République Algérienne Démocratique et Populaire

الجمهورية الجزائرية الديمقراطية الشعبية

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

وزارة التعليم العالي والبحث العلمي

Département d'informatique, Faculté des Sciences Université M'hamed Bougara
Boumerdes



Mémoire de fin d'études

Pour l'obtention du diplôme de Licence en Informatique

Filière : Informatique

Spécialité : Systèmes Informatiques

Intitulé

Classification Automatique de Textes Médicaux utilisant NLP

Présenté par :

Narimane Tair	/	**Meddahi Idir**
Benkaci ali Doua	/	**Chikhi Aya**

Mémoire Soutenu Devant L'honorable Jury Composé de

Examinateur	:	Dr. TOUAZI Fayçal
Encadreur	:	Dr . BOUSTIL Amel

Année Universitaire : 2025/2026

SESSION : Juin



Remerciement

Louange à Allah, le Tout Miséricordieux et le Tout-Puissant, pour avoir accordé la force, la patience et la persévérance nécessaires à l'accomplissement de ce travail, ainsi que pour avoir guidé chaque étape de ce parcours.

Une profonde gratitude est adressée à nos chers parents, dont le soutien inconditionnel, les sacrifices et les encouragements constants ont constitué une source essentielle de motivation et de détermination tout au long de ces années d'études.

De sincères remerciements sont également exprimés à nos copines pour leur présence, leur écoute, leur soutien moral et leur aide précieuse, qui ont grandement contribué à surmonter les difficultés rencontrées durant la réalisation de ce mémoire.

Une reconnaissance particulière est témoignée à notre encadrante pour son accompagnement attentif, sa disponibilité, ainsi que pour la qualité de ses conseils et de ses orientations, qui ont permis de mener ce travail dans de bonnes conditions.

À travers ces quelques lignes, une pensée reconnaissante est adressée à toutes les personnes qui, de près ou de loin, ont contribué à la réalisation de ce projet et à la réussite de ce parcours universitaire.

Table des matières

Introduction Générale **1**

1 Contexte général et concepts fondamentaux **3**

1.1 Introduction 3

1.2 Les maladies rares et les ressources ontologiques 3

1.2.1 Définition des maladies rares 3

1.2.2 Orphanet : référence des maladies rares 3

1.2.3 HPO : Human Phenotype Ontology 3

1.2.4 Jeu de données basé sur symptômes et maladies rares 4

1.2.5 Notes cliniques 4

1.3 Intelligence artificielle dans le domaine médical 5

1.3.1 Définition de l'intelligence artificielle 5

1.3.2 Machine Learning 5

1.3.3 Réseaux de neurones 6

1.3.4 Deep Learning 6

1.4 Transformers 7

1.5 BERT 7

1.6 Traitement automatique du langage naturel (NLP) 7

1.7 Large Language Models (LLM) 8

1.8 Conclusion 8

2 Méthodologie **9**

2.1 Introduction 9

2.2 Préparation du dataset 10

2.2.1 Dataset de notes cliniques 10

2.2.2 Base de connaissances 10

2.3 Correspondance sémantique et embedding 11

2.3.1 BioBERT 11

2.3.2 Sentence-BERT (SBERT) 11

2.3.3 ClinicalBERT 12

2.3.4 SapBERT 12

2.4 Filtrage textuel par RapidFuzz 14

2.4.1 Présentation de la bibliothèque RapidFuzz 14

2.4.2 Principe de la comparaison textuelle approximative 14

2.4.3 Calcul du score de similarité et réduction vers le Top 7 14

2.4.4 Construction et validation du dataset d'entrée pour le LLM 15

2.5 Diagnostic des maladies rares par LLM 15

2.5.1 Approche naïve — Prompt et dataset complet 15

2.5.2 Modèles testés dans l'approche naïve 15

2.5.3 Approche améliorée par prompt engineering — Llama 3.1 16

2.6 Conclusion 17

3 Évaluation du Pipeline et Analyse des Performances **18**

3.1 Protocole d'évaluation 18

3.1.1 Accuracy 18

3.1.2 F1-score 18

3.1.3 Top-K Accuracy 18

3.2 Résultats de la correspondance sémantique 18

3.3 Résultats du filtrage RapidFuzz 19

3.4 Résultats de l'approche naïve LLM 19

3.5 Résultats de l'approche améliorée — Llama 3.1 19

3.6 Discussion générale 20

3.7 Conclusion 20

Conclusion Générale **21**

Table des figures

1.1 Exemples de maladies rares – code Orpha et phénotype HPO 4

1.2 Exemple de dataset médical [11] 4

1.3 Les domaines de l'IA [3] 5

1.4 Apprentissage supervisé, non supervisé, et par renforcement [9] 5

1.5 Architecture de réseau de neurones profonds [5] 6

1.6 Vector Embedding [33] 7

2.1 Processus de diagnostic des maladies rares à partir des notes cliniques . . . 9

2.2 Extraction des notes cliniques à partir du dataset RareArena 10

2.3 Processus de fusion et de regroupement des données HPO et Orphanet . . 11

2.4 Pipeline d'embedding biomédical basé sur BioBERT, SBERT et Clinical-BERT 12

2.5 Correspondance sémantique par SapBERT — Top-5 phénotypes HPO vers Top-10 maladies candidates 13

2.6 Tableau de comparaison des modèles d'embeddings médicaux 13

2.7 Prompt naïf utilisé pour le diagnostic des maladies rares par LLM 15

2.8 Modèle MiniMax-M2.7 16

2.9 Modèle Claude Sonnet 4.6 16

2.10 Scores de similarité RapidFuzz et sélection du Top 7 des maladies candidates 16

2.11 Prompt structuré utilisé pour le raisonnement diagnostique du modèle Llama 3.1 16

3.1 Comparaison des performances des modèles d'embeddings et des modèles LLM pour le diagnostic automatique des maladies rares 20

Introduction Générale

Le diagnostic des maladies rares constitue aujourd'hui un enjeu majeur de santé publique. Bien que chaque maladie rare affecte un nombre restreint de personnes, l'ensemble de ces pathologies touche plus de 300 millions d'individus dans le monde. La complexité du diagnostic s'explique par plusieurs facteurs, notamment la faible fréquence des cas, l'hétérogénéité des manifestations cliniques ainsi que la diversité et la complexité des données médicales contenues dans les dossiers des patients. Cette situation entraîne souvent une « errance diagnostique », durant laquelle les patients peuvent attendre plusieurs années avant qu'un diagnostic précis ne soit établi [12].

Les notes cliniques et les dossiers médicaux contiennent une grande quantité d'informations textuelles importantes pouvant aider à identifier ces maladies. Cependant, ces données sont généralement non structurées, hétérogènes et difficiles à exploiter manuellement. Cette problématique a encouragé l'utilisation de l'Intelligence Artificielle et particulièrement des techniques de Traitement Automatique du Langage Naturel (Natural Language Processing – NLP) afin d'analyser automatiquement les textes médicaux.

