Intelligent Help Center

An intelligent conversational agent capable of understanding the context of questions and providing relevant answers based on an existing knowledge base.

Our Team

Alexandre Laroudie

Tom Delahaye

Pierre Briand

Gabriel Carlotti

Kentin Guillemot

Aymane Sfouli

We are a team of six students studying for a Master's degree in Data and AI at ESILV.

Our Solution

Main Objective :

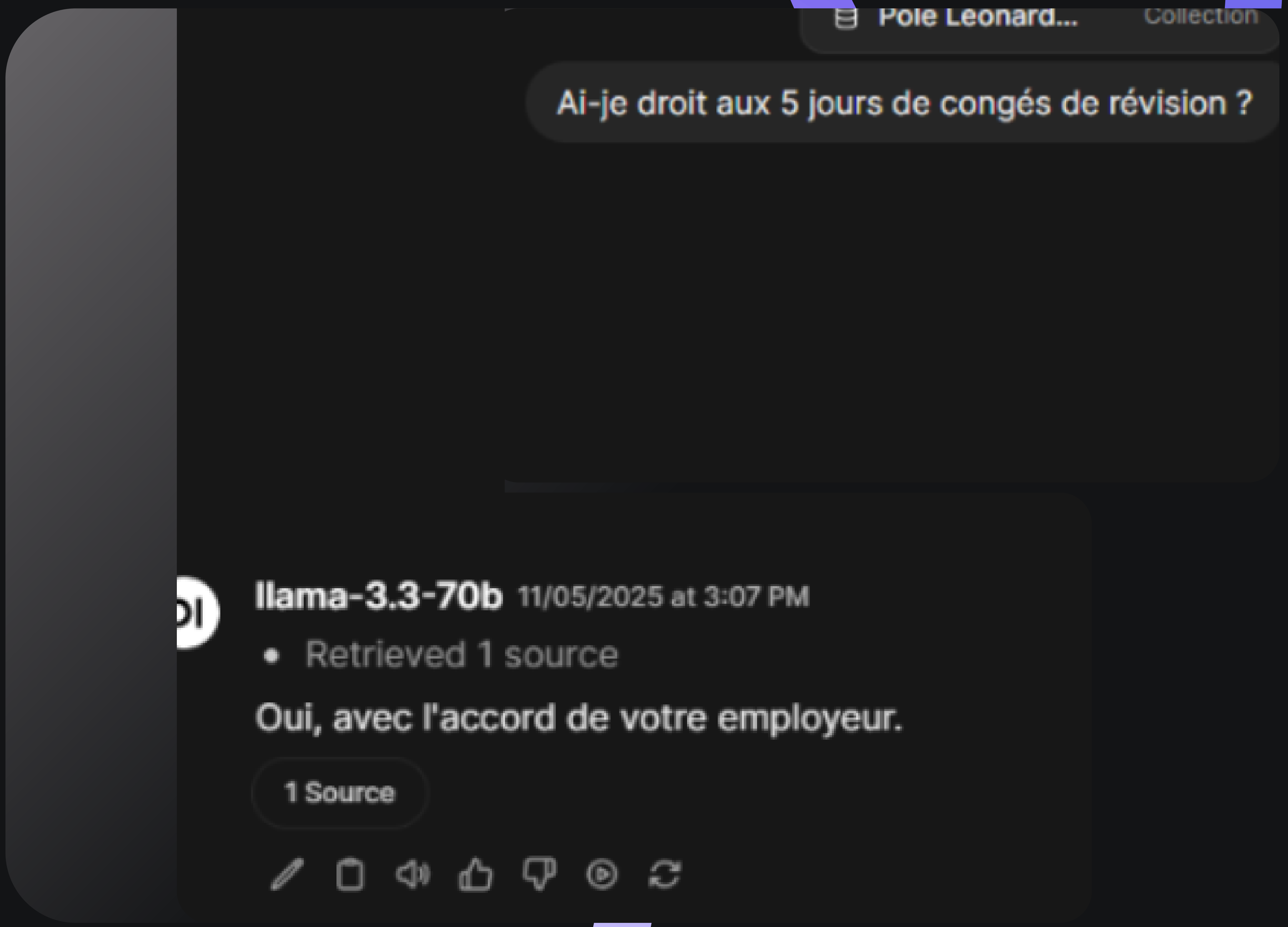
Improve the Help Center user experience and make information more accessible to students of the Pôle Léonard de Vinci.

We developed an Intelligent Help Center using the open-source platform OpenWebUI, enabling us to build a Retrieval-Augmented Generation (RAG) environment capable of:

- Understanding natural language queries.
- Retrieving the most relevant answer from a knowledge base of around 400 question–answer pairs.
- Generating clear, structured, and enriched responses in HTML.

La solution, hébergée sur une machine virtuelle et accessible via une URL sécurisée, est fiable, évolutive et intégrable au futur portail étudiant.

Elle analyse le contexte, fournit la meilleure réponse et, si besoin, redirige vers le bon canal (formulaire ou e-mail).



