OFFLINE MULTIMODAL LARGE LANGUAGE MODELS FOR DECISION SUPPORT IN AIR OPERATIONS

Joao P. A. Dantas[1], Jelton A. Cunha[1] & Gabriel Dietzsch[1]

[1]Instituto de Estudos Avançados, São José dos Campos, São Paulo, Brazil, 12.228-001

Abstract

Air operations rely on complex rules, established procedures, and time-critical analysis under limited connectivity and strict security constraints. In such environments, analysts must combine written doctrine with images, often without access to external computing resources. This paper studies offline large language models as decision support tools, deployed in isolated and restricted environments to give analysts access to doctrinal knowledge that remains traceable to its original sources, through natural language interaction. We describe a modular retrieval-augmented architecture suitable for operation without Internet connectivity, supporting both text and image input from technical manuals. As a first step toward evaluating this architecture, we report a pilot study with four image analysts of the Brazilian Air Force, combining (i) a doctrinal knowledge assessment based on their electronic-target identification doctrine, comparing human and proposed system performance on the same test, and (ii) a measurement of the cognitive workload involved in manually producing a reconnaissance target report (*Relatório de Missão de Reconhecimento* — REMIR) without AI assistance. The results show a demanding manual task, especially in terms of mental demand (6.0/7) and effort (5.0/7), while the proposed system matches the human score (8/10) and completes the assessment in 7.1 minutes (compared to a human average of 26.5 minutes), establishing a baseline for future AI-assisted evaluation. Finally, we describe a future evaluation protocol to systematically compare manual and AI-assisted workflows.

Keywords: air operations, decision support systems, offline artificial intelligence, large language models, multimodal AI, human-in-the-loop, aerospace systems

1 Introduction

Modern air operations involve large amounts of information, strict rules, and time-critical decisions, placing heavy demands on operators' situation awareness and mental workload [1, 2]. Command and control (C2) and intelligence, surveillance, and reconnaissance (ISR) tasks require analysts to interpret doctrine, operational guidance, and a wide range of intelligence products, often under limited connectivity and strict security rules that prevent access to external computing resources [3, 4].

One such intelligence product is image analysis, where analysts must compare visual features with exact details described in technical manuals that can be hundreds of pages long. When performed manually under time pressure, this comparison is mentally demanding and increases the risk of errors or poorly supported judgments [5, 6]. Effective decision support in this setting must therefore make it faster to find reliable information, while keeping the analyst responsible and able to trace every conclusion back to its source in the doctrine [7]. A clear example of this challenge is electronic target intelligence: to identify a radar, antenna, or command post in a reconnaissance image, an analyst must match observed features, such as reflector size and antenna layout, with details described in doctrine, such as frequencies and functions [8].

Recent progress in large language models (LLMs) has enabled natural language interaction with large collections of text, and retrieval-augmented generation (RAG) further connects model outputs

to real documents, reducing the risk of fabricated content [9, 10]. However, the most capable systems run on cloud infrastructure, which is not compatible with air-gapped intelligence environments, where transmitting classified images or doctrine over external networks is prohibited [3]. This gap between the capabilities of modern LLMs and the security limits of operational settings is the main motivation for this work.

In this work, the LLM is treated as a cognitive support tool, not as a decision maker: it retrieves relevant parts of the doctrine, explains rules and procedures, and supports the analyst's reasoning, while the final decision stays with the analyst, following well-known principles of human-centered AI [11, 5]. Given this framing, this paper makes three main contributions:

- We identify the requirements for using offline LLMs as decision support tools in restricted aerospace environments, such as air-gapped operation, traceability to sources, and human accountability.
- We propose a modular RAG architecture that meets these requirements, running fully offline and producing answers that can be traced back to their sources.
- We report a pilot study with image analysts of the Brazilian Air Force (FAB) that compares human and AI performance on a doctrinal knowledge test and measures the cognitive workload of the manual electronic-target reporting task, establishing a baseline for future AI-assisted evaluation and guiding the design of a full within-subjects evaluation protocol.

This study is framed as a feasibility-oriented pilot, and its empirical results should be interpreted descriptively rather than as confirmatory evidence of operational effectiveness.

The rest of this paper is organized as follows. Section 2 reviews related work on LLM-based decision support in defense and on multimodal retrieval-augmented generation. Section 3 describes the proposed offline architecture, including the document processing pipeline, two-phase retrieval, and the human-in-the-loop feedback mechanism. Section 4 presents the case study and evaluation methodology based on FAB's doctrine manual for electronic target identification. Section 5 reports the experimental results, and Section 6 concludes and describes future work.

2 Related Work

Decision support in defense and aerospace has historically relied on expert systems, structured workflows, and deterministic databases [3, 7]. While such systems provide consistency, they are often rigid and have trouble handling unstructured narrative data, such as tactical debriefs, mission reports, and multi-page doctrinal manuals [4]. Large language models introduce conversational interfaces capable of parsing natural language and producing coherent summaries [9]. This research line has explored several related directions in this area: offline language-based support for tactical decision-making in air combat [12], radar-based reconstruction of target behavior to support situational awareness [13], and simulation tools that bring a human operator into the loop of air combat scenarios [14, 15]. In line with human-automation frameworks, such models are best understood as support tools that improve human understanding, consistency, and situational awareness while keeping accountability and control with the human, rather than as autonomous decision-makers [5, 11, 7].