Ces dernières années, les techniques de Machine Learning et de Deep Learning ont fortement progressé dans le domaine biomédical. Des modèles comme **BioBERT**, **ClinicalBERT**, **Sentence-BERT** et **SapBERT** améliorent la compréhension des textes médicaux et les performances des systèmes de classification. De plus, les modèles de **langage de grande taille (LLM)** facilitent l'analyse automatique des données cliniques non structurées et contribuent à l'aide au diagnostic, notamment pour les maladies rares, en réduisant l'errance diagnostique des patients [23].

Dans le cadre de ce projet, nous nous intéressons à la classification des textes médicaux liés aux maladies rares. Pour cela, nous avons exploité d'abord des ressources médicales spécialisées telles que **HPO (Human Phenotype Ontology)** [16] et **ORDO (Orphanet Rare Disease Ontology)** [34] afin de construire un dataset regroupant les maladies rares et leurs symptômes associés.

Afin d'améliorer le processus de classification, nous avons proposé une approche hybride en plusieurs niveaux. Dans une première étape, des modèles d'embedding biomédicaux, notamment **ClinicalBERT** [1], **Sentence-BERT** [31], **BioBERT** [17] et **SapBERT** [19] ont été utilisés pour représenter les textes médicaux sous forme vectorielle et sélectionner les maladies les plus proches sémantiquement. Ensuite, une méthode de similarité textuelle basée sur **RapidFuzz** [4] a été appliquée afin d'affiner les résultats et de conserver les maladies les plus pertinentes. Finalement, un **Large Language Model (LLM)**, basé sur Llama 3, a été utilisé comme assistant diagnostique pour choisir la maladie la plus probable parmi les meilleures prédictions obtenues.

Les résultats expérimentaux obtenus montrent que l'intégration des modèles de langage biomédicaux, des techniques de similarité textuelle et des LLMs permet d'améliorer significativement les performances de classification des maladies rares.

Problématique

Le diagnostic des maladies rares reste difficile en raison de la rareté des cas, de la similarité des symptômes et de la faible qualité des datasets médicaux disponibles [12]. Les données provenant de **Human Phenotype Ontology** [16] et **Orphanet** [27] contiennent souvent des descriptions limitées et peu enrichies, ce qui complique la classification automatique des textes médicaux.

De plus, les notes cliniques sont non structurées et difficiles à analyser avec les méthodes classiques de traitement de texte.

- Comment améliorer la classification automatique des textes médicaux liés aux maladies rares malgré la faiblesse des données médicales disponibles ?

Solution

Pour répondre à cette problématique, nous avons proposé une approche hybride basée sur plusieurs niveaux.

Tout d'abord, les datasets **HPO** et **ORDO** ont été fusionnés afin de construire un dataset médical des maladies rares. Ensuite, des modèles biomédicaux tels que **ClinicalBERT**, **Sentence-BERT**, **BioBERT** et **SapBERT** ont été utilisés pour calculer la similarité sémantique entre les notes médicales et les **maladies rares**.

Par la suite, une méthode de similarité textuelle utilisant **RapidFuzz** a permis de sélectionner les maladies les plus pertinentes. Enfin, les modèles **Llama 3.1** et **Llama 3** ont été utilisés comme assistant diagnostique afin de choisir la maladie la plus probable parmi les meilleures prédictions.

Organisation du mémoire

Nous avons structuré notre mémoire en trois (03) chapitres :

- **Chapitre 1 : Contexte général et concepts fondamentaux.**
  Ce chapitre présente les maladies rares, les bases de données Human Phenotype Ontology et Orphanet, ainsi que les concepts de l'intelligence artificielle, du NLP, des LLM, des embeddings et des outils utilisés dans ce travail, notamment BioBERT, Sentence-BERT, ClinicalBERT, SapBERT et RapidFuzz.
- **Chapitre 2 : Méthodologie.**
  Description de l'approche méthodologique adoptée, incluant la fusion des jeux de données médicaux, l'utilisation des modèles d'embeddings BioBERT, ClinicalBERT, Sentence-BERT et SapBERT pour la représentation des données textuelles, ainsi que l'application de RapidFuzz pour le filtrage des maladies et l'intégration des modèles de langage à grande échelle (LLM).
- **Chapitre 3 : Évaluation du Pipeline et Analyse des Performances.**
  Présentation de l'utilisation de **Llama 3** et **Llama 3.1** pour le diagnostic des maladies rares ainsi que l'évaluation et l'analyse des résultats obtenus.

Chapitre 1

Contexte général et concepts fondamentaux

1.1 Introduction

Ce chapitre présente les concepts fondamentaux nécessaires à la compréhension de notre travail dans le domaine des maladies rares. Il aborde les difficultés liées à leur diagnostic ainsi que les notions d'intelligence artificielle, notamment le machine learning, le deep learning et le traitement automatique du langage naturel (NLP), qui constituent la base de notre approche. Nous présentons également les datasets médicaux utilisés, en particulier Human Phenotype Ontology et Orphanet Rare Disease Ontology.

1.2 Les maladies rares et les ressources ontologiques

1.2.1 Définition des maladies rares

Les maladies rares sont des pathologies caractérisées par leur faible prévalence au sein de la population générale. L'Organisation Mondiale de la Santé (OMS) les définit comme des maladies dont la fréquence est très faible et dont le diagnostic est souvent complexe en raison de leur diversité clinique [37].

Ces maladies se caractérisent par une grande diversité de symptômes, souvent similaires à ceux d'autres pathologies, ce qui rend leur diagnostic particulièrement complexe. Le manque d'expertise spécialisée et la rareté des cas cliniques disponibles entraînent fréquemment des retards diagnostiques, impactant la prise en charge des patients.

1.2.2 Orphanet : référence des maladies rares

Orphanet est une base de données et un portail de référence international dédié aux maladies rares et aux médicaments orphelins. Il fournit des informations structurées et validées sur les maladies rares, incluant leur description clinique, leur classification, leurs synonymes ainsi que les gènes associés.

Orphanet joue un rôle essentiel dans la standardisation des connaissances médicales relatives aux maladies rares. Il permet également de faciliter l'identification des pathologies et l'harmonisation des données entre différentes sources médicales [34].

1.2.3 HPO : Human Phenotype Ontology

Human Phenotype Ontology (HPO) est une ontologie biomédicale qui fournit une représentation standardisée, structurée et hiérarchique des phénotypes humains.

Elle permet de décrire de manière précise les signes cliniques et symptômes observés chez un patient à l'aide de termes normalisés.

Cette ontologie est largement utilisée dans les approches d'intelligence artificielle pour la représentation structurée des symptômes cliniques [16].

1.2.4 Jeu de données basé sur symptômes et maladies rares

Dans le cadre de ce travail, un fichier de données a été utilisé, contenant des associations entre des symptômes (représentés par des termes HPO) et des maladies rares (référencées via Orphanet). Cette structuration permet de relier les phénotypes observés aux pathologies correspondantes.

