Martin Expert : Une architecture RAG pour l'assistance réglementaire dans le BTP

Léna Gaubert

Directeur de mémoire : Damien Nouvel (INALCO)

Encadrant de stage : Corentin Gartner (INEX)

Université Sorbonne Nouvelle

Sommaire

Contexte

BTP

- Secteur du **Bâtiment et des Travaux Publics** (BTP) : forte réglementation (RT, RE2020, Décret Tertiaire...).
  - *évolution constante* des documentations et réglementations (contexte : urgence climatique)
  - pas ou peu d'outils intelligents disponibles (TAL/*computer vision*)

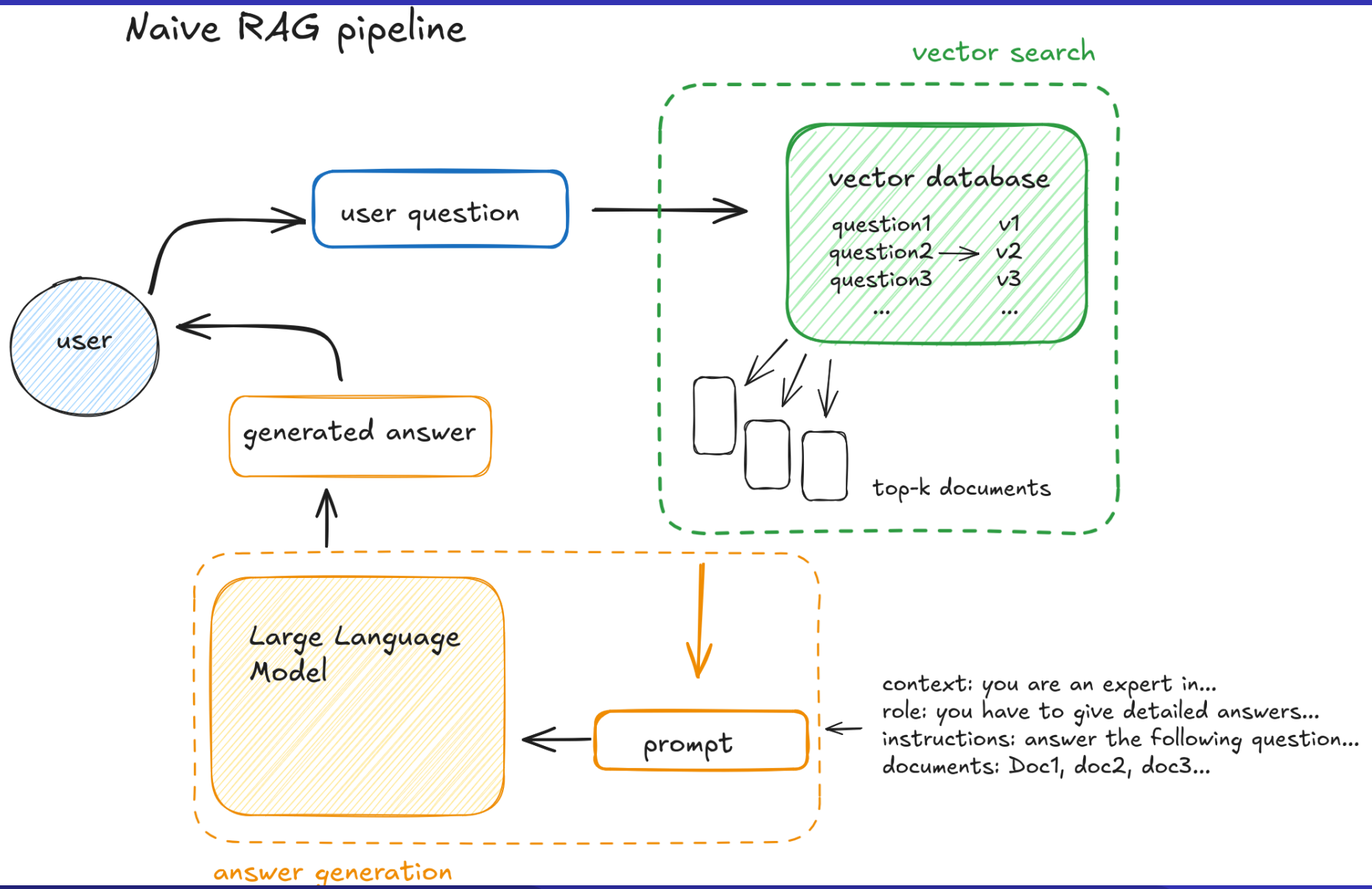
INEX BET

- Bureau d'études techniques : *ingénierie thermique et environnementale*
  - Eco-conception des bâtiments (existants et nouveaux)

Martin Expert

- Un assistant *intelligent* pour faciliter la compréhension et la prise de décision dans l'application du **Décret Tertiaire**.
- Basé sur une architecture **Retrieval Augmented Generation (RAG)**.

