Mise en place d'un workflow d'évaluation de RAG avec intégration d'interface : Cas d'usage dans le domaine juridique

Présenté par : Aminata THIOUNE

Encadrants : Mustapha LEBBAH, Kamel MESBAHI

IUT Villetaneuse
Université Sorbonne Paris Nord

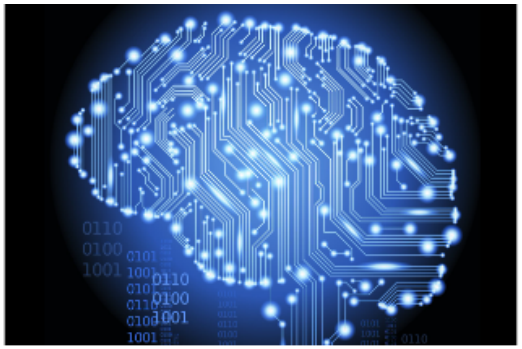
99 Av. Jean Baptiste Clément,
93430 Villetaneuse

HephIA

30 rue de Gramont,
75002 Paris

Table of contents

Présentation de l'entreprise



L'intelligence Artificielle

Pour l'automatisation

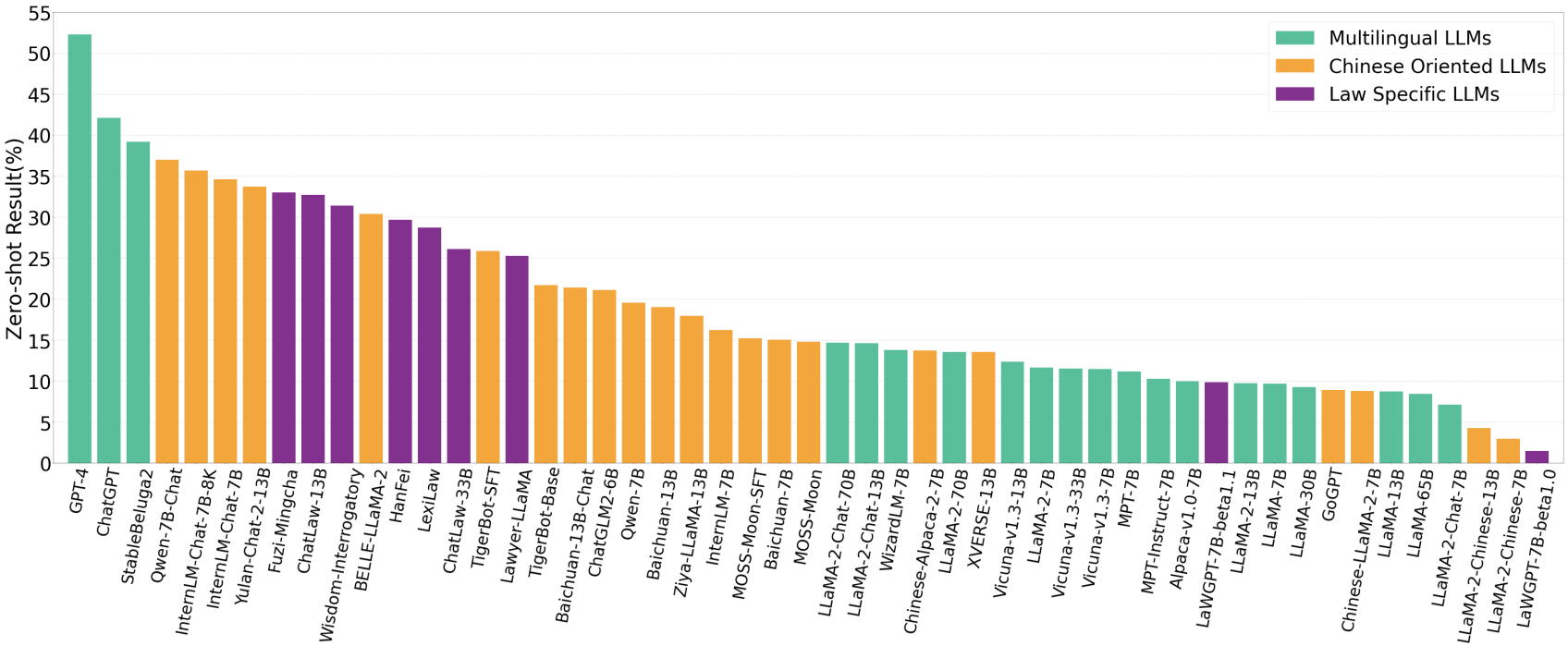
Des processus métiers

Sélection des meilleurs LLMs pour le domaine juridique

- Sélection basée sur l'état de l'art
- Sélection basée sur des métriques de validation

Sélection basée sur l'état de l'art

La recherche de LLMs s'est concentrée sur la revue des articles récents portant sur des LLMs ou les benchmarks (Lawbench) des modèles open source.



