



Innovation and AI in Intermodal

Developing an AI Chatbot prototype for a train-related use case

21.10.2025

Aldo Puglisi, Head IT & Digital

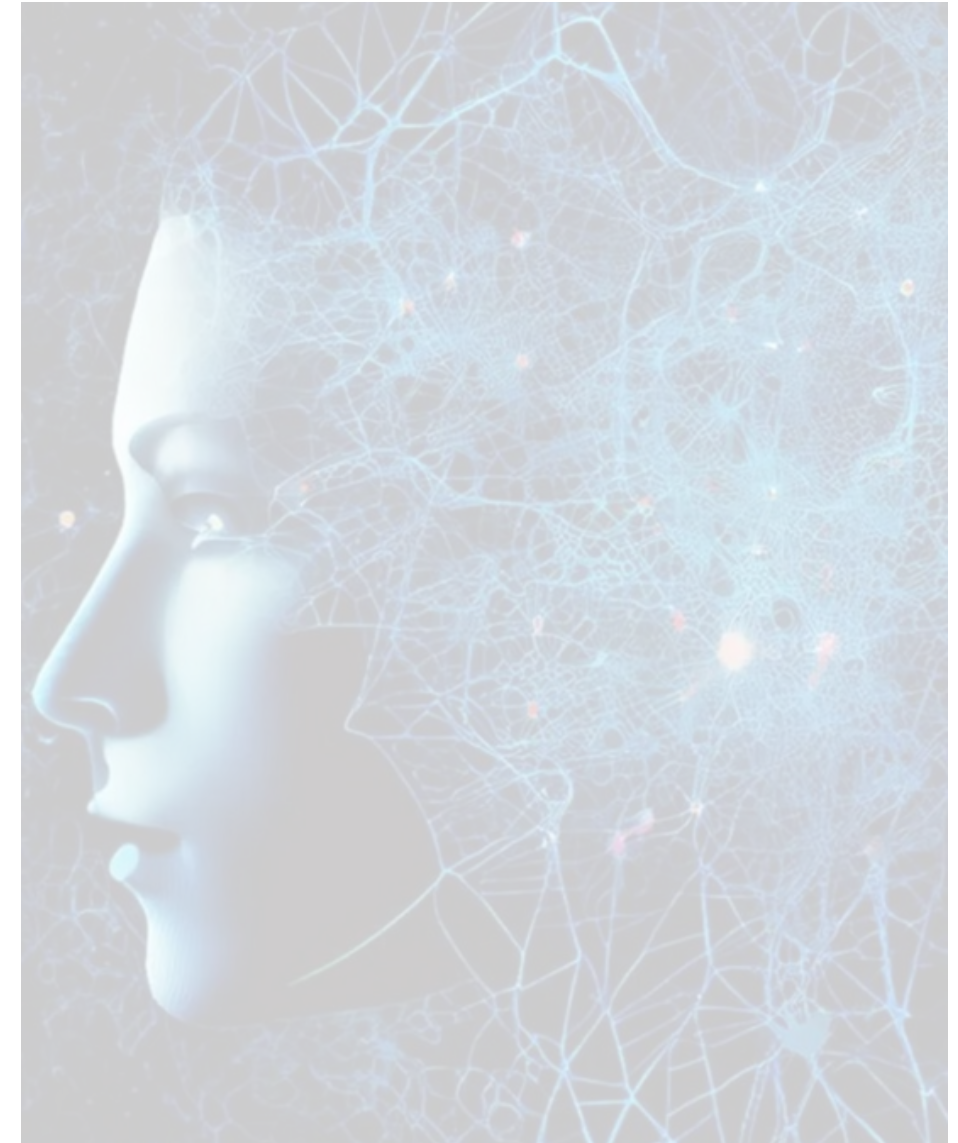


Agenda

- Introduction
- Current Scenario
- Proposed Solution
- Advantages and Challenges of an AI RAG
- Solution Details
- Chatbot Examples Demo
- Evaluation of First Results
- Conclusions

