CAN IMAGES IMPROVE RAG? INSIGHTS FROM MULTIMODAL EXPERIMENTS IN INDUSTRY

Monica Riedler

2nd Bavaria, Advancements in SEarch Development (BASED) Meetup

12.06.2025

QUICK INTRO

- Hi, I'm Monica Riedler
- PhD student @ TU Munich
- Working on **Multimodal Learning** to bridge text, image and audio modalities
- Today: How **text + image models** can boost **search and QA systems**

Beyond Text: Optimizing RAG with Multimodal Inputs for Industrial Applications

Monica Riedler, Stefan Langer

SIEMENS

Paper

GitHub

OVERVIEW

1. Why care about multimodal RAG?
2. Unimodal vs. multimodal systems
3. How do we measure performance?
4. What did we learn?
5. What's next?

1. WHY CARE ABOUT MULTIMODAL RAG?

WHY CARE ABOUT MULTIMODAL RAG?

Challenges with LLMs:

- **Latest Knowledge:** LLMs are limited to their training data, recent events are not covered.
- **Domain-specific Knowledge:** specialized areas such as the **industrial sector** are often underrepresented in the training data.

Possible solution: **Retrieval Augmented Generation (RAG)**

- Use an external database that contains specific expertise to fill the information gap.
- Two-step process:
  - **Retrieval:** Identification and retrieval of documents relevant to the user query.
  - **Generation:** The user query and the relevant documents are passed to an LLM to generate an answer.

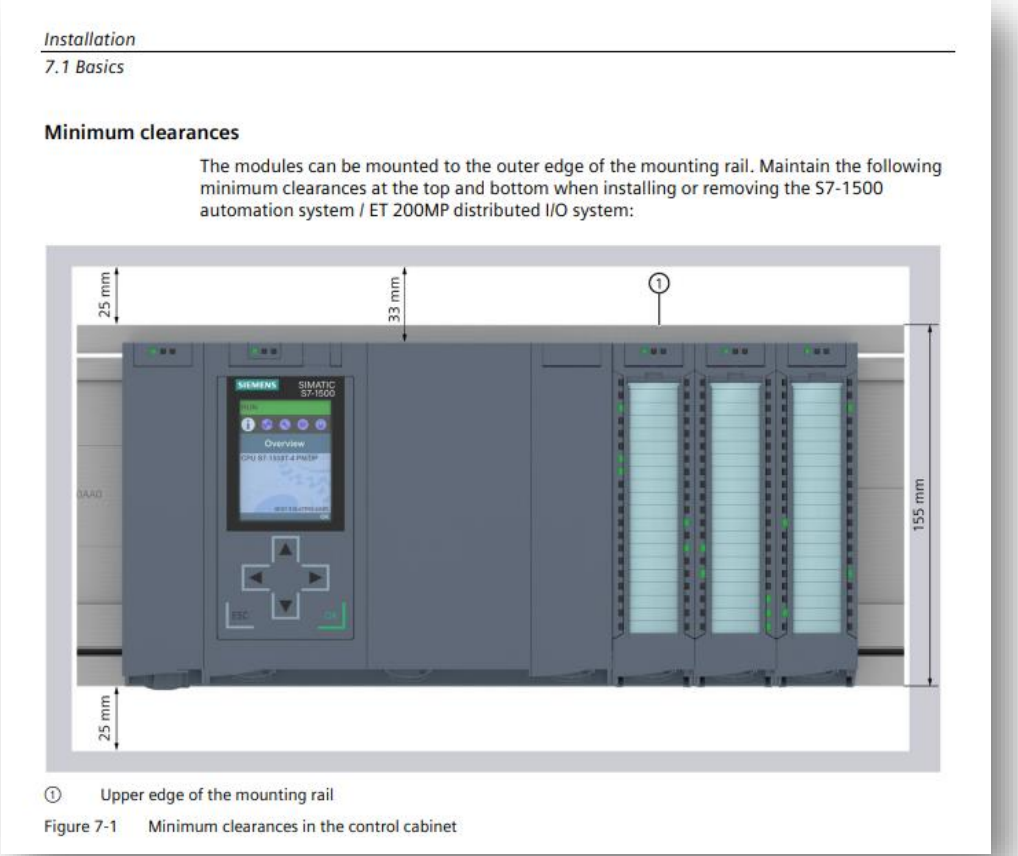
WHY CARE ABOUT MULTIMODAL RAG?