Sélection basée sur l'état de l'art

Les meilleurs modèles de langage à grande échelle ont été sélectionnés après une analyse approfondie des articles. Les modèles choisis sont les suivants :

- Modèles multilingues :
  - GPT-4,
  - Mistral,
  - LLama-3,
  - SauLM,
  - AdaptLLM
- Modéles orientés résumé de textes juridiques :
  - RoBerta-BART-Fixed,
  - RoBerta-BART-Dependent,
  - LongFormer-BART,
  - T5-EUR

Sélection basée sur des métriques de validation

Après avoir identifié les LLMs les plus avancés, trois méthodes d'évaluation ont été appliquées :

- Validation par métriques standards de "Hugging Face"
- Validation par un LLM appelé "LLM Judge"
- Validation hamaine

Validation par métriques standards de "Hugging Face"

Les métriques standards sont des outils de mesure utilisés pour évaluer et comparer la performance des modèles.

Métrique	**Définition**
Perplexity	Mesure l'incertitude d'un modèle.
Precision	Proportion de résultats pertinents parmi ceux retournés par le modèle.
Recall	Proportion de résultats pertinents correctement identifiés par le modèle parmi tous les résultats pertinents disponibles.
F1 Score	Moyenne harmonique de la précision et du rappel (Recall), fournissant une mesure équilibrée de performance.
ROUGE	Métrique de comparaison de texte incluant plusieurs versions : ROUGE-1 (chevauchement des unigrams), ROUGE-2 (chevauchement des bigrams), ROUGE-L (plus longue sous-séquence commune), ROUGE-Lsum (version de ROUGE-L pour les résumés).
SacreBLEU	Évalue la qualité des traductions automatiques en comparant les traductions générées avec des traductions de référence, en utilisant des scores de précision basés sur des n-grammes.

Validation par métriques standards de "Hugging Face"

Les résultats obtenus avec les métriques standard sont présentés dans le tableau suivant :