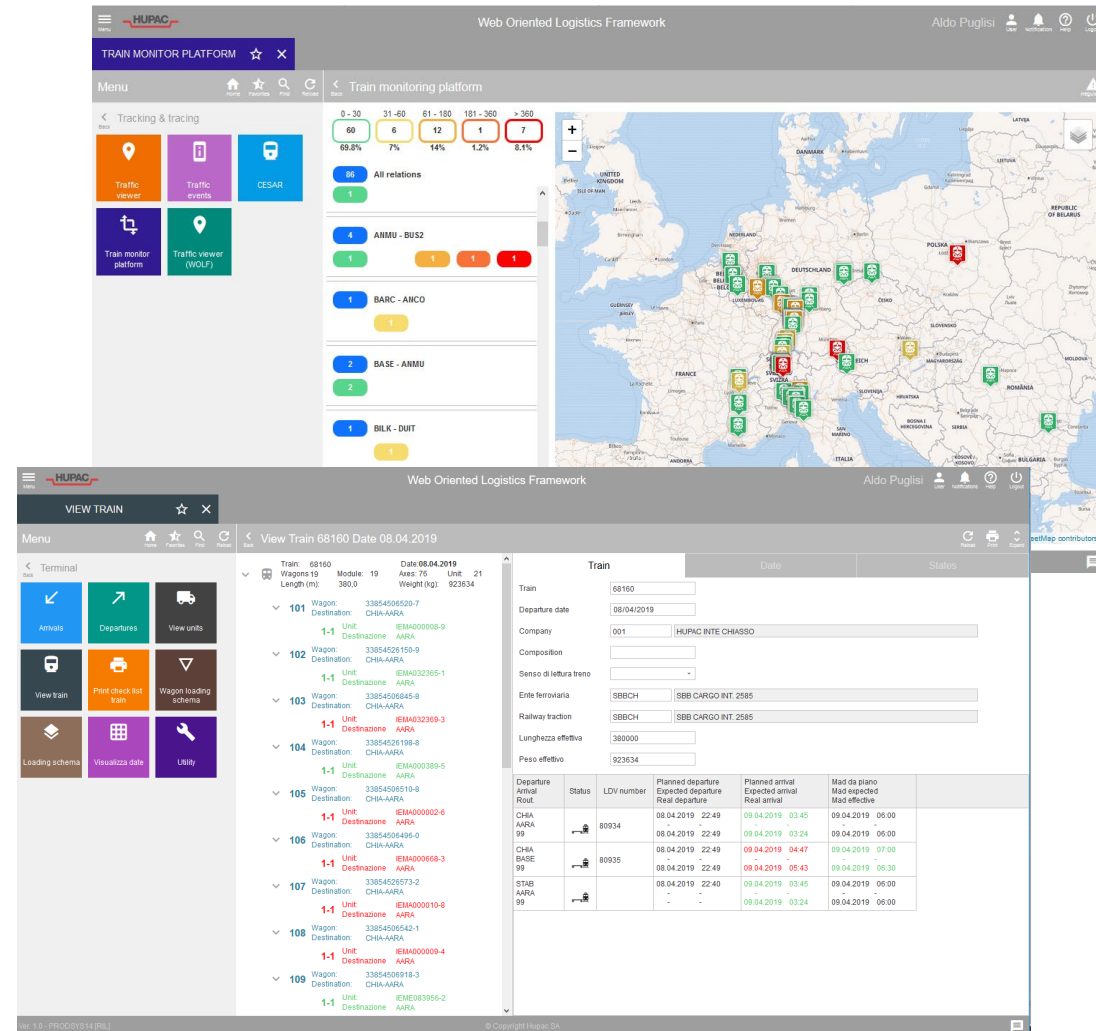


- With **logistics becoming increasingly digital and data-driven**, the demand for **intelligent systems** that can interact with users in **natural language** has grown
- **Hupac AI RAG Project**
 - **Support operations** with intuitive and efficient **access to real-time train data**
 - Leveraged **two Thesis works with SUPSI University**
 - Design and deliver a **prototype of an AI chatbot**



Current Scenario

- Today the users can **access train monitoring data using different sources** (WOLF TMP, web platforms, mobile, ...)
 - Train status (location, ETA/ETP, delays, ...)
 - Train circulation (real time conditions, line information, ...)
 - Irregularities (train, wagon and unit)
- In addition, **large number of documents** are also available to the users (emails, reports, images, EDI streams, ...)