Architecture RAG



Architecture RAG

Briques	Explication
Indexation	Préparation des données (nettoyage, chunking, vectorisation...) & stockage (*vector store*)
Extraction	Recherche sémantique (similarité)
Génération	Injection des documents dans le prompt + génération de la réponse (LLM)

Lewis et al. (2020), Facebook AI Lab

Architecture RAG : problématiques

- **granularité** : unité d'information utile (*chunk*, segment)
- top_k : combien de documents pertinents extraire ? (fenêtre de contexte du modèle)
- ***prompt-engineering*** : quelles instructions pour une meilleure compréhension des documents par le modèle ? méthodes ?

Mise en oeuvre : Données disponibles

Intitulé	Données	Type
Décret & arrêté méthode	Texte de loi	Textes courts
Arrêtés Valeurs Absolues I à V	Tableaux de seuils par zone climatique, catégorie/sous-catégorie	Tableaux en format PDF (plusieurs centaines)
FAQ OPERAT (ADEME)	162 Q&R	Texte

Mise en oeuvre : Enrichissement de la FAQ

- *Web scraping* des paires Q&R OPERAT

Type	Nombre	Moyenne	Écart-type	Min / Max
Question	162	22,86	15,75	5 / 83
Réponse	162	333,31	325,75	7 / 2608

Statistiques des longueurs (en nombre de tokens) des questions et réponses dans la FAQ OPERAT extraite en juillet 2024

Mise en oeuvre : Enrichissement de la FAQ

- Segmentation manuelle des paires de Q&R OPERAT

Type	Nombre	Moyenne	Écart-type	Min / Max
Question	465	18,96	10,26	5 / 102
Réponse	465	107,37	88,26	11 / 988

Statistiques des longueurs (nombre de tokens) des questions et réponses *après enrichissement* de la FAQ OPERAT

Mise en oeuvre : RAG

Indexation & Extraction

- `multi-qa-mpnet-base-dot-v1` (`SentenceTransformers`)
- `Qdrant` (*vector store* : question, réponse, métadonnées)

Génération

- `Phi3.1-mini-128k-instruct` (Microsoft)
  - *quantized model* (optimisation inférence)
- **prompt** : *chain of thought*, *few-shot*

Mise en oeuvre : prompt engineering

persona

processus de réflexion

instruction de réponse

vocabulaire

instructions utilisateurs

analyse & réponse à la question sous format JSON.

Résultats

Jeu d'évaluation

Affaire/usage	Question	Réponse	FAQ

- ~35 questions

Evaluation du RAG

- évaluation de l'**extraction**
- évaluation de la **génération**

Résultats : extraction

Précision au rank k

- $P@k = \frac{\text{Documents pertinents} \cap \text{Top k documents}}{k}$
- $P@3 = 0.368$

Rappel au rang k

- $R@k = \frac{\text{Documents pertinents} \cap \text{Top k documents}}{\text{Nb total de docs pertinents pour répondre à la question}}$
- $R@3 = 0.605$

Résultats : génération

Observations

- Format de sortie toujours **respecté**
- Réponses **sourcées**
- **Questions fermées** : "bonne" réponse, mais raisonnement erroné
- **Définitions** : bien comprises dans l'ensemble
- Réflexions sur les **méthodes à appliquer** : plus complexe
- Modèle **sensible au vocabulaire** de la question (désambiguïsation)

Limites

- tester différents modèles d'embeddings/LLMs
  - améliorer extraction, raisonnement LLM...
- comparer méthodes de prompt engineering
  - différentes structures, comment présenter les documents...
- réponses aux questions interprétatives uniquement
  - intégration des autres documents (tableaux !)

Conclusion

Martin Expert

- un **prototype** avec des pistes d'*améliorations futures*
- seul chatbot autour d'une réglementation dans le BTP aujourd'hui

Remerciements

Merci pour votre attention !