haute école
neuchâtel berne jura
arc ingénierie
www.he-arc.ch

RAG-oriented chatbot to help level-1 hotliners

Certifié ISO 9001

Hes·so
Haute Ecole Spécialisée
de Suisse occidentale

Contexte

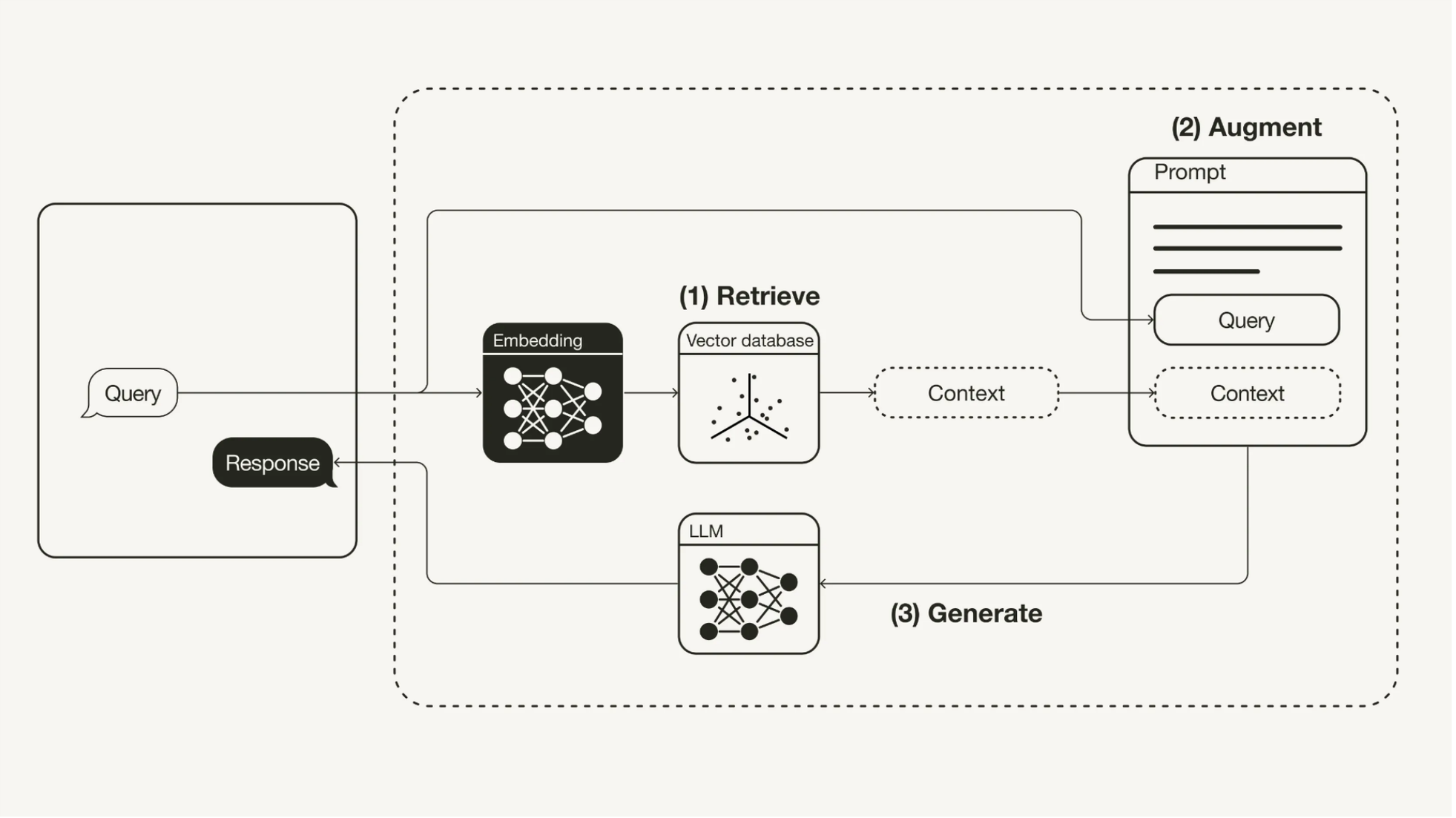
- Chèque de prototypage Swiss AI Center
- Collaboration avec l'entreprise Tornos
- Team de recherche et développement HE-Arc