Ai-je droit aux 5 jours de congés de révision ?

llama-3.3-70b 11/05/2025 at 3:07 PM

- Retrieved 1 source

Oui, avec l'accord de votre employeur.

Real Impact and Added Value

Our solution enhances the Student Help Center by improving information access, response relevance, and user autonomy.

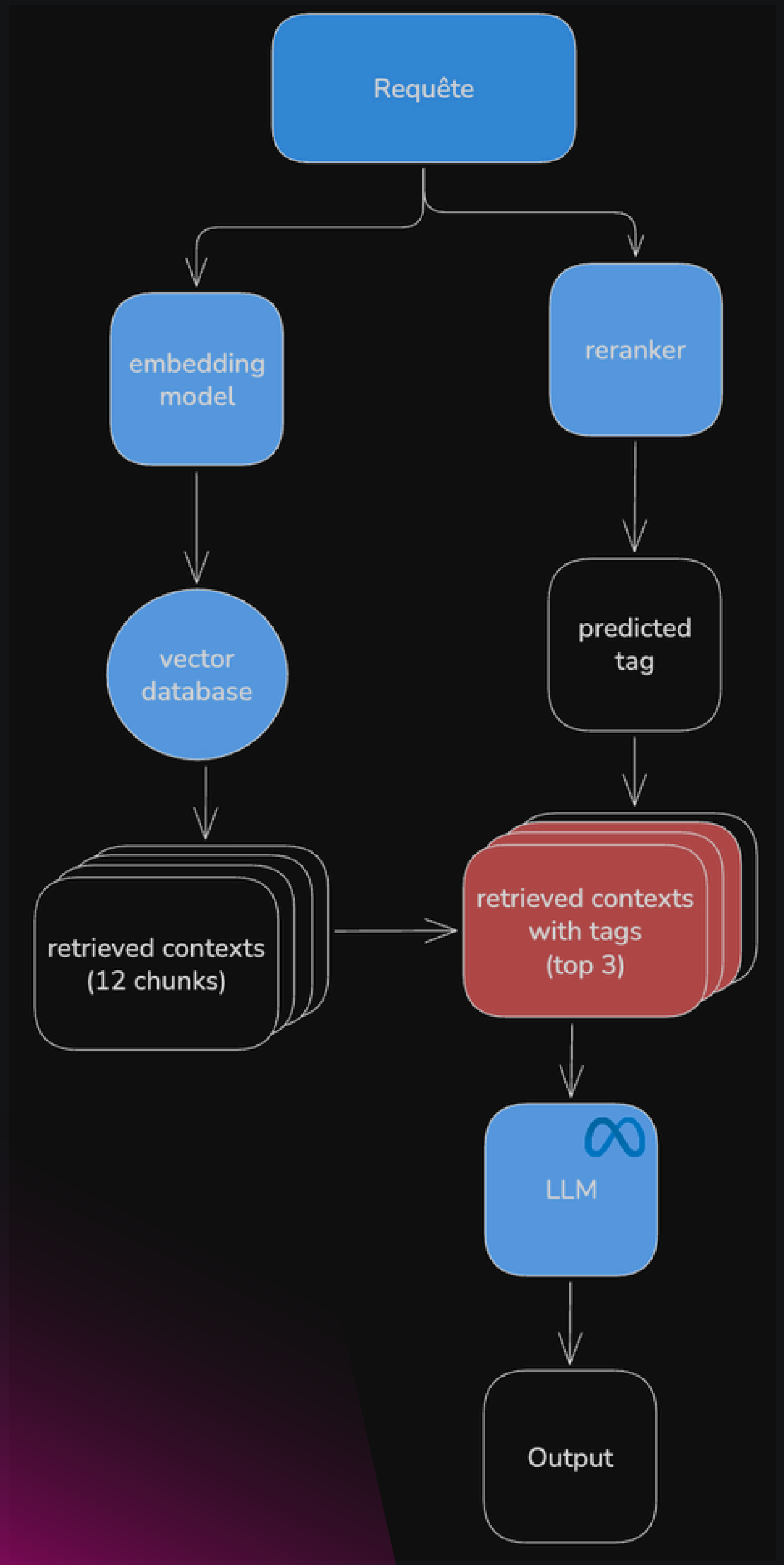
Real Impacts

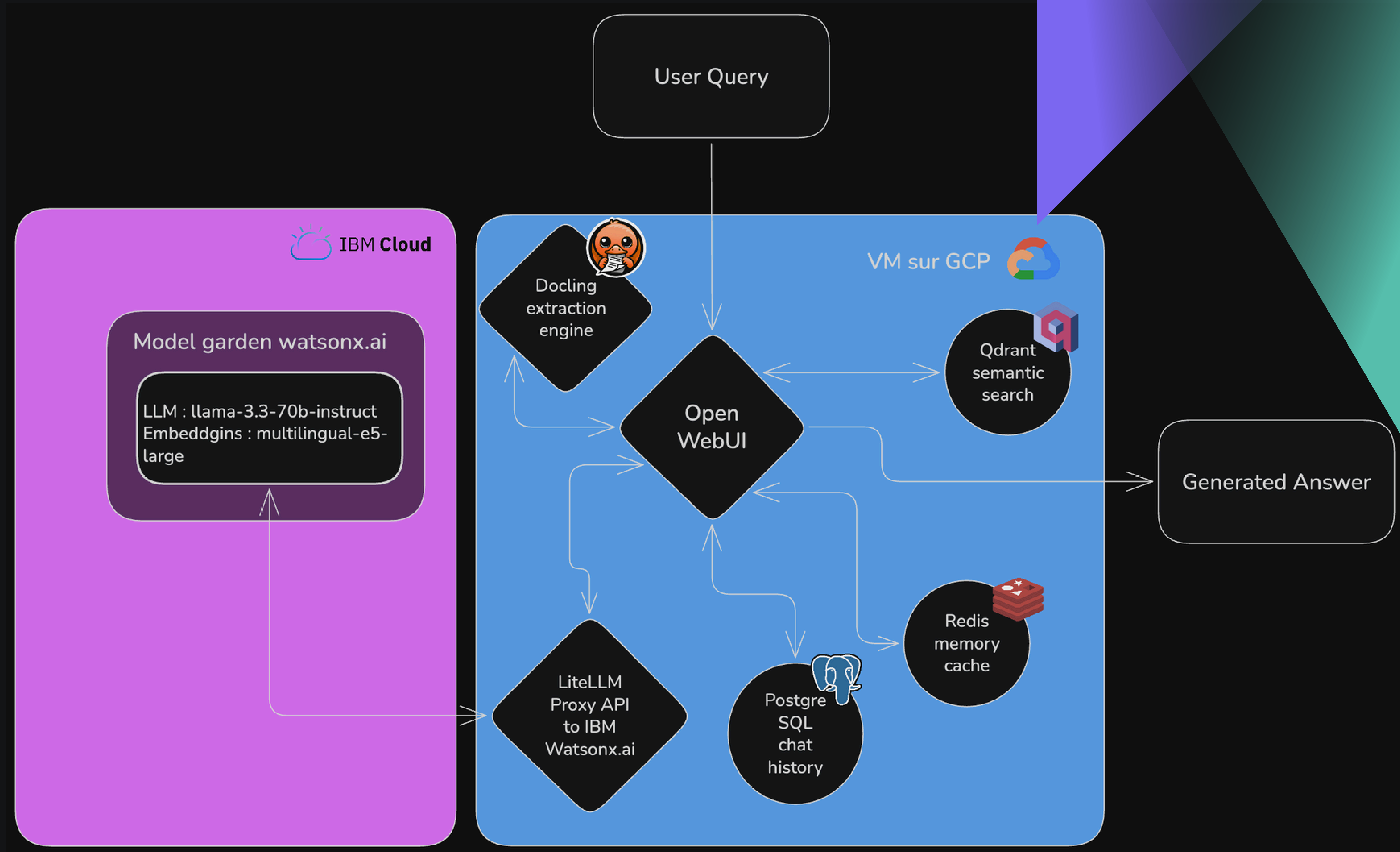
- Significant time savings: students receive immediate answers without having to manually search through dozens of resources.
- Improved user experience: natural language interaction, accessible from any device.
- Centralization and enhancement of existing knowledge: the original Excel knowledge base becomes a structured and intelligent knowledge hub.
- Reduced workload for IT and administrative teams: fewer repetitive requests and better routing to the correct support channels.

Added Value

- Deployment of an intelligent conversational agent using RAG, capable of explaining, contextualizing, and formatting responses in HTML.
- Secure VM-based deployment, accessible through a private URL and ready to be integrated into the student portal.
- Continuous improvement: the system learns from usage logs to refine and expand its knowledge over time.

Technical Architecture





Challenges Encountered

IBM Cloud Environment Constraints

- First-time use of Watsonx.ai and its Model Garden, requiring quick onboarding of APIs and IBM authentication system.