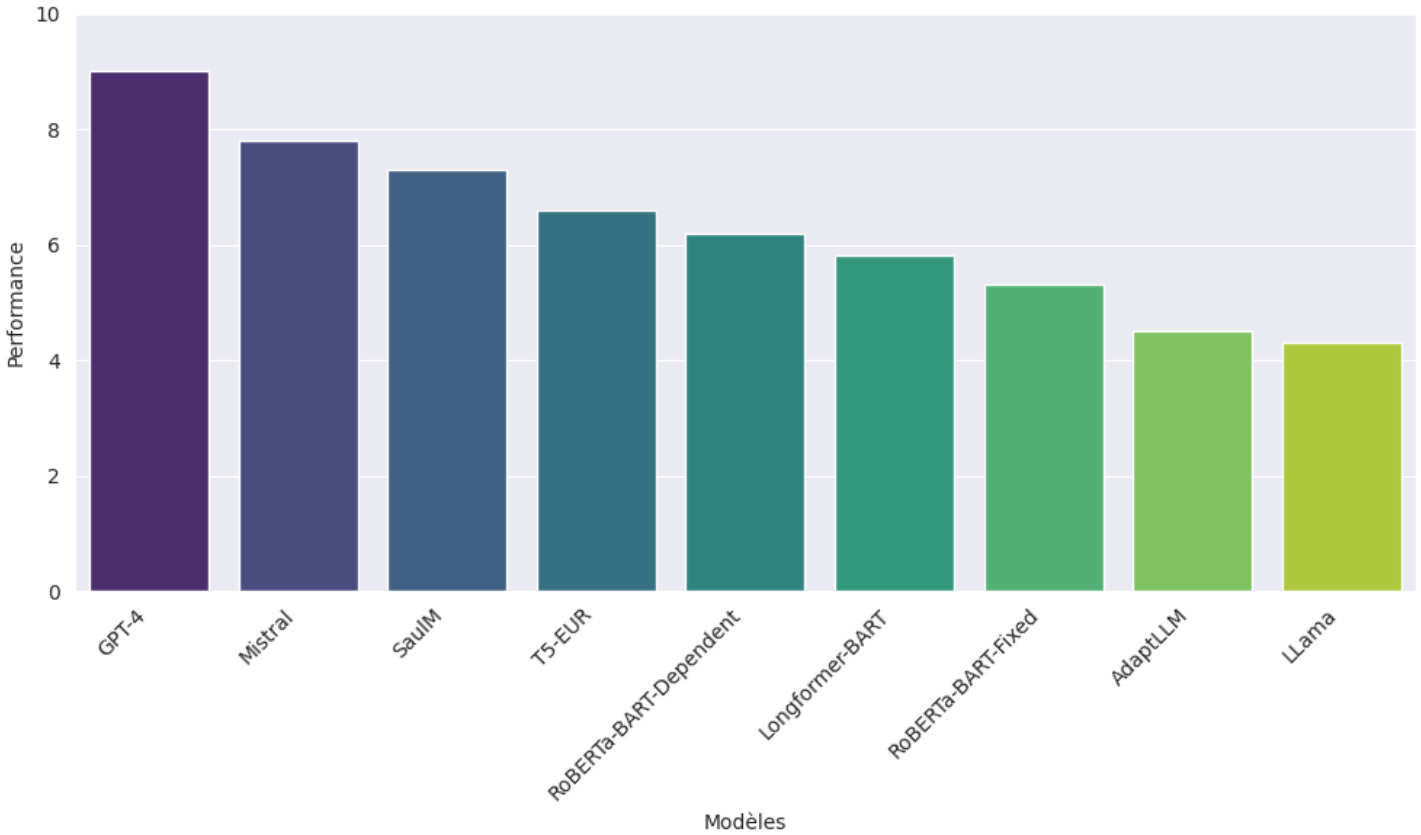
Modèle	Perplexity	Precision	Recall	F1	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-Lsum	SacreBLEU
GPT-4	64.52	**0.733**	**0.774**	**0.753**	**0.419**	**0.191**	**0.308**	**0.307**	**11.054**
Mistral	28.98	0.651	0.755	0.699	0.232	0.094	0.153	0.169	3.338
LLama3	30.11	0.693	0.734	0.712	0.251	0.090	0.164	0.181	4.471
SaulM	22.44	0.634	0.665	0.648	0.136	0.031	0.086	0.093	0.963
AdaptLLm	**19.50**	0.685	0.685	0.684	0.090	0.027	0.062	0.063	1.194
RoBERTa-BART-Fixed	44.17	0.651	0.642	0.647	0.134	0.010	0.062	0.082	0.471
Roberta-BART-Dependent	34.91	0.639	0.643	0.641	0.018	0.000	0.018	0.018	1.219
Longformer-BART	59.25	0.659	0.651	0.655	0.155	0.049	0.077	0.097	1.656
T5-EUR	111.09	0.685	0.677	0.681	0.282	0.062	0.132	0.132	1.955

Validation par un LLM appelé "LLM Judge"

Un "LLM Judge" est un modèle robuste qui évalue objectivement les sorties des modèles selon des critères de qualité prédéfinis.
Des outils comme DeepEval, un framework open source basé sur un "LLM Judge", offrent diverses métriques d'évaluation.

Validation humaine

La validation humaine consiste à faire évaluer les réponses des LLMs par des évaluateurs humains. Les résultats de cette évaluation sont les suivants :



Sélection basée sur des métriques de validation

En nous appuyant sur les résultats des trois méthodes de validation, nous avons sélectionné les modèles open-source dont les performances sont proches de celles de GPT-4 :

- Modèles Multilingues :
  - Mistral
  - SaulM
- Modèles Orientés résumé de textes juriques :
  - RoBerta-BART-Fixed,
  - RoBerta-BART-Dependent,
  - LongFormer-BART,
  - T5-EUR