Why **Multimodal** RAG?

- Many documents contain both text and images.
- Sometimes text might not be enough.
- Sometimes a piece of information is contained only in an image or a combination of text and image.

- Example question:
  “What are the minimum clearances at the top and bottom when installing the S7-1500 automation system?”

- Correct answer:
  “The minimum clearances are 25 mm at the bottom, 25 mm at the top left and 33 mm at the top center”



Installation
7.1 Basics

Minimum clearances

The modules can be mounted to the outer edge of the mounting rail. Maintain the following minimum clearances at the top and bottom when installing or removing the S7-1500 automation system / ET 200MP distributed I/O system:

① Upper edge of the mounting rail

Figure 7-1 Minimum clearances in the control cabinet

2. UNIMODAL VS. MULTIMODAL SYSTEMS

Text-only RAG

Image-only RAG

Multimodal RAG

DATA & MODELS

- Data:
  - Document Collection: 8540 texts, 8377 images from 20 PDFs in the industrial sector with a focus on automation, software documentation and equipment manuals.
  - Test Set: 100 questions with annotated gold standard answers.

- Models:

Multimodal LLMs	Embedding Models
GPT-4Vision	text-embedding-3-small
LLaVA	CLIP

UNIMODAL VS MULTIMODAL RAG SYSTEMS

- **Upper and Lower Performance:**
  - **Baseline (Lower Bound):**
    - Directly prompting the LLM with the user query without additional context (no RAG).
  - **Gold Standard Context Prompting (Upper Bound):**
    - For the test set the correct context (text and/or image) to be retrieved from the document store is already known.
    - The correct context and the query are sent to the LLM, skipping the retrieval step.
    - Tests how well models perform in a setting with a perfect retrieval.
- **Text-Only RAG:**
  - Use only text from the document collection for retrieval, ignoring images.

UNIMODAL VS MULTIMODAL RAG SYSTEMS

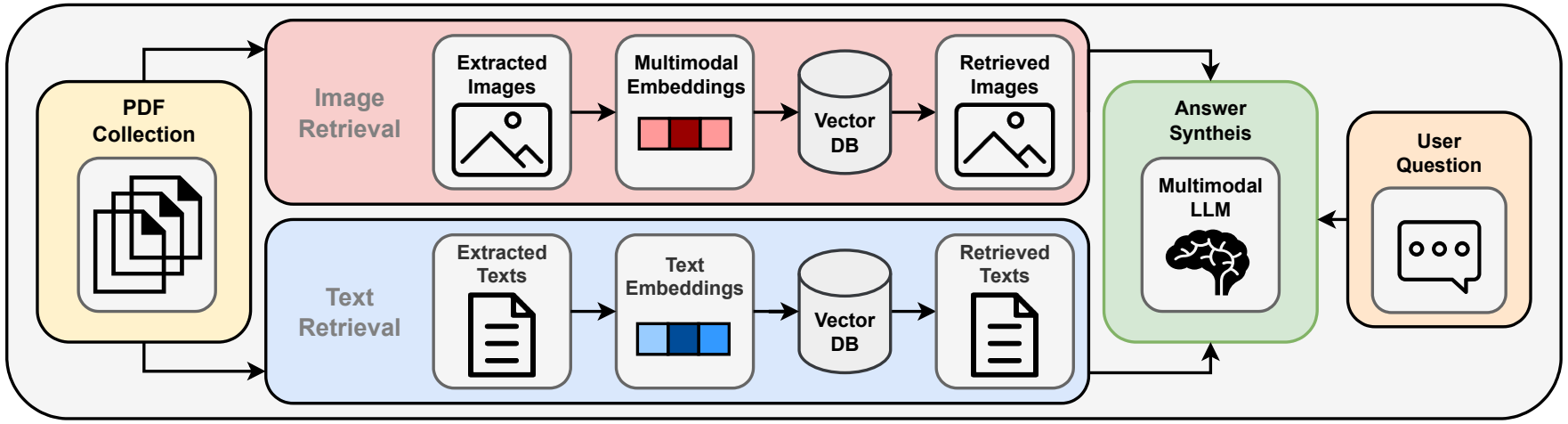
- **Image-Only RAG:**
  - Use only images from the document collection for retrieval, ignoring text.
  - How can we perform RAG on images?

