Figure 1: Extraction Model Architecture

2 Table Extraction Models

AnyParser is powered by specific-trained billions parameters vision language models (VLMs). As illustrated in Figure 1, it includes an visual encoder, a text encoder, an adapter, and a decoder-only LLM. AnyParser leverages advanced VLMs learning techniques to accurately identify and extract relevant information from diverse data sources. Traditional table extraction models typically rely on OCR followed by text generation. This approach has significant limitations, particularly its dependence on OCR accuracy and potential loss of structural information, which hampers the connection between the text and its surrounding context. Additionally, many existing VLMs only support input resolutions of 224 or 336 pixels. These resolutions are inadequate for the precision required in financial or medical document extraction, as they fail to capture tiny or small texts accurately. Our approach overcomes these limitations by integrating visual and text-based encoders, inspired by LLaVA, allowing for more comprehensive and precise data extraction. Our visual encoder is designed to handle higher resolutions, supporting input image resolutions up to 1344 pixels, ensuring that even the smallest text details are captured accurately, thus maintaining the integrity of the data structural and contextual information.

3 Retrieval-Augmented Generation (RAG) Pipeline

Epsilla is built upon a robust and flexible Retrieval-Augmented Generation (RAG) infrastructure designed to support and optimize various RAG pipelines, from simple implementations to complex, modular systems. This infrastructure addresses the primary challenges of RAG—indexing, retrieval, and response generation—through a series of sophisticated optimization strategies.

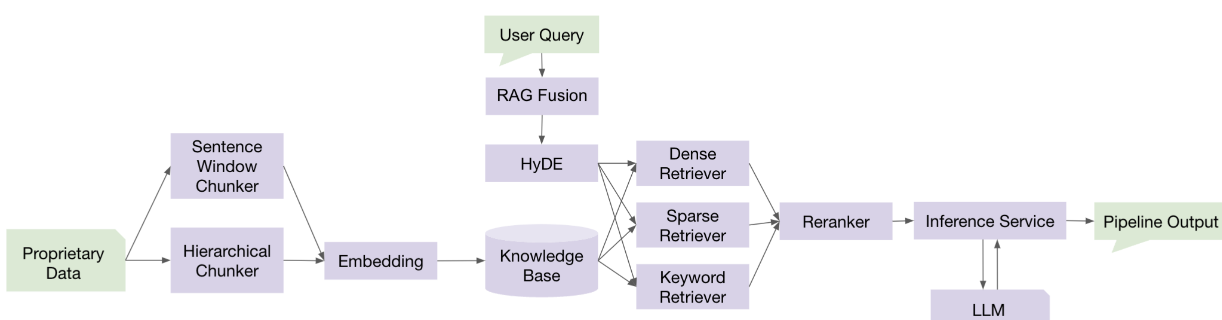


Figure 2: An Example Modular RAG pipeline

Epsilla employs advanced chunking and indexing strategies to enhance the efficiency and accuracy of knowledge retrieval. This includes semantic chunking, which breaks down large documents into semantically coherent chunks to ensure each segment is meaningful and contextually relevant. Additionally, hypothetical question generation anticipates potential user queries and generates corresponding hypothetical questions to improve the indexing process and ensure comprehensive coverage.

Epsilla refines user queries and enhances retrieval methods through hybrid search techniques, combining the strengths of keyword-based searches and semantic searches to deliver more accurate and relevant results. Query refinement dynamically adjusts and refines user queries to align better with the indexed content, ensuring precise and relevant document retrieval.

Once the relevant documents are retrieved, Epsilla focuses on further optimizing the content for response generation. Document reranking utilizes heuristic methods and advanced models to rerank documents based on their relevance to the user query. Prompt compression and context condensation summarize and condense the retrieved information to fit within the constraints of the model's input, ensuring the generated responses are both accurate and contextually rich.

In addition to the predefined modular RAG pipeline, Epsilla supports no-code agentic RAG and graph RAG, encapsulating advanced retrievers as tools for LLM use. Epsilla's micro-service RAG infrastructure not only supports plug-and-play integration of any advanced and modular RAG techniques but is also forward-compatible, allowing for the inclusion of any new RAG techniques developed in the future.

