

Vers une classification multi-labels en domaines de connaissances du Trévoux 1743

Nicolas Neila

Sabine Loudcher, Julien Velcin, Denis Vigier, Nicolas Gutehrlé, Ludovic Moncla



Fédération
Informatique
de Lyon



Contexte du stage

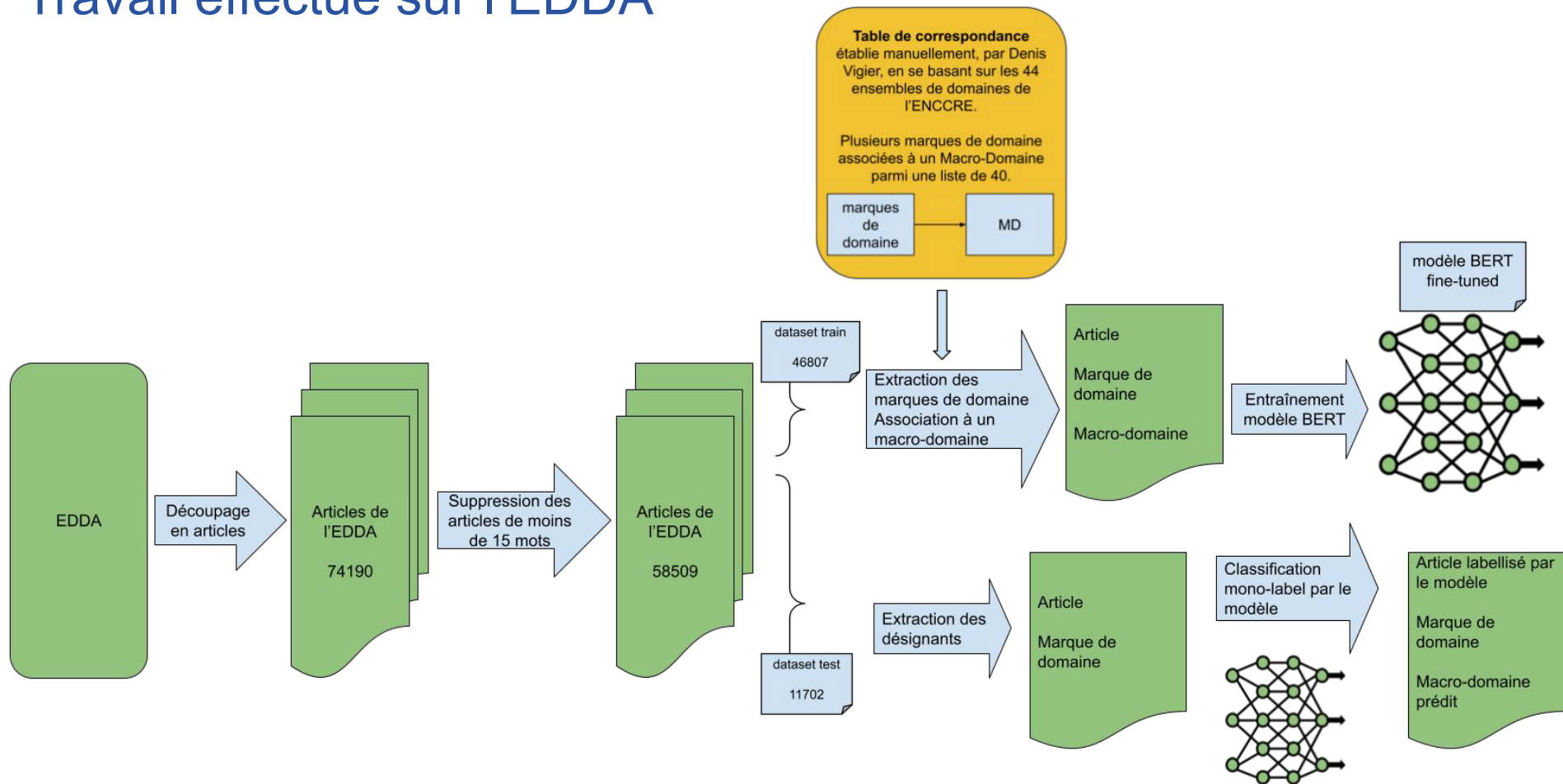
Projet ECoDa

Poursuite du projet GEODE¹ sur l'Encyclopédie de Diderot et d'Alembert (EDDA) sur la question de la classification en domaine de connaissances

Objectif : adapter la méthodologie au Dictionnaire Universel Français-Latin de Trévoux (DUFLT) et étendre à une classification multi-labels.

1: Alice Brenon, Ludovic Moncla, Katherine McDonough. Classifying encyclopedia articles: Comparing machine and deep learning methods and exploring their predictions. Data and Knowledge Engineering, 2022, 142, pp.102098. [10.1016/j.datak.2022.102098](https://doi.org/10.1016/j.datak.2022.102098). [hal-03821073](https://hal.archives-ouvertes.fr/hal-03821073)

Travail effectué sur l'EDDA



Classification en domaines de connaissances

À quelle fin ?

Peu de marques de domaine :

101 119 paragraphes dans le DUFLT

→ dont 10 934 avec marque de domaine

Mais présentes avec une grande variété (1 006)

ASYMMÉTRIE

ASYMMÉTRIE , s. f. Terme d'Arithmétique . *Asymmetria*. C'est lorsque dans un nombre proposé l'on ne peut pas trouver une racine telle qu'on la demande : comme la racine quarrée de 10.

