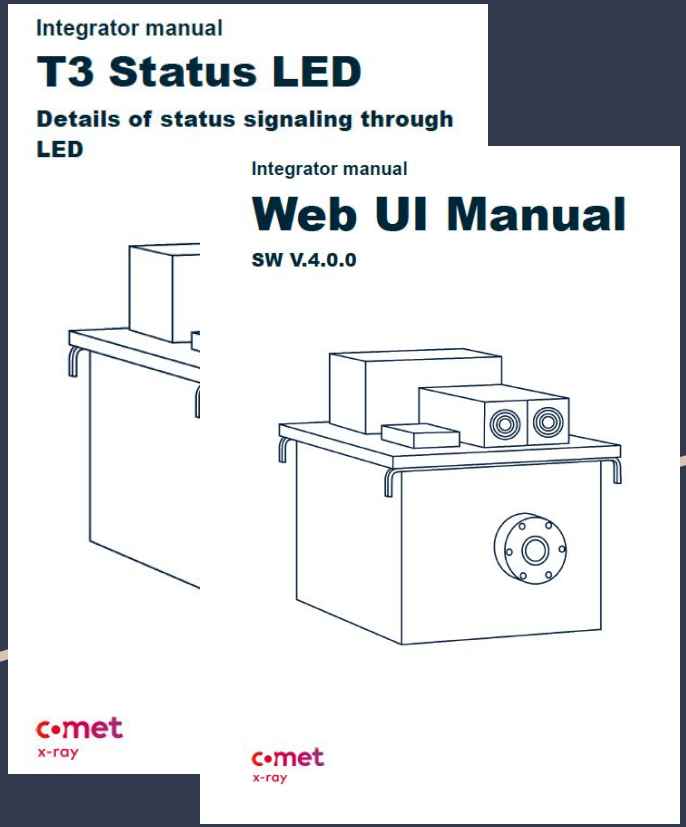
QA with Multimodal RAG: Document Retrieval and Vision Language Models