Integrating LLMs into military environments, however, raises strict security and reliability challenges. Cloud-based LLM APIs are not compatible with air-gapped networks because of the risk of transmitting sensitive data over public channels [3]. Furthermore, generic LLMs are prone to "hallucinations", meaning they can generate plausible but factually incorrect statements, which is unacceptable where decisions must comply with specific rules of engagement and established doctrine [10]. These constraints have motivated the local deployment of smaller, quantized models running entirely on physical servers, combined with grounding mechanisms that restrict responses to verified source materials.

Among these grounding mechanisms, RAG reduces hallucination by retrieving document passages semantically similar to the user query and adding them to the model context [9]. While text-only RAG is mature, air operations involve both textual doctrine and visual intelligence products, such as target photographs, structural diagrams, and maps, which motivates multimodal retrieval that aligns visual content with textual context [16, 17]. Importantly, the goal is not to automate image analysis or target

recognition, but to assist analysts by retrieving relevant doctrinal criteria and supporting structured reasoning that combines textual procedures with visual observations [4, 5].

Multimodal RAG architectures generally follow two approaches: direct cross-modal embedding models that map images and text into a shared vector space [17], or hybrid pipelines in which visual assets are first translated into textual descriptions prior to indexing [16]. For technical manuals, the hybrid approach is frequently preferred, because visual elements such as tables and schematics encode topological relationships that are difficult to capture with generic zero-shot image embeddings. Combining Optical Character Recognition (OCR) with local vision-language captioning converts such diagrams into descriptive text that can be indexed alongside the textual chapters. To improve retrieval precision over single-stage dense search, recent systems add a neural re-ranking stage based on cross-encoders [10]; and to align outputs with domain-specific formats without costly fine-tuning, in-context prompt adaptation has emerged as a practical alternative in constrained settings. We build on the hybrid OCR-based approach, neural re-ranking, and prompt adaptation, detailing our implementation in Section 3.

Beyond accuracy, evaluating decision support tools in safety-critical domains also requires methods that capture the human side of the interaction. Recent evidence from multimodal AI applied to aerial monitoring indicates that accuracy alone is insufficient for operational assessment, since models may produce overconfident errors, instruction-following inconsistencies, or unsupported semantic interpretations [18]. This motivates the inclusion of behavioral criteria (such as uncertainty communication, confidence calibration, instruction adherence, and source-grounded reasoning) alongside subjective human-centric metrics. While the former focuses on verifying the AI's output directly (as discussed in Section 6), the latter captures the human experience during the interaction. To evaluate this human component, the NASA Task Load Index (NASA-TLX) is a widely used instrument for measuring perceived cognitive workload across operational tasks [19], while short instruments such as the Usability Metric for User Experience (UMUX-Lite) provide compact measures of perceived usability [20]. Trust in automation, a key factor in the adoption of AI-based tools, is commonly assessed with short scales, such as the Short Trust in Automation Scale (S-TIAS), derived from established trust frameworks [21, 22], and behavioral intention to adopt new technology is often captured through models such as the Unified Theory of Acceptance and Use of Technology (UTAUT) [23]. Together, these instruments form the basis of the evaluation approach and the full evaluation protocol outlined in Section 6.

3 System Architecture

The proposed decision support system is designed to operate in a strictly disconnected, offline military environment, keeping sensitive data secure while enabling natural language and multimodal interaction with operational documentation. The high-level modular architecture, shown in Figure 1, consists of two main phases: *offline knowledge preparation* (detailed in Section 3.1) and *runtime decision support* (comprising retrieval, re-ranking, and user interaction in Sections 3.2 and 3.3). This architecture implements a RAG framework, reducing model hallucinations and ensuring that every response generated by the system can be traced back to authoritative doctrinal publications.

3.1 Offline Document Processing and Indexing

The document preparation pipeline processes various technical manuals, regulations, and doctrinal texts (such as PDFs, DOCXs, and TXT files) into a structured knowledge base. To optimize semantic retrieval, documents are processed through a structured pipeline:

1. **Text Segmentation (Chunking):** Documents are parsed and split into overlapping chunks to preserve local semantic context. Based on empirical testing, a chunk size of 1024 tokens with an overlap of 128 tokens is used. This specific configuration balances vector information density with the context window constraints of local LLMs, ensuring that paragraphs and critical tables are not severed across chunks, thereby maintaining semantic coherence.

2. **Deduplication:** Chunks are normalized by stripping excess whitespace and filtering duplicates using MD5 hashing. This prevents redundant storage and indexing, reducing the search space and avoiding token waste in the LLM's context window.

