Rapport de Veille Technologique

Projet FAQ Intelligent pour Collectivité Territoriale

Étudiant : Arnaud Rambourg

Date : 13/02/2026

Version : 1.0

1. OBJECTIF DE LA VEILLE

Contexte

Le projet vise à créer un assistant FAQ intelligent pour une collectivité territoriale. La veille technologique a pour but d'identifier les solutions d'IA les plus adaptées à ce contexte.

Questions clés

- Quels modèles de langage sont accessibles gratuitement ?
- Comment garantir des réponses fiables (sans hallucination) ?
- Quelle architecture pour un service public 24h/24 ?

2. PANORAMA DES LLM (2024-2026)

Modèle	Éditeur	Taille	Accès
GPT-4o	OpenAI	~1.8T	Payant
Llama 3.1	Meta	8B-70B	Gratuit
Mistral 7B	Mistral AI	7B	Gratuit
Qwen 2.5	Alibaba	7B	Gratuit
Gemma 2	Google	9B	Gratuit

Choix retenu : Qwen 2.5 7B (stabilité API + qualité FR)

3. APPROCHES DE Q&A PAR IA

A. LLM Seul (Génératif)

Le modèle répond uniquement avec ses connaissances internes.

- \+ Simple à mettre en œuvre
- \- Hallucinations fréquentes

B. RAG (Retrieval-Augmented Generation)

Recherche sémantique + génération. Le LLM s'appuie sur des documents récupérés.

- \+ Fiable, traçable, à jour
- \- Plus complexe, dépend de l'indexation

C. Extractif (Q&A)

Extraction de la réponse exacte depuis un document source.

- \+ Zéro hallucination
- \- Réponses parfois incomplètes

4. MODÈLES D'EMBEDDINGS

Les embeddings transforment le texte en vecteurs numériques pour la recherche sémantique.

Modèle	Dimensions	Vitesse
all-MiniLM-L6-v2	384	Rapide
all-mpnet-base-v2	768	Moyen
e5-large-v2	1024	Lent

Choix retenu : all-MiniLM-L6-v2 (meilleur ratio vitesse/qualité)

5. INFRASTRUCTURE & DÉPLOIEMENT

Technologie	Rôle	Choix
FastAPI	API REST	Retenu
Docker	Conteneurisation	Retenu
GitHub Actions	CI/CD	Retenu
Pytest	Tests automatisés	Retenu
HF Inference API	Accès aux modèles	Retenu

6. CONCLUSION & RECOMMANDATION

Synthèse

L'approche **RAG** est la plus adaptée au contexte d'un service public :

- Elle garantit des réponses **sourcées** (pas d'invention)
- Elle s'adapte à la mise à jour de la base documentaire
- Elle offre un bon compromis **fiabilité / coût / latence**

Sources principales

- Hugging Face Documentation (huggingface.co)
- Sentence-Transformers (sbert.net)
- Lewis et al., "RAG for Knowledge-Intensive NLP Tasks" (2020)
- FastAPI Documentation (fastapi.tiangolo.com)