The Epsilla RAG as a Service platform delivers high-quality, context-aware responses while maintaining exceptional flexibility and scalability. By continuously refining each stage of the RAG pipeline, Epsilla ensures efficient and effective knowledge retrieval and generation, making it the ideal solution for diverse LLM applications across various industries.

4 Experiment & Evaluation

To evaluate the knowledge retrieval capability of AnyParser and Epsilla RAG system, we demonstrate a common use case in financial services: question answering (QA) on 10-K documents. A 10-K is an annual report filed by public companies in the U.S. to provide a comprehensive summary of their financial performance. 10-K documents usually have a richer trove of information such as tables, charts and figures. As illustrated in Figure 3, the experiment pipeline contains below components:

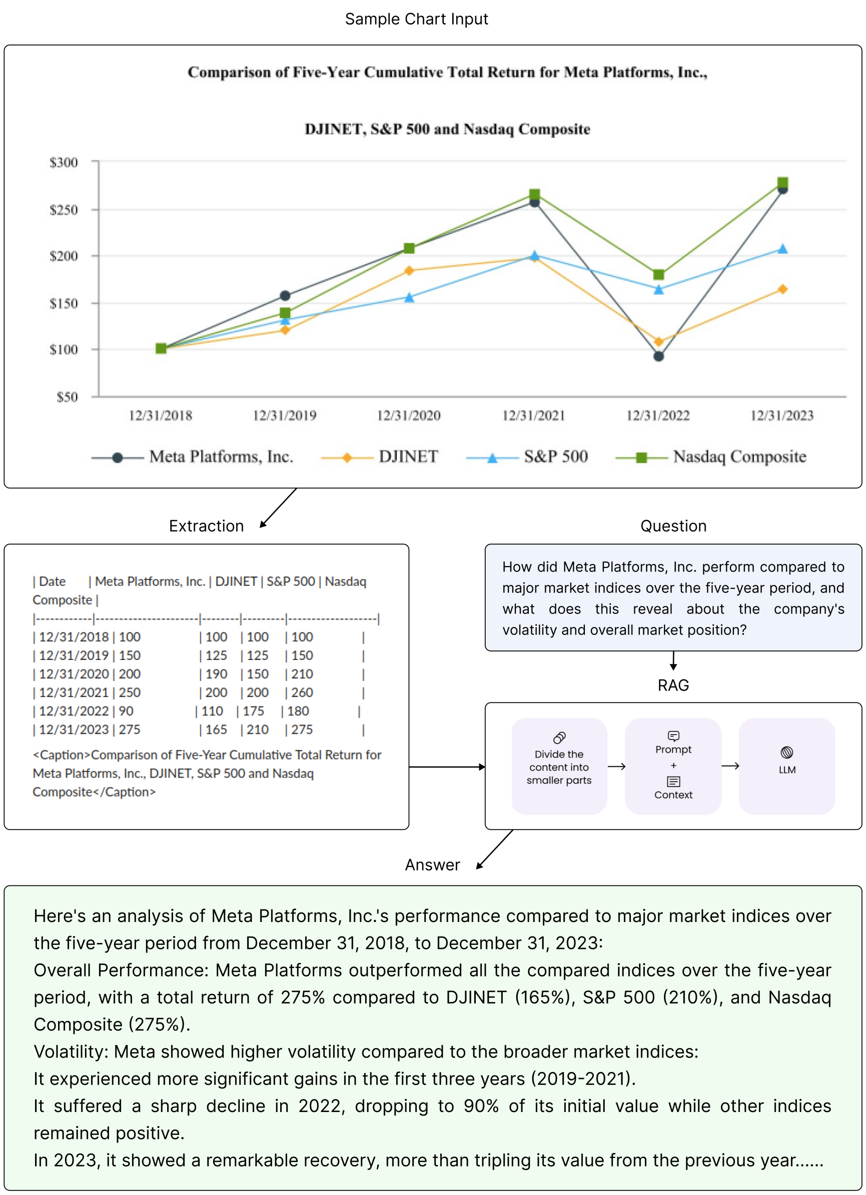
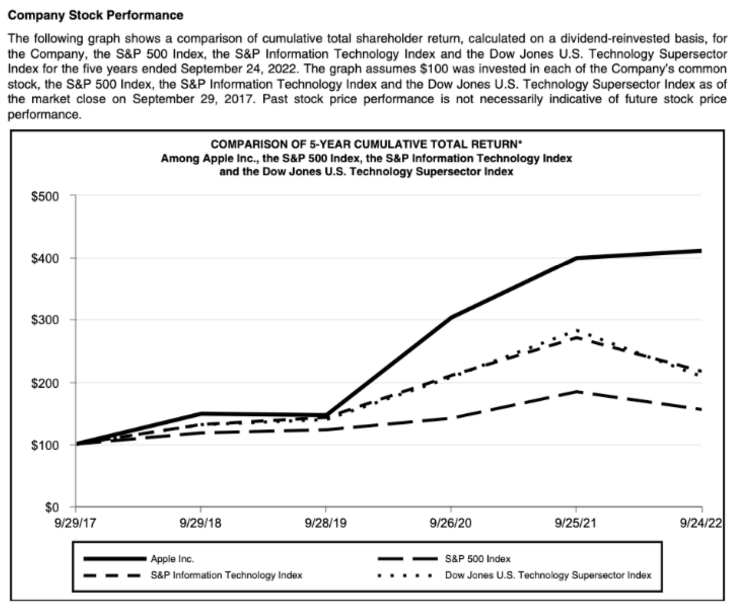