  - Requirements:
    - **Multimodal model** for image and text inputs.
    - Generation of embeddings from images.
    - Image embeddings must be transferred to a **semantic space compatible with text queries** to enable effective similarity searches.

  - **Possible options:**
    - Multimodal embeddings.
    - Text embeddings from image summaries.

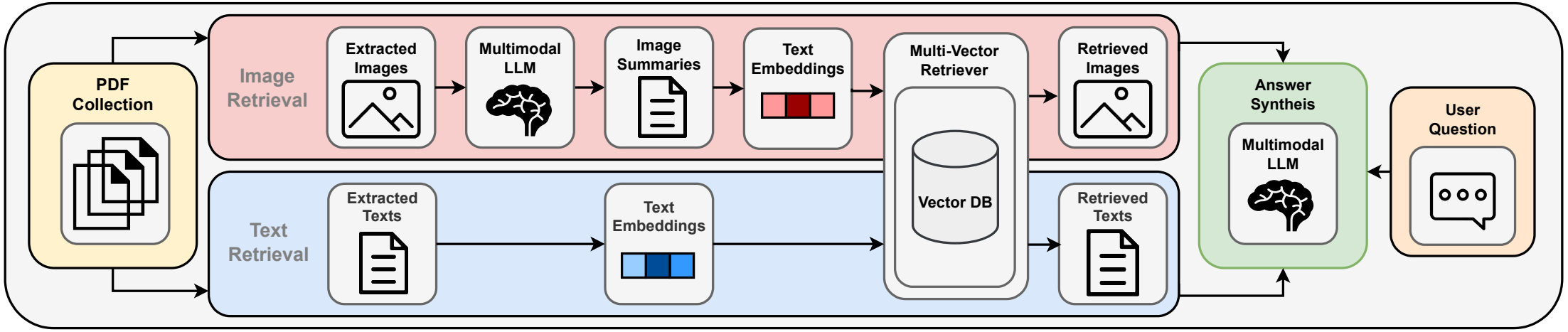
UNIMODAL VS MULTIMODAL RAG SYSTEMS

- **Multimodal RAG with CLIP Embeddings:**
  - Texts are converted into embeddings via text embedding model.
  - Images are converted into embeddings via multimodal model (CLIP).
  - Separate vector store and similarity search for each modality.



UNIMODAL VS MULTIMODAL RAG SYSTEMS

- **Multimodal RAG with Image Summaries:**
  - Images are converted into textual summaries by a multimodal LLM.
  - Both texts and textual summaries from images are embedded with a text embedding model.
  - Text and image embeddings are stored in a single vector store.
  - Single similarity search to the query for both modalities.