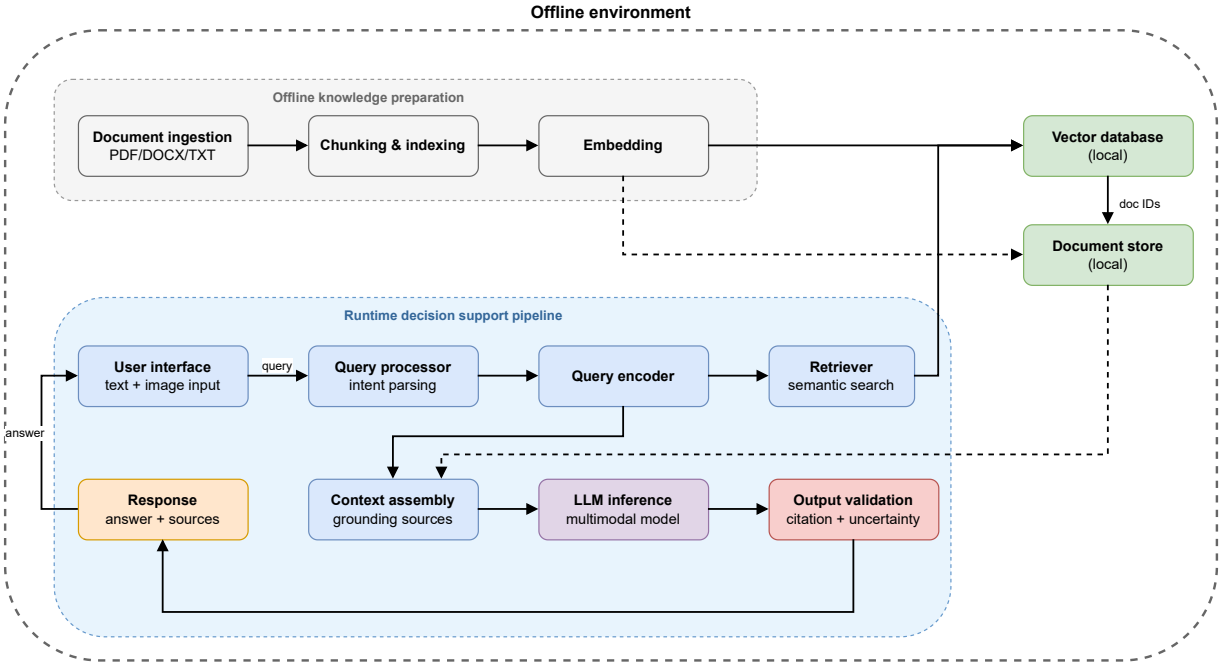

Figure 1 – Architecture for offline multimodal language model deployment in air operations decision support.

3. **Multimodal Element Extraction:** Visual elements such as diagrams, tables, and charts are extracted using PyMuPDF [24]. Rather than discarding these elements, which contain essential schematic and tactical descriptions, the document processing pipeline processes them through a multimodal sub-pipeline. First, OCR is performed via EasyOCR [25] to extract textual annotations, numbers, and labels embedded within figures. Second, semantic captions are generated locally using a lightweight vision-language model (VLM), MiniCPM-V 8B [26], describing system configurations, connections, and topologies. The OCR-extracted text and semantic VLM descriptions are concatenated with their surrounding text context to form unified multimodal chunks, preserving the link between visual assets and text.

4. **Vector Indexing:** Text and multimodal chunks are vectorized using a local instance of the multilingual BGE-M3 embedding model (`BAAI/bge-m3`) [27]. This model produces 1024-dimensional embeddings, capturing rich semantic relationships across multiple languages, which is essential for translating Portuguese air operations doctrine. The resulting embeddings are indexed in a local, persistent vector database client (ChromaDB) [28] alongside metadata (document title, UUID, page number, and chunk type), ensuring strict traceability.

3.2 Two-Phase Retrieval and Re-ranking

To ensure high-precision retrieval under strict latency constraints, a two-phase retrieval pipeline is implemented:

1. **Phase 1: Broad Vector Search (Bi-Encoder Phase):** The user's query is vectorized in real-time using the local BGE-M3 embedding model [27] and matched against the ChromaDB vector database using cosine similarity. The system retrieves the top-20 candidate chunks ($K_{initial} = 20$). Although bi-encoders are computationally efficient and enable rapid searches over millions of documents, they calculate query and document embeddings independently, which can limit their capacity to capture fine-grained semantic nuances or exact technical manual jargon.

2. **Phase 2: Neural Re-ranking (Cross-Encoder Phase):** The query and the retrieved candidate chunks are re-evaluated using a local cross-encoder model, `BAAI/bge-reranker-v2-m3` [27], running with GPU acceleration (or CPU fallback). Unlike the bi-encoder, the cross-encoder

processes the query and each candidate chunk concatenated together, performing full joint cross-attention across all tokens. This allows token-to-token comparison, producing a significantly more precise semantic alignment score at the expense of higher latency per pair.

3. **Filtering and Constraint:** Chunks with a re-ranking score below a minimum threshold ($S_{min} = 0.1$) are discarded to prevent introducing noise into the prompt. A maximum of the top-5 final chunks ($K_{final} = 5$) is selected to build the context. This constraint prevents overloading the generative LLM's context window, mitigating the risk of the model overlooking relevant information due to "lost-in-the-middle" attention degradation.

4. **Query Caching:** To minimize computational overhead and response latency, retrieval results are cached in a local in-memory cache with a Time-To-Live (TTL) of 1 hour. This cache is automatically invalidated when new manual versions are uploaded or when documents are modified, guaranteeing that analysts always retrieve the latest authorized tactical guidelines.

3.3 Runtime Interaction and Feedback Loop

During runtime, the system acts as an interactive assistant. The user interacts via a web interface running as a single-page application (SPA) (as illustrated in Figure 2), submitting queries and optional images to a FastAPI [29] backend server. If the user uploads a reconnaissance image, the backend routes the request to a local multimodal VLM (such as MiniCPM-V 8B [26] or Moondream run via Ollama [30]). For textual queries, the system uses a local generative model, such as Qwen-2.5 14B [31]. To provide a highly responsive user experience at the edge, response tokens are streamed back to the frontend in real time using Server-Sent Events (SSE).
The backend dynamically coordinates the retrieval of manual context and builds the system prompt using five distinct layers:

1. **System Identity & Guidelines:** Global instructions setting the assistant's persona, enforcing strict grounding, and demanding source attribution.