- API call quota limitations (10 requests per project), making testing and iteration difficult.
- Integration issues when connecting Llama 3.3-70B-Instruct and multilingual-e5-large embeddings via LiteLLM due to limited documentation.
- Inter-cloud communication latency (IBM ↔ GCP), increasing response time.

Very Short Development Timeline (2 Days)

- Need to prioritize core features only (dataset ingestion, answer generation, minimal RAG).
- Limited time for load testing and performance optimization.
- Rapid coordination within the team to deliver a functional POC on time.

What We Learned

Technical Skills Gained

- Discovery and integration of Watsonx.ai: understanding Model Garden, embeddings, and the LiteLLM API.
- Implementation of a distributed RAG architecture: connecting the model hosted on IBM Cloud with components deployed on GCP (Qdrant, Redis, PostgreSQL, Docling).
- Full VM deployment: configuration, access management, and performance optimization in a real cloud environment.
- Handling API limitations and request optimization (maximum of 10 calls), requiring careful planning and step-by-step testing.

Human and Project Skills

- Ability to quickly adapt to new tools and technical environments.
- Importance of data quality and prompt design to ensure relevant answers.
- Strengthened teamwork in terms of role distribution and time management under tight deadlines.
- Pride in delivering a functional and accessible solution in just 2 days, secured behind a private URL.