Nécessité de les regrouper, proposition de 40 macro-domaines

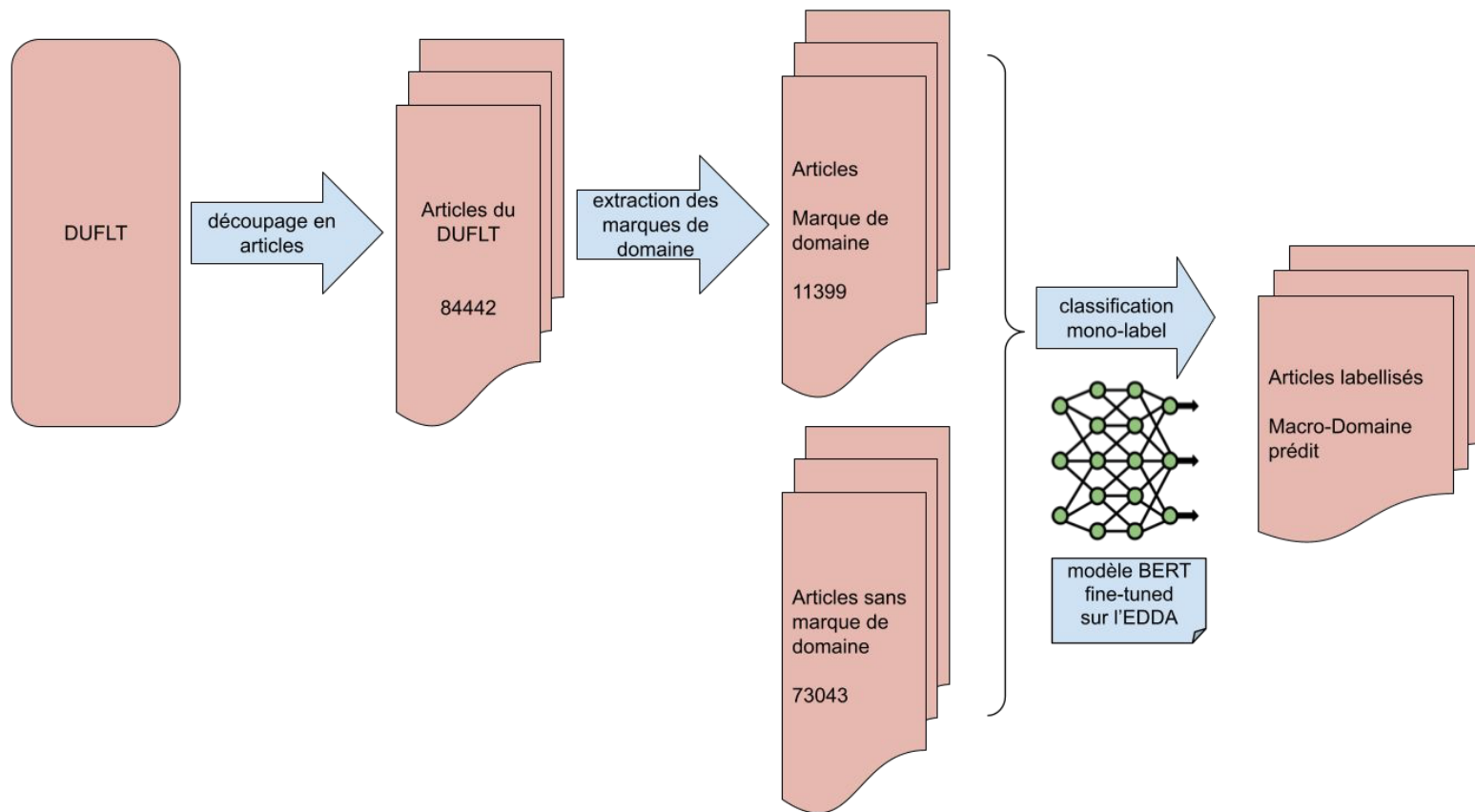
Classification en domaines de connaissances

Macro-domaines

Agriculture	Commerce	Médaille	Arts mécaniques - Manufactures
Anatomie	Coutume Droit Jurisprudence	Médecine	Militaire
Antiquité	Géographie	Mesure	Monnaie
Architecture	Grammaire	Métiers [alimentation]	Musique et Danse
Beaux-Arts	Histoire	Métiers [autres]	Pharmacie
Belles-Lettres	Jeu	Métiers [bois]	Philosophie
Blason	Lexique général	Métiers [cuir, peaux]	Religion
Botanique	Maréchalerie - Manège	Métiers [métal, minéral, dérivés]	Sciences Physico-mathématiques
Chasse-Pêche	Marine	Métiers [papier]	Superstition
Chimie	Mathématiques	Métiers [tissu, habit]	Voyage

Exploration des données du DUFLT

Utilisation du modèle entraîné sur l'EDDA



Exploration des données du DUFLT

Utilisation du modèle entraîné sur l'EDDA

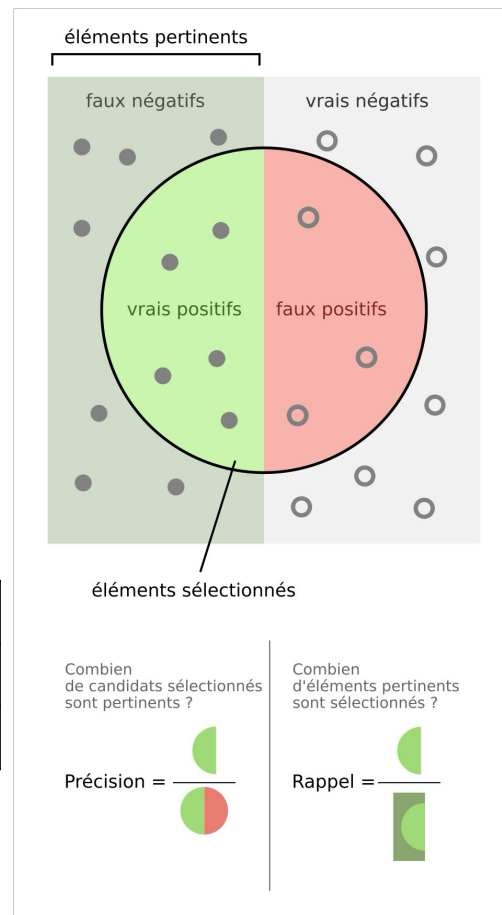
Résultats

Indicateurs des performances du modèle :

