Offline Multimodal Large Language Models for Decision Support in Air Operations

ICAS PAPER 2026_0450

João P. A. Dantas, Jelton A. Cunha and Gabriel Dietzsch

Instituto de Estudos Avançados (IEAv)

Brazilian Air Force, São José dos Campos, Brazil

Outline

1. **Motivation and the task**
   why offline support, and what the analyst does today
2. **The proposed system**
   an offline multimodal retrieval architecture
3. **Pilot study and results**
   four FAB analysts, knowledge test and workload
4. **Conclusions and next steps**
   takeaways and the evaluation protocol ahead

Motivation and the task

Air operations: information rich, connectivity poor

Doctrine is long

Manuals run to hundreds of pages and must be matched against what the analyst sees.

No cloud access

Security rules block external services. Classified imagery cannot leave the environment.

Faster decisions

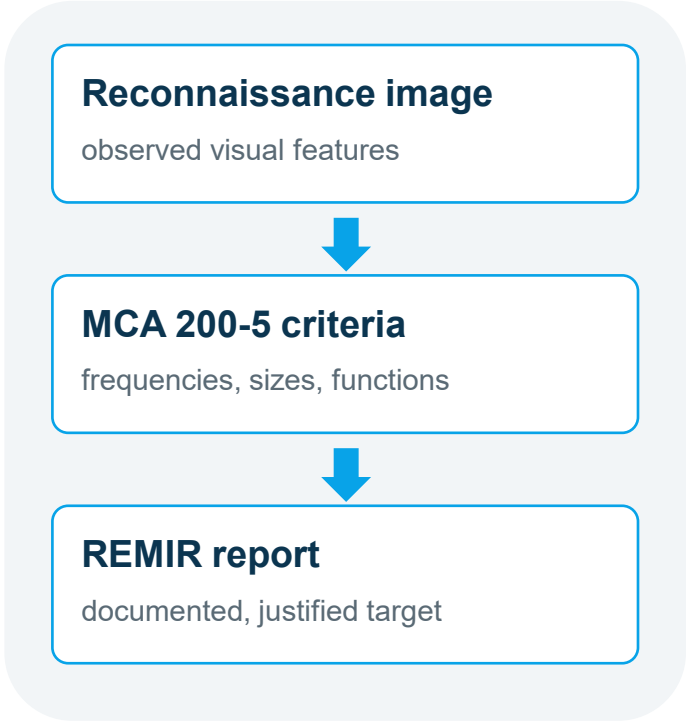
Manual cross-referencing is slow, time-consuming and raises the risk of weakly supported judgments.

The most capable language models are in the cloud. The work that needs them happens offline.

The task: identifying electronic targets

Governed by the FAB manual MCA 200-5

- Analysts match visual features, such as reflector size and antenna layout, to doctrinal criteria such as frequency, power and function.
- The product is a reconnaissance mission report, the REMIR.
- Today the whole comparison is done by hand, page by page.



Contributions

1

Requirements

What an offline decision aid must satisfy: air-gapped operation, traceability to the source, and accountability that stays with the analyst.

2

Architecture

A modular retrieval-augmented system that meets those requirements and runs entirely on local hardware, with text and image input.