CAS NLP, Module 4 Project

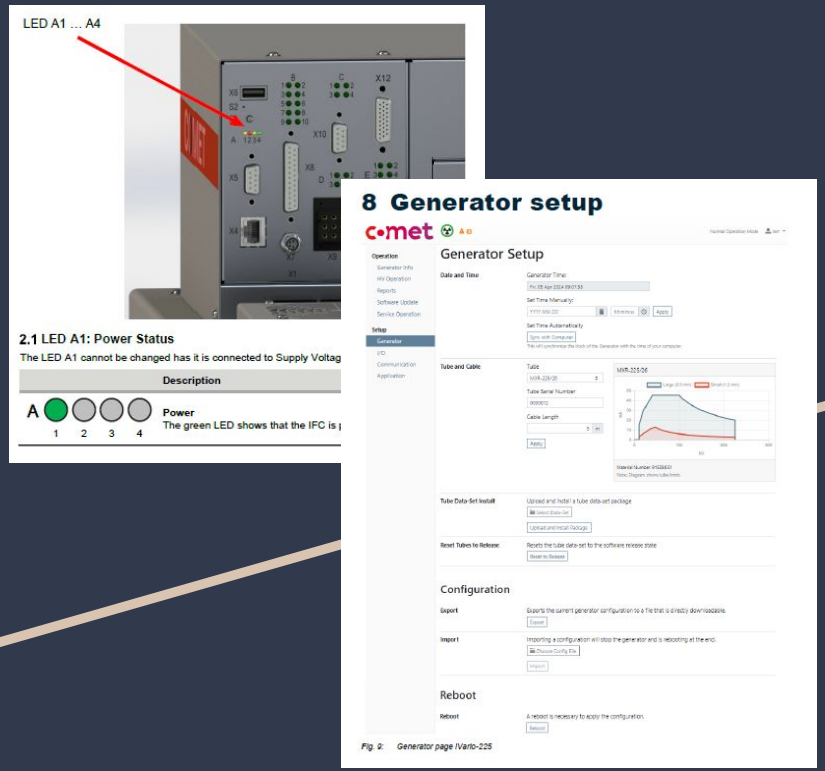
Sandro Wiedmer, 28 Jan'25

Problem



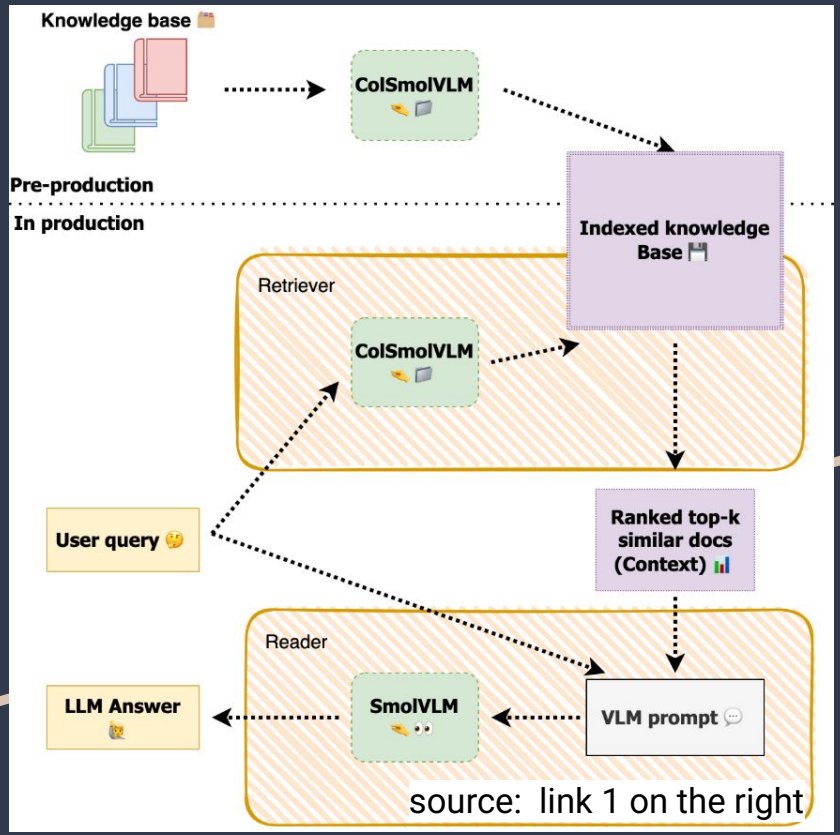
- Domain-/product-specific QA: RAG document retrieval with several manuals of an X-ray generator product family
- Beside text, information is also included in images and tables
- Use case: Could be helpful for Technical Customer Support, service personnel, customers (OEMs) and end-users, to quickly find information by asking questions in natural language about product usage, debugging, maintenance etc. which by classical text search does not work to the same degree of satisfaction

Dataset



- “Complex” documents: PDFs with text, tables, images
- Choice of 6 different manuals made available (10…259 pages)
- Only 2 of them connected to RAG, to run with low resources
- Queries chosen in such a way that answers should be found in the connected documents (validation of approach for this use case)

Approach



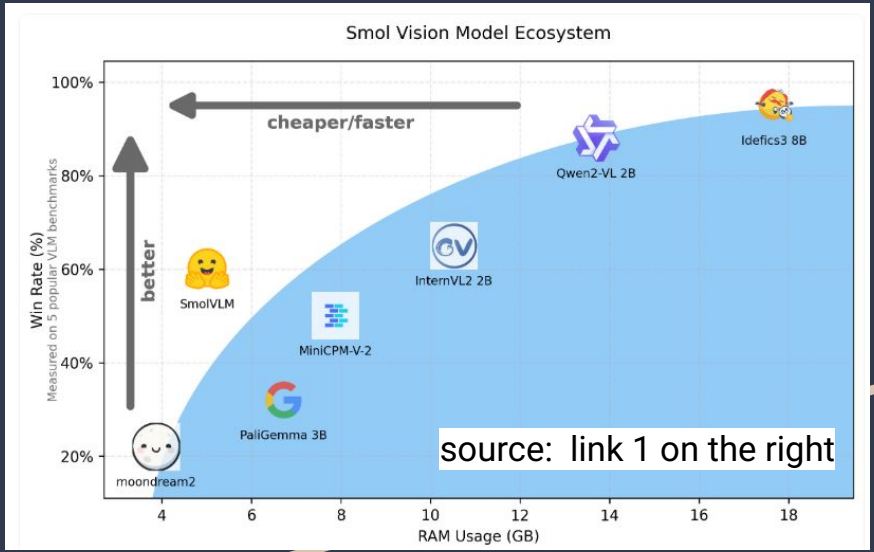
Multimodal RAG, built with small vision language models
(lightweight models as an actual trend: efficient & low-cost)

- Powerful RAG system capable of enhancing query responses with both text-based documents and visual data
- No complex document processor pipeline needed for extracting data through OCR (optical character recognition)
- ColSmolVLM-alpha: Visual document retriever based on SmolVLM-Instruct with ColBERT strategy (encoder)
- SmolVLM-Instruct - 'small yet mighty' Vision Language Model, instruction fine-tuned, QA part (encoder) - input: text and images; output: text; Idefics3 architecture

Links:

- https://huggingface.co/learn/cookbook/multimodal_rag_using_document_retrieval_and_smol_vlm
- https://huggingface.co/vidore/colsmolvlm-alpha
- https://huggingface.co/HuggingFaceTB/SmolVLM-Instruct
- https://huggingface.co/HuggingFaceM4/Idefics3-8B-Llama3
- https://arxiv.org/pdf/2408.12637

