Development and Evaluation of a Chatbot to Support Pre-Mission Planning in a Launch and Re-entry Coordination Center Using Retrieval-Augmented Generation (RAG)

Artificial intelligence (AI) in the context of safe and efficient air traffic management [1]

Jens Hampe M.Sc.

German Aerospace Center (DLR), Institute of Flight Guidance, 38108 Braunschweig, jens.hampe@dlr.de

Abstract

The growing number of orbital launch and re-entry operations demands precise and well-coordinated planning.
DLR is developing a Launch and Re-entry Coordination Center Software Suite.
An AI-based chatbot can support planning activities by providing relevant information.

SpaceTracks Suite integrated in the Airport and Control Center Simulator (ACCES) as part of the DLR Air Traffic Validation Center [2]

Objectives

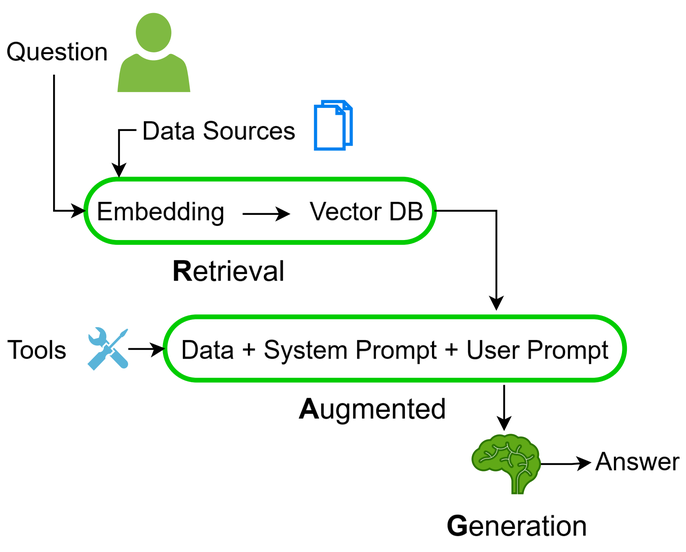
Assist and automate:
LLM-based assistance and automation in pre-mission planning for space launches.

Generate answers:
Context-specific responses grounded in domain-specific resources: procedures, mission documentation, requirements.

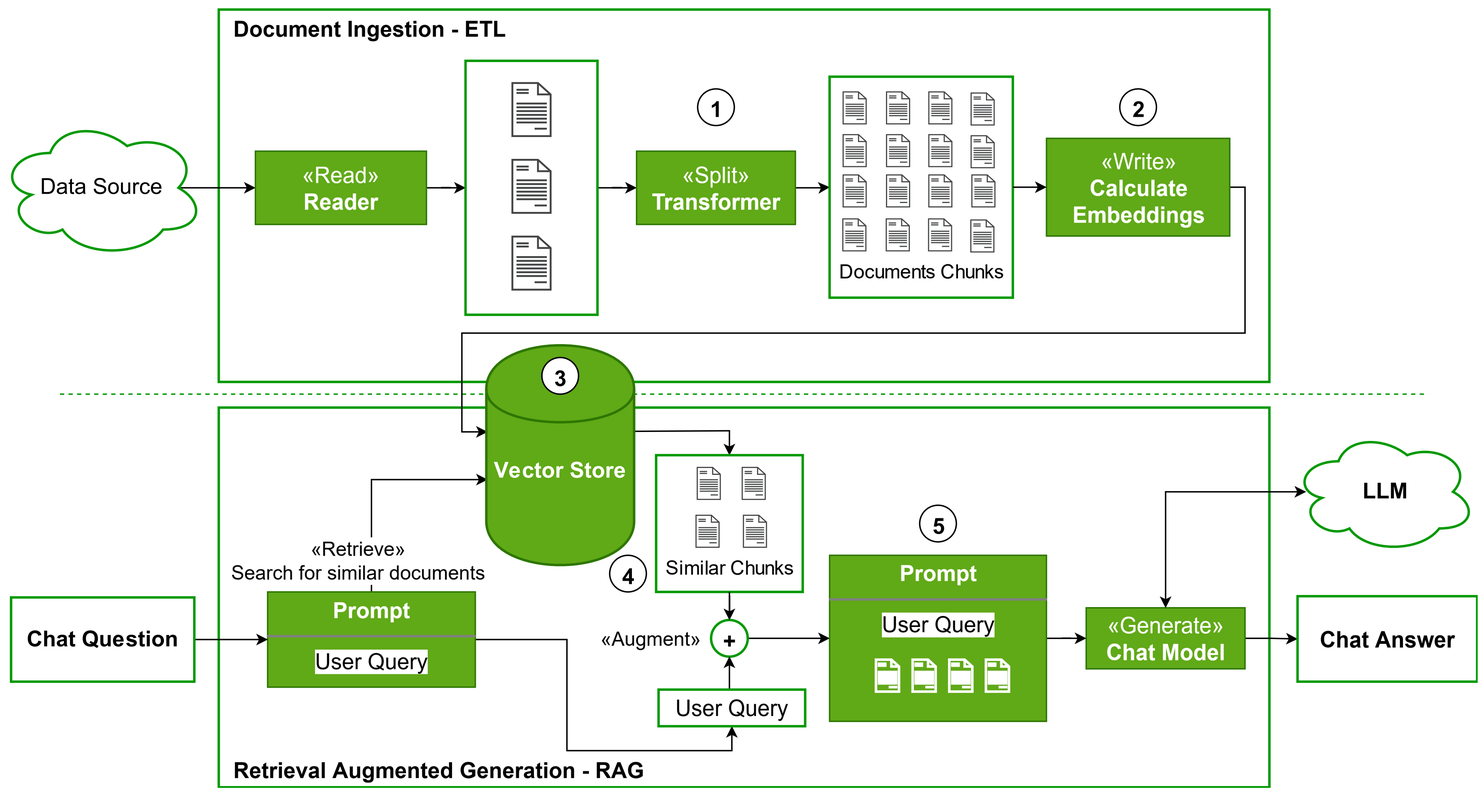
Update and customize:
Model pre-training is generic and not specific to the user's context and needs. Update and Customize domain-specific resources.