Maladie rare	Code ORPHA	Symptômes principaux	Code HPO
Syndrome de Marfan	ORPHA:558	Grande taille, arachnodactylie, dilatation aortique	HP:0001519 HP:0004942
Phénylcétonurie (PKU)	ORPHA:716	Déficience intellectuelle, eczéma, hyperactivité	HP:0001249 HP:0000708
Maladie de Gaucher	ORPHA:77259	Splénomégalie, hépatomégalie, douleurs osseuses	HP:0001744 HP:0002653
Mucoviscidose	ORPHA:586	Insuffisance pulmonaire, infections bronchiques	HP:0006536 HP:0002099
Syndrome de Rett	ORPHA:778	Régression neurologique, crise d'épilepsie	HP:0000252 HP:0001250

Fig. 1.1 : Exemples de maladies rares – code Orpha et phénotype HPO

1.2.5 Notes cliniques

Les symptômes médicaux constituent la principale source d'information permettant d'orienter le diagnostic d'une maladie. Dans le contexte des maladies rares, ils se manifestent de manière hétérogène et variable d'un patient à l'autre, ce qui complique considérablement leur interprétation clinique. En effet, un même symptôme peut être associé à plusieurs maladies différentes, et une même maladie rare peut se manifester différemment selon les patients, rendant ainsi toute tentative de classification particulièrement délicate.

Les notes cliniques sont des documents textuels rédigés par les professionnels de santé lors des consultations médicales. Elles regroupent les symptômes observés, les antécédents du patient ainsi que les hypothèses diagnostiques formulées par le médecin. Ces notes représentent une ressource précieuse pour tout système automatique de classification des maladies rares.

Dans ce travail, le dataset utilisé est issu du benchmark RareArena [11], un ensemble de cas cliniques construit à partir de la littérature biomédicale. Il contient des descriptions de patients incluant les symptômes, les antécédents médicaux et les évolutions cliniques, associées à des diagnostics de maladies rares. L'ensemble couvre plusieurs pathologies rares afin de tester la performance des systèmes sur des cas complexes et peu fréquents.

	1	2	3	4	5	6	7	8
1	_id	case_report	diagnosis	Orpha_name	Orpha_id	age	gender	pub_date
2	8289685-1	A 38-year-old	Gastric Neur	Gastric Neuroen	100075	[[38.0, 'year'	M	2021-07
3	2683450-1	A three-mon	Aicardi syndr	Aicardi syndrom	50	[[3.0, 'month	F	2009
4	7443986-1	A 25-year-old	Epithelioid m	Malignant perip	3148	[[25.0, 'year'	F	21/08/2020

Fig. 1.2 : Exemple de dataset médical [11]

1.3 Intelligence artificielle dans le domaine médical

1.3.1 Définition de l'intelligence artificielle

L'Intelligence Artificielle (IA) est une discipline au croisement de l'informatique, des mathématiques et des sciences cognitives, visant à concevoir des systèmes capables de simuler et d'améliorer les capacités intellectuelles humaines. Elle repose sur des modèles mathématiques et des algorithmes d'apprentissage automatique permettant aux machines d'exploiter les données, d'en extraire des connaissances et de prendre des décisions dans des contextes complexes. Au-delà de l'automatisation, l'IA se distingue par sa capacité à transformer de grandes quantités de données en informations pertinentes, ce qui en fait un élément clé des systèmes intelligents [7].

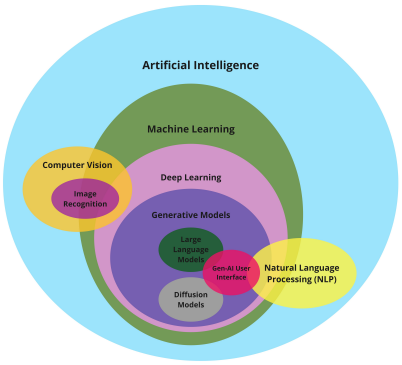


Fig. 1.3 : Les domaines de l'IA [3]

1.3.2 Machine Learning

Le machine learning est une branche de l'intelligence artificielle qui permet aux systèmes informatiques d'apprendre automatiquement à partir des données afin d'identifier des patterns et de produire des prédictions [22]. Dans le contexte des maladies rares, ces algorithmes permettent d'extraire des informations pertinentes à partir de données limitées et hétérogènes, facilitant ainsi la classification des maladies et l'identification des relations entre les symptômes et les pathologies [25, 28].

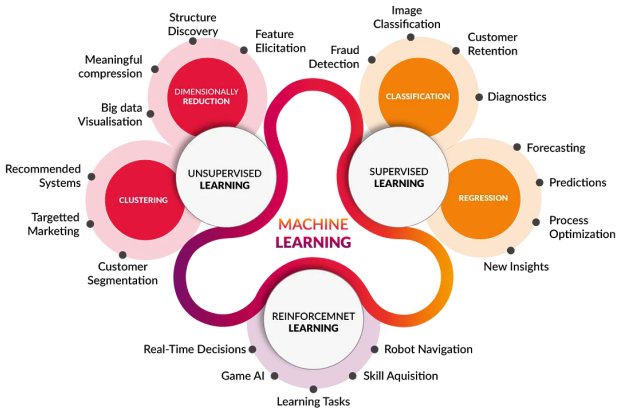


Fig. 1.4 : Apprentissage supervisé, non supervisé, et par renforcement [9]

1.3.3 Réseaux de neurones

Les réseaux de neurones artificiels sont des modèles informatiques inspirés du fonctionnement du cerveau humain. Ils sont composés de couches de neurones interconnectés : une couche d'entrée, des couches cachées et une couche de sortie. Chaque connexion entre les neurones possède un poids qui s'ajuste automatiquement lors de la phase d'apprentissage afin de minimiser les erreurs de prédiction.

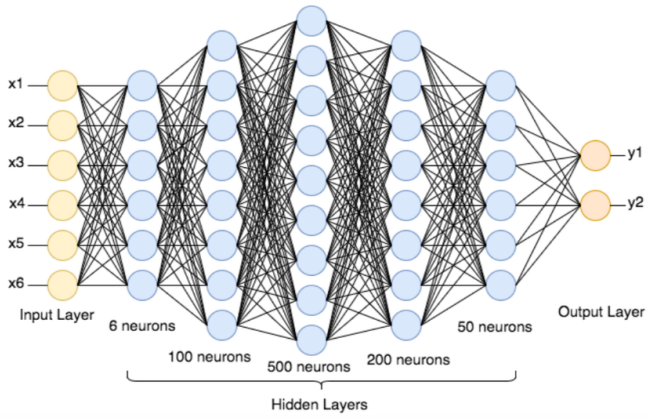


Fig. 1.5 : Architecture de réseau de neurones profonds [5]

1.3.4 Deep Learning

Le deep learning représente une évolution du machine learning basée sur l'utilisation de réseaux de neurones artificiels comportant plusieurs couches. Ces modèles sont capables d'apprendre des représentations hiérarchiques des données, ce qui leur permet de capturer des relations complexes et abstraites [14].

Dans le domaine médical, le deep learning permet d'analyser efficacement les textes cliniques et les notes médicales. Il facilite l'extraction des informations importantes ainsi que l'identification des relations entre les symptômes afin d'améliorer la précision du diagnostic.

1.4 Transformers

L'architecture Transformer est une architecture de Deep Learning basée sur le mécanisme d'attention, permettant de comprendre le contexte des textes et de transformer les données textuelles en vecteurs numériques (embeddings). Cette architecture améliore les performances des systèmes de traitement automatique du langage naturel [35].