3. HOW DO WE MEASURE PERFORMANCE?

Evaluation Setup

Metrics

EVALUATION SETUP & METRICS

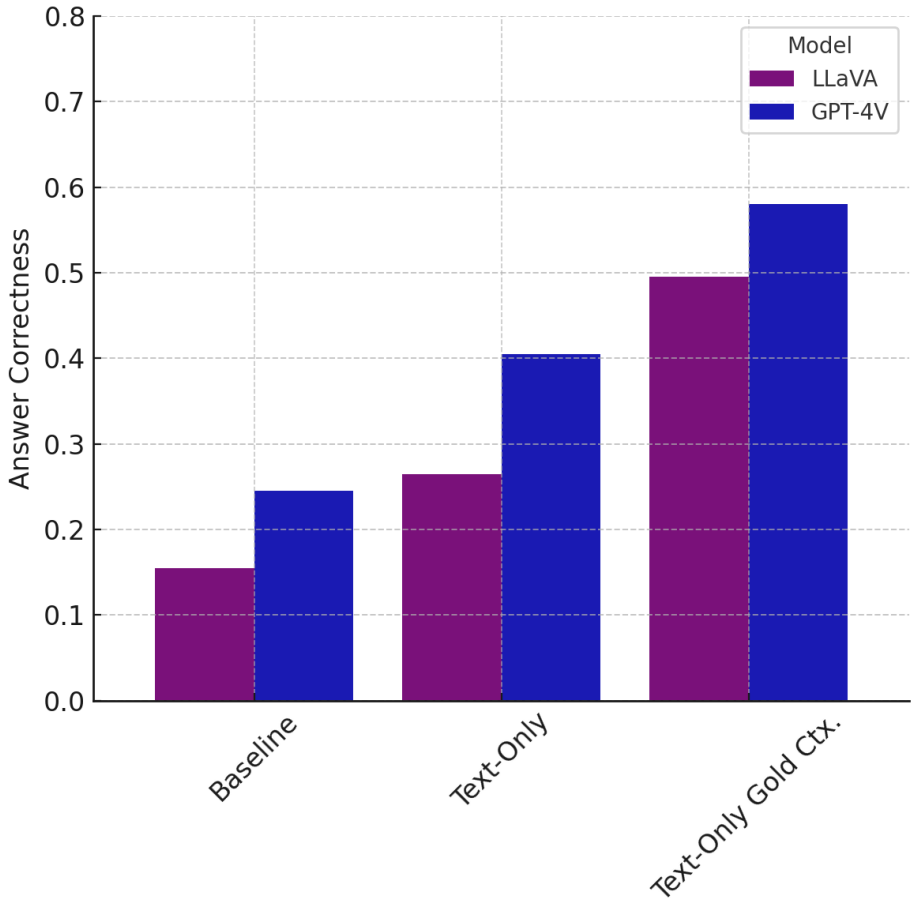
- **LLM-as-a-Judge:**
  LLMs serve as evaluators, assessing the quality of outputs produced by another LLM.
- **Evaluation Process:**
  - **Binary Evaluation:** For each answer generated by an LLM, the judge model assigns a binary score based on a specific metric. The score is either 1 (metric is met) or 0 (metric is not met).
  - **Final Score Calculation:** Average of all binary scores for each metric across the dataset.
  - **Cross-evaluation:** Each output from a RAG pipeline is evaluated by two models (GPT-4 and LLaVA).