2. **Dynamic Learned Rules:** Up to 15 operational constraints compiled dynamically from user feedback to adapt to specific analyst preferences.

3. **Retrieved RAG Context:** The top-K_{final} text passages and image caption summaries relevant to the query.

4. **Message History:** Up to 100 previous turns to retain conversational context during a multi-turn dialogue.

5. **Current Query:** The analyst's latest input, supplemented by raw image data if applicable.

To support continuous workflow alignment, a feedback service captures user thumbs-up/down actions. While positive ratings (thumbs-up) are stored to serve as few-shot exemplars of desired response styles, any comment submitted alongside a negative rating (thumbs-down) is logged and used directly as a prompt correction rule, adapting the model's behavior without resource-intensive fine-tuning. As an illustration of this adaptation mechanism, during system testing a reviewer rated a response containing an unstructured target description with a thumbs-down, submitting the comment: *"Always structure target descriptions using numbered lists to facilitate rapid reading"*. The feedback service automatically registered this comment as an active correction rule, which the prompt builder injected into subsequent queries. Consequently, the generative model began formatting all subsequent target descriptions using structured numbered lists, resolving the layout issue within the next two dialogue turns. This shows that adjusting prompts with feedback rules is a practical way to adapt model outputs in isolated environments without the need for model fine-tuning.

4 Case Study: Electronic Equipment Target Identification

To evaluate the operational utility and cognitive impact of the offline RAG assistant, we designed a case study centered on a highly specialized air operations task: the identification of electronic equipment targets.

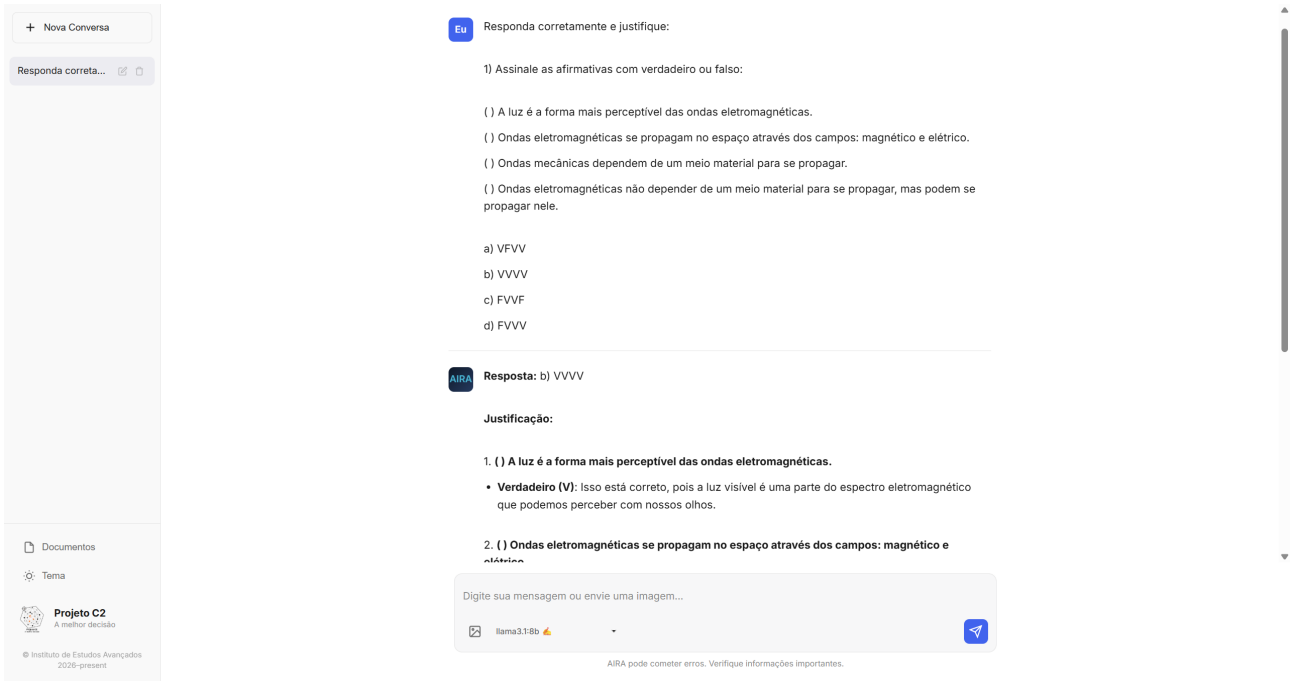

Figure 2 – User interface mockup illustrating a decision support session where the analyst submits a technical query and receives a structured response grounded in the MCA 200-5 doctrine, complete with source citations.

4.1 Doctrinal Context: MCA 200-5

In air operations, military analysts must identify and document relevant tactical assets. For electronic systems, this process is governed by the Brazilian Air Force manual **MCA 200-5 (Manual do Comando da Aeronáutica: Descrição de Alvo, Equipamentos Eletrônicos)** [8]. This technical manual establishes visual parameters, frequencies, power metrics, and functional descriptions for electronic targets, including early-warning radars (antennas, command cabins, power generators), height-finder radars (vertical scanning reflectors), and telecommunication and command posts (tower layouts, antenna configurations, feed horns, and signal distribution nodes). Identifying these targets in aerial or satellite reconnaissance imagery requires analysts to map visual features, such as reflector diameters and array configurations, to the classification criteria defined in the manual.

