RENATER
CONNECTEUR DE SAVOIRS

Local LLMs and Semantic Similarity (Proxi-tickets)

6th SIG-AI for NRENs Meeting

Vincent Mariller, vincent.mariller@renater.fr

Context & objective

- Increasing interest in **generative AI** across research and education institutions
- Need for **internal tools** that respect
  - data confidentiality
  - infrastructure control
  - operational requirements
- At RENATER, we explored two complementary directions :
  1. **Local LLM platform** for internal use
  2. **Proxi-tickets**, a semantic search tool for support tickets

RENATER
CONNECTEUR DE SAVOIRS

Local LLMs

Deploying sovereign generative AI for internal use

Why local LLMs at RENATER?

- **Growing internal demand for generative AI tools**
  writing assistance, documentation analysis, troubleshooting, coding
- **Confidentiality of internal data**
  operational, HR, administrative and technical information
- **Digital sovereignty**
  avoiding dependence on external commercial AI platforms
- **European ecosystem**
  collaboration with the ILaaS federation

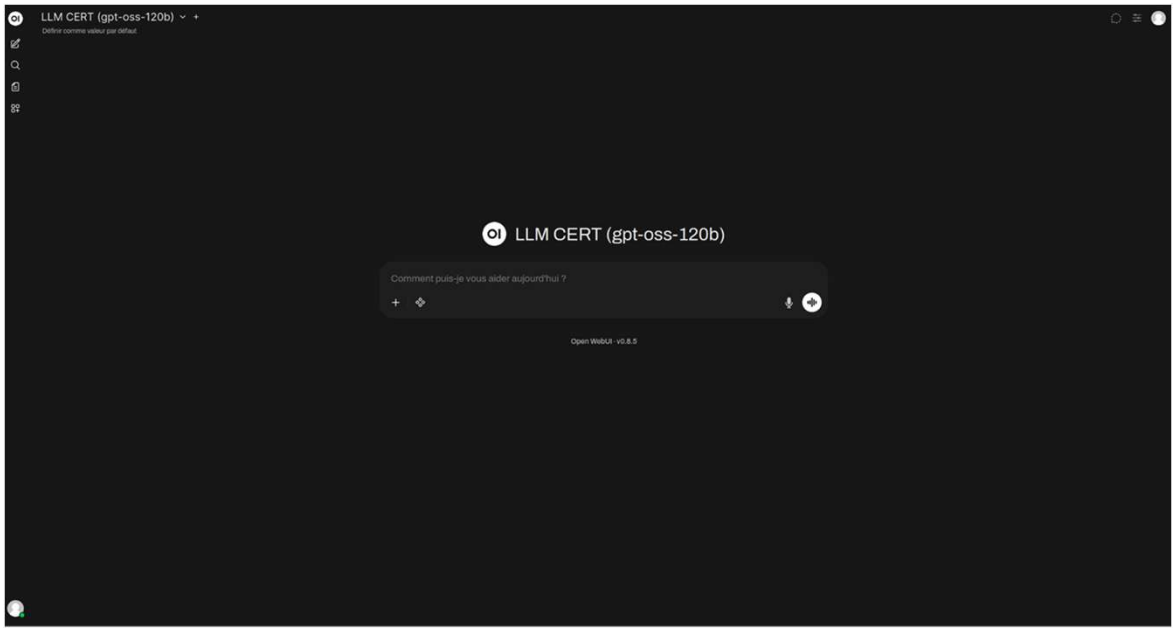
Internal LLM platform

- **Internal access to multiple open-weight LLMs**
  (Llama, Mistral, Qwen, GPT-OSS…)

- **Models can be interacted with through a web interface**
  via Open WebUI

- **Production deployment**
  available to all RENATER staff (~70 users)

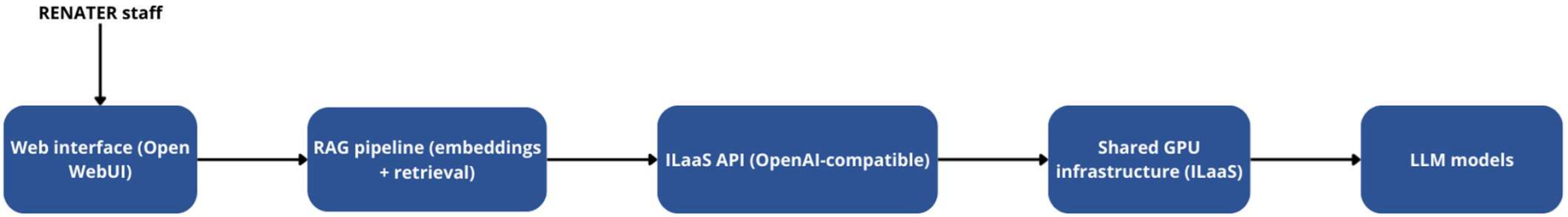
- **Flexible usage**
  writing assistance, technical analysis, coding, troubleshooting



Team knowledge assistants (RAG)

