Offline Language-Based Support System for Tactical Decision-Making in Air Combat Operations

Joao P. A. Dantas*
Decision Support Systems Subdivision
Institute for Advanced Studies
São José dos Campos, São Paulo, Brazil, 12228-001
jpdantas@ita.br

Abstract

An offline language-based support tool is presented to assist pilot training and tactical decision-making in air combat operations. The system applies retrieval-augmented generation (RAG) to answer natural language questions using content extracted from doctrinal and tactical manuals. Built entirely with open-source components, it performs text preprocessing, embedding generation, and semantic search locally, ensuring full offline functionality. Current capabilities are demonstrated through use cases involving air combat manuals, and potential applications are discussed in onboard mission systems, flight simulators, and training environments. The source code is modular and intended for public release to support future extensions and integration.

1 Introduction

Air combat missions often require fast and reliable access to tactical information that is usually found in lengthy and complex manuals. During planning or even in flight, pilots and mission planners may need to consult this content quickly. However, doing so manually is often slow and impractical, especially in environments where internet access is not available or allowed. The ability to interact with technical documents in a natural and efficient way is increasingly valuable in defense contexts [Möbius et al., 2023].

Recent progress in natural language processing (NLP), especially with large language models (LLMs), has made it possible to build systems that answer user questions in plain language using information from documents [Brown et al., 2020, OpenAI, 2023]. However, most of these tools depend on cloud services, which limits their use in sensitive or disconnected environments such as military aircraft or secure simulators.

To address this gap, we present an offline language-based support tool that uses retrieval-augmented generation (RAG) [Lewis et al., 2020] to provide helpful answers based on doctrinal manuals. The system combines document search with a local language model, allowing users to ask questions in natural language and receive relevant, context-aware answers. It is built entirely with open-source tools and works fully offline, making it suitable for use in onboard systems, simulators, and training environments. The system is modular and can be adapted to different types of documents.

The main contribution of this work is the design and implementation of a tactical tool that runs completely offline and helps users access sensitive documents that should not be processed online. We show how it works using real examples from air combat manuals and explain how it can support decision-making and training in air combat situations. We also propose a validation process with experts to check the quality and usefulness of the answers provided by the tool.

*www.joaopadantas.com

Proceedings of the Latinx in AI Workshop @ NeurIPS-25.

2 Related Work

NLP has been increasingly applied in military contexts to enhance decision-making, intelligence analysis, and operational efficiency [Teze and Nazaruka, 2024]. Early initiatives, such as DARPA's TIPSTER program, aimed to improve information retrieval and extraction for defense applications, laying the groundwork for subsequent advancements in military NLP systems [Defense Advanced Research Projects Agency, 1991].

Recent developments have seen the integration of NLP in various military operations. For instance, the U.S. Army's TransApps program developed mobile applications that leverage NLP for tasks like speech recognition, cultural training, and mission planning, demonstrating the versatility of NLP in field operations [Defense Advanced Research Projects Agency, 2010]. Similarly, the MiTAP system, developed by the Mitre Corporation, utilizes NLP to gather, translate, and present information for monitoring global events, aiding in biosecurity and defense intelligence [Damianos et al., 2002].

In the field of intelligence analysis, NLP has been employed to process vast amounts of unstructured data. Primer, an AI company, has developed NLP tools for the U.S. Special Operations Command to analyze and identify disinformation campaigns, showcasing the role of NLP in modern information warfare [Wired, 2020]. Additionally, the COA-GPT model employs large language models to generate courses of action in military operations, highlighting the potential of NLP in strategic planning [Goecks and Waytowich, 2024].

Specific applications of NLP in air combat have also been explored. For example, the Air Combat Evolution (ACE) program by DARPA aims to develop collaboration between AI and human pilots, allowing pilots to act as "battle managers" for unmanned aircraft [The New Yorker, 2022]. Moreover, speech recognition systems have been implemented in advanced fighter aircraft, such as the F-35 Lightning II, to manage cockpit functions like adjusting radio frequencies and autopilot commands, thereby reducing pilot workload [Schutte, 2007].

3 Methodology

The proposed system is designed to provide offline, natural language access to tactical knowledge by combining document retrieval and generative language modeling. It follows a modular architecture with five key elements: document indexing, semantic retrieval, language model interaction, user interface, and offline deployment.

Document Indexing: The pipeline starts by preprocessing and indexing doctrinal and tactical manuals stored locally. Supported formats (.txt, .pdf, .doc, .docx) are converted into raw text. The `all-MiniLM-L6-v2` model from SentenceTransformers [Reimers and Gurevych, 2019] generates fixed-size embeddings, stored in a FAISS index [Johnson et al., 2019] for efficient similarity search. A metadata file links document names to their embeddings.