Dans le domaine médical, les Transformers sont utilisés pour analyser les textes cliniques, extraire les symptômes et améliorer la classification ainsi que le diagnostic des maladies à partir des données médicales non structurées.

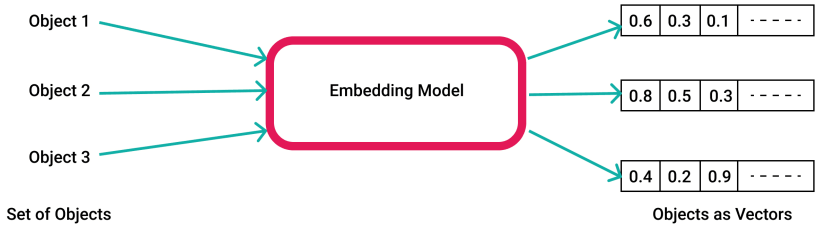


Fig. 1.6 : Vector Embedding [33]

1.5 BERT

BERT (Bidirectional Encoder Representations from Transformers) est un modèle de langage pré-entraîné développé par Google en 2018. Il analyse le texte de manière bidirectionnelle, c'est-à-dire en tenant compte du contexte à gauche et à droite de chaque mot simultanément, ce qui lui permet de capturer une compréhension plus fine et plus précise du langage naturel.

BERT a établi des performances remarquables sur de nombreuses tâches de NLP, notamment la classification de textes, la reconnaissance d'entités nommées et la question-réponse, constituant ainsi une base solide pour les modèles spécialisés qui en sont dérivés, tels que ClinicalBERT et Sentence-BERT utilisés dans notre projet [10].

1.6 Traitement automatique du langage naturel (NLP)

Le Traitement du Langage Naturel (**NLP**) est une branche de l'intelligence artificielle qui vise à permettre aux machines de comprendre, analyser et traiter le langage humain sous forme de texte ou de parole.

Il permet ainsi d'analyser et de comprendre les données textuelles en s'appuyant sur plusieurs étapes essentielles, notamment le nettoyage (suppression du bruit et des caractères inutiles) et la tokenisation, qui consiste à découper le texte en unités appelées *tokens* [24, 13].

1.7 Large Language Models (LLM)

Les Large Language Models (LLM) sont des modèles d'intelligence artificielle basés sur l'architecture des transformeurs, conçus pour comprendre, analyser et générer du langage naturel. Entraînés sur de grands volumes de données textuelles, ils sont capables de capturer le contexte et les relations sémantiques entre les mots afin d'effectuer diverses tâches de traitement automatique du langage naturel. Grâce à leurs performances élevées, les LLM sont largement utilisés dans plusieurs domaines, notamment l'analyse de textes, les systèmes conversationnels et l'aide à la décision médicale [20, 6].

1.8 Conclusion

Ce chapitre a présenté les concepts fondamentaux nécessaires à la compréhension de notre travail, notamment les maladies rares, les bases de connaissances HPO et Orphanet, ainsi que les techniques d'intelligence artificielle utilisées, en particulier le NLP, les modèles Transformers, BERT et les grands modèles de langage. Ces notions constituent le socle théorique sur lequel repose l'ensemble de l'approche méthodologique développée dans le chapitre suivant.

Chapitre 2

Méthodologie

2.1 Introduction

Ce chapitre présente la méthodologie mise en œuvre pour le diagnostic automatique des maladies rares à partir de notes cliniques textuelles, à travers un pipeline multi-étapes combinant NLP, représentation vectorielle sémantique et grands modèles de langage (LLM) [20].

Le pipeline s'articule autour de trois phases principales :

- La première couvre la préparation des données : exploitation du dataset de notes cliniques, construction d'une base de connaissances à partir des ontologies HPO et Orphanet, et extraction des phénotypes pertinents pour chaque note clinique.
- La deuxième est dédiée à la correspondance sémantique via les modèles BioBERT, ClinicalBERT, Sentence-BERT et SapBERT, permettant de réduire l'espace de recherche vers un Top 10, affiné ensuite vers un Top 7 par filtrage textuel avec RapidFuzz.
- La troisième exploite les LLM pour identifier la maladie finale : une approche naïve testée avec MiniMax-M2.7 et Claude Sonnet 4.6, et une approche améliorée par prompt engineering avec Llama 3.1.

L'ensemble forme un pipeline progressif allant de la donnée brute jusqu'au diagnostic final, dont la figure 2.1 présente une vue d'ensemble.

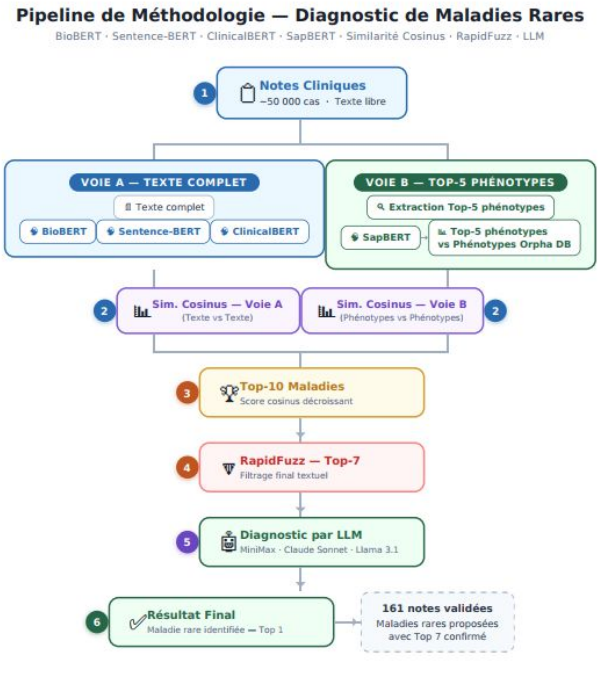


Fig. 2.1 : Processus de diagnostic des maladies rares à partir des notes cliniques

2.2 Préparation du dataset

2.2.1 Dataset de notes cliniques

Dans le cadre de ce travail, nous avons utilisé le dataset RareArena qui est un benchmark public dédié au diagnostic des maladies rares, regroupant près de 50 000 patients et couvrant plus de 4 000 maladies rares, publié dans *The Lancet Digital Health* (2026) [30]. Construit à partir de cas cliniques issus de PubMed Central et structuré via GPT-4o, il offre pour chaque entrée un identifiant unique, un rapport clinique, un diagnostic ainsi que des métadonnées Orphanet et démographiques. Dans le cadre de ce projet, seules les colonnes *_id* et *case_report* ont été retenues comme entrée du pipeline NLP. Le dataset est disponible sur GitHub :
https://github.com/zhao-zy15/RareArena.