Proposed Solution (1/2)

- **AI Chatbot prototype** that allows operators to **submit requests in natural language** (Italian, English, or German) about train **circulation and disruptions**
 - The user requests may concern train delays, reasons for disruptions, and other information
 - The AI system automatically
 1. Detects the user's intent
 2. Determines the relevant source (query, API, ...)
 3. Returns a concise, structured answer
- The AI chatbot's core logic is powered by **Retrieval-Augmented Generation (RAG)**, enabling the underlying **Generative AI (LLM)** to deliver **grounded answers** based on core and Hupac knowledge data

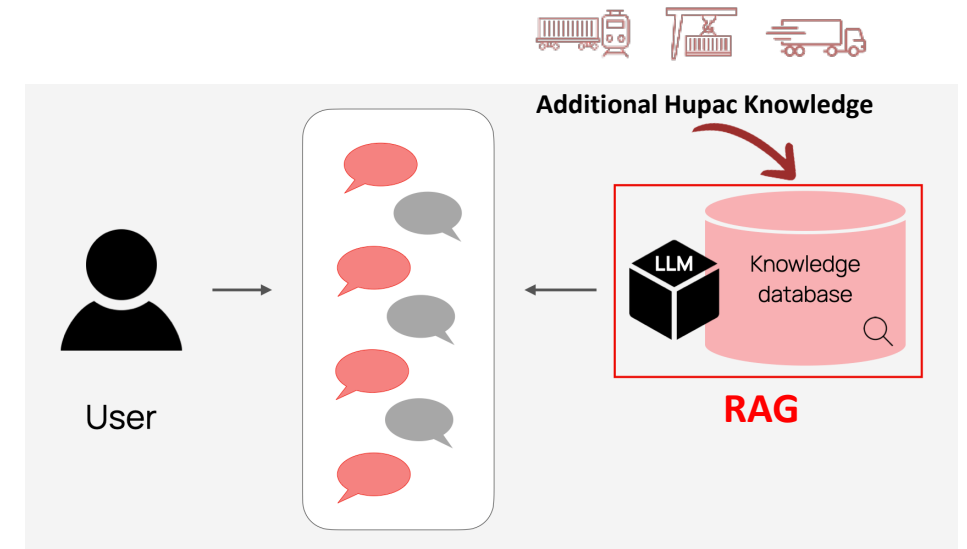
Proposed Solution (2/2)

- **What is LLM?**

- A **Large Language Model** (LLM) is a type of **artificial intelligence program** that can recognize & generate text, among other tasks
- LLM are **very large models** pre-trained on vast amount of data (ChatGPT, Gemini, ...)
- Use a **neural network architecture**

- **What is a RAG?**

- RAG stands for **Retrieval-Augmented Generation** and it is an advanced technique used LLMs
- RAG empowers LLMs to **leverage additional knowledge** for improved performances in various natural language processing rules



Advantages and Challenges of an AI RAG

Advantages

- **Improved accuracy and relevance**
 - Incorporates specific information from a database or dataset, typically in real time
 - Ensures that the output is based on both the model's pre-existing knowledge
- **Customization**
 - Adapt the model's knowledge to a specific domain or use case
- **Flexibility**
 - Possible apply RAG to various structured and unstructured data
 - Update or swap out the data sources as needed, enabling the model to adapt

Challenges

- **Dependency on data quality**
 - Effectiveness of RAG is directly tied to the quality of the data it retrieves
 - If you give a model bad data, it will give you bad results
- **Need for data manipulation and cleaning**
 - Data in the recesses of the company often has a lot of value to it, but it is not often in good, accessible shape
- **Computational overhead**
 - RAG pipeline introduces new computational steps into the response generation process
 - Additional costs have to be considered