Semantic Retrieval: When a user submits a natural language question, the system computes its embedding using the same `all-MiniLM-L6-v2` model and performs a nearest-neighbor search over the FAISS index. The top-k most relevant documents (default $k = 3$) are selected, and their contents are concatenated to form a textual context. This context is passed as input to the language model. All retrieval logic is implemented in Python using the FAISS and SentenceTransformers libraries.

Language Model Interaction: The context and user query are forwarded to a locally deployed large language model via the Ollama API [Ollama, 2023]. The model used is a fine-tuned variant of Mistral called `mistral:instruct`, which operates entirely offline. Mistral is an open-weight transformer-based model designed for efficient inference and strong performance on language tasks [Mistral AI, 2023]. A FastAPI [Ramírez, 2018] backend handles requests from the interface or external applications and formats the prompt as a simple concatenation of the retrieved context and the user's question. The response is generated locally and returned to the interface.

User Interface: The interface is built using Tkinter [Python Software Foundation, 2024], a lightweight GUI library for Python. It allows users to type questions in natural language and receive AI-generated answers. The GUI includes support for keyboard shortcuts, error handling, and customization options such as a splash screen and application icon. All interactions with the backend occur through HTTP requests to the local FastAPI server. Figure 1 shows a screenshot of the user interface, where the assistant answers a tactical question using doctrinal content offline.

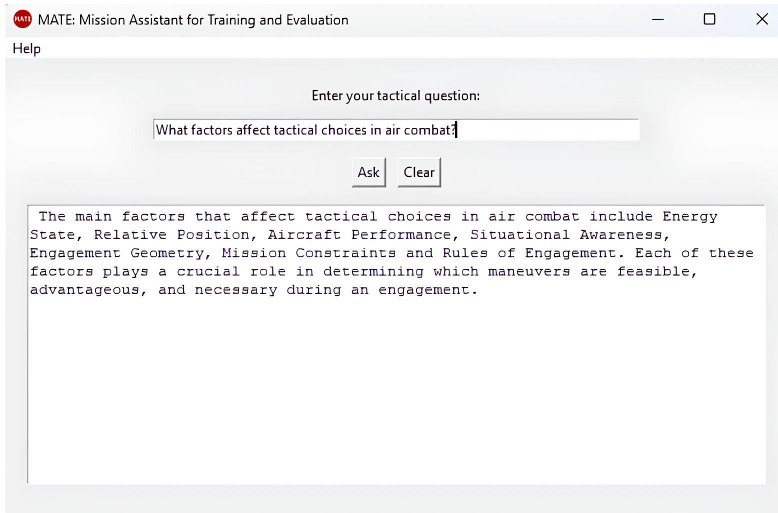

Figure 1: Screenshot of the offline assistant's graphical interface.

Offline Deployment: All components are designed to run entirely offline, supporting use in air-gapped environments. The embedding model, FAISS index, and the language model are all stored locally. The system can be packaged using tools like PyInstaller [PyInstaller Development Team, 2025] to create a standalone executable, and no internet connection is required at runtime. This ensures suitability for sensitive or classified military applications. Figure 2 summarizes the overall architecture and data flow of the system, showing how user queries are processed through semantic retrieval and local language generation.

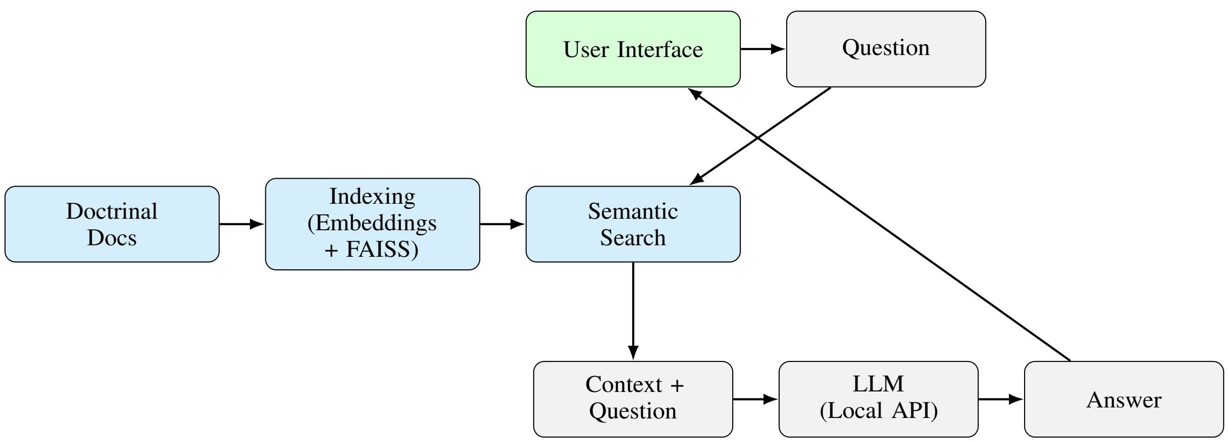

Figure 2: Overview of the offline tactical assistant system architecture.

