Veille technique

Méthodologies d’extraction de données sur du texte utilisant les « modèles de langues » (LLM)

Geoffroy Daumer

# Préambule

Qu’est-ce qu’un grand modèle de langage (LLM) ?

Les grands modèles de langage (LLM) sont des modèles d’apprentissage automatique capables de comprendre et de générer des textes en langage humain. Ils fonctionnent en analysant des ensembles de données linguistiques massives. (1)

# Table des matières

[Préambule 2](#_Toc167814121)

[Table des matières 2](#_Toc167814122)

[1. Présentation des documents et analyse de fiabilité 3](#_Toc167814123)

[2. Analyse des documents 4](#_Toc167814124)

[2.1 Xebia – How to extract structured data from unstructured text using LLMs 4](#_Toc167814125)

[2.2 Langchain – Use cases : Extraction 5](#_Toc167814126)

[2.3 Medium - Harness The Power of LLMs : How to extract data from legacy documents using LLMs 7](#_Toc167814127)

[2.4 Towardsai - Demystifying Information Extraction using LLM. 8](#_Toc167814128)

[3. Conclusion 9](#_Toc167814129)

[Bibliographie 10](#_Toc167814130)

# 1. Présentation des documents et analyse de fiabilité

Les documents ayant servi au recueil d’informations proviennent du site Cloudflare, Xebia (article de septembre 2023), Langchain (date inconnue mais copyrigt 2024 en bas du site), Medium et Towardsai.

Cloudflare est une entreprise de cyber-sécurité américaine, ils ont une section de vulgarisation des termes relatifs à l'IA sur leur site, ce sont des spécialistes en informatique.

Xebia est une entreprise de consultation en transformation digitale, il s’agit de professionnels du domaine, leurs informations peuvent s’avérer utiles et donner une vue du sujet traité.

Langchain est une librairie open source, c’est la référence pour de développement d’applications autour des LLM. La documentation à travers cette librairie est indispensable pour savoir comment traiter le sujet en pratique.

Medium est une plateforme de partage de connaissance, où les gens peuvent voter pour populariser les articles utiles.

Towardsai est un site de publications relatives à l'IA.

Les sources proposées sont donc de différents types : certains articles ont été écrits par des entreprises spécialistes du domaine, d’autres sont écrits par les développeurs des outils de pratique.

Quant à d’autres ce sont des sites spécialisés dans l’IA. Le fait d’inclure un système de vote comme dans Medium permet de populariser les articles utiles.

# 2. Analyse des documents

## 2.1 Xebia – How to extract structured data from unstructured text using LLMs

L’intérêt des LLM est qu’ils permettent de compléter une tâche sans entraîner de modèles.

L’extraction de données utilisant les LLM se déroule en 4 étapes :

* Créer une demande
* Donner le texte non structuré au LLM
* Récupérer la sortie structurée du LLM
* La sauvegarder dans une base de données

On peut demander une sortie structurée au modèle (comme couleur = noir ou blanc) avec un fichier « .yaml ». La structure demandée sera chargée par des classes « Pydantic BaseModel ».

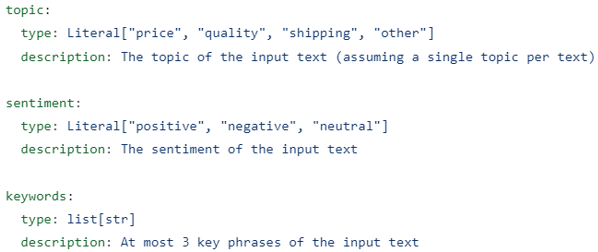


Figure 1 : Exemple de structure de fichier yaml

On peut indiquer au modèle les indications de recherche et les sorties possibles pour chaque caractéristique.

La sortie du modèle est validée par un modèle pydantic. Si on a une erreur, on peut tenter de s’adapter à cette erreur en adaptant le prompt.