- précision : proportions de bonnes réponses parmi les prédictions
- rappel : proportions d'éléments attendus correctement trouvés

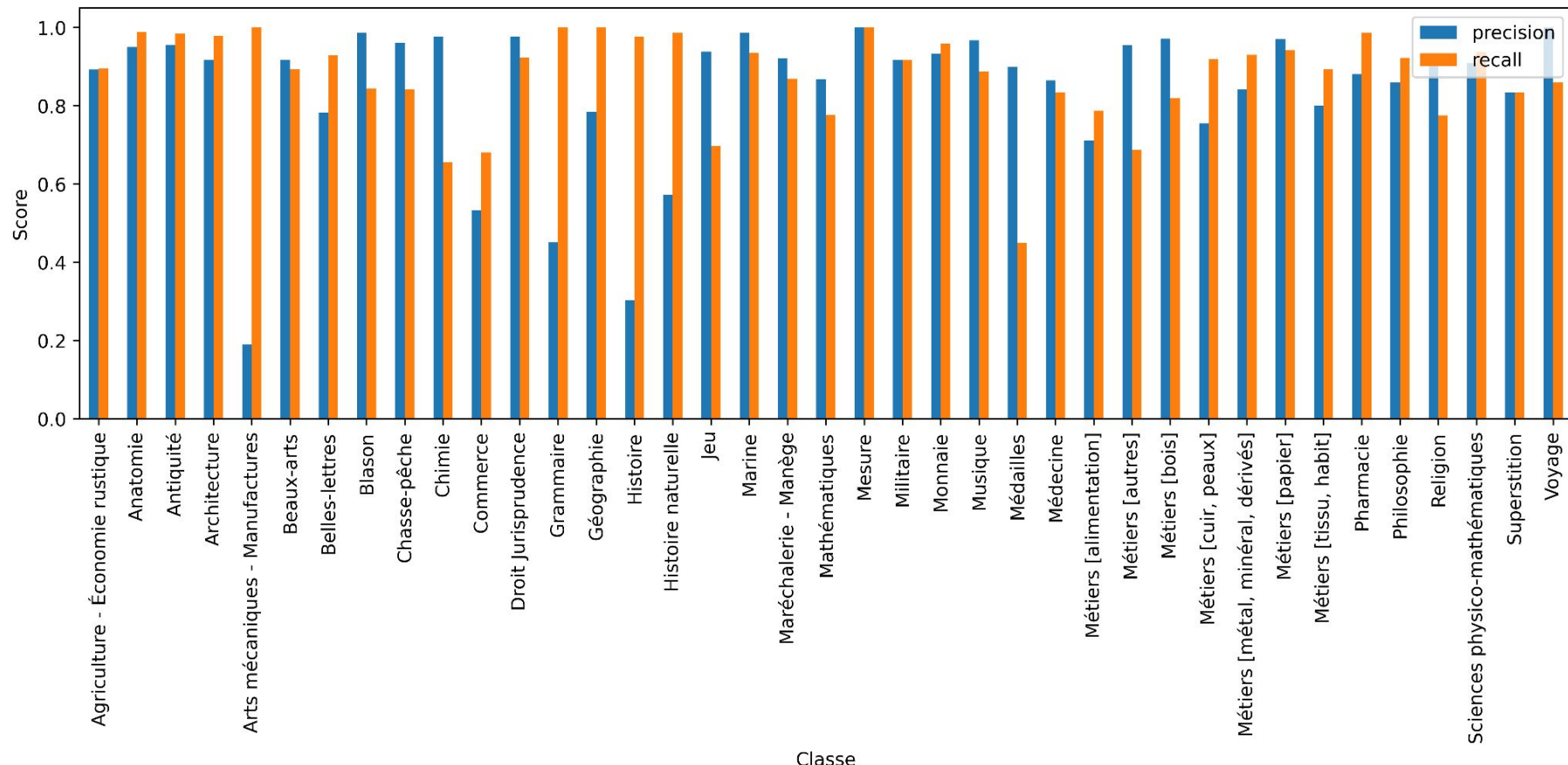
MD prédits	MD réel
Guerre	Marine
Guerre	Marine
Marine	Marine
Guerre	Guerre

Label	Précision	Rappel
Guerre	$1/3 = 0.33\dots$	$1/1 = 1$
Marine	$1/1 = 1$	$1/3 = 0.33\dots$



Exploration des données du DUFLT

Utilisation du modèle entraîné sur l'EDDA



Exploration des données du DUFLT

Prédiction au niveau du paragraphe

ABANDON

ABANDON , s. m. Mépris, délaissement de quelque chose. *Derelictio, destitutio. Neglectus rei alicujus.* Il n'est point du bel usage. On ne le trouve guère que dans MOLIERE, lequel dit, en parlant des coquettes qui renoncent par nécessité au monde qui les quitte :

Dans un tel abandon leur sombre inquiétude.