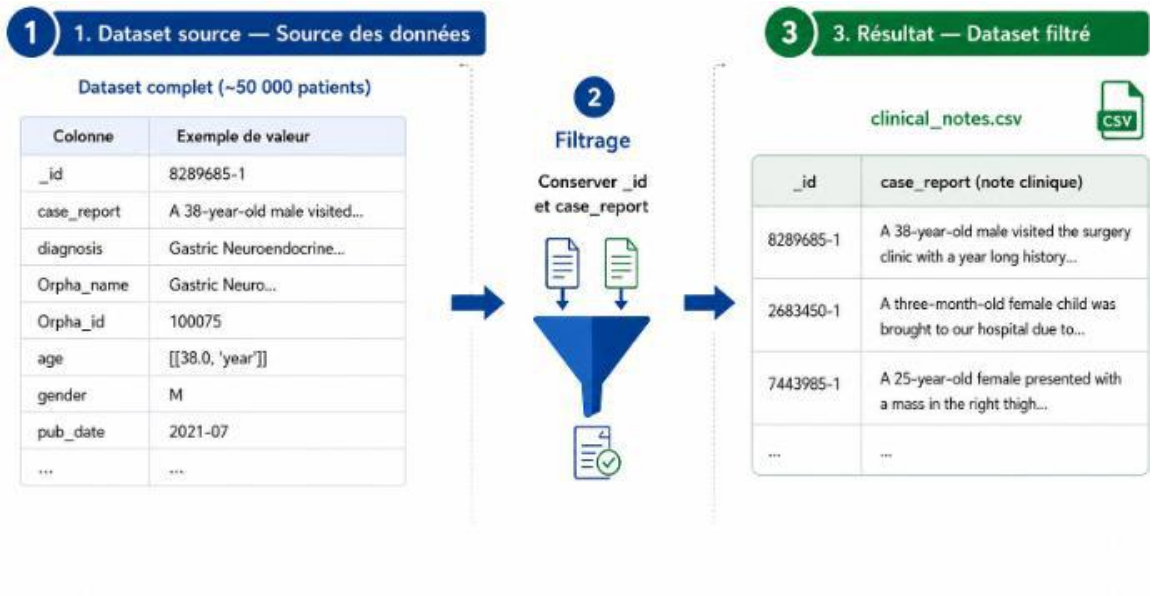


Fig. 2.2 : Extraction des notes cliniques à partir du dataset RareArena

2.2.2 Base de connaissances

Dans ce projet, RapidFuzz compare les mots-clés médicaux des notes cliniques aux phénotypes des maladies rares de notre base de connaissances, afin d'affiner les résultats de ClinicalBERT et de réduire le Top-10 vers un Top-7 plus précis.

Ensuite, afin d'optimiser l'organisation des données, nous avons appliqué un traitement de regroupement. Cette opération consiste à regrouper les phénotypes (*Phenotype_Name*) appartenant à une même maladie et au même *orpha_id*, puis à les concaténer dans une seule ligne séparée par des points-virgules. Le fichier obtenu après cette étape de prétraitement est nommé ***matrice_finale_regroupee_finale.csv***, et constitue la base de données finale utilisée dans notre projet.

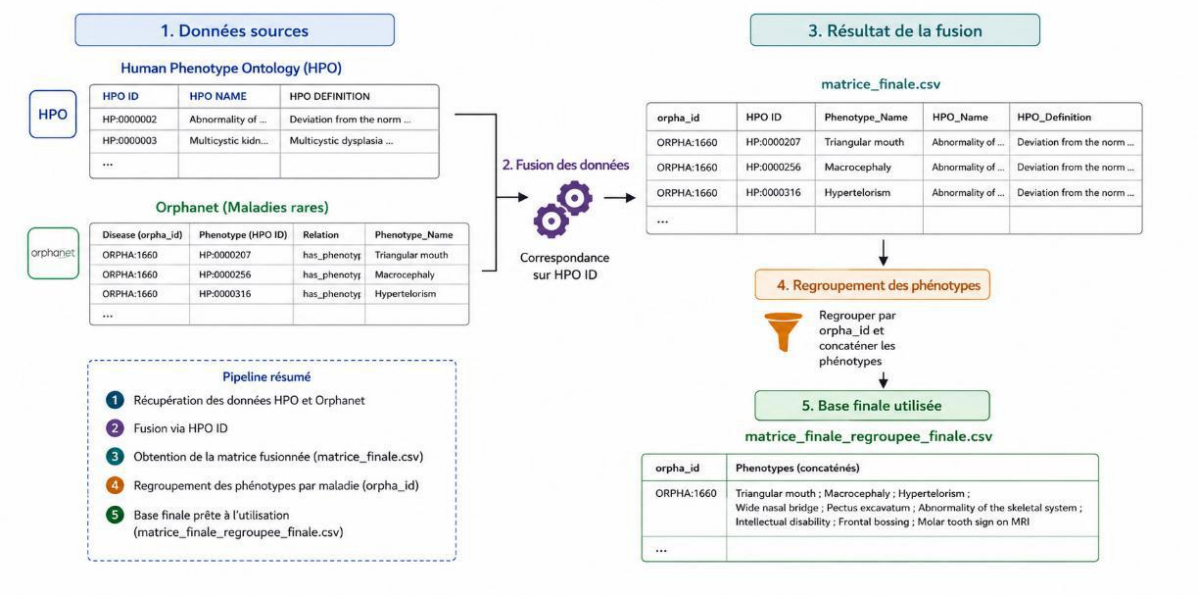


Fig. 2.3 : Processus de fusion et de regroupement des données HPO et Orphanet

2.3 Correspondance sémantique et embedding

Afin de générer les embeddings des notes cliniques et des maladies rares, plusieurs modèles biomédicaux ont été étudiés.

2.3.1 BioBERT

BioBERT est une version de BERT spécialisée dans le domaine biomédical et entraînée sur des articles scientifiques tels que PubMed. Il produit des vecteurs de dimension 768 et est particulièrement adapté au traitement de textes médicaux et biologiques [17].

2.3.2 Sentence-BERT (SBERT)

SBERT est une adaptation de BERT optimisée pour la comparaison sémantique de phrases. Il génère également des embeddings de dimension 768, efficaces pour mesurer la similarité entre deux textes [31].

2.3.3 ClinicalBERT

ClinicalBERT est une variante de BERT entraînée sur des données cliniques réelles. Il produit des vecteurs de dimension 768 et permet une meilleure représentation du vocabulaire médical et des notes cliniques [15].