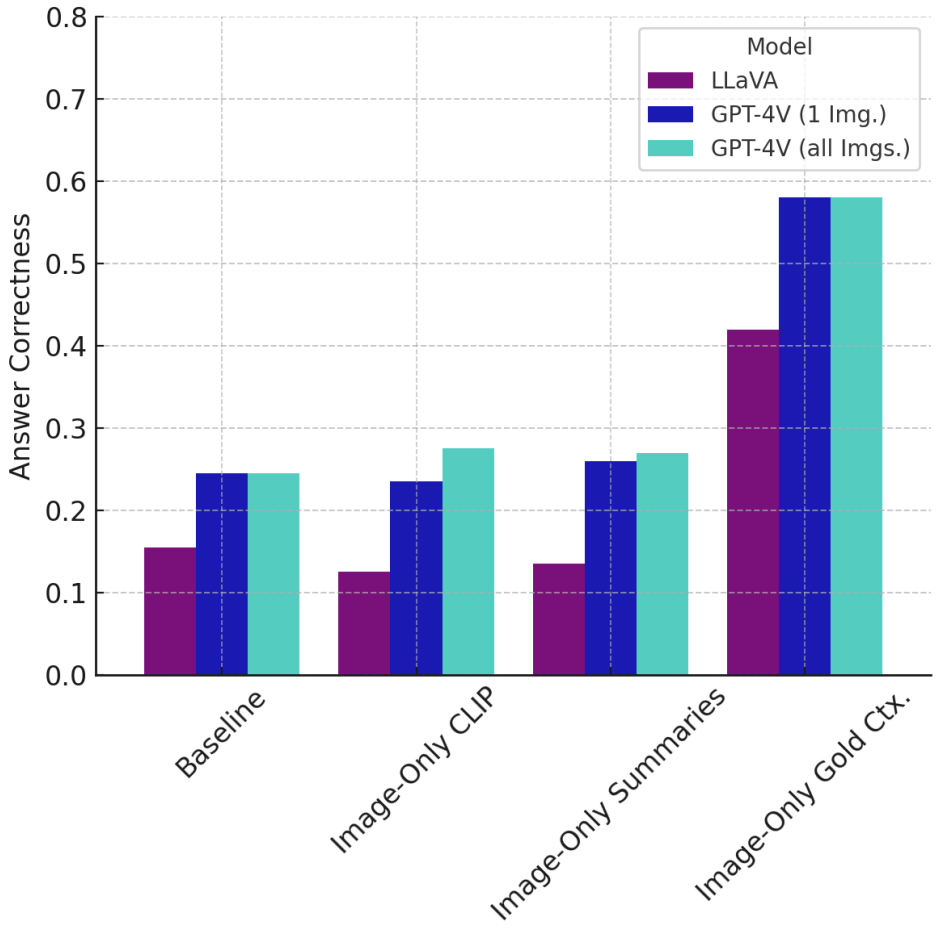
EVALUATION SETUP & METRICS

Metric	Description	Required Inputs
Answer Correctness	Is the generated answer correct?	User Query Generated Answer Reference Answer
Answer Relevancy	Is the generated answer relevant to the user query?	User Query Generated Answer
Text Faithfulness	Is the answer faithful to the context provided by the text, i.e., does it factually align with the context?	Generated Answer Text Context
Image Faithfulness	Is the answer faithful to the context provided by the image, i.e., does it factually align with the context?	Generated Answer Image Context
Text Context Relevancy	Is the context provided by the text relevant to the user query?	User Query Text Context
Image Context Relevancy	Is the context provided by the image relevant to the user query?	User Query Image Context