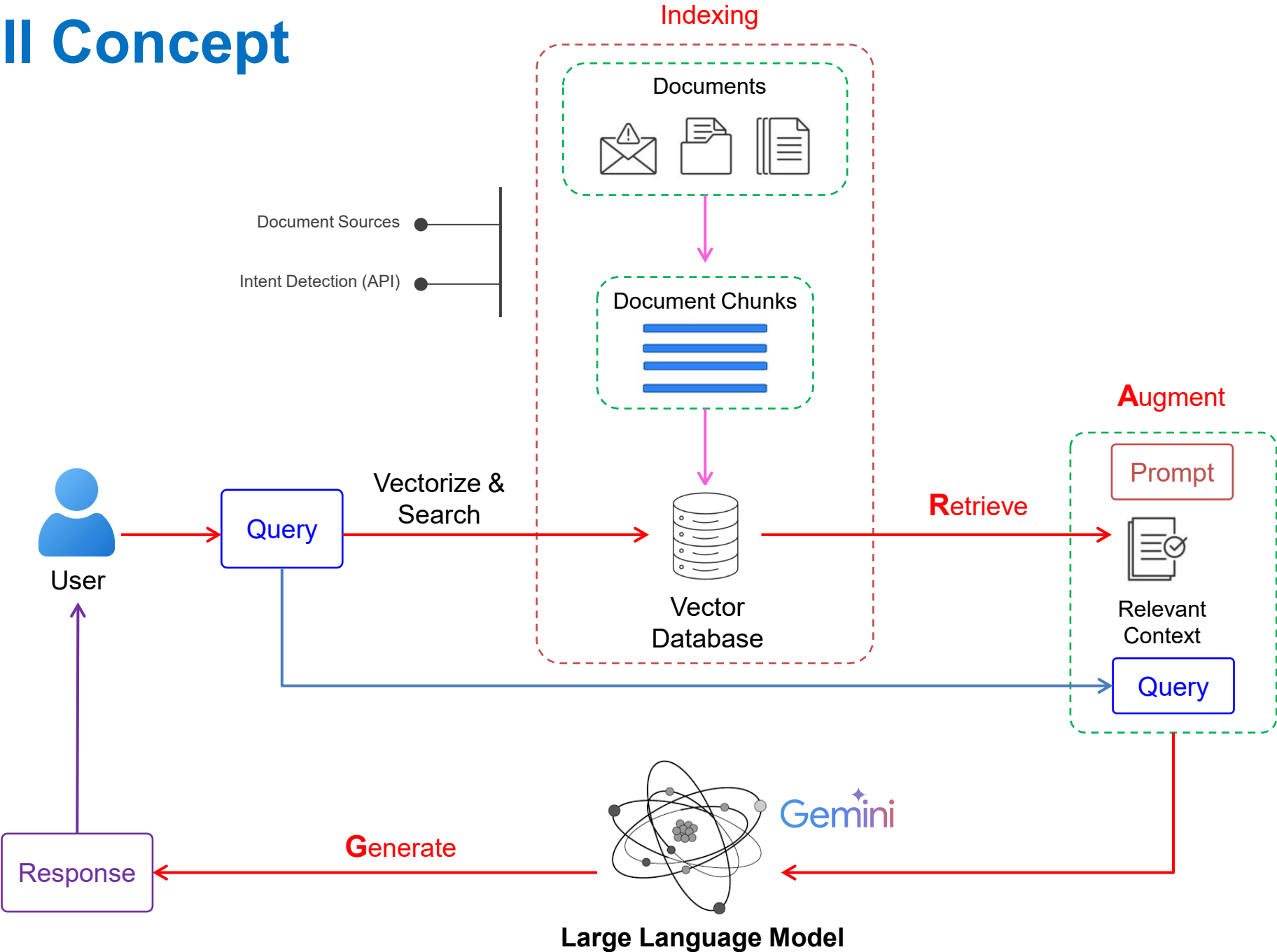
Solution Details

Datasets used for the AI RAG Prototype

- The dataset used for the prototype is based on a **small set of representative of data** provided by the Hupac experts
- The final dataset contains **different samples per language** (DE, EN and IT), distributed evenly across categories
- These initial queries reflected common **information needs encountered for train monitoring**, such as:
 - When is train XXXXX expected to arrive?
 - Where is unit XXXX located?
 - What is the status of train traffic in the group?
 - What is the status of delays on line 1?
 - How many units of dangerous goods is train XXXXX carrying?
 - How is train XXXXX traveling?