But et contraintes

Supporter les hotliners dans la recherche d'informations dans la documentation technique

- Multiples sources
- Documentations sensibles
- Fonctionnement uniquement en interne
- Solution valide en anglais et français

Architecture RAG



Source: https://towardsdatascience.com/retrieval-augmented-generation-rag-from-theory-to-langchain-implementation-4e9bd5f6a4f2

Mise en place du RAG

A la mise en place du LLM

1. Choix du corpus à traiter
2. Paramètres de chunks à définir
3. Vectorisation du corpus
4. Choix des paramètres de recherche par similarités

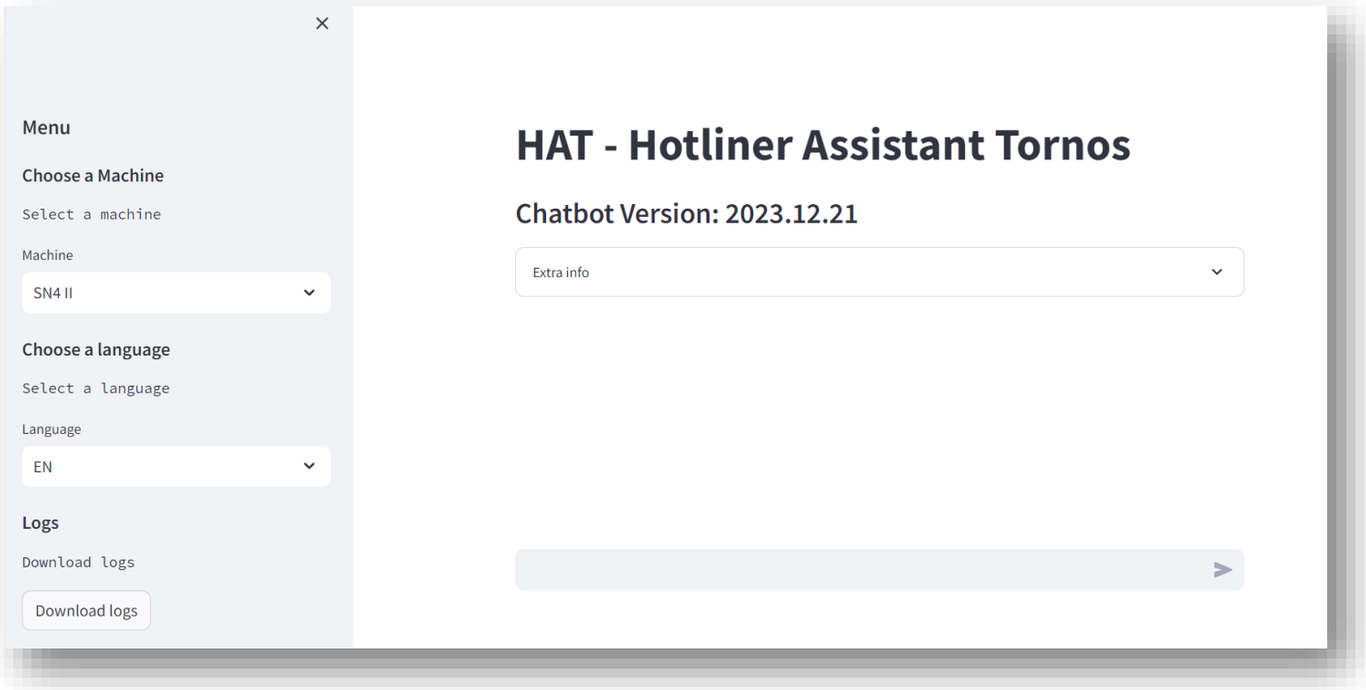
Au moment de l'interaction

1. Traitement de la requête
2. Recherche de similarités
3. Injection des chunks dans le contexte du prompt
4. Génération de la réponse du LLM

Solution technique

LLM	Dolphin2.2-mistral
Interfaçage LLM	Ollama, Langchain
Embeddings	BAAI/bge-large-en-v1.5
Mise en place	Docker, Docker compose
Code	Python
Interface (UI)	Streamlit

Résultats



- Résultats prometteurs – dépendant de l'information disponible
- Traitement et récupération de l'information est une étape critique du processus
- Réception positive
- Performances techniques acceptables

Merci pour votre attention

Disponible pour des questions sur cette présentation ou l'IA en général:

emmanuel.desalis@he-arc.ch