4 Validation Proposal

To assess the effectiveness of the proposed system in supporting tactical decision-making, we propose a validation procedure based on the analysis of publicly available air combat manuals. As we cannot include or disclose classified military documents in this study, the main reference will be chapter 5, "Section Tactics: Two-versus-One", of Robert L. Shaw's manual "Fighter Combat: Tactics and Maneuvering" [Shaw, 1985]. This chapter provides comprehensive descriptions of section-level tactics, two-versus-one engagements, and the tactical principles relevant to our validation.

The validation will involve extracting tactical concepts and scenarios from the manual and using them to formulate strategic questions that reflect typical queries a pilot or mission planner might have. These questions will serve as a reference to evaluate the tool's ability to retrieve and generate accurate and contextually appropriate responses. The validation process will include the following steps:

1) **Question Formulation**: Strategic questions will be developed, covering aspects of air combat tactics.
2) **System Query**: Each question will be submitted to the tool, and the generated responses will be recorded.
3) **Expert Review**: Subject matter experts (SMEs) in air combat will evaluate the responses.
4) **Scoring**: Responses will be rated based on predefined criteria: relevance, accuracy, and completeness.

The responses will initially be evaluated by a panel of 5 SMEs in air combat tactics. This number was chosen to provide a balance between the diversity of expert opinions and the practical feasibility of conducting a structured assessment. Prior studies suggest that panels of 5 to 10 experts are sufficient for exploratory evaluations involving domain-specific judgment [Okoli and Pawlowski, 2004]. Each SME will independently rate the responses using a 5-point Likert scale [Boone and Boone, 2012], assessing the relevance, accuracy, and completeness of each answer.

To analyze the results, the mean and standard deviation of the scores for each question will be computed. In addition, inter-rater agreement will be assessed using statistical measures such as the Intraclass Correlation Coefficient (ICC)[Koo and Li, 2016] and Kendall's coefficient of concordance (W)[Legendre and Legendre, 2012]. These analyses will help determine the consistency among expert evaluations and the overall effectiveness of the tool in generating doctrinally sound responses.

The following example questions were chosen to reflect important aspects of air combat. They include decisions made during the initial merge, pursuit tactics, offensive maneuvers, energy management, and defensive actions. This helps ensure the tool is tested on different types of tactical situations.

Question 1: What is the primary advantage of flying as a section instead of a single fighter?
Question 2: What are the main advantages of the Fighting Wing formation for new pilots?
Question 3: What are the main weaknesses of the Fighting Wing doctrine?
Question 4: What tactical innovation did the Double Attack introduce compared to Fighting Wing?
Question 5: How does the Loose Deuce differ from the Double Attack in offensive engagement?

The SMEs will evaluate each response using a 5-point Likert scale, ranging from 1 (poor) to 5 (excellent), according to three key criteria: **relevance** (how well the response addresses the question), **accuracy** (the correctness of the information based on the manual), and **completeness** (whether the response covers all key aspects of the question).

The aggregated scores and expert feedback will provide a basis for assessing the system's potential as a tactical decision support tool in air combat contexts. Results from this validation will guide further refinements in retrieval and response generation.

5 Conclusion and Future Work

This work presented an offline tactical assistant that combines document retrieval with a local language model to support decision-making and training in air combat operations. The tool allows users to ask questions in natural language and receive answers based on doctrinal manuals, even in environments without internet access. It was built using open-source tools and follows a modular architecture, making it suitable for integration into simulators and mission planning systems.

An initial set of tests using content from a widely known air combat manual showed that the tool has the potential to produce relevant and accurate answers to strategic questions. These promising results led to the design of a more complete validation process, which has already been planned and will include structured evaluation with SMEs.

Future work will focus on four main improvements: **validation with military personnel**, where the proposed procedure will be tested with fighter pilots and tactical instructors to assess the tool's usefulness in real scenarios; **dialogue history and context awareness**, adding support for follow-up questions and maintaining context across interactions; **integration with simulation environments**, enabling dynamic queries during training sessions; and **voice interface**, developing a voice-controlled version to support hands-free use during flight or simulator-based training.

Offline tools like the one presented in this work may eventually help pilots access relevant tactical knowledge during training, operational missions, or post-mission debriefings. Although still in its early stages, this type of system has the potential to serve as a valuable support tool in simulators or onboard aircraft, particularly in air combat operations that demand quick decision-making under high pressure and without reliable connectivity.