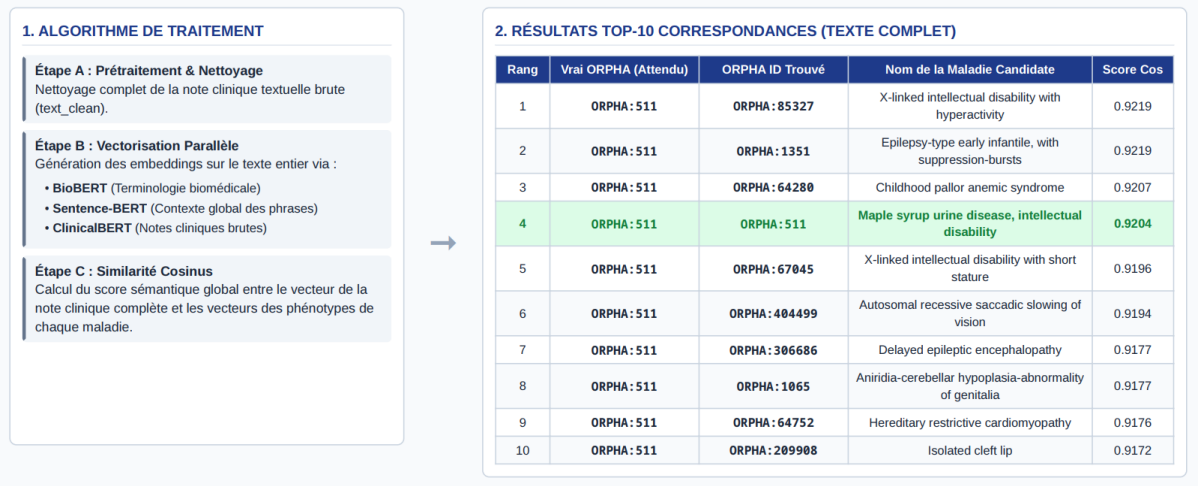


Fig. 2.4 : Pipeline d'embedding biomédical basé sur BioBERT, SBERT et ClinicalBERT

2.3.4 SapBERT

SapBERT (*Self-Alignment Pre-training for Biomedical Entity Representation*) est un modèle biomédical pré-entraîné sur les ontologies médicales UMLS. Il génère des embeddings de dimension 768 et est spécialement conçu pour capturer la similarité sémantique entre entités biomédicales telles que les maladies et les phénotypes [19].

Après comparaison, SapBERT a été retenu dans ce travail grâce à sa meilleure capacité à représenter les relations sémantiques entre les termes médicaux et à améliorer la correspondance entre les notes cliniques et les maladies rares candidates.

Préparation des données d'entrée

Deux fichiers sont utilisés comme entrées du pipeline :

- *RDC_final_organised.xlsx* : contient les notes cliniques enrichies, avec pour chaque note les Top-5 phénotypes les plus proches, identifiés par leurs codes HPO et leurs dénominations standardisées.
- *matrice_finale_regroupee_finale.csv* : base de référence des maladies rares issue d'Orphanet, qui associe à chaque identifiant Orphanet l'ensemble de ses phénotypes caractéristiques concaténés.

Ces deux fichiers constituent les deux côtés de la comparaison : d'un côté les phénotypes extraits des notes cliniques, de l'autre les phénotypes associés à chaque maladie rare.

Génération des embeddings via SapBERT

SapBERT est appliqué sur les deux fichiers afin de transformer chaque ensemble de phénotypes en un vecteur de représentation dense (embedding) dans un espace sémantique commun [33]. Concrètement, chaque note clinique est représentée par l'embedding de ses Top-5 phénotypes concaténés, et chaque maladie rare est représentée par l'embedding de l'ensemble de ses phénotypes Orphanet. Cette projection dans un espace vectoriel commun permet de comparer directement les phénotypes des notes avec ceux des maladies, indépendamment des différences de formulation ou de terminologie.

Calcul de la similarité cosinus

Une fois les embeddings générés, la similarité cosinus est calculée entre chaque note clinique et l'ensemble des maladies rares de la base Orphanet [31]. Cette métrique mesure l'angle entre deux vecteurs dans l'espace sémantique : plus l'angle est petit, plus les deux entités sont sémantiquement proches. Pour chaque note, les maladies sont classées par ordre décroissant de similarité.

Sélection du Top-10 et sauvegarde

Pour chaque note clinique, les 10 maladies ayant obtenu les scores de similarité les plus élevés sont sélectionnées et constituent le Top-10 des maladies candidates. Les résultats sont sauvegardés dans le fichier *predictions_top10_sapbert.csv* (fichier brut des prédictions) ainsi que dans le fichier *predictions_top10_sapbert_organised.xlsx* (fichier organisé avec les colonnes structurées pour les étapes suivantes).

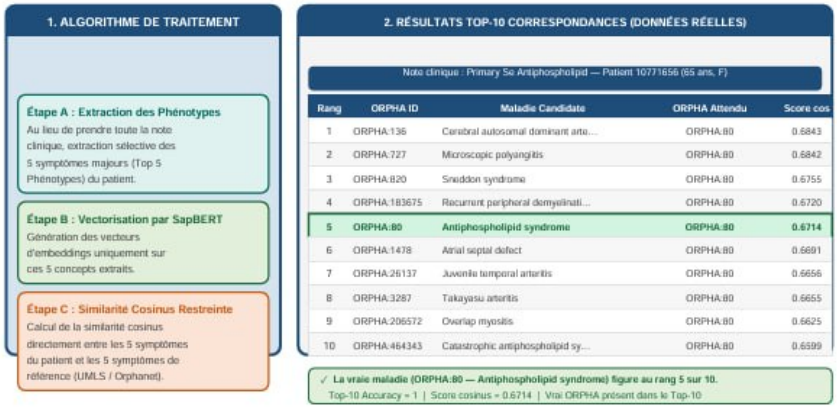


Fig. 2.5 : Correspondance sémantique par SapBERT — Top-5 phénotypes HPO vers Top-10 maladies candidates

Critère	BioBERT	ClinicalBERT	SentenceBERT	SapBERT
Modèle de base	BERT	BERT	BERT	BERT
Données d'entraînement	Littérature biomédicale (PubMed)	Notes cliniques (MIMIC-III)	Corpus général + paires de phrases	Ontologies biomédicales (UMLS, MeSH...)
Domaine d'application	Biomédical	Médical clinique	Similarité sémantique	Entités biomédicales
Type de représentation	Token	Token	Phrase entière	Entité / Phrase
Adapté aux notes médicales	Partiellement	Oui ✓	Partiellement	Oui ✓
Comparaison de phrases	Non	Non	Oui ✓	Oui ✓
Métrique de similarité	Cosinus	Cosinus	Cosinus	Cosinus
Utilisé dans notre projet	Oui ✓	Oui ✓	Oui ✓	Oui ✓
Source	Lee et al. (2020)	Alsentzer et al. (2019)	Reimers & Gurevych (2019)	Liu et al. (2021)

Fig. 2.6 : Tableau de comparaison des modèles d'embeddings médicaux

2.4 Filtrage textuel par RapidFuzz

2.4.1 Présentation de la bibliothèque RapidFuzz

RapidFuzz est une bibliothèque open-source Python implémentée en C++, conçue pour effectuer des comparaisons de chaînes de caractères de manière rapide et efficace. Elle constitue une alternative optimisée à **FuzzyWuzzy**, offrant des temps d'exécution nettement plus courts grâce à son implémentation bas niveau [4]. Elle repose sur des algorithmes de correspondance textuelle approximative (***fuzzy string matching***), qui mesurent le degré de similarité entre deux textes, même en présence de fautes orthographiques, d'abréviations ou de variations de formulation [8].

Dans le cadre de ce projet, **RapidFuzz** a été utilisé pour comparer les mots-clés médicaux extraits des notes cliniques avec les phénotypes des maladies rares issus de notre base de connaissances, afin d'affiner la sélection issue de **ClinicalBERT** et de réduire le **Top 10** des maladies candidates vers un **Top 7** plus précis et pertinent.