3

Pilot study

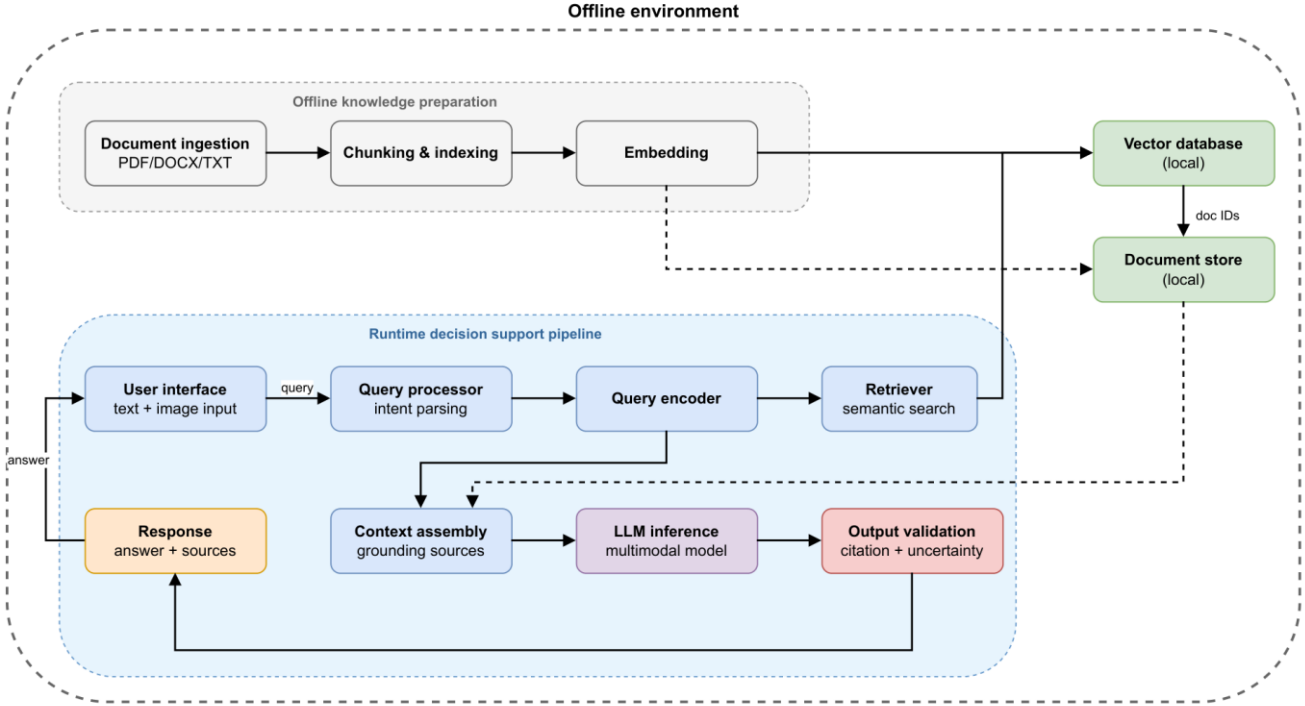
A first evaluation with FAB image analysts: a doctrinal knowledge test and a workload baseline for the manual task.

The model is a cognitive support tool. The decision stays with the analyst.

The proposed system

1. Motivation and the task
2. **The proposed system**
3. Pilot study and results
4. Conclusions and next steps

Everything runs inside the offline environment



Offline knowledge preparation feeds a local vector store; the runtime pipeline answers with sources attached.

Preparing the knowledge base offline

1

Chunking

1024-token chunks with 128-token overlap, so tables and paragraphs stay intact.

2

Deduplication

MD5 filtering removes repeated passages and saves context window.

3

Multimodal extraction

PyMuPDF pulls out figures; EasyOCR reads labels; a local VLM captions them.

4

Vector indexing

BGE-M3 embeddings in ChromaDB with document, page and chunk metadata.

Diagrams and tables become searchable text, linked to the page they came from.