Results and Evals



ViDoRe Benchmark (**Vi**sual **Do**cument **Re**trieval)

- Comprehensive Evaluation: ViDoRe Benchmark assesses document retrieval systems across diverse tasks, domains, languages, and settings, including visually rich documents
- Focus on Visual Understanding: It emphasizes the importance of visual content in document retrieval, going beyond traditional text-based approaches
- Real-world Relevance: ViDoRe includes datasets from various real-world applications, making it a valuable tool for evaluating the practical performance of retrieval models
- ColSmolVLM is ranked no. 8 on the ViDoRe Leaderboard, although one of the smallest VLMs

Links:

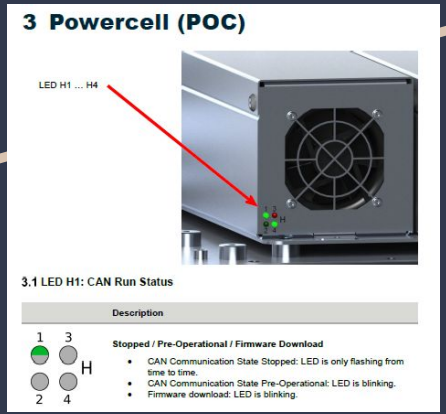
- https://huggingface.co/blog/smolvlm
- https://huggingface.co/vidore
- https://arxiv.org/pdf/2407.01449
- https://huggingface.co/spaces/vidore/vidore-leaderboard

Results and Evals

```
text_query = "What is the minimum kV set value of a mesofocus tube 225MF?"
```

```
print(output_text[0])
```

```
The minimum kV set value of a mesofocus tube 225MF is 50kV.
```



Manual Result Evaluation

- 100% correct answers (6 out of 6 queries)
- https://colab.research.google.com/drive/1E9x

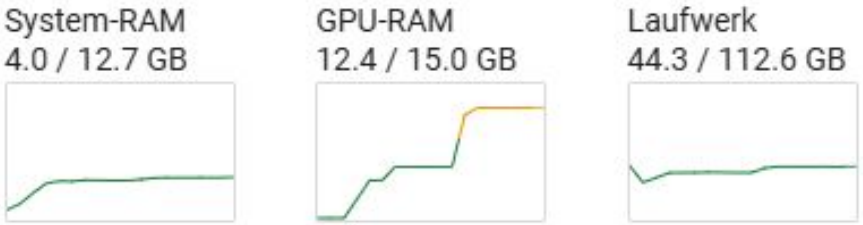
Text query	Output text	result
What is the minimum kV set value of a mesofocus tube 225MF?	The minimum kV set value of a mesofocus tube 225MF is 50kV.	correct
what's the default setting IP network address?	The default setting IP network address is 192.168.177.199.	correct
what does the green LED no IFC show	The green LED no IFC shows that the IFC is powered.	correct
what does the red LED on IFC blinking at 2 Hz mean	The LED A2 blinks red at 2 Hz when the system could boot correctly, but a general boot error is present.	correct
what does the LED H4 on PowerCell mean, and is it controlled by software?	The LED H4 on PowerCell shows that the Power of the PowerCell is on. The LED H4 cannot be changed by software as it is connected to auxiliary power.	correct
How many LEDs are there on the PowerCell, and how are they named?	There are four LEDs on the PowerCell, named LED H1, LED H2, LED H3, and LED H4.	correct

Discussion

- **Multimodal RAG** approach significantly improves information retrieval from PDFs compared to traditional text-based methods: By using visual language models, information presented in **tables, images, and diagrams**, is available => **more comprehensive and accurate answers**
- **Efficiency**: Utilizing **small** vision language models minimizes computational costs and resource requirements
- **Easy-to-use, straightforward setup** of Multimodal RAG pipeline using “Smol’ visual language models for retrieval from documents with text, tables, images: Thanks to **user-friendly models and helpful documentation** available
- **Good results with small, lightweight models** (in this case) => **No further fine-tuning** performed, since the available SmolVLM-Instruct model performs well
- As a nice-to-have feature, **explainability mechanisms** could be expanded: Additionally to showing the page(s) where the information was retrieved, also the specific area on the page(s) could be highlighted (see https://arxiv.org/pdf/2407.01449 as an example)
- Multimodal and vision language models: **A lot of potential use cases expected**

Limitations of Approach

- **Resource Constraints:** Although lightweight models used, the focus was on further resources minimization => System and GPU RAM (Google Colab free-tier): Number of document (pages) connected & retrieval relevance ranking top_k = 2
- Potential for getting smaller (even smaller GPU memory footprint): **Quantization**
- **Answer Generation Latency:** The current implementation has a relatively high latency (14…15 seconds) for generating answers



Questions?