- **Team-specific knowledge bases**
  technical documentation, procedures, regulations, internal guides
- **Retrieval-Augmented Generation (RAG)**
  answers grounded in internal documents
- **Access control by team**
  each team maintains its own knowledge base
- **Supported formats**
  PDF, DOCX, Markdown, TXT, CSV and other internal documents

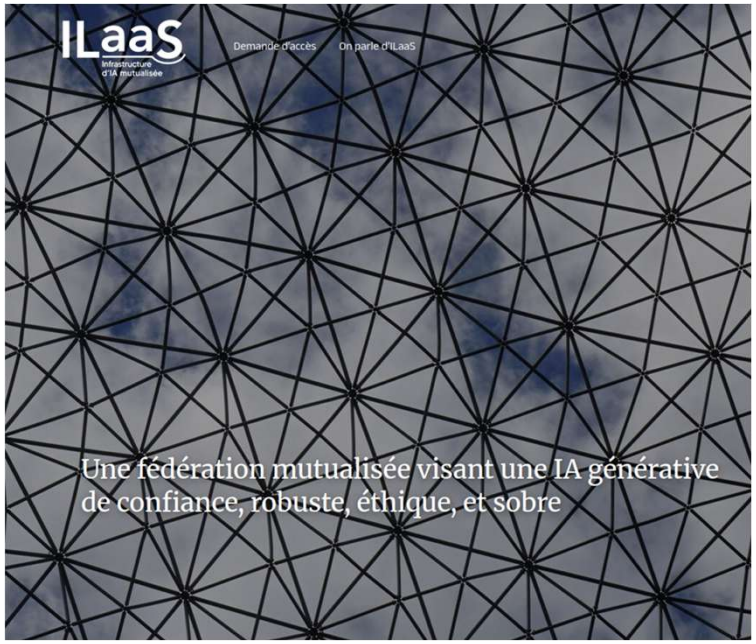
Architecture overview



ILaaS: shared inference infrastructure

https://www.ilaas.fr/

- **Consortium of french universities and research centers, ~1.5 years old**
- **The ILaaS federation provides LLM inference**
- **Shared GPU infrastructure across several institutions**
- **OpenAI-compatible API used by our platform, Mistral partnership**
- **Regular coordination meetings between partners to develop the platform and its services**



Université de Rennes · Université de Reims Champagne-Ardenne · Panthéon Sorbonne · Université de Lille · CentraleSupélec · Université Paris-Saclay · Université de Lorraine

A possible model for NRENs

- **Local LLM platform**
- **Team-level RAG assistants built on internal knowledge**
- **Shared inference infrastructure**
- **Benefits**
  - scalable inference
  - resource mutualisation (digital sobriety)
  - reduced infrastructure cost
  - digital sovereignty

Similar approaches could potentially be explored by other NRENs.

RENATER
CONNECTEUR DE SAVOIRS

Proxi-tickets

A tool for retrieving relevant tickets via semantic analysis

Context & objective

- **Tickets stored in our RT ticketing system**
  No structured database had previously been designed or exported

- **Objective**
  Export and preprocess tickets into a clean dataset that can be used to:

  - Perform business intelligence and in-depth analysis
  - Quickly compare tickets and detect similar issues

Why the need for a dedicated tool?

- **Support tickets contain noisy free text**
  - signatures and automated messages
  - quoted previous emails
  - names, URLs and contact information

→ **Keyword search becomes unreliable**, **semantic similarity is required**

Expected benefits

- **Time savings for support teams**
  Faster identification of similar incidents
- **Analytical insights**
  Identify clusters of related tickets and patterns
- **Structured ticket database**
  Ready for further analysis and BI
- **Issue lifecycle analysis**
  Track how problems evolve over time

Data processing

- **Data cleaning**
  Aggressive removal of noise: URLs, addresses, proper names, etc.

- **Ticket metadata tagging**
  Institution, Applicant, Date, QueueName, Subject, Content, TicketID