Two-phase retrieval and re-ranking

Broad search

BGE-M3 over ChromaDB
cosine similarity
top 20 candidates

Scans the whole corpus

the whole manual

Neural re-ranking

BGE-Reranker-v2-m3
query and chunk together
full cross-attention

Judges true relevance

20 candidates

Filter and cache

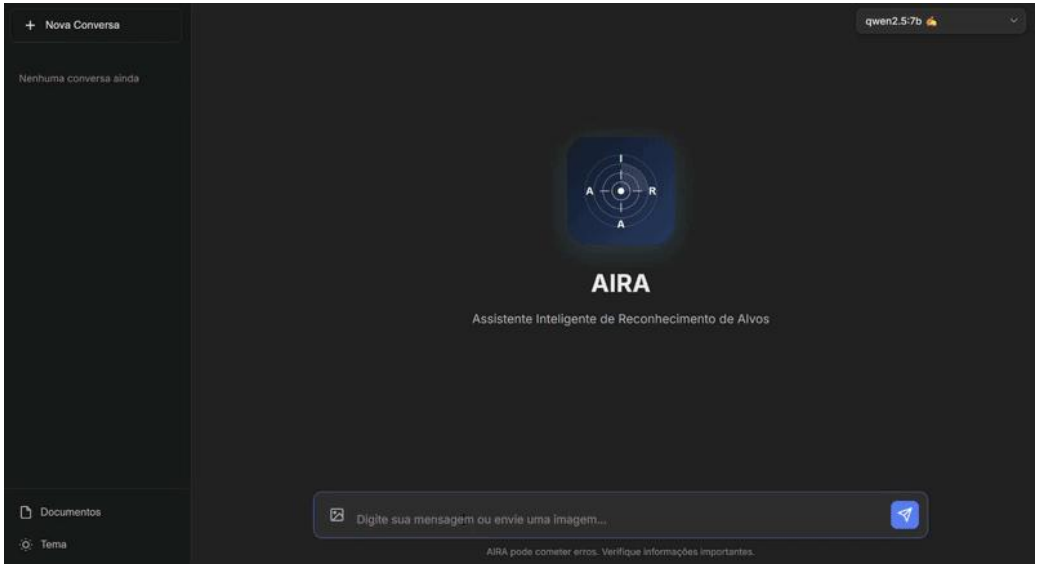
score below 0.1 discarded
top 5 chunks kept
1-hour query cache

Keeps only the best five

5 passages

Cheap search over everything, expensive comparison over the few candidates that matter.

Runtime interaction and the feedback loop



Local web interface: answer, justification and sources, in Portuguese.

Grounded answers

Retrieved passages, learned rules and history are assembled into the prompt.

Thumbs up

The answer is stored as an example of the style analysts want.

Thumbs down

The comment becomes an active correction rule for later queries.

Analyst feedback changes behaviour through prompts, with no fine-tuning.

How the prompt is assembled

1

The rules

How to answer, always naming the source. Plus the corrections the analysts asked for themselves.

2

The evidence

The five passages from the manual that survived the search, figures included.

3

The question

What was asked earlier in the conversation, and what is being asked right now.

The model never changes. What changes is the package we hand it.

Pilot study and results

1. Motivation and the task
2. The proposed system
3. **Pilot study and results**
4. Conclusions and next steps

How the pilot was run

Participants

Analyst	Role	Experience	MCA 200-5
01	Officer	over 7 years	High
02	Officer	4 to 7 years	Moderate
03	Enlisted	over 7 years	Very high
04	Enlisted	4 to 7 years	Moderate

Very few analysts hold the required credentials, so this is a pilot study.

Two blocks, manual condition

Block A

Ten-item test on MCA 200-5 fundamentals, using only the manual. Time self-recorded.