4.2 Participants

The pilot involved $N = 4$ image analysts from the FAB: two officers and two enlisted personnel, all trained in tactical image interpretation (Table 1). Self-reported experience in image/target analysis ranged from 4 to over 7 years, and self-reported familiarity with the MCA 200-5 ranged from moderate to very high. This sample size is representative of the actual operational environment, as the total pool of military analysts who possess the required target-analysis courses, visual credentials, and operational experience to perform electronic-target interpretation is known to be extremely limited. Given this restricted operational setting and the highly specialized nature of the task, the study is framed as a **pilot evaluation** to characterize baseline behavior and operational feasibility, rather than as a large-scale confirmatory experiment.

Table 1 – Profile of the image analysts participating in the pilot study.

Participant	Role	Experience in image analysis	Familiarity with MCA 200-5
Analyst 01	Officer	>7 years	High
Analyst 02	Officer	4 – 7 years	Moderate
Analyst 03	Enlisted	>7 years	Very high
Analyst 04	Enlisted	4 – 7 years	Moderate

4.3 Pilot Design and Procedure

The pilot study was organized into two sequential blocks administered individually to the four FAB image analysts. This initial phase does not yet include the AI-assisted condition for the human participants; instead, it establishes baseline metrics under a manual workflow, which are compared against the offline language models:

- **Block A: Doctrinal Knowledge Assessment.** Analysts individually answered a ten-item multiple-choice test on the MCA 200-5 fundamentals (electromagnetic waves and antennas) consulting only the printed/PDF manual. Start and end times were self-recorded to capture completion time. To explore the operational trade-offs of deploying different local models, the same ten-item assessment was run independently on three LLM configurations (`qwen2.5:14b`, `deepseek-r1:14b`, and `llama3.1:8b`), enabling a direct comparison of accuracy and speed.
- **Block B: Manual REMIR Construction.** Analysts produced a target-reconnaissance report (*Relatório de Missão de Reconhecimento* – REMIR) for a representative electronic-target image using standard image-analysis tools and the MCA 200-5 manual. Start and end times of the task were self-recorded. Immediately after completing the report, analysts filled out the Raw NASA Task Load Index (RTLX) to characterize the task's cognitive workload.

Since the MCA 200-5 manual and the assessment questions are written in Portuguese, all evaluations (including human tests, model prompts, and RAG outputs) were conducted in Portuguese, testing the models' multilingual capabilities in a specialized domain. The AI-assisted condition, in which analysts interact with the offline RAG assistant, is detailed as part of the future evaluation protocol in Section 6. All model evaluations were conducted locally on a representative tactical edge workstation equipped with an Intel Core Ultra 7 155H CPU, 32 GB of RAM, and an NVIDIA GeForce RTX 4060 Laptop GPU (8 GB VRAM) running Windows 11 and Ollama. Each model was evaluated under two prompt regimes: a *Standard Regime* that required a detailed 2-to-4 line textual justification alongside the answer (essential for auditability), and an *Optimized Regime* that strictly demanded only the letter of the correct alternative (aimed at maximum speed). The standard regime was evaluated using the system's default conversational temperature of 0.3, whereas the optimized regime utilized deterministic greedy decoding (temperature set to 0.0) to maximize consistency on objective multiple-choice selections.

4.4 Evaluation Instruments

This pilot phase relies on two baseline instruments. The first is a **Doctrinal Knowledge Assessment**: the ten-item test (Block A) measures accuracy (number of correct answers out of ten) and completion time. The second is the **Raw NASA Task Load Index (RTLX)** [19], which captures perceived workload during the manual REMIR-construction task (Block B) across six dimensions: Mental Demand, Physical Demand, Temporal Demand, Performance, Effort, and Frustration, each rated on a 7-point scale. The unweighted (Raw) variant was adopted to reduce administration time; no pairwise weighting procedure was used. In this pilot, the instrument was administered once, referenced to the manual condition, to confirm the operational expectation that manually producing a REMIR is a demanding task and to provide a baseline workload measurement for future comparison with the AI-assisted condition.

5 Results

We conducted a pilot exercise with $N = 4$ FAB image analysts (two officers and two enlisted personnel), combining a doctrinal knowledge assessment and a characterization of the cognitive workload associated with manually producing a REMIR under the MCA 200-5. Because of the small, specialized sample, results are reported descriptively (individual scores, means, and standard deviations) and should be interpreted as indicative trends from a pilot study rather than as confirmatory inferential findings.

5.1 Doctrinal Knowledge Assessment: Human vs. Proposed System Results

Table 2 reports the Block A results for the four FAB analysts on the ten-item MCA 200-5 fundamentals test compared with the proposed system running the `qwen2.5:14b` model in its optimized fast-response mode, which serves as the baseline configuration (a comparative evaluation of all three candidate models is detailed in Section 5.2). Although the standard regime is essential for auditability and verification, this initial comparison evaluates the system's baseline accuracy and speed under the optimized regime, with trade-offs between execution speed and explainability analyzed in the subsequent section.

Table 2 – Knowledge test results: Human analysts vs. Proposed System.