⚠ **Limitation**
Message quotes or duplicates cannot be automatically captured

- **Data pipeline**
  Enables easy future updates to the ticket database

Semantic representation

- **Semantic vector representation**

  Each ticket is transformed into a numerical embedding

- **Similarity computation**

  Similarity between tickets is measured using cosine similarity

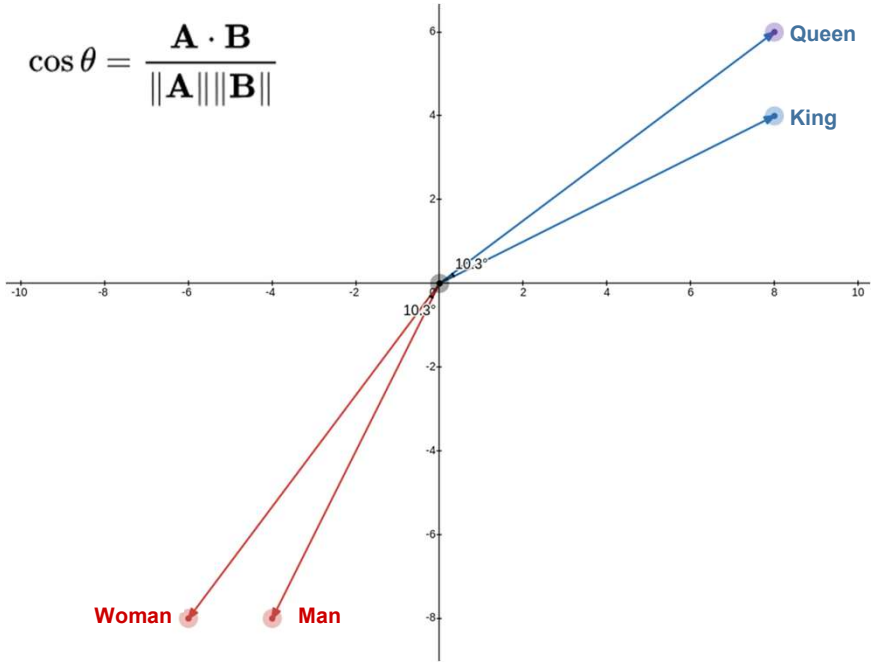
- **Modern embedding models**

  Same type of models used in many NLP systems

A ticket becomes a point in a space with several hundred dimensions.

Tickets that are close together in this space are close together in meaning.

$$\cos\theta = \frac{\mathbf{A}\cdot\mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|}$$



How semantic similarity works

The tool: Proxi-tickets

- **Python application**

  Interface built with Streamlit

- **Semantic search**

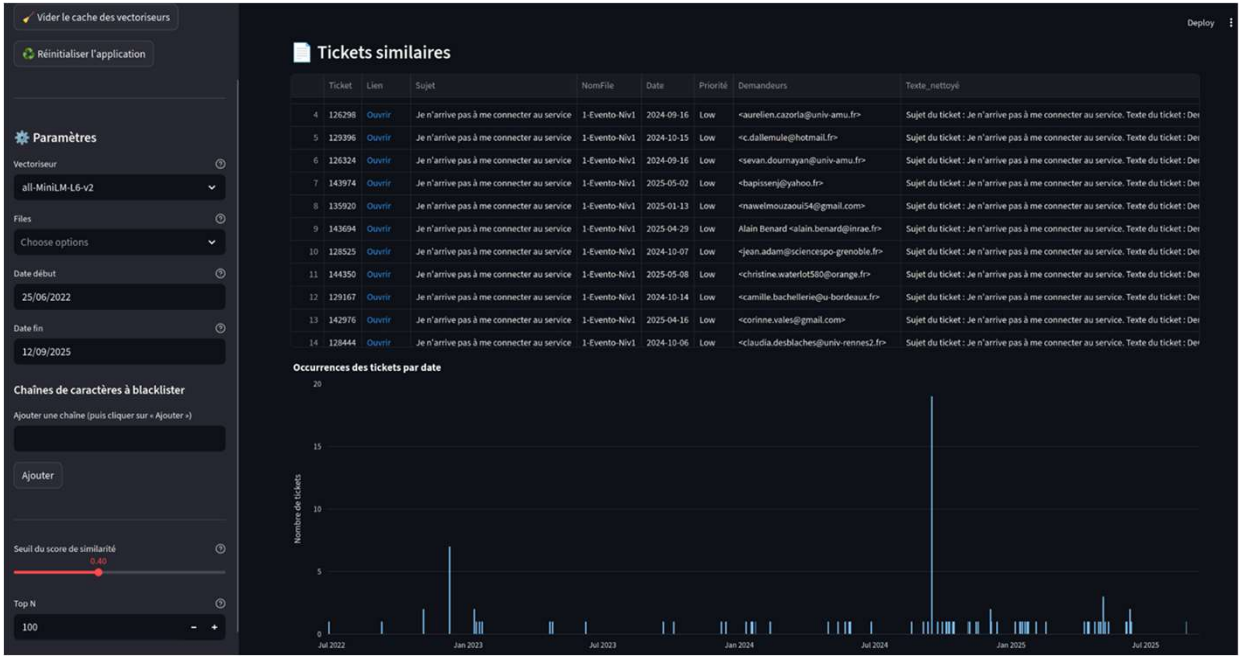
  Compare tickets using four embedding models

- **Flexible filtering**

  Period, queue, similarity score, number of results, blacklisted terms

- **Data exploration**

  Visualization of ticket occurrences and optional AI-assisted analysis



What’s next

- **Local LLMs:**
  - Continuous improvement of internal RAG assistants
  - Exploration of additional internal AI use cases
  - Evolution of available capabilities depending on the ILaaS platform roadmap

- **Proxi-tickets:**
  - Improved AI-assisted analysis using local LLMs
  - Trend analysis over time
  - Automatic ticket clustering
  - Integration with support workflows

Key takeaways

- Local LLM platforms enable **safe internal adoption of generative AI**
- Shared infrastructures such as ILaaS make deployment easier
- Practical use cases like **Proxi-tickets** demonstrate operational value

Q&A

Questions ?

Vincent Mariller

Data Scientist - RENATER

vincent.mariller@renater.fr