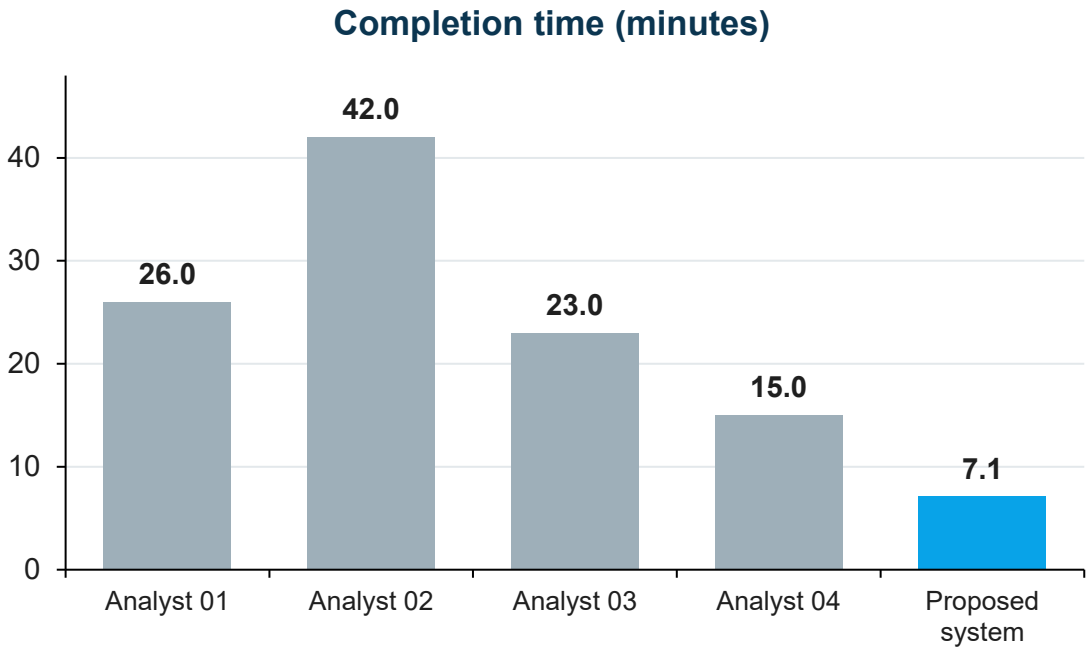
Block B

Manual REMIR for a real electronic target, followed by the Raw NASA-TLX.

The same ten-item test was run on three local models on an edge workstation (RTX 4060 Laptop, 8 GB VRAM), under two prompt regimes.

Knowledge test: analysts and the system

BLOCK A



8 / 10

system score, matching one of the analysts

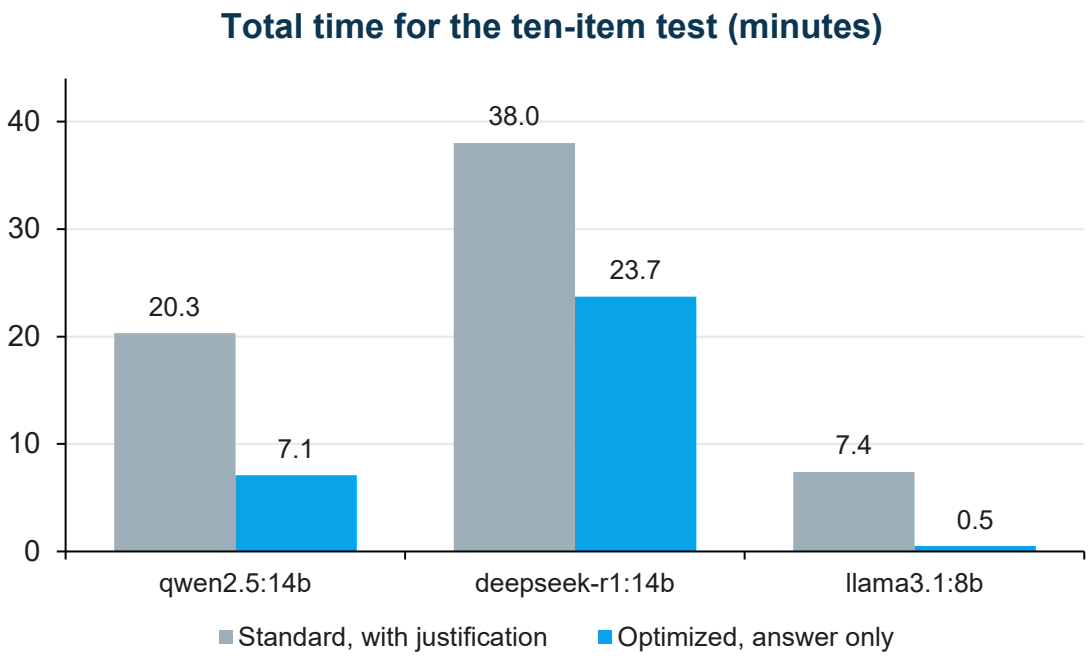
7.1 min

against a human average of 26.5 minutes

Three of the four analysts scored 10/10. Accuracy is established in this group; search time is not.

Prompt design matters more than model size

BLOCK A



15x faster

llama3.1:8b fits in VRAM. Accuracy also rose, from 5/10 to 7/10.

Offloading costs

The 14B models exceed 8 GB and pay to move weights between RAM and VRAM.

Best balance

qwen2.5:14b: 8/10 in 7.1 minutes, the baseline configuration.