Participant	**Score (correct / 10)**	**Time (min)**
Analyst 01	10	26.0
Analyst 02	10	42.0
Analyst 03	10	23.0
Analyst 04	8	15.0
Proposed System	8	7.1

Three of the four analysts answered all ten items correctly, and the fourth scored 8/10, showing that the underlying doctrinal fundamentals are well established within this group. Completion time, however, varied considerably (15 to 42 minutes, with a mean of 26.5 minutes), reflecting differences in individual search strategy rather than differences in accuracy. The proposed system scored 8/10, matching the score of Analyst 04 and completing the exam in 7.1 minutes, which is faster than all four human analysts. This highlights the capacity of offline RAG architectures to deliver high-quality doctrinal grounding while maintaining competitive execution speeds.

5.2 Model Evaluation and Prompt Optimization

The results of the local model evaluation under the different prompting regimes are summarized in Table 3 and visualized in Figure 3.

Table 3 – Local model evaluation under different prompting regimes.

Model & Regime	**Score (correct / 10)**	**Avg Time (s)**	**Total Time (min)**
Standard Regime (with Justification)			
`qwen2.5:14b`	8	121.91	20.3
`deepseek-r1:14b`	8	227.90	38.0
`llama3.1:8b`	5	44.28	7.4
Optimized Regime (Answer Only)			
`qwen2.5:14b`	8	42.86	7.1
`deepseek-r1:14b`	8	141.97	23.7
`llama3.1:8b`	7	2.94	0.5

The experimental data reveals that prompt structure is a critical factor for local inference speed. In the standard regime, model response times are highly dominated by the length of the generated response. For the larger 14B models, which exceed the graphics card's memory (8 GB VRAM), processing is heavily constrained by the slow process of transferring the AI model's data back and forth between the computer's main memory (RAM) and the graphics card (VRAM), a process known as memory offloading. Under the optimized regime, the output size is restricted to a single letter option (or limited to internal reasoning steps), bypassing this transfer bottleneck and yielding a substantial speedup across all configurations.

For `llama3.1:8b`, which fits fully within the graphics card memory, the optimized regime dropped the average retrieval-to-response latency to just 2.94 seconds (a 15x speedup). Interestingly, its accuracy also improved from 5/10 to 7/10. This accuracy gain suggests that smaller models are highly susceptible to reasoning drift or self-distraction when forced to generate verbose justifications before

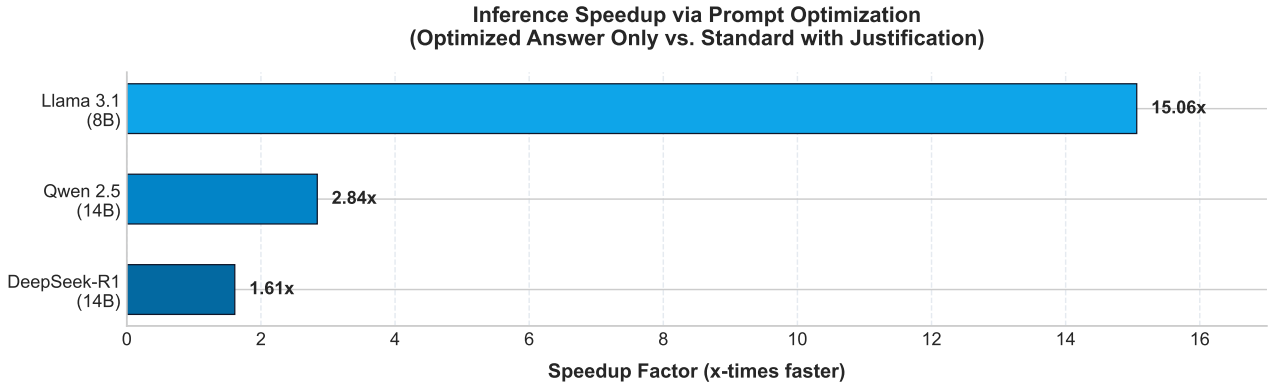

Figure 3 – Inference acceleration (speedup factor) achieved by the optimized prompting regime relative to the standard prompting regime across the three local models.

selecting the final option. In an autoregressive sequence (where the AI generates its response word-by-word, building on what it has already written), an early logical error in the justification text tends to bias the model towards selecting an incorrect option to maintain consistency with its own previous text. Bypassing this step keeps the output focused on option selection. While the temperature difference between the regimes (0.3 vs. 0.0, as detailed in Section 4.3) does not impact execution latency (which depends on prompt-enforced token generation lengths and RAM-to-VRAM transfer overhead), deterministic decoding (greedy decoding, where the model always picks the most likely next word) prevents the sampling of low-probability reasoning paths. Under the optimized regime, the reasoning model `deepseek-r1:14b` maintained its accuracy of 8/10 but was slower than the other two models under the same regime (averaging 141.97s) due to its mandatory reasoning process (the internal "chain-of-thought" generation) which cannot be bypassed at runtime. Overall, `qwen2.5:14b` represents the most balanced trade-off between accuracy and speed for air operations decision support on constrained edge hardware.

5.3 Cognitive Workload of the Manual REMIR Task

Immediately after manually producing the REMIR for a representative electronic target (Block B), each analyst completed the Raw NASA-TLX. Table 4 reports the per-dimension means and standard deviations across the four analysts.

Table 4 – Subjective workload of the manual REMIR task (Raw NASA-TLX, $N = 4$, scale 1–7; SD = Standard Deviation).