Ne voit d'autre recours que le métier de Prude.

pratique > Coutume
Droit Jurisprudence

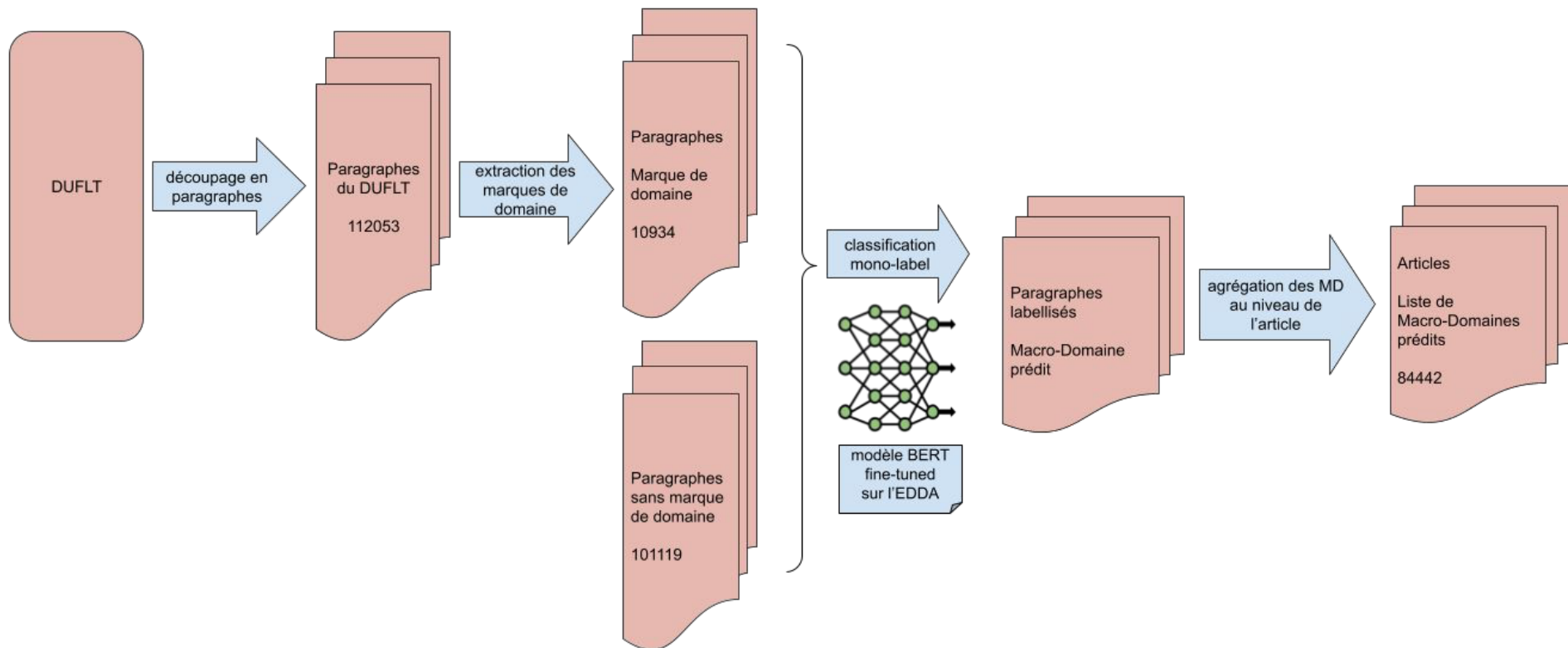
Il n'est supportable en ce sens qu'en termes de pratique . Le débiteur a fait l'*abandon* de tout son bien à ses créanciers. *Abandonnement* vaux mieux.

Dieu, Chrétien >
Religion

Les Mystiques ont nommé *abandon*, la sainte indifférence d'une ame désintéressée, qui s'abandonne totalement & sans réserve à Dieu. Cet *abandon* n'est que l'abnégation ou renoncement de soi-même FENEL. Les Quiétistes ont abusé de ce terme dans un sens impie & très justement condamné. Tout bon Mystique est toujours bon Chrétien, & par conséquent son *abandon* ne va jamais jusqu'à l'indifférence pour son salut, ni au mépris des bonnes oeuvres, ni à la négligence dans le service de Dieu, &c. Les excès sont détestables, & bien éloignés du louable *abandon* que les saints Mystiques ont enseigné & pratiqué.