References

H. N. Boone and D. A. Boone. Analyzing Likert Data. *Journal of Extension*, 50(2):Article 2TOT2, 2012. URL https://archives.joe.org/joe/2012april/tt2.php.

T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020. URL https://arxiv.org/abs/2005.14165.

L. Damianos, D. Day, J. Ponte, F. Reeder, G. Wilson, S. Wohlever, and L. Hirschman. Mitap for bio-security: A case study. https://www.mitre.org/news-insights/publication/mitap-bio-security-case-study, 2002. Accessed on 3 May 2025.

Defense Advanced Research Projects Agency. Tipster text program. https://www-nlpir.nist.gov/related_projects/tipster/, 1991. Accessed on 3 May 2025.

Defense Advanced Research Projects Agency. Transformative apps (transapps) program. https://www.darpa.mil/research/programs/transformative-apps, 2010. Accessed on 3 May 2025.

V. G. Goecks and N. Waytowich. COA-GPT: Generative Pre-trained Transformers for Accelerated Course of Action Development in Military Operations. *arXiv preprint arXiv:2402.01786*, 2024. URL https://arxiv.org/abs/2402.01786.

J. Johnson, M. Douze, and H. Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 2019. doi: 10.1109/TBDATA.2019.2921572.

T. K. Koo and M. Y. Li. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine*, 15(2):155–163, 2016. doi: 10.1016/j.jcm.2016.02.012. URL https://www.sciencedirect.com/science/article/pii/S1556370716000158.

P. Legendre and L. Legendre. *Numerical Ecology*. Elsevier, 3rd edition, 2012. ISBN 978-0-444-53868-0. URL https://shop.elsevier.com/books/numerical-ecology/legendre/978-0-444-53868-0.

P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474, 2020. URL https://arxiv.org/abs/2005.11401.

Mistral AI. Mistral 7B: A High-Performance Open-Weight Language Model. https://mistral.ai/news/announcing-mistral-7b, 2023. Accessed: 2025-05-03.

M. Möbius, D. Kallfass, T. Doll, and D. Kunde. Natural Language AI for Military Decision Support and Swarm Control for Autonomous UAS Trained in a Combat Simulation. In *Proceedings of NATO MSG-207: Simulation—Going Beyond the Limitations of the Real World*, 2023. URL https://www.sto.nato.int/publications/STO%20Meeting%20Proceedings/STO-MP-MSG-207/$MP-MSG-207-INV02.pdf.

C. Okoli and S. D. Pawlowski. The Delphi Method as a Research Tool: An Example, Design Considerations and Applications. *Information & Management*, 42(1):15–29, 2004. doi: 10.1016/j.im.2003.11.002. URL https://www.sciencedirect.com/science/article/pii/S0378720603001794.

Ollama. Ollama. https://ollama.com, 2023. Accessed on 3 May 2025.

OpenAI. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*, 2023. URL https://arxiv.org/abs/2303.08774.

PyInstaller Development Team. PyInstaller: Freeze Python Applications into Stand-Alone Executables. https://pyinstaller.org, 2025. Accessed on 3 May 2025.

Python Software Foundation. Tkinter Documentation. https://docs.python.org/3/library/tkinter.html, 2024. Accessed on 3 May 2025.

S. Ramírez. FastAPI. `https://fastapi.tiangolo.com`, 2018. Accessed on 3 May 2025.

N. Reimers and I. Gurevych. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 3982–3992. Association for Computational Linguistics, 2019. URL `https://aclanthology.org/D19-1410`.

J. Schutte. Researchers fine-tune F-35 pilot-aircraft speech system. `https://www.afmc.af.mil/News/Article-Display/Article/154961/researchers-fine-tune-f-35-pilot-aircraft-speech-system/`, 2007. Accessed on 3 May 2025.

R. L. Shaw. *Fighter Combat: Tactics and Maneuvering*. Naval Institute Press, 1985.

V. Teze and E. Nazaruka. Future Directions in Defence NLP: Investigating Research Gaps for Low-Resource Languages. In *International Baltic Conference on Digital Business and Intelligent Systems*, pages 93–105. Springer, 2024.

The New Yorker. The Rise of A.I. Fighter Pilots. *The New Yorker*, 2022. URL `https://www.newyorker.com/magazine/2022/01/24/the-rise-of-ai-fighter-pilots`.

Wired. The AI Company Helping the Pentagon Assess Disinfo Campaigns. *WIRED*, 2020. URL `https://www.wired.com/story/ai-helping-pentagon-assess-disinfo-campaigns`.