Workload Dimension	**Mean $\pm$ SD**
Mental Demand	6.0 ± 1.2
Physical Demand	4.0 ± 1.8
Temporal Demand	4.3 ± 1.5
Performance	2.0 ± 0.8
Effort	5.0 ± 2.2
Frustration	4.0 ± 1.4
Overall (RTLX)	$\mathbf{4.2 \pm 1.1}$

Mental Demand was the highest-rated dimension (6.0 out of 7, with low variability across analysts), and Effort was also rated highly (5.0 out of 7, but with the largest spread, $SD = 2.2$). The Performance dimension uses a scale where lower values indicate better perceived performance (1 = perfect, 7 = total failure). Consequently, a lower score (such as the observed mean of 2.0) reflects a high sense of success, translating to a low workload contribution. Because a higher numerical score on this scale represents poorer performance (i.e., greater failure and thus higher workload), the dimension naturally aligns with the other five (where higher values represent greater demand and workload). Therefore, the overall RTLX is computed as the direct unweighted mean of all six raw scores. The

resulting overall RTLX score (4.2 out of 7) lies above the midpoint of the scale, even for a group with moderate-to-very-high self-reported familiarity with the MCA 200-5 and several years of operational experience. Taken together, these results show that manually cross-referencing visual target features against the MCA 200-5 to produce a REMIR is a cognitively demanding task even for experienced analysts. This strongly supports the case for offline RAG-based decision support; by automating the search and retrieval of complex doctrinal text, the system is positioned to directly reduce mental demand and effort, while maintaining or enhancing the high performance levels (reflected in the low Performance score of 2.0) required in air operations.

6 Conclusions and Future Work

This work presented a modular, secure architecture for offline retrieval-augmented language models as decision support tools in air operations. The system combines local text segmentation (1024-token chunks, 128-token overlap), indexing with the multilingual BGE-M3 model, ChromaDB vector storage, neural re-ranking with BGE-Reranker-v2-m3, and dynamic prompt customization through a local user-feedback service, ensuring that every response is traceable to authoritative doctrinal sources while operating entirely offline.

As an initial step toward evaluating this architecture, we conducted a pilot with four FAB image analysts combining a ten-item doctrinal knowledge assessment based on the MCA 200-5 and a Raw NASA-TLX administration referenced to the manual construction of a REMIR. Three of the four analysts answered all ten knowledge items correctly (the fourth scored 8/10), while completion time varied widely (15 to 42 minutes), suggesting that accuracy is largely established within this group but that retrieval efficiency is not. The workload results showed Mental Demand (6.0/7) and Effort (5.0/7) as the highest-rated dimensions, with an overall RTLX of 4.2/7 above the scale midpoint. This shows that, even for experienced analysts, manually cross-referencing visual target features against the MCA 200-5 is a demanding task. Given the small, specialized analyst pool, these results should be read as indicative evidence from a pilot evaluation rather than as confirmatory findings.

These results directly motivate the next phase of this work, for which a within-subjects evaluation protocol has already been designed and is ready for execution:

1. **AI-Assisted Condition:** repeating the REMIR-construction task (Block B) with the same analysts using the offline RAG assistant, with a second Raw NASA-TLX administration to assess the change in cognitive workload relative to the manual baseline reported in Section 5.3.

2. **Perceived Usability, Trust, and Adoption:** administering the UMUX-Lite, S-TIAS, and UTAUT Behavioral Intention instruments (using a 7-point Likert scale), which were introduced in Section 2, to assess perceived usefulness and ease of use, trust in the assistant's doctrinal grounding, and intention to adopt the tool operationally.

Beyond this evaluation protocol, future work will focus on five key system-level improvements. First, we plan to develop cross-modal search capabilities by integrating image-matching AI models, such as Contrastive Language-Image Pre-training (CLIP) [32] or Sigmoid Language-Image Pre-training (SigLIP) [33], which connect text and visual data. This will enable direct image-to-image queries, allowing analysts to search for technical diagrams using reconnaissance photos rather than text alone. Second, we will pursue tactical hardware optimization by reducing the model's computational size through quantization and grouping processing tasks through batching, which allows the AI to run efficiently on low-power, portable field servers without requiring high-end workstation hardware. Third, we will introduce automated target description sheets using templates to extract technical details from the analyst's chat logs and compile them directly into standard military reconnaissance reports, eliminating the need to write them manually from scratch. Fourth, we intend to evaluate the porting of this modular architecture to other operational doctrines, such as airspace control or mission planning regulations, to demonstrate its generalizability across different military publications. Fifth, we plan to implement automated verification metrics to mathematically assess faithfulness and relevance, ensuring that the assistant's outputs do not deviate from the retrieved manual text before they are presented to the operator.

7 Acknowledgments

This work was supported by the C2 Project — Command and Control from Massive Aerospace Data Fusion in High-Performance Computing Environments (Agreement No. 01/IEAv/2020), funded by the Department of Aerospace Science and Technology (DCTA), Brazilian Air Force. The authors would also like to thank the anonymous military analysts who volunteered to participate in this study.

8 Contact Author Email Address

Joao P. A. Dantas: jpdantas@ita.br

9 Copyright Statement

References

[1] Mica R Endsley. Toward a theory of situation awareness in dynamic systems. In *Situational awareness*, pages 9–42. Routledge, 2017.

[2] Christopher D Wickens. Multiple resources and mental workload. *Human Factors*, 50(3):449–455, 2008.