Methodology

ETL – Extract Transform Load
RAG – Retrieval Augmented Generation
MCP – Model Context Protocol



Architecture Solution



Architecture Solution using the Spring AI[3] framework, with Qdrant [4] as vector store and Ollama [5] to host the Models

1 Splitter

- To optimize query efficiency, documents of varying sizes are divided into smaller sub-documents (chunks) through an algorithmic splitting strategy. This approach typically enforces a maximum token limit per chunk to ensure manageable data structures.

2 Preparing a Dataset with Embeddings

- Each document chunk is then assigned a set of multidimensional coordinates, known as embeddings, based on its attribute characteristics. These embeddings serve as a numerical representation of the chunk's content in high-dimensional space.

3 Vector Store

- The individual chunks are subsequently stored in a vector database (vector store), which enables fast and efficient searching of documents based on their embeddings. This allows for rapid retrieval of relevant document chunks during query execution.

4 Finding Relevant Data with Unsupervised Machine Learning

- When a query is initiated, embeddings are computed for the query itself and sent to the vector store as query parameters. The resulting embeddings are used to locate the most similar document chunks in high-dimensional space, thereby reducing the search space.
- Semantic search of our dataset using a measure called cosine similarity.

$$cosine\ similarity = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

$$consine\ distance = 1 - cosine\ similarity$$

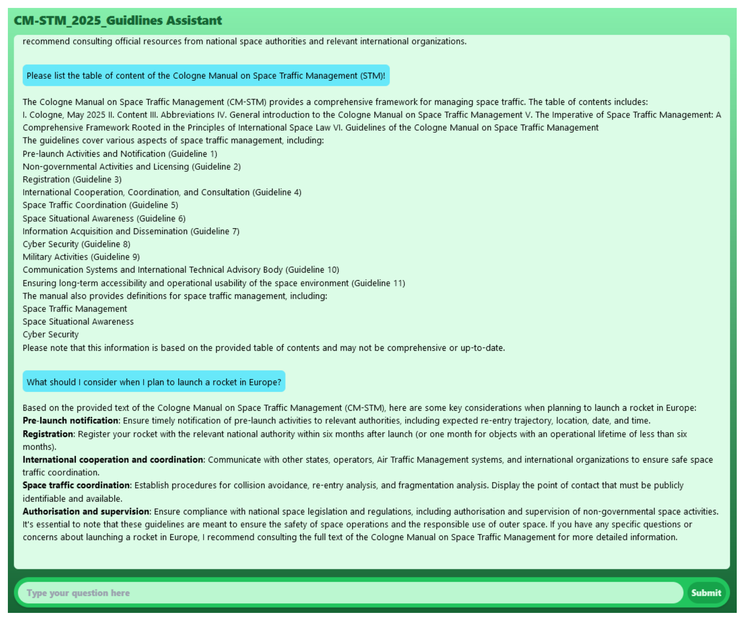
The cosine similarity does not depend on the magnitudes of the vectors, but only on their angle. The cosine similarity interval [-1,+1].

5 Composing a Custom Text Prompt

- The top-ranked document chunks retrieved from the vector store are then selected as context for constructing the prompt. This approach enables efficient information retrieval by leveraging the spatial relationships between query and document content.

Verification Validation

- Spring AI, Qdrant DB, Ollama system.
- The stochasticity of generative AI models makes traditional testing challenging.
- Spring AI's evaluators provide probabilistic assessments of response quality and accuracy by analyzing LLM outputs in real-time, allowing for prompt retries if an unsatisfactory response is generated.
- Metrics via Micrometer [6] gain visibility into all of the Generative AI activity.



Space Guidelines Assistant for CM-STM [7]

Conclusion & Future Work

- LLM-based assistance and automation in planning for space launches.
- RAG enables applications to submit prompts that ask about information that the LLM isn't trained on and can reduce "hallucinations".
- Activating more tool-driven (MCP), multimodal and agentic generation.

References

(1) J. Hampe, Potential Methods and Applications of Artificial Intelligence (AI) in the Context of Safe and Efficient Air Traffic Management, Deutscher Luft- und Raumfahrtkongress 2023, Stuttgart, https://publikationen.dglr.de/?tx_dglrpublications_pi1%5bdocument_id%5d=610387
(2) DLR - Air Traffic Validation Center, https://www.dlr.de/fl/en/desktopdefault.aspx/tabid-1140/
(3) Spring AI, https://spring.io/projects/spring-ai
(4) Qdrant, https://qdrant.tech/documentation/
(5) Ollama, https://ollama.com/ ; (6) Micrometer observability systems, https://micrometer.io/
(7) CM-STM - The Cologne Manual on Space Traffic Management, https://ilwr.jura.uni-koeln.de/cologne-manual/manual

Deutsches Zentrum für Luft- und Raumfahrt e.V. (DLR)
Institut für Flugführung
Lilienthalplatz 7, 38108
0531 295-2588 | jens.hampe@dlr.de
Braunschweig, Deutschland