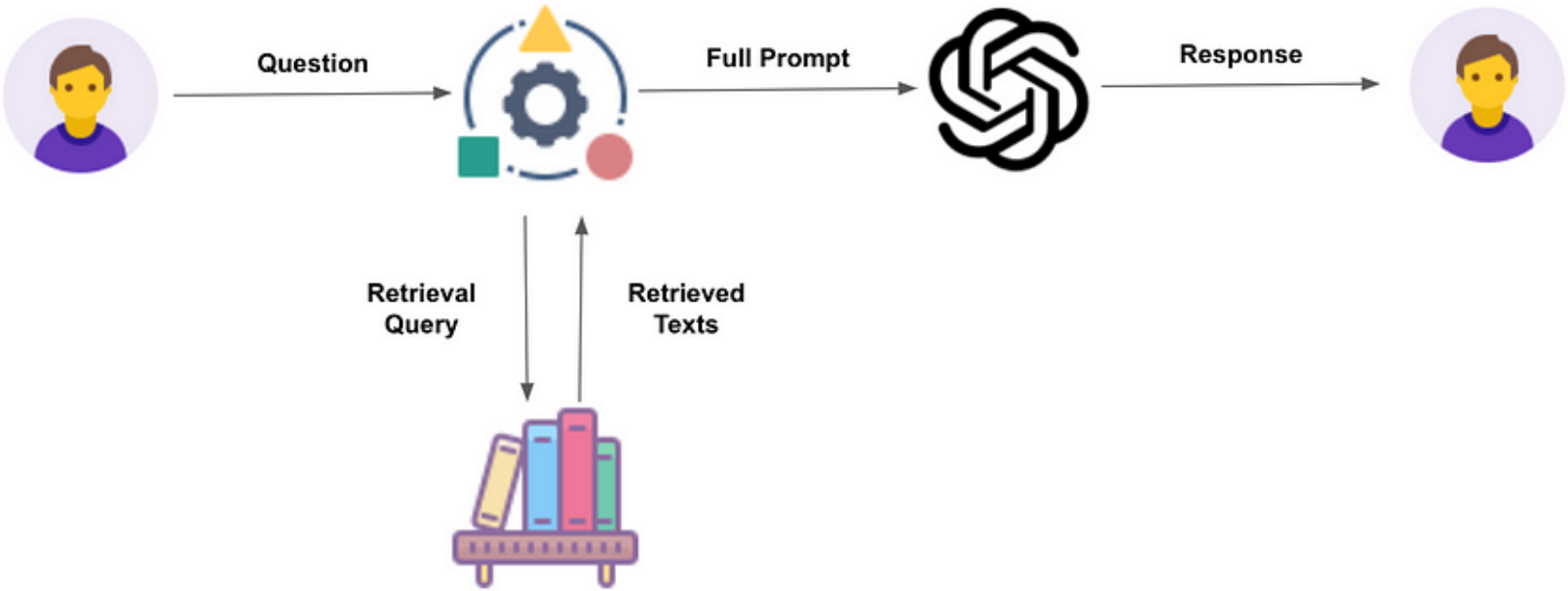
Développement d'une interface utilisateur pour le résumé de textes juridiques

Streamlit

Transformers

LangChain

Application de l'approche RAG pour améliorer la fiabilité des réponses et résoudre les problèmes d'Hallucination



Algorithme du système RAG

Algorithm 1 Construction du VectorStore pour RAG avec FAISS

Require: Documents, LLM, Embedder, Splitter
Ensure: VectorStore

Initialisation :
Créer une liste vide *VectorStore*
Initialiser FAISS index
Étape 1 : Prétraitement des Documents
for chaque document *doc* dans *Documents* **do**
    Utiliser le *Splitter* pour diviser *doc* en segments
    for chaque segment *seg* dans *segments* **do**
        Utiliser l'*Embedder* pour obtenir l'embedding du segment *seg*
        Ajouter l'embedding au *VectorStore*
    end for
end for
Étape 2 : Construction du FAISS Index
Convertir *VectorStore* en matrice de vecteurs
Ajouter les vecteurs à l'index FAISS
Entraîner l'index FAISS pour l'optimisation des requêtes
Retourner l'index FAISS comme *VectorStore*

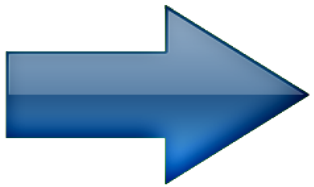
Algorithme du système RAG

Algorithm 2 Répondre aux Questions avec RAG

Require: Prompt (Question), VectorStore, Embedder, LLM
Ensure: Réponse

1: **Étape 1 : Recherche du Contexte**
2: Utiliser l'*Embedder* pour obtenir l'embedding du *Prompt*
3: Chercher les segments similaires dans le *VectorStore* en utilisant FAISS
4: Sélectionner les segments les plus pertinents comme *Contexte*
5: **Étape 2 : Génération de la Réponse**
6: Combiner le *Prompt* et le *Contexte*
7: Fournir cette combinaison au *LLM*
8: Obtenir la réponse générée par le *LLM*
9: **Retourner** la réponse générée par le *LLM*

Demo de l'application



MERCI DE VOTRE ATTENTION