4. WHAT DID WE LEARN?

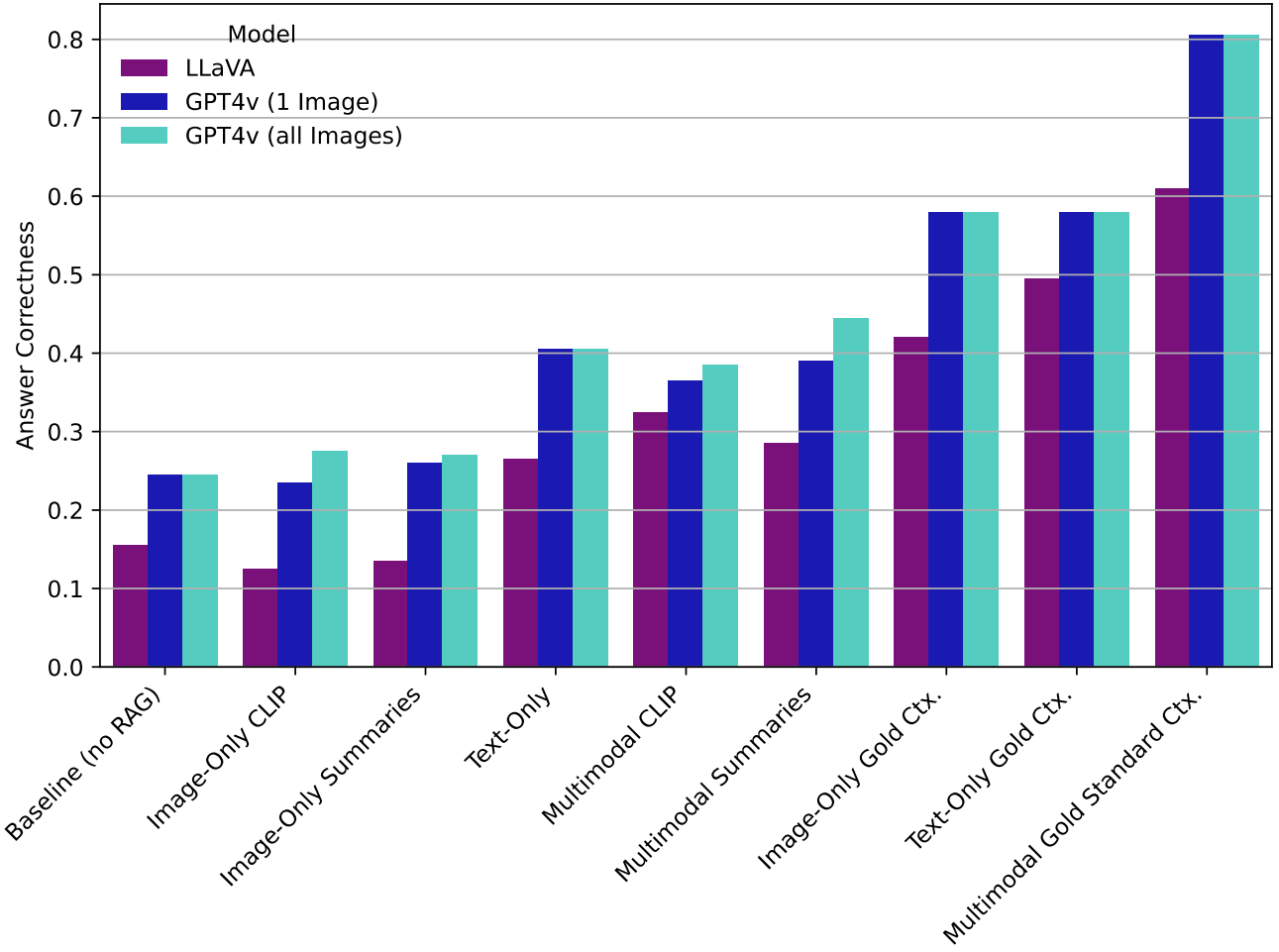
RESULTS: TEXT-ONLY VS IMAGE-ONLY RAG





MULTIMODAL RAG

- **Multimodal RAG outperforms single-modality** setups, provided correct, high-quality text and image context is available.
- **Gold standard context shows the full potential**: combining text + image boosts Answer Correctness from ~60% to ~80%.
- **Image-only RAG lags behind** highlighting current weaknesses in image retrieval.
- GPT-4v generally outperforms LLaVA.



WHAT DID WE LEARN?

- **Multimodal RAG works - when retrieval works**: combining text and image context improves answer quality.
- **Image retrieval is still tough:** way harder than getting the right text.
- Using **image summaries** works about as well as **multimodal embeddings**, but gives us more to tweak and improve.
  - Summarization prompt can be tailored to focus on specific aspects or include few-shot examples
  - More choices for the text embedding model compared to multimodal ones

5. WHAT'S NEXT?

WHAT'S NEXT?

- **Many ways to improve performance:** more advanced retrieval strategies, smarter chunking, better embeddings, and higher-quality image summaries.
- **Fix retrieval:** Look at each part of the pipeline separately to find where things break, especially image handling.
- Try **fine-tuned multimodal models** that know the industrial domain better.
- Combine **RAG** with **fine-tuning** approaches.
- Tweak the way we summarize images (better prompts, few-shot examples, etc.)

QUESTIONS?

Beyond Text: Optimizing RAG with Multimodal Inputs for Industrial Applications

Monica Riedler, Stefan Langer

SIEMENS

Paper

GitHub