**Hybrid Multi-Modal RAG Coding Challenge**

**Goal**

Build a Retrieval-Augmented Generation (RAG) pipeline that processes unstructured, multi-modal documents, including PDFs with German and English text, tables, invoices, PowerPoints, and more, to answer the questions provided in the Q&A excel.

Your solution should include a two-part agentic approach:

1. Agentic OCR Framework: An intelligent agent determines when OCR is required (e.g., for scanned or image-based documents) and when standard text extraction is sufficient.
2. Agentic RAG Pipeline: Orchestrate the full retrieval and answering workflow. The system should perform robust ETL, support multilingual queries, and include traces and RAG evaluation.

**Key Requirements**

* **Data Ingestion & OCR**Use at least one open-source framework like PyMuPDF and one commercial solution to extract text and tables from diverse PDFs. The OCR agent should efficiently detect and process only those pages or files that require OCR and VLMs, skipping or filtering irrelevant or already machine-readable content.
* **ETL Pipeline**Clean and structure the extracted data. Preserve key metadata, such as page numbers, for accurate source citation.
* **RAG Pipeline**
  + Index extracted content in a vector and relational database for hybrid search *(alternatively you might also use a VectorDB, which supports both sparse and dense embeddings like Qdrant or Postgres).*
  + Build a knowledge Graph and embed it into your retrieval system to capture relationships between your documents.
  + Implement a multi-modal retriever for text, tables, and (optionally) images.
  + Always cite the source.
* **Agentic Orchestration**
  + Use AutoGen and LangGraph (one for the data extraction in your ETL pipeline and one framework to coordinate filtering, retrieval and answering).
* **Evaluation**Integrate Arize Phoenix to trace and evaluate your system’s retrievals and answers. Measure these key metrics below based on the Q&A excel, expectation is for you to reach these scores and on all questions:
  + **Precision:** High precision (≥0.8)
  + **Recall:** High recall (≥0.8)
  + **Accuracy:** High accuracy (≥0.8)
  + **F1 Score:** F1 score of 0.85

Here you can find a list of [sample questions](https://aparavi-my.sharepoint.com/:x:/p/hendrik_krack/EWZdYRkLMh5BmCS1d8WxH7YB5SoKmo46aIDG7AV3HLo7EQ?e=RIyh5f) and their respective ground truth answers. You will be graded on how well your system answers the questions that have missing ground truth answers.

**Dataset**: [SampleDataSet](https://aparavi-my.sharepoint.com/:f:/p/hendrik_krack/EgLcDgP80PpJlH_p7J4ha5wBrB090tLq2DVaqX4VTBFo3A?e=Ciyrxn)

**Submission:**

* Link to a public GitHub repository within 24h after receiving this challenge
* Well designed architecture diagram
* Short, precise but comprehensive README.md file
* 5-10 min demo video of your end2end solution working and answering at least 2-3 questions from the excel