Exploration des données du DUFLT

Prédiction au niveau du paragraphe



Exploration des données du DUFLT

Utilisation du modèle entraîné sur l'EDDA

Résultats

Le multilabel augmente le nombre de prédictions au niveau de l'article

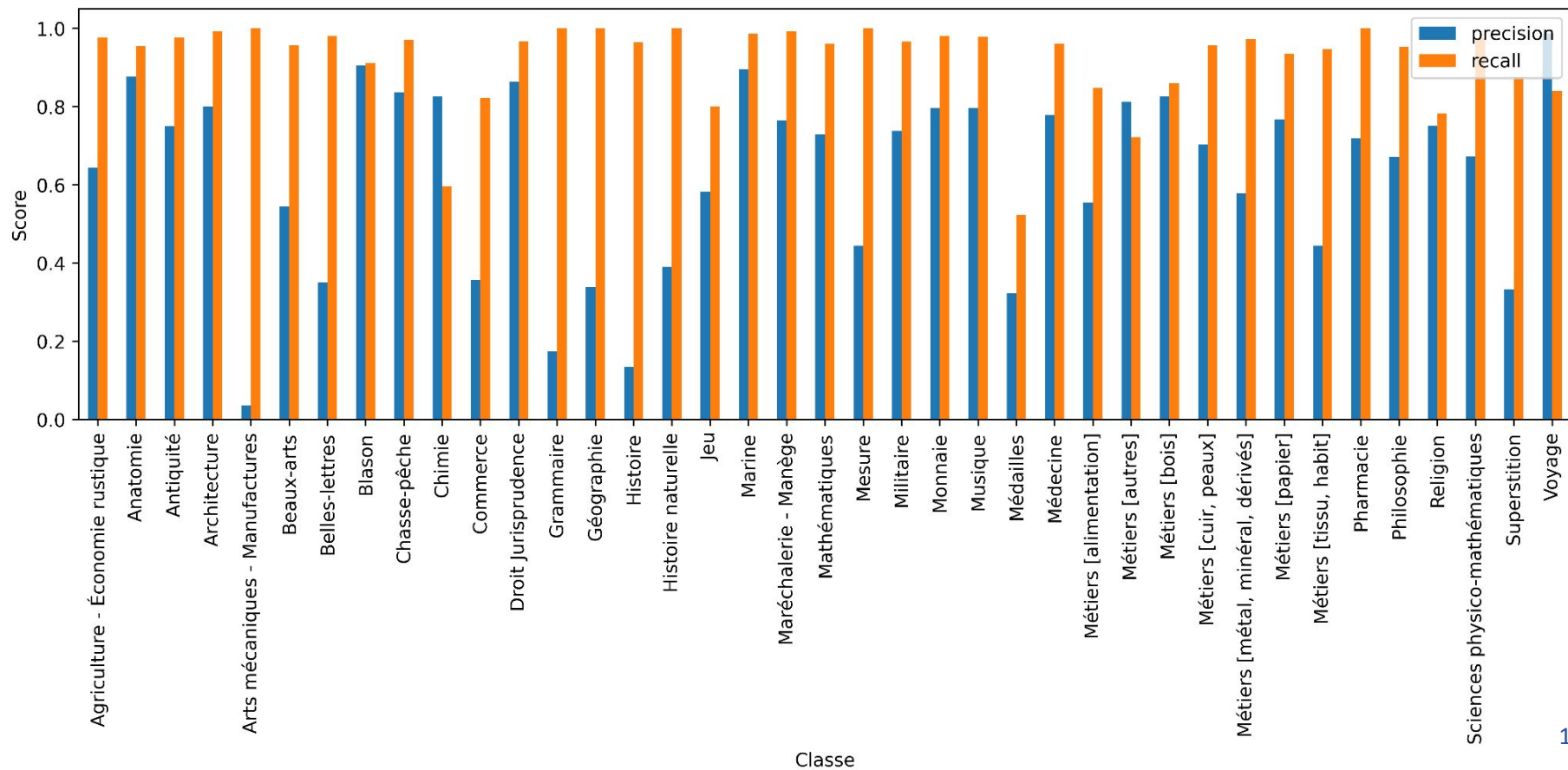
- La précision va être plus faible, car + de faux positifs
- Le rappel devrait augmenter, car sur ce nombre accru de prédition il devrait y avoir la bonne

MD prédits (label unique)	MD prédits (multilabel)	MD réel
Guerre	Guerre ; Agriculture	Marine
Guerre	Guerre ; Marine	Marine
Marine	Marine	Marine
Guerre	Guerre ; Marine	Guerre

Label	Précision	Rappel
Guerre	$1/3 = 0.33...$	$1/1 = 1$
Marine	$2/3 = 0.66...$	$2/3 = 0.66...$
Agriculture	0	0