Figure 3: End-to-end example of AnyParser and Epsilla RAG system

In this experiment, we use the AAPL and META 10-K filings as examples to illustrate how to effectively retrieve insights from tables and charts using AnyParser and Epsilla.

1. **AnyParser Extraction**: AnyParser uses advanced VLM techniques to accurately identify and extract relevant information from a variety of data sources. For example, as illustrated at the top of Figure 3 above, AnyParser can analyze a line chart comparing the performance of Meta Platforms, Inc. with three major stock market indices. It systematically converts each data point on the chart into a table format, detailing year-end dates alongside the corresponding normalized values for Meta Platforms, Inc., DJINET, S&P 500, and Nasdaq Composite.

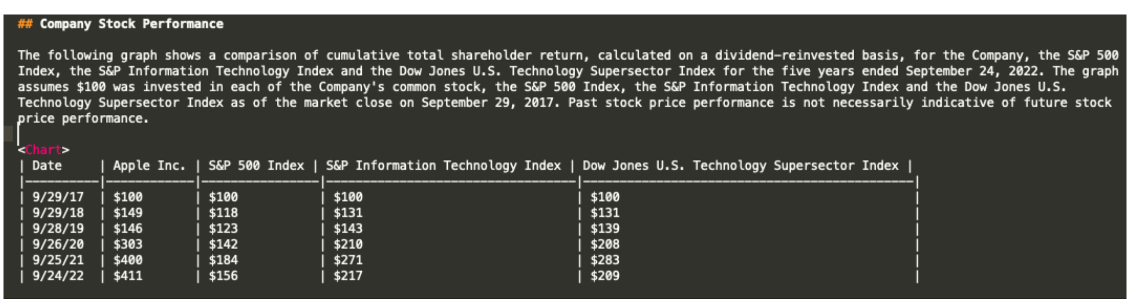


Company Stock Performance

The following graph shows a comparison of cumulative total shareholder return, calculated on a dividend-reinvested basis, for the Company, the S&P 500 Index, the S&P Information Technology Index and the Dow Jones U.S. Technology Supersector Index for the five years ended September 24, 2022. The graph assumes $100 was invested in each of the Company's common stock, the S&P 500 Index, the S&P Information Technology Index and the Dow Jones U.S. Technology Supersector Index as of the market close on September 29, 2017. Past stock price performance is not necessarily indicative of future stock price performance.

