# Présentation Stratégique

## Assistant FAQ Intelligent

**Étudiant :** Arnaud Rambourg

**Date :** 12/01/2026

**Version :** 2.0

## 1. CONTEXTE & ENJEUX

La communauté de communes reçoit de nombreuses questions récurrentes (déchets, urbanisme).

### Objectifs

- Réduire la charge de demandes et de réponses identiques.
- Offrir une réponse 24/7 aux citoyens.
- Moderniser l'image de la collectivité.

## 2. ANALYSE DU BESOIN

**Cible :** Citoyens (tout âge).

### Contraintes

- Langage clair
- Fiabilité (pas d'erreurs légales).
- Souveraineté des données (hébergement local possible).

## 3. STRATÉGIE A - LLM SEUL

**Principe :** Prompt Système + Meta-Llama-3-8B-Instruct (⚠️ **Fiabilité : Faible)**

### Points Clés

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

## 4. STRATÉGIE B - RAG

**Principe :** Prompt Système + Embeddings + Meta-Llama-3-8B-Instruct

### Points Clés

- Recherche vectorielle puis génération.
- "Avantages :" Fiable, sourcé.
- "Inconvénients :" Latence, complexité.

## 5. STRATÉGIE C - Q&A EXTRACTIF

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

### Processus

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

## 6. STACK TECHNIQUE

| Rôle | Technologie |
|---|---|
| Langage | Python 3.11 |
| Framework | Hugging Face |
| LLM | Meta-Llama-3-8B-Instruct |
| Embeddings | all-MiniLM-L6-v2 (384 dims) |
| Base Vect. | ChromaDB (Local) |
| Interface | Streamlit / API |

## 7. MODÈLES IA

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

## 8. RISQUES & MITIGATIONS

### Hallucinations
Solution : RAG (Sources obligatoires).

### Latence
Solution : Modèles quantifiés (4-bit).

### Confidentialité
Solution : Modèles locaux (Off-cloud).

## 9. PLANNING PRÉVISIONNEL

- **Jours 1-2 :** Cadrage & Données.
- **Jours 3-5 :** Dév (RAG + API).
- **Jours 6-7 :** Benchmark & Déploiement.
- **Jours 8+ :** Tests utilisateurs.

## 10. RESSOURCES & LIENS

**Documentation :**

- Hugging Face Inference
- SBERT Docs
- FastAPI Tuto