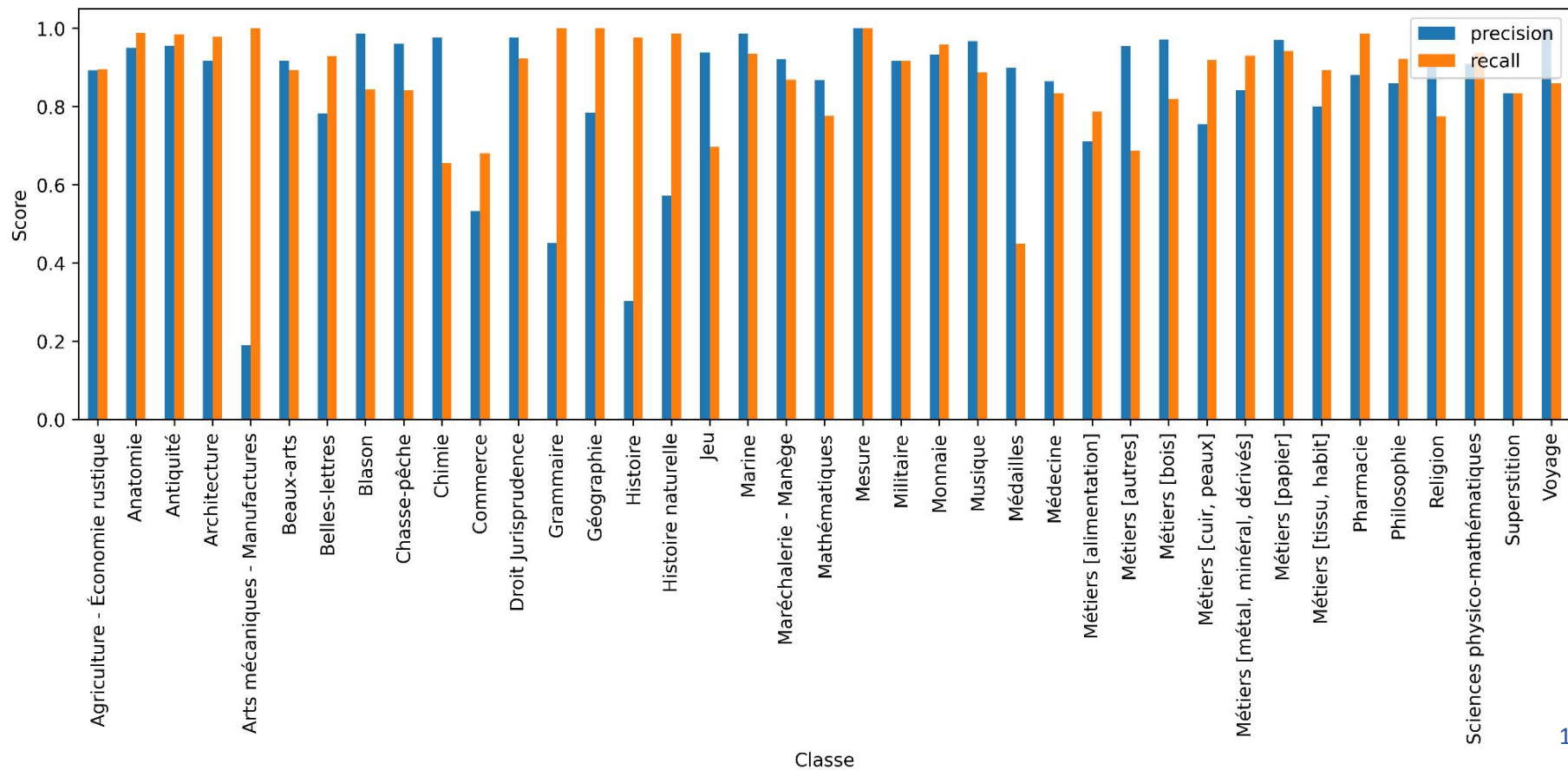
Exploration des données du DUFLT

Prédiction au niveau du paragraphe



Exploration des données du DUFLT

Utilisation du modèle entraîné sur l'EDDA



Le multilabel dans le DUFLT

Quelles données ?

Plusieurs paragraphes ont
une marque de domaine
multiple (contient une “&”)

DÉTACHER

Détacher , en termes de Guerre , se dit en parlant des soldats, dont on en choisit quelques-uns de diverses compagnies, afin de les commander pour quelque expédition. *Subducere*. Il y eut cent hommes *détachés* qui emportèrent ce retranchement.

En termes de Marine *détacher* se dit dans le même sens, en parlant des vaisseaux. On *détacha* six vaisseaux pour aller en garde à la tête de l'armée.

Se *détacher*, en termes de Guerre & de Marine , signifie se *séparer*. Quatre régimens de troupes auxiliaires se *détacherent*. Deux frégates, deux vaisseaux se *détacherent*.

Le multilabel dans le DUFLT

Quelles données ?

Comparaison avec l'EDDA

DUFLT : 101 119 paragraphes

- dont 10 934 avec marque de domaine
- dont 472 avec marque de domaine "&" "*eaux & forets*", "*jardinier & d'agriculture*"
- dont 200 avec plusieurs macro-domaines

EDDA : 77 085 articles

- dont 55 248 avec marque de domaine
- dont 5 890 avec plusieurs marques de domaine
- dont 2 094 avec plusieurs macro-domaines

Le multilabel dans le DUFLT

Quelles données ?

Pertinence à classifier en multilabel

200 paragraphes annotés en multilabel

- 189 où les marques de domaine concernent le seul paragraphe
- dans 93% des cas la définition est à l'intersection des marques de domaine

Classifier en multilabel est donc pertinent.

Permet aussi de mieux capturer les liens entre les macro-domaines.

BOUSSOIR

BOUSSOIR , s. m. Terme de Marine & de Charpenterie . Les *boussoirs* sont deux pièces de bois, dont une partie est posée & attachée au-dessus du château d'avant vers la proue, tant à tribord qu'à bas-bord, & le reste saillant hors du navire sert à lever les ancres. *Tignum tollendis ancoris aptum.* CARON.