Overall Concept



Chatbot Examples Demo

Example 1: chatbot interaction for unit irregularity

HUPAC Chatbot

What happened to unit PV[REDACTED]22-5?

Unit PV[REDACTED]22-5, associated with wagon 33 85 [REDACTED] 750-8, experienced a technical problem with its UTI at Dottikon Umspannanlage. Specifically, its door was not completely secured, and the cam was not engaged at the top and bottom. An inspector fixed these defects without discarding the wagon. Photos, protocols, and a WTU-Protokoll were attached for registration of the irregularities.

Cosa è successo all'unità PV[REDACTED]22-5?

L'unità PV[REDACTED]22-5, sul carro 33 85 [REDACTED] 750-8 del treno 40 [REDACTED] 3 del 2025-06-06 a Dottikon Umspannanlage, ha presentato un problema tecnico alla/e UTI. In dettaglio, la porta non era completamente assicurata e la camma non era agganciata in alto e in basso. Il verificatore ha fissato i difetti senza scartare il carro. Sono stati allegati foto, protocolli e WTU-Protokoll.

Was the verifier involved in the problem with unit PV[REDACTED]22-5?

Yes, the verifier was involved. The verifier fixed the defects without discarding the wagon.

Scrivi la tua domanda...

Invia

Example 2: chatbot interaction for train irregularity

HUPAC Chatbot

What happened to train 40-09 on January 19, 2025?

On January 19, 2025, train 40-09 experienced an irregularity at Taulov. The tarpaulins of semi-trailers, specifically 20 / trailers, loaded on the train were cut. This incident was discovered at 12:00. The irregularity was categorized as VANDALISM / GRAFFITI. Surveillance video confirmed there was no theft. There was no damage to the wagons or injuries. The trailers are being repaired by Taulov and will be able to depart on train 40-09. A police report has been requested.

Was ist mit dem Zug 40-09 vom 19.01.2025 passiert?

Mit dem Zug 40-09 vom 19.01.2025 ist ein Vorfall in Taulov geschehen. Es wurde gemeldet, dass die Planen der auf dem Zug geladenen Sattelaufleger zerschnitten wurden. Dieser Vandalismus wurde um 12:00 Uhr entdeckt, während der Zug um 14:30 Uhr abfahrbereit war. Es gab keine Schäden an den Waggonen oder Verletzte. Taulov repariert die Anhänger, und ein Polizeibericht wurde angefordert. Die war bereits über die Nichtabfahrt des Zuges informiert.

Gibt es Beweise für Unregelmäßigkeiten im Zusammenhang mit dem Zug 40-09 vom 19. Januar 2025?

Ja, es gibt Beweise für Unregelmäßigkeiten im Zusammenhang mit dem Zug 40-09 vom 19. Januar 2025. Der Grund für die Unregelmäßigkeit wird als VANDALISMI / GRAFFITI am Ort TAULOV angegeben. Das Terminal von Taulov hat gemeldet, dass die Planen der auf den Zug 40-09 vom 19.01.2025 geladenen Sattelaufleger zerschnitten wurden. Der Zug befand sich im Terminal und war um 14:30 Uhr abfahrbereit; der Vorfall wurde um 12:00 Uhr entdeckt. Taulov repariert die Trailer. Ein Polizeibericht wurde angefordert und wird gesendet, sobald er vorliegt. Die Unregelmäßigkeiten für jeden Trailer werden so bald wie möglich erstellt.

Gibt es Überwachungsvideos zu Unregelmäßigkeiten im Zug 40-09 vom 19.01.2025?