Figure 4: Input Document: Document with line chart

In another example involving Apple's 10-K document as shown in Figure 4, AnyParser transforms a line chart into a markdown table in Figure 5 without needing of any extensive prompt engineering.



```
## Company Stock Performance

The following graph shows a comparison of cumulative total shareholder return, calculated on a dividend-reinvested basis, for the Company, the S&P 500 Index, the S&P Information Technology Index and the Dow Jones U.S. Technology Supersector Index for the five years ended September 24, 2022. The graph assumes $100 was invested in each of the Company's common stock, the S&P 500 Index, the S&P Information Technology Index and the Dow Jones U.S. Technology Supersector Index as of the market close on September 29, 2017. Past stock price performance is not necessarily indicative of future stock price performance.

<Chart>
| Date     | Apple Inc. | S&P 500 Index | S&P Information Technology Index | Dow Jones U.S. Technology Supersector Index |
|----------|------------|---------------|----------------------------------|---------------------------------------------|
| 9/29/17  | $100       | $100          | $100                             | $100                                        |
| 9/29/18  | $149       | $118          | $131                             | $131                                        |
| 9/28/19  | $146       | $123          | $143                             | $139                                        |
| 9/26/20  | $303       | $142          | $210                             | $208                                        |
| 9/25/21  | $400       | $184          | $271                             | $283                                        |
| 9/24/22  | $411       | $156          | $217                             | $209                                        |
```

Figure 5: Output Markdown: Extracted table from line chart

2. **Epsilla Knowledge Retrieval**: once all the data got extracted, we index them using Epsilla RAG services and tested with 20 manually crafted questions by experienced financial analysts, covering a variety of structured and unstructured knowledge form the given 10-K.

   Example question: How did Meta Platforms, Inc. perform compared to major market indices over the five-year period, and what does this reveal about the company's volatility and overall market position?

   Epsilla accurately retrieves the document chunk with the Stock Performance Graph:

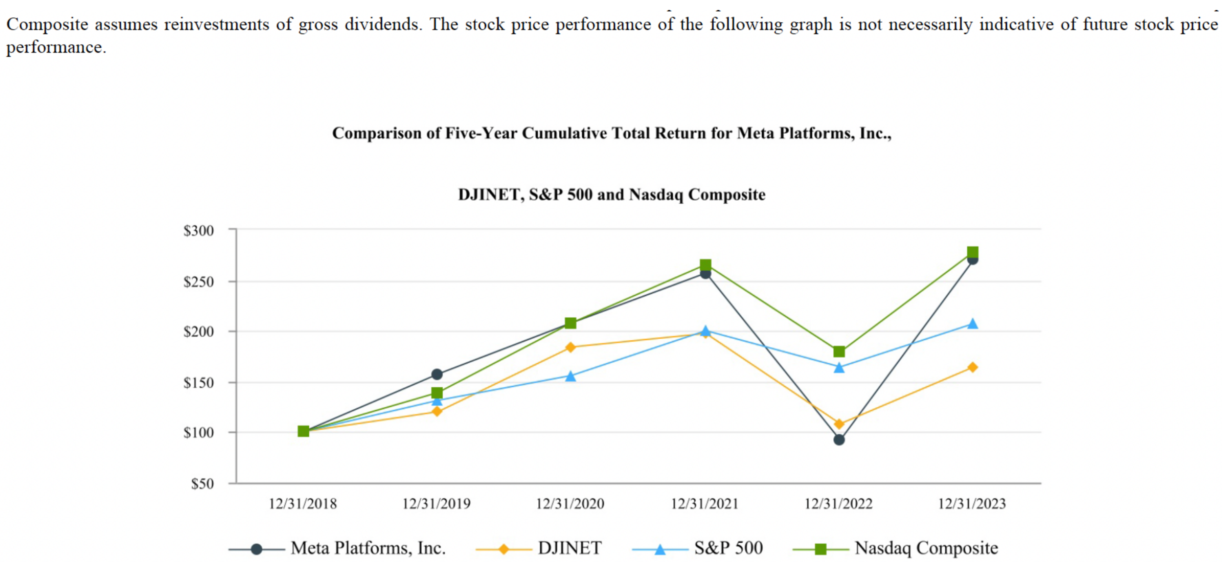


Figure 6: Meta Stock Performance Graph

```
The stock price performance of the following graph is not necessarily
indicative of future stock price performance.

| Date       | Meta Platforms, Inc. | DJINET | S&P 500 | Nasdaq Composite |
|------------|----------------------|--------|---------|------------------|
| 12/31/2018 | 100                  | 100    | 100     | 100              |
| 12/31/2019 | 150                  | 125    | 125     | 150              |
| 12/31/2020 | 200                  | 190    | 150     | 210              |
| 12/31/2021 | 250                  | 200    | 200     | 260              |
| 12/31/2022 | 90                   | 110    | 175     | 180              |
| 12/31/2023 | 275                  | 165    | 210     | 275              |

<Caption>Comparison of Five-Year Cumulative Total Return for Meta Platforms,
Inc., DJINET, S&P 500 and Nasdaq Composite</Caption>
```

