

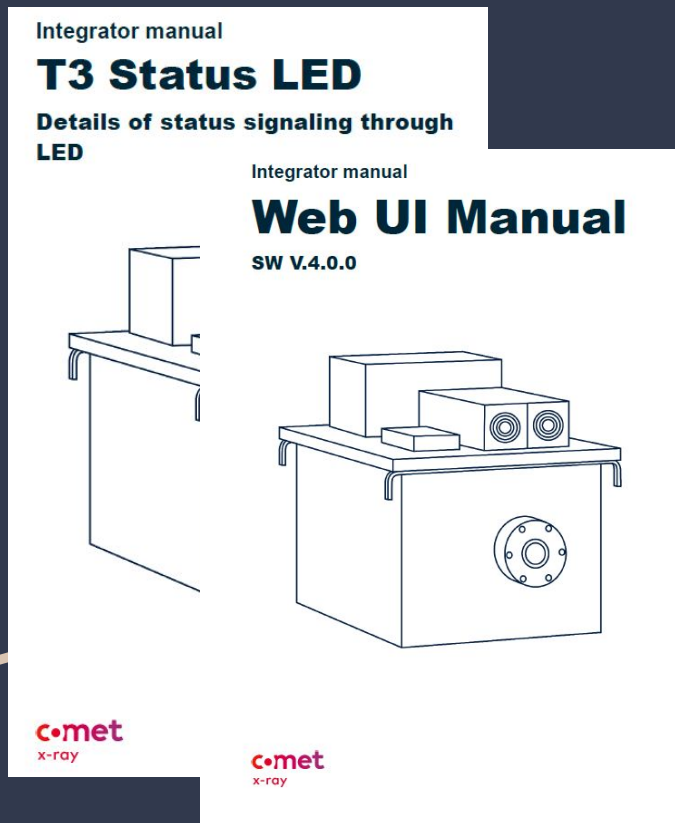
# QA with Multimodal RAG: Document Retrieval and Vision Language Models

CAS NLP, Module 4 Project

Sandro Wiedmer, 28 Jan'25

A dark blue diagonal gradient bar that starts from the bottom left and extends towards the top right, covering the lower half of the slide.

# 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



## 8 Generator setup

**c-met**   Home Operation Mode

**Generator Setup**

**Operation**

- Generator Info
- Web Operation
- Reports
- Software Update
- Service Operation

**Setup**

- Generator
- IPC
- Communication
- Apparatus

**Date and Time**

Generator Time:

Set Time Manually:

Set Time Automatically:

**Tube and Cable**

Tube:

Tube Serial Number:

Cable Length:

**Tube Data Set Install**

Upload and install a tube data set package

**Reset Tubes to Release**

Resets the tube data set to the software release state

**Configuration**

**Export**

Exports the current generator configuration to a file that is directly downloadable

**Import**

Importing a configuration will stop the generator and is reconnecting at the end.