Le multilabel dans le DUFLT

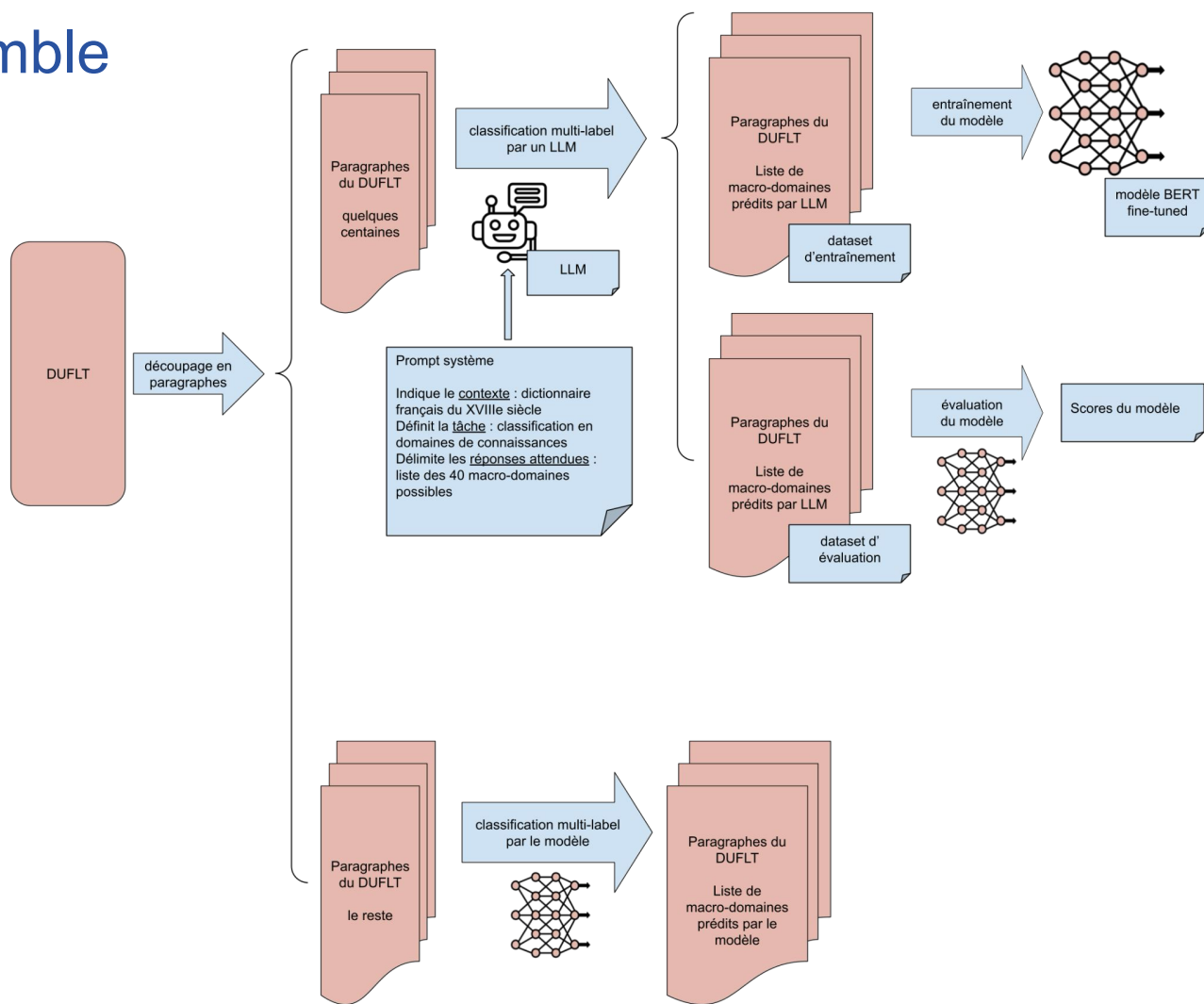
Stratégie d'entraînement du modèle

On fait l'hypothèse qu'un modèle supervisé sera plus efficace qu'un LLM en few-shots

1. Utilisation d'un LLM pour constituer un dataset d'entraînement
2. Fine-tuning d'un modèle BERT en classification multilabel
 - a. Framework Set Fit² : efficace avec peu de données, se base sur les différences entre classes plutôt que les similitudes au sein d'une classe
 - b. Active Learning : tester sur des données non annotées après entraînement, et incorporer au dataset les données pour lesquelles le modèle est le moins sûr

2: Tunstall, L., Reimers, N., Jo, U. E. S., Bates, L., Korat, D., Wasserblat, M., & Pereg, O. (2022). Efficient Few-Shot Learning Without Prompts. arXiv:2209.11055. <https://doi.org/10.48550/arXiv.2209.11055>

Plan d'ensemble



Constitution d'un dataset d'entraînement

Choix du LLM

1ère étape : Création d'un dataset de test

- Composé de paragraphes avec tous types de macro-domaines :
 - lexique Général
 - multilabel
 - monolabel
- Support entre 10 et 14 pour chacun des macro-domaines
- Prompt système pour définir la tâche du LLM (classification multilabel) ainsi que les 40 macro-domaines possibles
- Diverses métriques permettent de faire le choix du LLM
 - Classiques : rappel, précision, mesure F1
 - Propres au multilabel : exact match, score de Jaccard, distance de Hamming

Constitution d'un dataset d'entraînement

Résultats

Métriques globales

Scores globaux par modèle



Hallucination des labels

GPT-5.4-Mini	Gemma-4	Deepseek-4
Métier [tissu, habit]	Métiers [monnaie]	Astronomie
Métier [métal, minéral, dérivés]		
Métaux [métal, minéral, dérivés]		

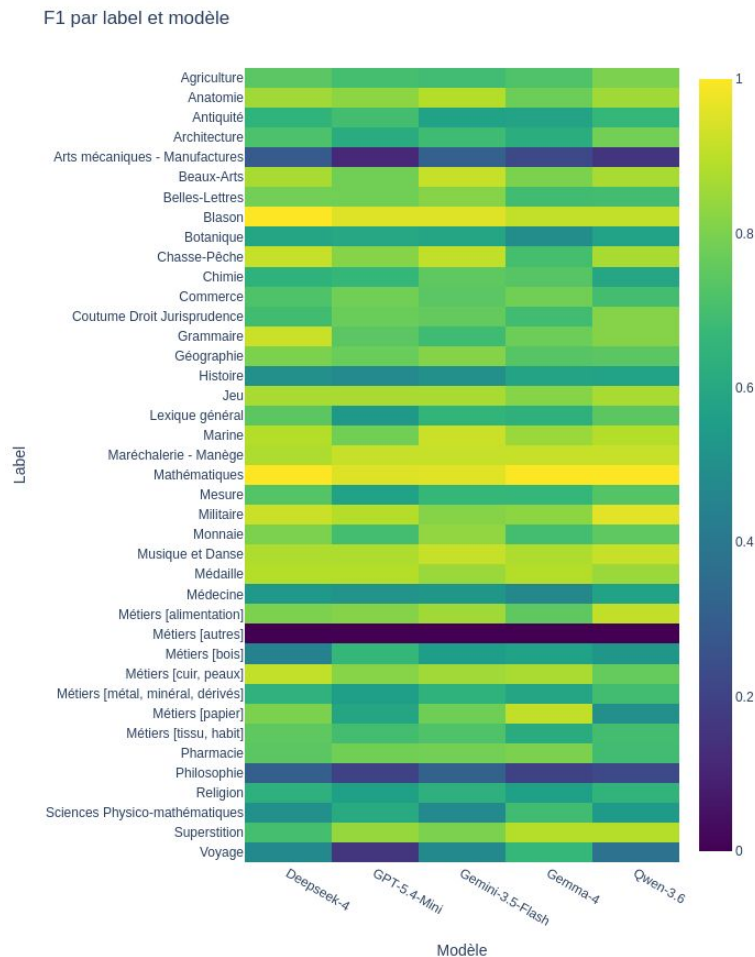
Constitution d'un dataset d'entraînement

Résultats

Mesure F1

3 macro-domaines en difficulté :

