# Projet FAQ

Steve Dos Santos, Jonathan Caillaux

# But du Projet

*Mettre en place un assistant IA pour automatiser la réponse à des questions de citoyens*

***Solutions***

- Stratégie A : LLM (Mistral 7B v0.2 Instruct)
- Stratégie B : RAG (Mistral 7B v0.2 Instruct)
- Stratégie C : Q&A extractif (camembert-base-squadFR-fquad-piaf)

*Stratégies B et C utilisent une base FAQ de référence. La stratégie A uniquement les connaissances internes du LLM*

# Evaluation des stratégies

*Evaluation automatisé avec le framework RAGAS (LLM-as-a-judge)*

## Critères d'évaluation

Pondération des métriques pour l'évaluation du système RAG

| CRITÈRE | DESCRIPTION | POIDS |
|---|---|---|
| **Exactitude** | % de réponses correctes | **30%** |
| **Pertinence** | Qualité de la réponse | **20%** |
| **Hallucinations** | % de réponses avec infos inventées | **20%** |
| **Latence** | Temps de réponse moyen | **15%** |
| **Complexité** | Facilité de maintenance | **15%** |

# Evaluation des stratégies

*Les métriques sont transformées puis agrégées un un score global*

FIDÉLITÉ

$$\text{Fidélité} = 1 - \text{Hallucination}$$

LATENCE

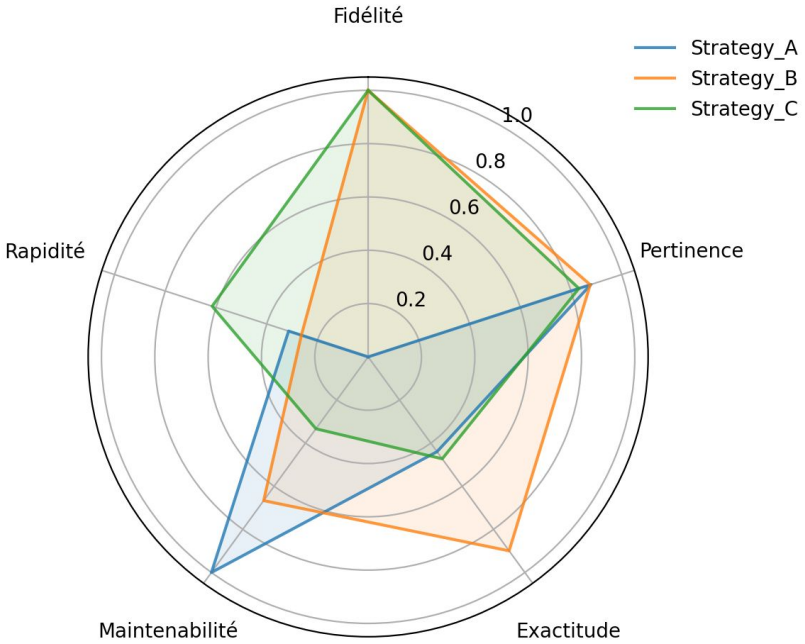
$$S_{\text{latence}} = \min\left(\max\left(\frac{\text{latence}_{\text{cible}}}{\text{latence}_{P75}},\ 0\right),\ 1\right)$$