Ja, die Überwachungsvideos von Taulov bestätigen die Abwesenheit von Diebstählen.

Scrivi la tua domanda...

Invia

Example 3: chatbot interaction for wagon irregularity

HUPAC Chatbot

What happened to the wagon 338-439-4?

The wagon 338-439-4 of train 42-12 had an irregularity at Biasca on 2025-07-18 due to an imbalance/load shift of the UTI cargo. It was subsequently discarded ("scartato") at Biasca on 2025-07-18 at 23:36 due to an axle load ratio alarm (Radlastverhältnis), classified as R3. Following a GICA inspection, the goods inside the unit VZ-477-3 need to be reordered.

Che unità era collegata all'irregolarità del vagone 338-439-4?

L'unità collegata all'irregolarità del vagone 338-439-4 era la VZ-477-3.

Where was the cargo from wagon 338-439-4 placed?

The cargo was not placed elsewhere. It remained within its unit (VZ-477-3) on wagon 338-439-4 at Biasca, where the irregularity was detected and the wagon was rejected. The cargo inside the unit required reordering due to imbalance/displacement.

Scrivi la tua domanda...

Invia

Example 4: asking for additional context

HUPAC Chatbot

What happened to train 4116 11.10.2024?

Train 4116 on 2024-10-11 experienced an irregularity due to FAULT/LACK OF TERMINAL RESOURCES at ANTWERPEN-HAVEN.

Do you have more details about the problem for the train 4116 11.10.2024?

For train 4116 on 2024-10-11, the irregularity was due to a FAULT/LACK OF TERMINAL RESOURCES at ANTWERPEN-HAVEN.

Do you have more details about the problem for the train 4116 11.10.2024?

For train 4116 on 11.10.2024, the irregularity is due to 'GUASTO/MANCANZA RISORSE TERMINAL' (FAULT/LACK OF TERMINAL RESOURCES) at ANTWERPEN-HAVEN. While the specific details are mentioned for train 4117/11, they describe an internet problem at the [redacted] terminal, which prevented the creation of waybills for wagons, forcing the train to be stopped. As of the last update, the internet service provider indicated that the issue would not be resolved before 12:00 today.

Is the Internet problem resolved for the train 4116 11.10.2024?

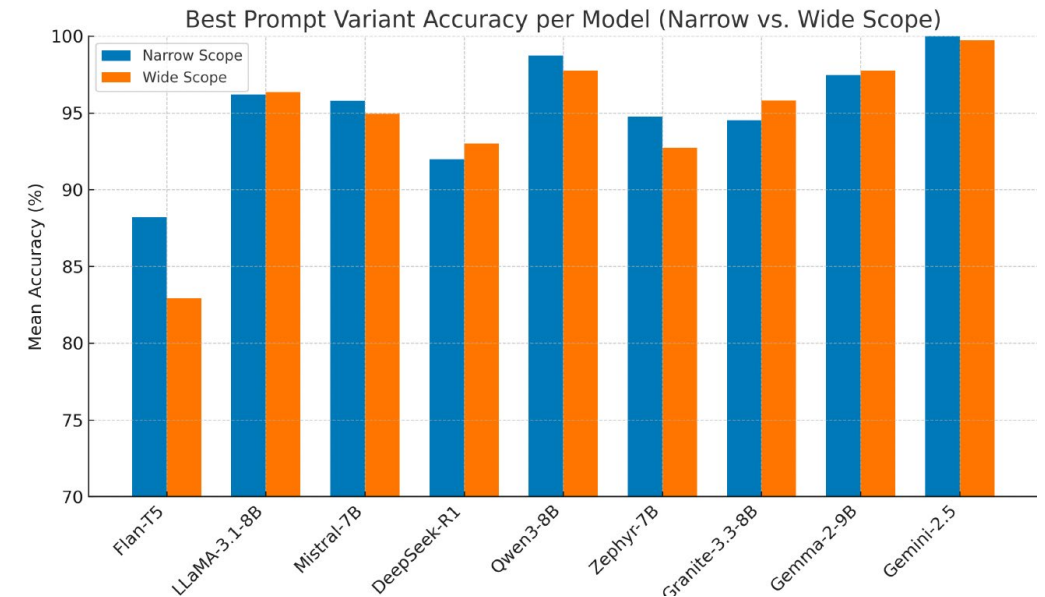
Not yet. The internet service provider informed us that it will be resolved no earlier than 12 today.

