Note de Cadrage

Projet FAQ Intelligent

Étudiant : Arnaud Rambourg

Date : 16/01/2026

Version : 2.0 (Finale)

1. CONTEXTE & OBJECTIFS

Contexte

- **Client :** Val de Loire Numérique
- **Problème :** Agents surchargés (60% temps)
- **Solution :** Assistant intelligent via API REST

Objectifs

- API d'assistante FAQ avec LLM
- Comparer 3 stratégies (Benchmark sur 30 questions)
- Industrialisation (Docker, CI/CD)

2. PÉRIMÈTRE

✅ In Scope

- 3 scripts de stratégie fonctionnels
- Rapport benchmark (Golden Set 25 questions)
- API FastAPI Dockerisée

❌ Out of Scope

- APIs payantes (OpenAI)
- Bases de données Cloud

3. STRATÉGIE A - LLM SEUL

Principe : Prompt Système + Zephyr-7B **(⚠️ Fiabilité : Faible)**

Points Clés

- Répond avec connaissances internes
- Avantages : Simplicité, rapide
- Inconvénients : Hallucinations élevées

4. STRATÉGIE B - RAG

Principe : Prompt Système + Embeddings + Zephyr-7B

Points Clés

- Recherche vectorielle puis génération
- Avantages : Fiable, contextuel
- Inconvénients : Latence, complexe

5. STRATÉGIE C - Q&A EXTRACTIF

Principe : Prompt Système + Recherche Sémantique + roberta-base-squad2

Processus

"Copie-colle" le passage exact du document source.

Points Clés

- Avantages : Extrêmement précis, zéro hallucination
- Inconvénients : Réponses robotiques

6. STACK TECHNIQUE

Langage	Python 3.10+
API	FastAPI (Swagger)
LLM	Zephyr-7B-Beta (HF)
Embeddings	all-MiniLM-L6-v2 (384 dims)
Tests	Pytest
CI/CD	GitHub Actions

7. MODÈLES IA

Génération : Mistral-7B-Instruct
Embeddings : all-MiniLM-L6-v2 (384 dims)
Extraction : roberta-base-squad2

8. PLANNING RÉALISÉ

- **J1-J2 :** Cadrage & Veille (✅ Fait)
- **J3-J5 :** Implémentation IA (✅ Fait)
- **J6 :** Benchmark complet (✅ Fait)
- **J7-J9 :** API & Industrialisation (En cours)
- **J10 :** Soutenance finale

9. RISQUES

- ⚠️ **API Down :** Modèle local ou cache
- ⚠️ **Hallucinations :** Prompt strict
- ⚠️ **Latence > 5s :** Embeddings légers

10. QUESTIONS & RESSOURCES

Questions

Délai max ? Multilingue ?

Ressources

- HF Inference API
- SBERT Guide
- FastAPI Docs