[3] David S Alberts and Richard E Hayes. *Understanding command and control*. CCRP Publication Series, 2006.

[4] David A Deptula. Effects-based operations: Change in the nature of warfare. Technical report, Aerospace Education Foundation, 2001.

[5] Raja Parasuraman, Thomas B Sheridan, and Christopher D Wickens. A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics*, 30(3):286–297, 2000.

[6] Robert R. Hoffman, Shane T. Mueller, Gary Klein, and Jordan Litman. Metrics for Explainable AI: Challenges and Prospects, 2018.

[7] Ramesh Sharda, Dursun Delen, and Efraim Turban. *Analytics, Data Science, and Artificial Intelligence: Systems for Decision Support*. Pearson, 2018.

[8] Força Aérea Brasileira. Manual do Comando da Aeronáutica (MCA) 200-5: Descrição de Alvo - Equipamentos Eletrônicos. Technical report, Comando da Aeronáutica, Comando de Operações Aeroespaciais, Brasília, DF, Brazil, 2020.

[9] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, 2020.

[10] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 2023.

[11] Saleema Amershi, Daniel Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul Bennett, Kori Inkpen, et al. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2019.

[12] Joao P. A. Dantas. Offline Language-Based Support System for Tactical Decision-Making in Air Combat Operations. In *Conference on Neural Information Processing Systems (NeurIPS), LatinX in AI (LXAI) Workshop*, 2025.

[13] Joao Dantas, Paulo Silva Filho, Jelton Cunha, and Gabriel Dietzsch. HARBOR: Heading Analysis and Reconstruction from Behavioral Observation and Radar. In *AIAA AVIATION 2026 Forum*, 2026.

[14] Samara R. Silva, Vitor C. F. Gomes, Alessandro O. Arantes, Andre F. M. Caetano, Victor L. D. B. Costa, Adrisson R. Samersla, Felipe L. L. Medeiros, Yuri D. Ferreira, Marcia R. C. Aquino, and Joao P. A. Dantas. AsaFG: A Human-in-the-Loop Integration Module for Air Combat Simulations. *IEEE Access*, 13:155821–155834, 2025.

[15] João Paulo de Andrade Dantas. *Simulation and Machine Learning for Decision Support and Autonomy in Air Combat Operations*. PhD thesis, Instituto Tecnológico de Aeronáutica, 2025.

[16] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Eliza Rutherford, Serkan Cabi, Junlong Han, Zhaoming Gong, Tengda Xu, Aaron van den Oord, Koray Kavukcuoglu, Andrew Zisserman, and Karen Simonyan.

Flamingo: a Visual Language Model for Few-Shot Learning. In *Advances in Neural Information Processing Systems*, 2022.

[17] Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, Qiang Liu, Kriti Aggarwal, Zewen Chi, Johan Bjorck, Vishrav Chaudhary, Subhojit Som, Xia Song, and Furu Wei. Language Is Not All You Need: Aligning Perception with Language Models. *arXiv preprint arXiv:2302.14045*, 2023.

[18] Gabriel Dietzsch and Elcio Hideiti Shiguemori. Beyond Accuracy: Performance and Behavioral Evaluation of Multimodal AI for Suspicious Aerial Traffic Monitoring. *The International FLAIRS Conference Proceedings*, 39(1), 2026.

[19] Sandra G. Hart and Lowell E. Staveland. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In Peter A. Hancock and Najmedin Meshkati, editors, *Human Mental Workload*, volume 52 of *Advances in Psychology*, pages 139–183. North-Holland, 1988.

[20] James R. Lewis, Brian S. Utesch, and Deborah E. Maher. UMUX-LITE: when there's no time for the SUS. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*, pages 2099–2102, New York, NY, USA, 2013. Association for Computing Machinery.

[21] Jiun-Yin Jian, Ann M. Bisantz, and Colin G. Drury. Foundations for an Empirically Determined Scale of Trust in Automated Systems. *International Journal of Cognitive Ergonomics*, 4(1):53–71, 2000.

[22] Melanie J. McGrath, Oliver Lack, James Tisch, and Andreas Duenser. Measuring trust in artificial intelligence: validation of an established scale and its short form. *Frontiers in Artificial Intelligence*, 8:1582880, 2025.

[23] Viswanath Venkatesh, Michael G. Morris, Gordon B. Davis, and Fred D. Davis. User Acceptance of Information Technology: Toward a Unified View. *MIS Quarterly*, 27(3):425–478, 2003.

[24] Artifex Software, Inc. PyMuPDF: Python bindings for the MuPDF library. `https://pymupdf.io`, 2026.

[25] Jaided AI. EasyOCR. `https://github.com/JaidedAI/EasyOCR`, 2026.

[26] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. MiniCPM-V: A GPT-4V Level MLLM on Your Phone. *arXiv preprint arXiv:2408.01800*, 2024.

[27] Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. M3-Embedding: Multi-Linguality, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation, 2024.

[28] Chroma Core. Chroma: The open-source embedding database. `https://www.trychroma.com/`, 2024.

[29] Sebastián Ramírez. FastAPI. `https://github.com/fastapi/fastapi`, 2018.

[30] Ollama. Ollama. `https://ollama.com`, 2026.

[31] Qwen Team. Qwen2.5 Technical Report. *arXiv preprint arXiv:2412.15115*, 2024.

[32] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 8748–8763. PMLR, 2021.

[33] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid Loss for Language Image Pre-Training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11975–11986, October 2023.