Scrivi la tua domanda...

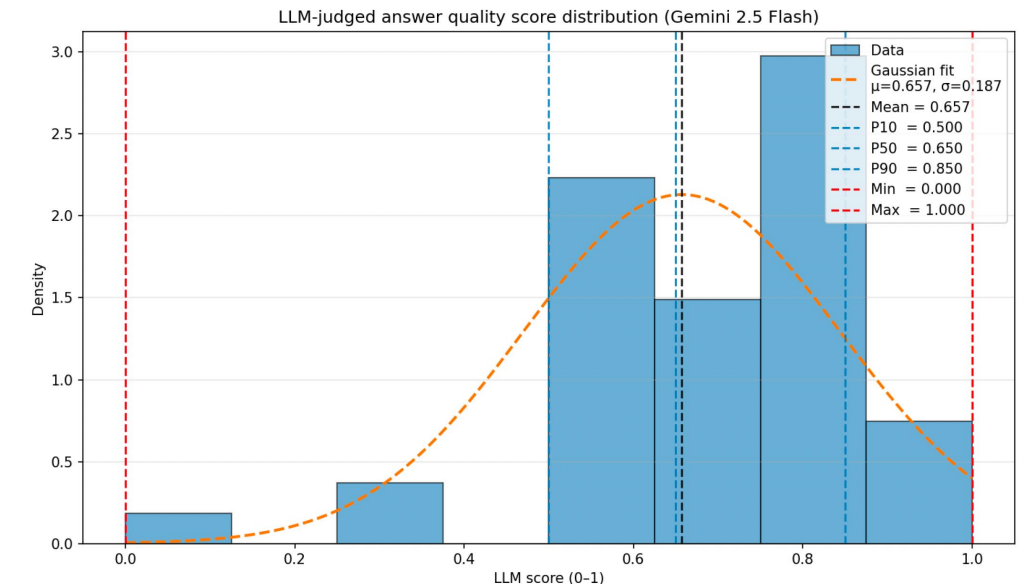
Invia

Evaluation of First Results

Model	Parameters	Context Window (tokens)	Lingual Scope
Gemini-2.5 Flash	N/A	1 M	Multilingual
Flan-T5-XL	3B	512-1024	Multilingual
Meta-LLaMA 3.1 8B	8 B	128 K	Multilingual
Mistral 7B Instruct v0.2	7 B	32 K	English-only
DeepSeek R1 8B	8B	128 K (base)	Multilingual
Qwen3 8B	8B	32 K	Multilingual
Zephyr 7B	7 B	8 K	English-focused
IBM Granite 3.3 8B	8 B	8-16 K	Multilingual
Google Gemma 2-9B	9 B	8 K	Multilingual



- Chatbot tested with different LLM, both proprietary and open source
- Clear dominance of Gemini-2.5, which reaches very good results in both the narrow scope and wide scope
- Qwen3-8B emerges as the strongest competitor



- This thesis set out to design and implement a **prototype of a Retrieval-Augmented Generation (RAG)** architecture tailored to Hupac's train monitoring
- **Main results**
 - RAG system was capable of retrieving and generating information from heterogeneous documentation related to train operations
 - Modular design: ingestion, vector storage, retrieval, language model and the query interface
- **Limitations**
 - Accuracy was satisfactory overall, occasional errors can undermine the retrieval phase
 - The latency, although acceptable, is not optimal for real-time applications evaluations were performed on a limited dataset and primarily with one LLM
- **Future works**
 - Expanding with additional document types, such as PDFs, scanned images and notes
 - Include additional sources for train monitoring and operations

Questions?



Thank you