**Reboot**

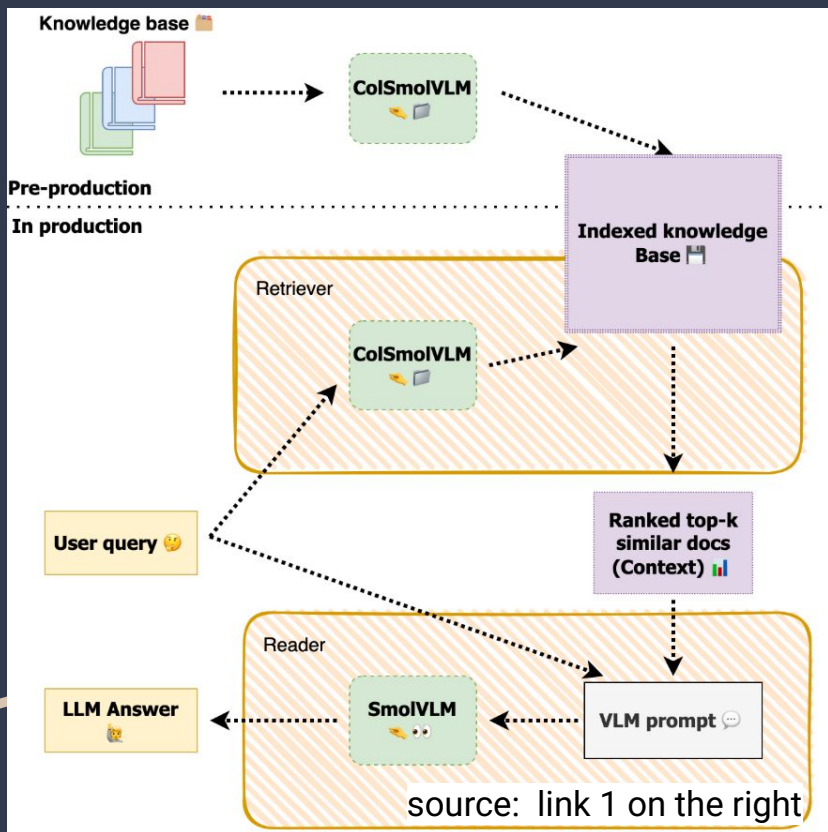
Reboot

A reboot is necessary to apply the configuration

Fig. 8 Generator page IVario-225

- “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)

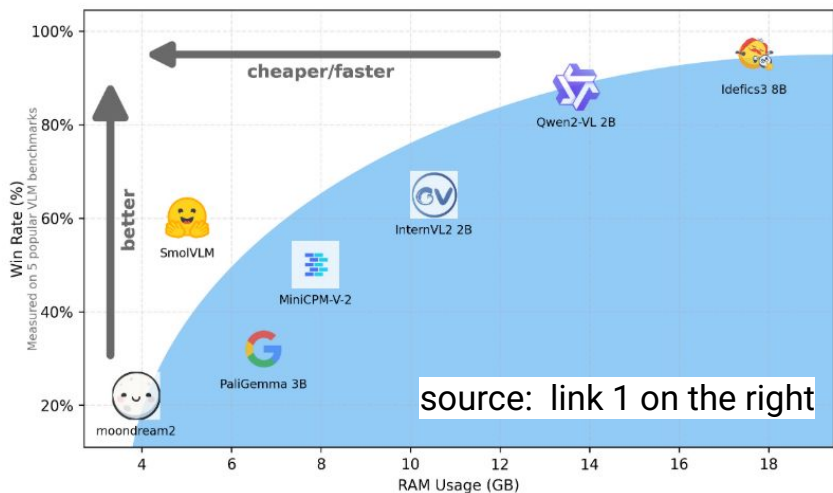
- Powerful RAG system capable of enhancing query responses with both text-based documents and visual data
- No complex document processor pipeline needed for extracting data through OCR (optical character recognition)
- ColSmoIVLM-alpha: Visual document retriever based on SmoIVLM-Instruct with ColBERT strategy (encoder)
- SmoIVLM-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/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/Idfics3-8B-Llama3>
- <https://arxiv.org/pdf/2408.12637>

# Results and Evals

Smol Vision Model Ecosystem



## ViDoRe Benchmark (Visual Document Retrieval)

- 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

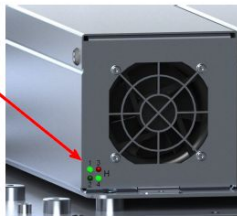
```
text_query = "What is the minimum kV set value of a mesofocus tube 225MF?"
```

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

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

## 3 Powercell (POC)

LED H1 ... H4



3.1 LED H1: CAN Run Status

Description



Stopped / Pre-Operational / Firmware Download

- CAN Communication State Stopped: LED is only flashing from time to time.
- CAN Communication State Pre-Operational: LED is blinking.
- Firmware download: LED is blinking.

## 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 <u>mesofocus</u> tube 225MF?	The minimum kV set value of a mesofocus tube 225MF is 50kV.	correct
<u>what's</u> the default setting IP network address?	The default setting IP network address is 192.168.177.199.	correct
what does the green LED <u>no</u> IFC show	The green LED <u>no</u> 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
<u>what</u> 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

System-RAM  
4.0 / 12.7 GB



GPU-RAM  
12.4 / 15.0 GB



Laufwerk  
44.3 / 112.6 GB





Questions?