Rapport de Veille Technique

ASSISTANT FAQ INTELLIGENT

Étudiant : Arnaud Rambourg

Date : 16/01/2026

Domaine : NLP & IA Générative

1. ÉTAT DE L'ART : ARCHITECTURES

1.1 Le RAG (Retrieval-Augmented Generation)

Connecte un LLM à des données externes sans ré-entraînement

- **Fonctionnement :** Récupère des extraits (Retrieval) pour contexte
- **Utilité :** Réduit les hallucinations

1.2 La Recherche Sémantique (Embeddings)

Comprend le sens des questions

- **Vecteurs :** Transformation du texte en nombres
- **Calcul :** Similarité cosinus

2. SÉLECTION DES MODÈLES

2.1 Modèles de Langage (LLM)

Modèle : Zephyr-7B-Beta

- **Pourquoi :** Basé sur Mistral, optimisé pour le chat, excellent en FR
- **Accès :** Via librairie "huggingface_hub"

2.2 Modèles d'Embeddings

Modèle : all-MiniLM-L6-v2

- **Pourquoi :** Compact (384 dimensions), rapide

3. OUTILS & ÉCOSYSTÈME

Modèle Extractif (Stratégie C)

Modèle : roberta-base-squad2

Référence pour l'extraction de réponse exacte

Écosystème de Développement

- **API :** FastAPI (Rapide, Swagger)
- **Tests :** Pytest
- **Indus :** Docker & GitHub Actions

4. CONCLUSION DE LA VEILLE

L'approche retenue privilégie les composants **open source** et **légers**.

Ceci est conforme aux contraintes du client souhaitant une future **souveraineté** sur son hébergement.