Figure 7: Extracted table from Meta Stock Performance Graph

Here is the final answer provided by Epsilla RAG pipeline:

```
Over the five-year period ending December 31, 2023, Meta Platforms,
     Inc. (META) showed significant fluctuations in its stock price
```

compared to major market indices such as the DJINET, S\&P 500, and Nasdaq Composite. Here is a summary of the performance:

- Meta Platforms, Inc.: Started at 100 in 2018, peaked at 250 in 2021, dropped to 90 in 2022, and then surged to 275 in 2023.

- DJINET: Started at 100 in 2018, reached 200 in 2021, dropped to 110 in 2022, and ended at 165 in 2023.

- S\&P 500: Started at 100 in 2018, reached 200 in 2021, dropped to 175 in 2022, and ended at 210 in 2023.

- Nasdaq Composite: Started at 100 in 2018, reached 260 in 2021, dropped to 180 in 2022, and ended at 275 in 2023.

This data reveals that Meta Platforms, Inc. experienced higher volatility compared to the major indices. The significant drop in 2022 followed by a substantial recovery in 2023 indicates that while Meta's stock can be highly volatile, it also has the potential for strong rebounds. This volatility could be attributed to various factors including market conditions, company-specific news, and broader economic trends.

Overall, Meta's performance over the five-year period shows that it can outperform major indices during periods of growth but also experience sharper declines during downturns, reflecting a higher risk and potentially higher reward profile.

4.1 Evaluation Metrics

To rigorously assess the performance of our system, we utilized the ragas scoring system to evaluate both the generation and retrieval aspects of our system. This evaluation framework provided us with four key metrics that offer insights into different facets of our system's performance for RAG, plus the final answer correctness assessment.

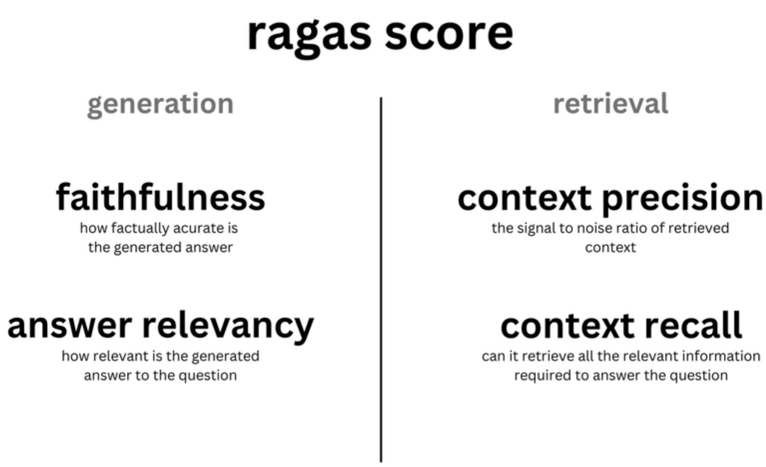


Figure 8: Evaluation Metrics from Ragas

4.1.1 Context Precision

Context Precision is a metric that evaluates whether all of the ground truth relevant items present in the `contexts` are ranked higher or not. Ideally, all the relevant chunks must appear at the top ranks. This metric is computed using the `question`, `ground_truth`, and the `contexts`, with values ranging between 0 and 1, where higher scores indicate better precision.

$$\text{Context Precision@K} = \frac{\sum_{k=1}^{K}(\text{Precision@k} \times v_k)}{\text{Total number of relevant items in the top } K \text{ results}} \tag{1}$$

$$\text{Precision@k} = \frac{\text{true positives@k}}{(\text{true positives@k} + \text{false positives@k})} \tag{2}$$

Where K is the total number of chunks in `contexts` and $v_k \in \{0, 1\}$ is the relevance indicator at rank k.