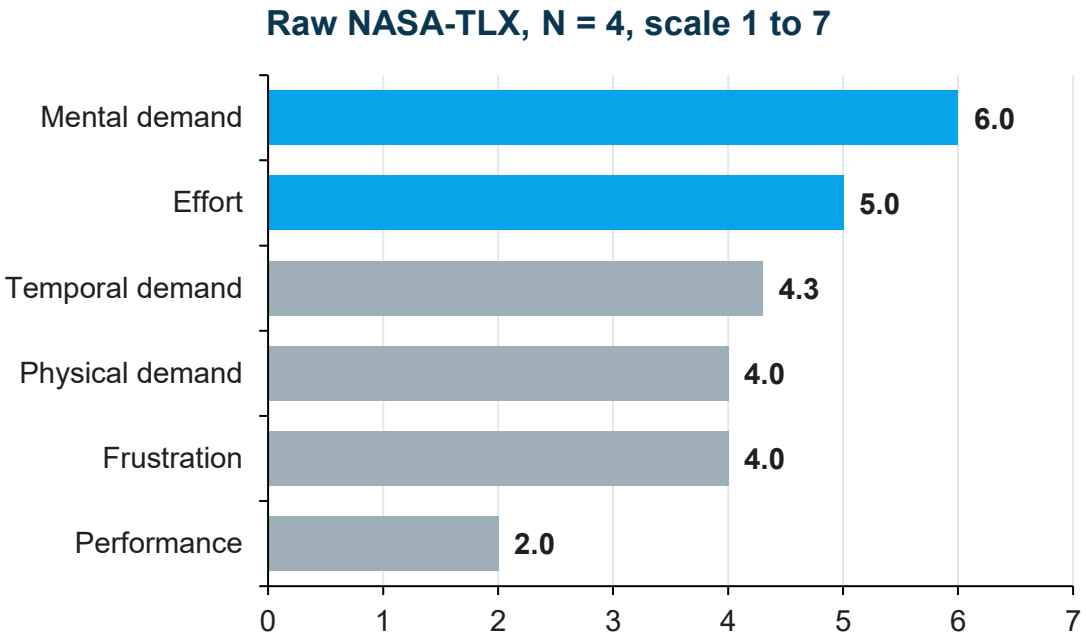
Full model evaluation: accuracy and time

BLOCK A

Model and regime	Score	Avg time (s)	Total (min)
Standard regime, with justification			
qwen2.5:14b	8 / 10	121.91	20.3
deepseek-r1:14b	8 / 10	227.90	38.0
llama3.1:8b	5 / 10	44.28	7.4
Optimized regime, answer only			
qwen2.5:14b	8 / 10	42.86	7.1
deepseek-r1:14b	8 / 10	141.97	23.7
llama3.1:8b	7 / 10	2.94	0.5

The manual task is cognitively demanding

BLOCK B



4.2 / 7

overall workload, above the midpoint of the scale

Mental demand and effort dominate, even for analysts with years of experience and high familiarity with the manual.

Performance is reverse scored: the low value of 2.0 means the analysts felt they did the job well.

Conclusions and next steps

1. Motivation and the task
2. The proposed system
3. Pilot study and results
4. **Conclusions and next steps**

Three things to take away

1

Useful support can run fully offline

A quantized 14B model on laptop-class hardware matched an analyst's score and finished in a quarter of the time.

2

Traceability is a requirement, not a feature

Retrieval, re-ranking and citation exist so that every claim can be checked against the doctrine page it came from.

3

Accuracy alone will not decide adoption

Doing this by hand scores 4.2 out of 7 on workload. A faster answer only counts if it lowers that number and the analyst trusts it.

What comes next

Evaluation protocol

designed and ready to run

- Repeat the REMIR task with the same analysts, this time with the assistant, and measure workload again.
- Add perceived usability, trust in automation and intention to adopt: UMUX-Lite, S-TIAS and UTAUT.

System improvements

three directions

- Image-to-image search: find manual figures that look like the photo.
- Draft the report automatically from the conversation.
- Measure automatically whether the answer really follows from the source.

www.joaopadantas.com

Thank you

João P. A. Dantas | jpdantas@ita.br

Instituto de Estudos Avançados (IEAv)
Brazilian Air Force