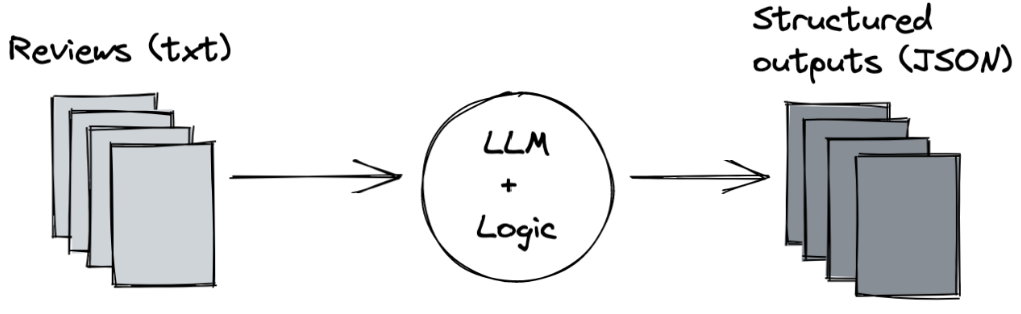


Figure 2 : Séparation de la logique modèle d’avec les données

## 2.2 Langchain – Use cases : Extraction

Il y a deux types d’approches pour extraire des données en utilisant les LLMs :

* **Fonctions** : Certains LLM ont des fonctions intégrées d’extraction arbitraires d’entités (peuvent donner des attributs que l’on n’a pas demandé).
* **Parsing** : On définit soit même les fonctions d'extraction (Pydantic comme vu plus haut).

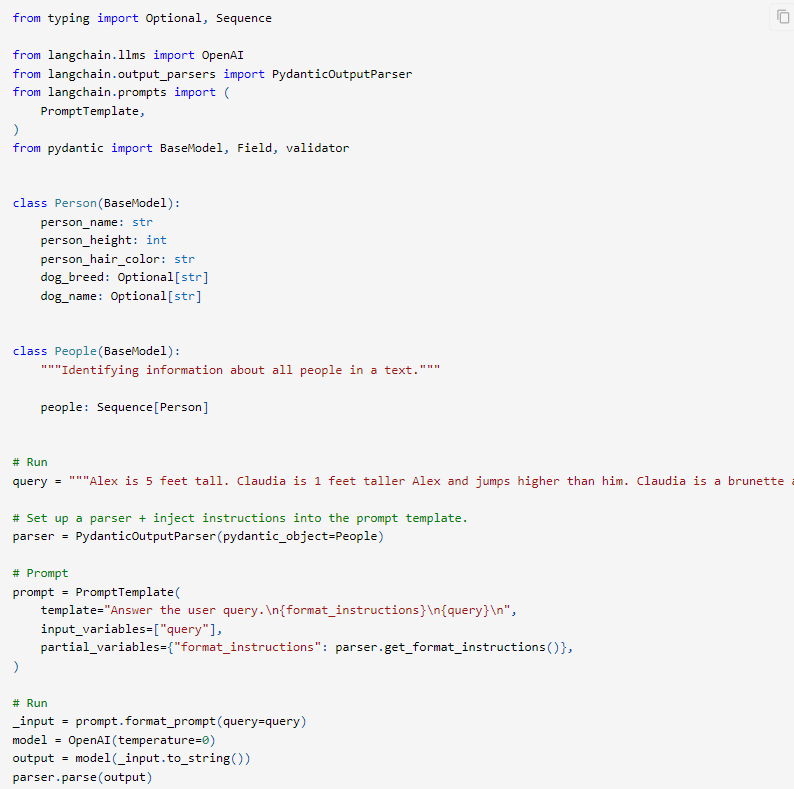
C’est le parsing qui nous intéresse car personnalisable, voici un exemple de parsing :

Figure 3 : Demande de parsing sur un prompt

La classe BaseModel (pydantic) est utilisée pour définir les attentes de sorties, ici, elle n’est pas chargée à partir d’un fichier yml comme avant.

Dans le prompt (langchain), définition d’un prompt général contenant toutes les informations, avec la variable d'entrée dans l'argument "input\_variables".

Chargement du modèle LLM, run du modèle avec le prompt, segmentation de la sortie avec l’objet Pydantic.