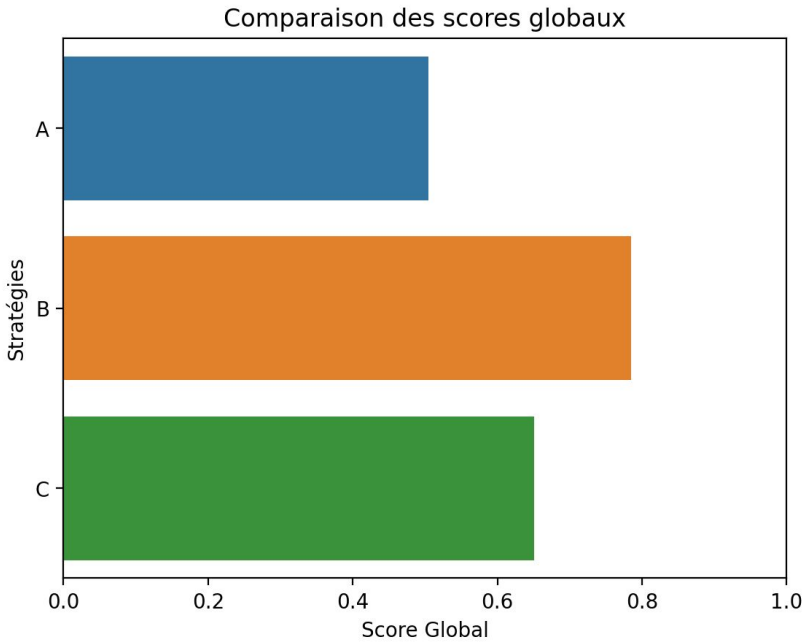
SIMPLICITÉ

$$\text{Simplicité} = 1 - \text{Complexité}$$

# Recommandation de stratégie

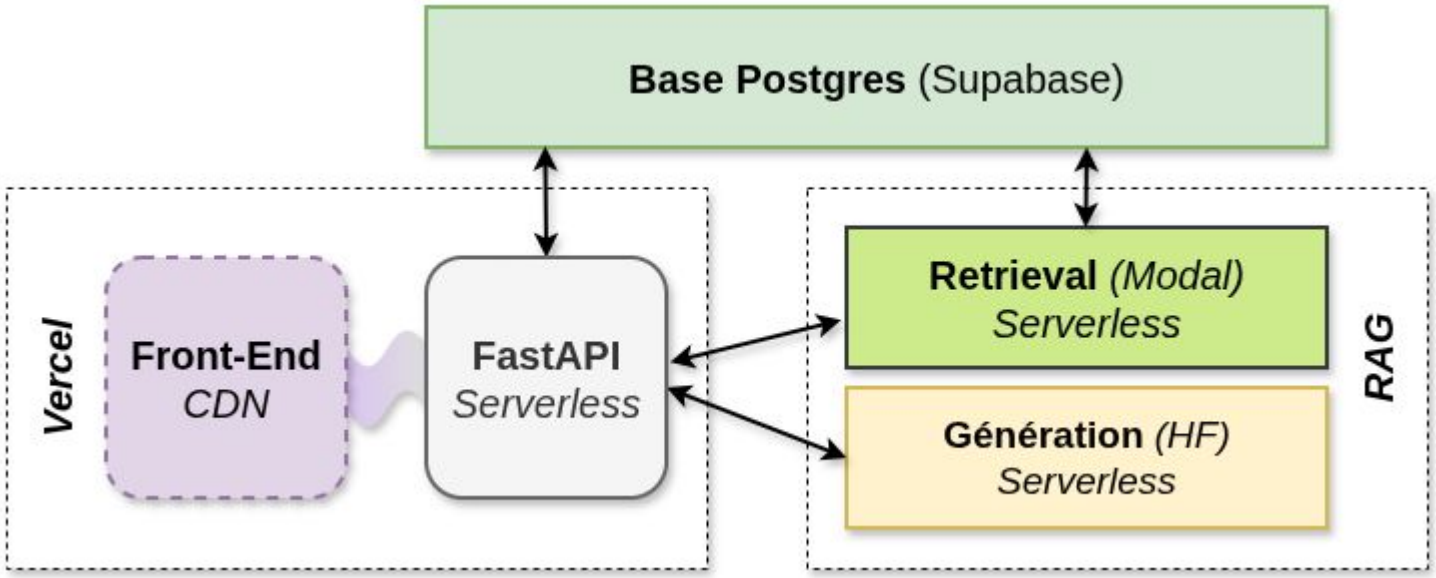
*La stratégie B est recommandée*



# Architecture

*Architecture à séparation de services orientée serverless*




*Chaque brique est isolée, scale-to-zero : coût nul au repos, montée en charge automatique.*

# URLs

*Frontend : projet-faq.vercel.app*

*API : api-projet-faq.vercel.app*

# Conclusion

- Evaluation automatique des Stratégie
- Stratégie B est recommandée
- RAG implémentée et déployée
- Approche serverless robuste (Scale-to-Zero)
- Tests
- Déploiement automatiques

*Perspectives:*

*Test automatisés*

*Mise en place du Monitoring (LogFire)*

*Amélioration de l'UX : adopter une stratégie de démarrage des services serverless*

# Choix des Embeddings

Les stratégies B et C nécessitent d'utiliser un modèle de vectorisation.

Le vocabulaire de la FAQ est spécialisé :

- Les modèles francophones sont plurilingues
- Quel est le modèle le plus adapté ?

| MODÈLE | ALIGNEMENT ↓ | UNIFORMITÉ ↓ | MRR ↑ | TEMPS (S) ↓ |
|---|---|---|---|---|
| intfloat/multilingual-e5-small BEST | **0.1638** | -0.6989 | **0.9900** | 1.302 |
| sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2 | 0.5295 | **-2.6313** | 0.8930 | 0.826 |
| sentence-transformers/all-MiniLM-L6-v2 | 0.5207 | -2.4255 | 0.9788 | **0.489** |

# Score de latence