4.1.2 Context Recall

Context recall measures the extent to which the retrieved context aligns with the annotated answer, treated as the ground truth. It is computed based on the `ground truth` and the `retrieved context`, and the values range between 0 and 1, with higher values indicating better performance. To estimate context recall from the ground truth answer, each sentence in the ground truth answer is analyzed to determine whether it can be attributed to the retrieved context or not. In an ideal scenario, all sentences in the ground truth answer should be attributable to the retrieved context. The formula for calculating context recall is as follows:

$$\text{context recall} = \frac{|\text{GT sentences that can be attributed to context}|}{|\text{Number of sentences in GT}|} \tag{3}$$

4.1.3 Faithfulness

This measures the factual consistency of the generated answer against the given context. It is calculated from the answer and retrieved context. The score is scaled to the (0,1) range, with higher values indicating better performance. The generated answer is regarded as faithful if all the claims made in the answer can be inferred from the given context. To calculate this, a set of claims from the generated answer is first identified. Then, each of these claims is cross-checked with the given context to determine if it can be inferred from the given context or not. The faithfulness score is given by the following formula:

$$\text{Faithfulness score} = \frac{\#\text{ of claims in the generated answer that can be inferred from given context}}{\#\text{ of claims in the generated answer}} \tag{4}$$

4.1.4 Answer Relevance

Answer Relevancy focuses on assessing how pertinent the generated answer is to the given prompt. A lower score is assigned to answers that are incomplete or contain redundant information, and higher scores indicate better relevancy. This metric is computed using the `question`, the `context`, and the `answer`. The Answer Relevancy is defined as the mean cosine similarity of the original `question` to a number of artificial questions, which were generated based on the `answer`:

$$\text{answer relevancy} = \frac{1}{N}\sum_{i=1}^{N}\cos(E_{g_i}, E_o) \tag{5}$$

$$\text{answer relevancy} = \frac{1}{N}\sum_{i=1}^{N}\frac{E_{g_i}\cdot E_o}{||E_{g_i}||\,||E_o||} \tag{6}$$

Where:

- E_{g_i} is the embedding of the generated question i.
- E_o is the embedding of the original question.
- N is the number of generated questions, which is 3 by default.

4.1.5 Answer Correctness

Answer Correctness involves gauging the accuracy of the generated answer when compared to the ground truth. This evaluation relies on the ground truth and the answer, with scores ranging from 0 to 1. A higher score indicates a closer alignment between the generated answer and the ground truth, signifying better correctness. Answer correctness encompasses two critical aspects: semantic similarity between the generated answer and the ground truth, as well as factual similarity. These aspects are combined using a weighted scheme to formulate the answer correctness score. Users also have the option to employ a 'threshold' value to round the resulting score to binary, if desired.

4.2 RAG Comparison

Using the metrics above, we compare the performance of our AnyParser + Epsilla RAG system against the a commonly used RAG system (baseline). Our system demonstrates significantly better results across all five metrics as shown in 9, highlighting the superiority of AnyParser + Epsilla state-of-the-art approach.