2.4.2 Principe de la comparaison textuelle approximative

Le fuzzy matching mesure la ressemblance entre deux textes sans exiger qu'ils soient identiques. Par exemple, si un médecin écrit « cardiomyopathie » face au phénotype « cardiomyopathie dilatée », le fuzzy matching détecte la proximité là où une recherche exacte échouerait

RapidFuzz utilise notamment la distance de Levenshtein, qui compte le nombre minimal d'opérations (insertion, suppression, substitution d'un caractère) nécessaires pour transformer un texte en un autre [18]. Le résultat est un score entre **0** et **1** : 1 indique une correspondance parfaite, 0 indique aucune similarité. Dans notre pipeline, ce score est calculé entre les mots-clés de chaque note clinique et les phénotypes des dix maladies candidates issues de ClinicalBERT, permettant ainsi de classer et filtrer ces maladies selon leur pertinence textuelle.

2.4.3 Calcul du score de similarité et réduction vers le Top 7

Après application de **RapidFuzz** sur les dix maladies candidates, un score de similarité entre 0 et 1 est attribué à chaque maladie pour chaque note clinique, reflétant le degré de correspondance entre la note et les phénotypes de la maladie. Les dix maladies sont ensuite triées par ordre décroissant selon ce score : les sept maladies les mieux classées sont conservées (**NIVEAU 2 : TOP 7 — Keep**), tandis que les trois maladies les moins pertinentes sont écartées du pipeline (**FAIBLE : ÉLIMINÉ — Drop 3**) [4, 8].

Cette réduction du **Top 10 vers le Top 7** constitue une étape clé de notre pipeline, car elle réduit l'espace de recherche tout en préservant les pathologies les plus proches sémantiquement et textuellement de chaque note clinique. Elle améliore ainsi la qualité des maladies candidates transmises à l'étape suivante, en éliminant les maladies les moins pertinentes et en concentrant l'analyse sur les pathologies les plus probables pour chaque note médicale.

2.4.4 Construction et validation du dataset d'entrée pour le LLM

Après le filtrage par **RapidFuzz**, une étape de validation finale est appliquée : pour chaque note clinique, on vérifie si la vraie maladie (**identifiée par son Orpha ID**) figure parmi les 7 maladies candidates du **Top 7**. Les notes pour lesquelles la vraie maladie est absente sont écartées, car elles rendraient toute expérimentation ultérieure non pertinente.

À l'issue de cette validation, le dataset final contient **161 notes** médicales validées, chacune accompagnée de ses **sept maladies** candidates avec leurs phénotypes, identifiants **Orpha ID** et scores de similarité. Ce dataset constitue l'entrée de l'étape suivante, où **Llama 3** sera utilisé pour identifier la vraie maladie parmi les sept candidates proposées.

2.5 Diagnostic des maladies rares par LLM

Le recours aux grands modèles de langage (LLM) constitue l'étape finale du pipeline de diagnostic. Deux approches distinctes ont été expérimentées : une **approche naïve** testée sur le dataset complet sans filtrage préalable, et une **approche améliorée** basée sur le prompt engineering appliquée aux 161 notes validées avec Top 7.

2.5.1 Approche naïve — Prompt et dataset complet

Dans la première approche, un prompt simple et direct est conçu pour demander au modèle d'identifier la maladie rare la plus probable à partir d'une note clinique, sans lui fournir de liste de maladies candidates. Le prompt est formulé comme suit : le modèle doit lire la note clinique et retourner uniquement le nom de la maladie, sans explication ni alternatives [32]. Cette approche est qualifiée de naïve car elle ne s'appuie sur aucun filtrage préalable ni contrainte structurelle : le modèle génère librement sa réponse. Ce prompt est appliqué sur le dataset complet de notes cliniques (∼50 000 notes), afin d'évaluer les capacités brutes des modèles sans aucune aide contextuelle.

2.5.2 Modèles testés dans l'approche naïve

Le même prompt naïf est appliqué sur deux modèles distincts afin de comparer leurs performances intrinsèques sur la tâche de diagnostic.

```
def load_jsonl(path):
    with path.open("r", encoding="utf-8") as f:
        return [json.loads(line) for line in f if line.strip()]


def prompt_model(note):
    prompt = f"""Read the clinical note and name the single most likely rare disease

Rules:
- Output only the disease name.
- Do not explain.
- Do not list alternatives.

Clinical note:
{note}
"""
```

Fig. 2.7 : Prompt naïf utilisé pour le diagnostic des maladies rares par LLM

MiniMax-M2.7 : modèle de type Mixture of Experts développé par MiniMaxAI, accessible via OpenRouter, reconnu pour ses capacités de raisonnement sur des tâches complexes [21].

```
MODEL = "minimaxai/minimax-m2.7"
```

Fig. 2.8 : Modèle MiniMax-M2.7

Claude Sonnet 4.6 : modèle développé par Anthropic, accessible via OpenRouter, reconnu pour sa compréhension contextuelle avancée et son raisonnement structuré [2].

```
OPENROUTER_MODEL = os.getenv("OPENROUTER_MODEL", "anthropic/claude-sonnet-4.6")
```

Fig. 2.9 : Modèle Claude Sonnet 4.6

Les deux modèles reçoivent uniquement la note clinique brute et doivent produire directement le nom de la maladie, ce qui permet une comparaison équitable de leurs capacités diagnostiques sans biais lié au prompt.

2.5.3 Approche améliorée par prompt engineering — Llama 3.1

Dans la deuxième approche, le pipeline est significativement amélioré à deux niveaux : le choix du dataset et la conception du prompt.

	A	B	C	D	E	F	G	H	I	J	K	L
1	_Id	clean_text	Orpha_Id	ng_Match	ore_Cos_S	aladie_no	_Phenotyp	Maladie_orpha_Id	e_Medical	tus_Final_N2		
2	10001893	case male	511	4	0,9204	Maple syr	Intellectua	ORPHA:511	0,52	NIVEAU 2 : TOP 7 (Keep)		
3	10001893	case male	511	7	0,9177	Delayed e	Decreasec	ORPHA:306686	0,509091	NIVEAU 2 : TOP 7 (Keep)		
4	10001893	case male	511	9	0,9176	Hereditar	Abnormal	ORPHA:64752	0,46	NIVEAU 2 : TOP 7 (Keep)		
5	10001893	case male	511	5	0,9196	X-linked in	Small post	ORPHA:67045	0,430556	NIVEAU 2 : TOP 7 (Keep)		
6	10001893	case male	511	8	0,9177	Aniridia-c	Abnormal	ORPHA:1065	0,423729	NIVEAU 2 : TOP 7 (Keep)		
7	10001893	case male	511	3	0,9207	Childhood	Pallor Anx	ORPHA:64280	0,423077	NIVEAU 2 : TOP 7 (Keep)		
8	10001893	case male	511	10	0,9172	Isolated cl	Abnormal	ORPHA:209908	0,416667	NIVEAU 2 : TOP 7 (Keep)		
9	10001893	case male	511	6	0,9194	Autosoma	Saccadic s	ORPHA:404499	0,411429	FAIBLE : ELIMINE (Drop 3)		
10	10001893	case male	511	2	0,9219	Epilepsy-t	Decreasec	ORPHA:1951	0,40625	FAIBLE : ELIMINE (Drop 3)		
11	10001893	case male	511	1	0,9219	X-linked in	Hyperacti	ORPHA:85327	0,393939	FAIBLE : ELIMINE (Drop 3)		