- Arts mécaniques - Manufactures
- Métiers [autres]
- Philosophie

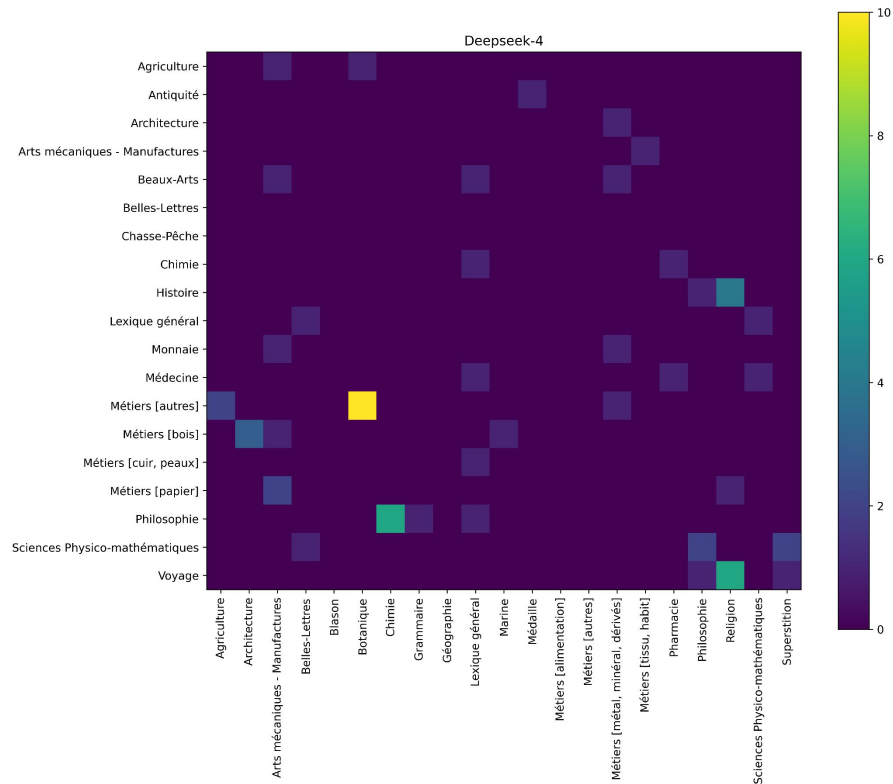


Constitution d'un dataset d'entraînement

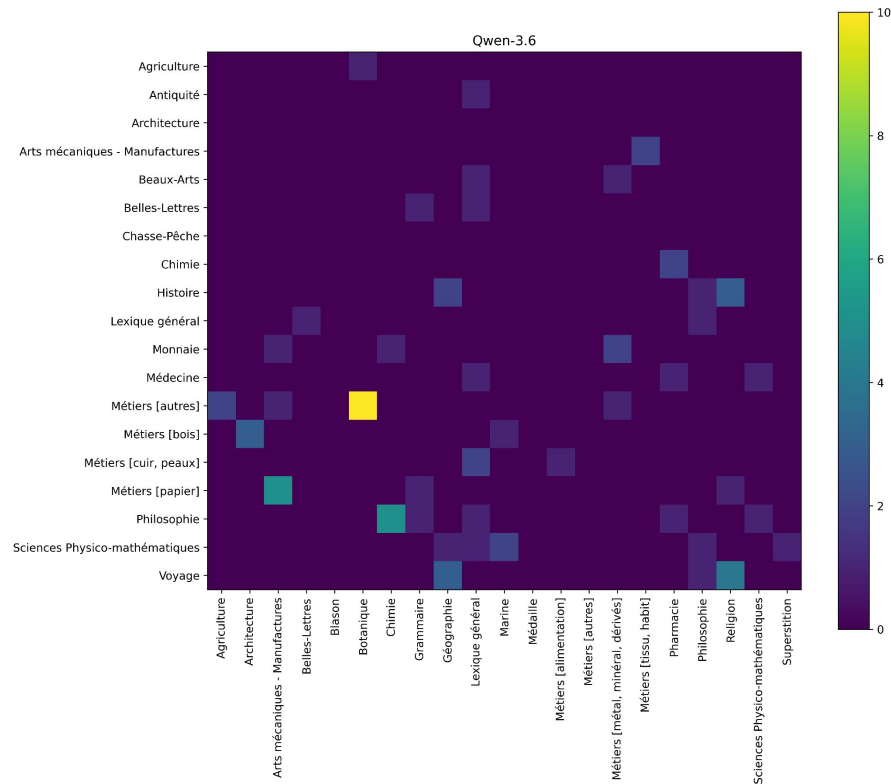
Résultats

Confusions les plus fréquentes

Deepseek-4



Qwen-3.6



Constitution d'un dataset d'entraînement

Résultats

Conclusions de ce premier passage

3 macro-domaines en difficulté

- Arts mécaniques - Manufactures
- Métiers [autres]

=> Font partie du grand ensemble "Métiers".

- Philosophie

=> Prédiction "Chimie" liée aux paragraphes sur la philosophie hermétique, soit l'alchimie.

ARGENTIFIQUE , adj. m. & f. Terme d'Alchymie & de Philosophie hermétique . Qui fait de l'argent, qui a la vertu de faire de l'argent, de changer en argent. *Argentificus*, . Le sel *argentifique*. A la fin du *Dictionnaire hermétique*, imprimé chez d'Hourry en 1695. il y a deux Traités, dont un est intitulé, de la droite & vraie manière de produire la Pierre Philosophique, ou le sel *argentifique* & aurifique.

Pour la suite

Fine-tuning d'un modèle BERT

- Utilisation du LLM pour créer un jeu d'entraînement
 - Ajouter dans le prompt des exemples des macro-domaines en difficulté
 - Corriger les erreurs de correspondance (“philosophie hermétique” -> “Chimie”)
- Entraînement du modèle
 - Framework Set Fit
 - Active Learning