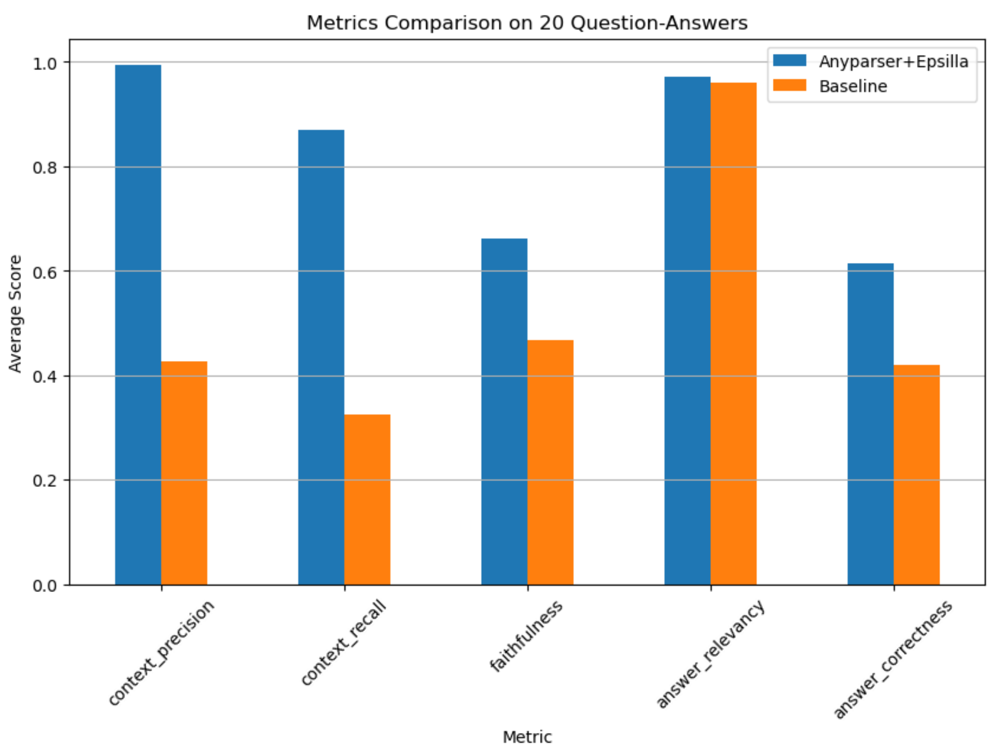


Figure 9: Evaluation Metrics from Ragas

To be specific,

- Context Precision: "AnyParser + Epsilla" scores 2 times higher (0.995) compared to "Baseline RAG" (0.425) on Context Precision. This indicates that "AnyParser + Epsilla" is much better at ranking relevant items higher in the context, ensuring that the most pertinent information appears at the top.

- Context Recall: "AnyParser + Epsilla" scores about 2.7 times higher (0.871) compared to "Baseline RAG" (0.325) on Context Recall, suggesting that "AnyParser + Epsilla" performs significantly better in terms of aligning the retrieved context with the ground truth annotations.
- Faithfulness: "AnyParser + Epsilla" has 1.4 times higher score (0.662) compared to "Baseline RAG" (0.468). This indicates that the answers generated by "AnyParser + Epsilla" are more factually consistent with the given contexts.
- Answer Relevance: "AnyParser + Epsilla" performs slightly better with a score of 0.973 versus 0.961 for "Baseline RAG". This metric suggests that "AnyParser + Epsilla" answers are more relevant and less redundant.
- Answer Correctness: "AnyParser + Epsilla" scores 0.614, and "Baseline RAG" scores 0.418. This metric evaluates the accuracy of the generated answers compared to the ground truth, and "AnyParser + Epsilla" demonstrates a 1.5 times performance in providing semantically and factually correct answers.

The results indicate that AnyParser along with Epsilla knowledge retrieval system excels in all evaluation metrics, achieving superior academic performance and practical effectiveness. These findings validate the robustness and efficiency of our approach in comparison to existing baseline RAG systems.

5 Use Case

Beyond building a question answering knowledge retrieval system using AnyParser and Epsilla system, we also list the other common use cases and key benefits below.

5.1 Common Use Cases

- **High-Resolution Knowledge Retrieval:** Extract text from image and PDF files, including smaller chunks or embedded pictures, with support for high-resolution images.
- **Diverse Document Extraction:** Enable text and tabular data extraction from a variety of documents, such as financial reports, medical documents, and research papers. Detect and extract tables and charts from documents while maintaining their structure and context.
- **Internal Workflow Integration:** Incorporate extraction capabilities into existing business workflows or AI applications, allowing for efficient processing of user-submitted data through forms.

5.2 Key Benefits

- **Accuracy:** The state-of-the-art AnyParser extraction models converts both structured and unstructured data into well-structured, usable formats with high precision.
- **Privacy:** Rather than general LLMs built by OpenAI or Anthropic, AnyParaser and Epsilla system can be easily deploying within a customers' data center, allowing full data security.
- **Scalability:** Rapidly process large volumes of documents in a few minutes, enabling faster decision-making and improved operational efficiency.

6 Conclusion

The AnyParser and Epsilla knowledge retrieval technology represents a significant advancement in RAG system. This technology not only significant improves content precision, recall, faithfulness and correctness, but also offers private host capability. As technology continues to evolve, the potential enterprise applications and benefits of this integrated solution are vast and promising.