## 2.3 Medium - Harness The Power of LLMs : How to extract data from legacy documents using LLMs

Organisation du script d'extraction des données :

- Chargement texte

- Créer prompt LLM

- Envoyer le prompt au LLM par l'API openAI

- Imprimer les résultats

Exemple d'un prompt :

Figure 4 : Exemple de prompt

Une image contenant texte, capture d’écran, Police

Description générée automatiquement

Le prompt est segmenté en 3 parties :

- Contexte : exemple : tu es un agent de génération de texte travaillant avec des contrats de location

- Task specification : tâche à exécuter, format attendu

- Texte source

"lease" est le texte brut dans lequel chercher les informations

Une image contenant texte, Police, capture d’écran, algèbre

Description générée automatiquement

Figure 5 : Envoi du prompt

Une image contenant texte, Police, ligne

Description générée automatiquement

Figure 6 : Résultat

## 2.4 Towardsai - Demystifying Information Extraction using LLM.

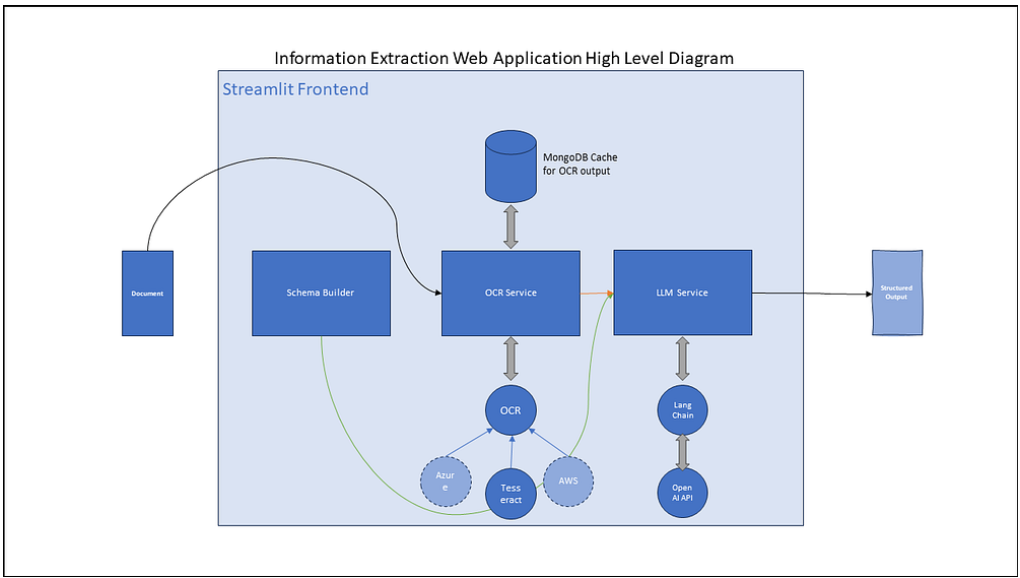
On peut utiliser des modèles OCR (Optical Character Recognition) couplés à des LLMs, avec des outils frameworks comme Streamlit. Ainsi on pourrait partir de documents scannés directement.

Figure 7 : Diagramme haut niveau d’un système d’extraction d’informations avec de l’OCR

# 3. Conclusion

Le fait d’analyser différents articles permet de compléter l’information. On retombe globalement sur la même information, mais avec des nuances.  
Chacun a sa façon de faire. L’article apportant le plus de détails techniques pour la réalisation est Langchain.

Medium, le site de publications d’articles d’IA où les gens peuvent voter, apporte également une certaine contribution, tout comme Xebia, le site d’entreprise de consultation digitale, et Towardai (site de publications d’articles d’IA).

Néanmoins, Xebia semble apporter moins de practicité pour résoudre ce genre de problèmes. Les sites d’entreprise pourraient donc être moins utiles pour les veilles techniques.

# Bibliographie

1. **Cloudflare.** [En ligne] https://www.cloudflare.com/fr-fr/learning/ai/what-is-large-language-model/.

2. **Langchain.** [En ligne] https://python.langchain.com/docs/use\_cases/extraction.

3. **Xebia.** [En ligne] https://xebia.com/blog/archetype-llm-batch-use-case/.

4. **Lights, Brightest.** Harness The Power of LLMs: How to extract data from legacy documents using LLMs. *Medium.* [En ligne] https://medium.com/@brightestlights/harness-the-power-of-your-how-to-extract-data-from-legacy-documents-using-llms-2841f5835359.

5. **Mohan, Aditya.** Demystifying Information Extraction using LLM. *Towardsai.* [En ligne] https://pub.towardsai.net/demystifying-information-extraction-using-llm-f1a551f01f66.