Fig. 2.10 : Scores de similarité RapidFuzz et sélection du Top 7 des maladies candidates

```
        prompt = f"""
You are a medical expert in rare diseases.

Clinical note:
{note}

Candidate diseases:
{diseases_text}

Task:
Choose exactly one disease from the candidate dideades that best matches the clinical note.
Return the selected disease name and its ORPHA ID exactly as shown in the candidate list.


Output:
Disease: ...
ORPHA: ...
Reason: ...
"""
```

Fig. 2.11 : Prompt structuré utilisé pour le raisonnement diagnostique du modèle Llama 3.1

Cette approche est appliquée exclusivement sur les 161 notes cliniques validées issues du fichier *Matched_Top7_FullGroups.xlsx*, pour lesquelles la vraie maladie est garantie d'être présente parmi les 7 maladies candidates du Top 7. Pour chaque note, le modèle reçoit un prompt structuré qui lui assigne explicitement le rôle d'un expert médical en maladies rares, lui fournit la note clinique complète ainsi que les 7 maladies candidates avec leurs noms, phénotypes et identifiants Orpha ID, et lui demande de sélectionner exactement une maladie en retournant son nom et son Orpha ID dans un format contraint [36].

Le modèle utilisé est Llama 3.1, exécuté localement via la bibliothèque Ollama [26]. La réponse du modèle est ensuite parsée par une expression régulière afin d'extraire l'Orpha ID prédit, qui est comparé à l'Orpha ID réel pour calculer le taux de réussite. Les performances sont évaluées via l'accuracy et le F1-score calculés avec la bibliothèque scikit-learn [29].

Cette approche structurée, combinant un prompt guidé et une base de connaissances filtrée, constitue l'expérience la plus rigoureuse du pipeline et produit les meilleures performances diagnostiques observées dans ce projet.

2.6 Conclusion

Ce chapitre a décrit en détail le pipeline méthodologique mis en œuvre pour le diagnostic automatique des maladies rares. De la préparation des données jusqu'à l'intégration des LLM, chaque étape a été conçue de manière progressive et complémentaire : extraction des phénotypes, correspondance sémantique par les modèles d'embeddings, filtrage textuel par RapidFuzz et diagnostic final par prompt engineering avec Llama 3.1. Le chapitre suivant présente l'évaluation quantitative de ce pipeline et l'analyse de ses performances.

Chapitre 3

Évaluation du Pipeline et Analyse des Performances

3.1 Protocole d'évaluation

Afin d'évaluer les performances du pipeline proposé, un protocole d'évaluation basé sur trois métriques complémentaires a été mis en place [29]. L'évaluation est réalisée sur un ensemble de cas cliniques réels issus de rapports médicaux publiés, dont le diagnostic confirmé représente la vérité terrain. Chaque cas clinique est traité par le pipeline afin de générer une liste ordonnée de maladies rares candidates, classées selon le score de similarité cosinus calculé entre les embeddings du cas clinique et ceux des maladies de la base Orphanet.

3.1.1 Accuracy

L'Accuracy mesure le pourcentage de cas où la maladie prédite en première position correspond à la vérité terrain.

3.1.2 F1-score

Le F1-score pondéré combine la précision et le rappel afin de mieux évaluer les performances dans un contexte de base de connaissances déséquilibrée.

3.1.3 Top-K Accuracy

La Top-K Accuracy vérifie si la vraie maladie apparaît parmi les 10 premières maladies proposées après calcul de la similarité cosinus. Ce choix permet ensuite au module RapidFuzz d'affiner le classement final.

3.2 Résultats de la correspondance sémantique

Cette section présente les résultats obtenus lors de l'étape de correspondance sémantique, réalisée à l'aide des modèles d'embeddings biomédicaux BioBERT, ClinicalBERT, Sentence-BERT et SapBERT. Pour chaque note clinique, un score de similarité cosinus a été calculé entre les phénotypes extraits de la note et les profils phénotypiques des maladies rares d'Orphanet, permettant de sélectionner un Top-10 des maladies candidates les plus proches sémantiquement.

L'évaluation repose sur un critère de couverture : une prédiction est considérée correcte si la vraie maladie associée à la note clinique figure parmi les dix maladies candidates retenues. Parmi les quatre modèles testés, SapBERT s'est révélé le plus performant, permettant de couvrir correctement 651 notes cliniques, soit le taux de couverture le plus élevé. Ce résultat confirme la supériorité de SapBERT pour la correspondance entre phénotypes cliniques et maladies rares, grâce à son pré-entraînement spécialisé sur les ontologies biomédicales, et valide le Top-10 obtenu comme base solide pour les étapes de filtrage suivantes.

3.3 Résultats du filtrage RapidFuzz

Cette section analyse l'impact du passage d'un Top-10 à un Top-7 candidats via l'algorithme RapidFuzz. En exploitant son score de similarité normalisé entre 0 et 1, ce filtrage lexical intermédiaire vise à épurer le contexte destiné aux LLM sans éliminer la pathologie correcte.

Les résultats valident cette stratégie : le taux de conservation de la vraie maladie reste excellent. Cette réduction confirme que les trois candidats exclus constituaient essentiellement du bruit sémantique, démontrant l'efficacité de ce filtre pour optimiser le pipeline en amont de la génération.

3.4 Résultats de l'approche naïve LLM

Cette section évalue les performances comparées des modèles MiniMax-M2.7 et Claude Sonnet 4.6 sur la tâche de diagnostic automatique, sans recours à un prompt engineering avancé. L'objectif était d'établir une base de référence (*baseline*) en mesurant leur capacité à identifier directement la bonne maladie rare à partir du dossier clinique brut.

Les résultats montrent que Claude Sonnet 4.6 surpasse largement son homologue avec une exactitude de 18 %, contre seulement 1,8 % pour MiniMax-M2.7. Ces scores particulièrement bas mettent en évidence les limites de la compréhension native des LLM face à la complexité des données médicales non structurées, confirmant la nécessité absolue d'optimiser le pipeline pour guider leur raisonnement clinique.

3.5 Résultats de l'approche améliorée — Llama 3.1

Cette section analyse l'impact de l'introduction d'un prompt engineering avancé sur le modèle Llama 3.1, et propose une comparaison directe avec l'approche naïve.

Les résultats démontrent une nette progression des performances : l'ajout d'instructions structurées et de consignes de raisonnement permet d'atteindre une exactitude remarquable de 70 %. En éliminant la quasi-totalité des hallucinations constatées lors de la phase précédente, cette approche optimisée surpasse largement la baseline de Claude Sonnet (18 %). Cette hausse spectaculaire valide la pertinence du guidage des LLM pour le traitement de cas cliniques complexes et confirme l'efficacité du pipeline final.

La figure 3.1 synthétise l'ensemble de ces résultats en comparant l'Accuracy Top-1 des modèles d'embeddings et des modèles LLM évalués dans ce projet.

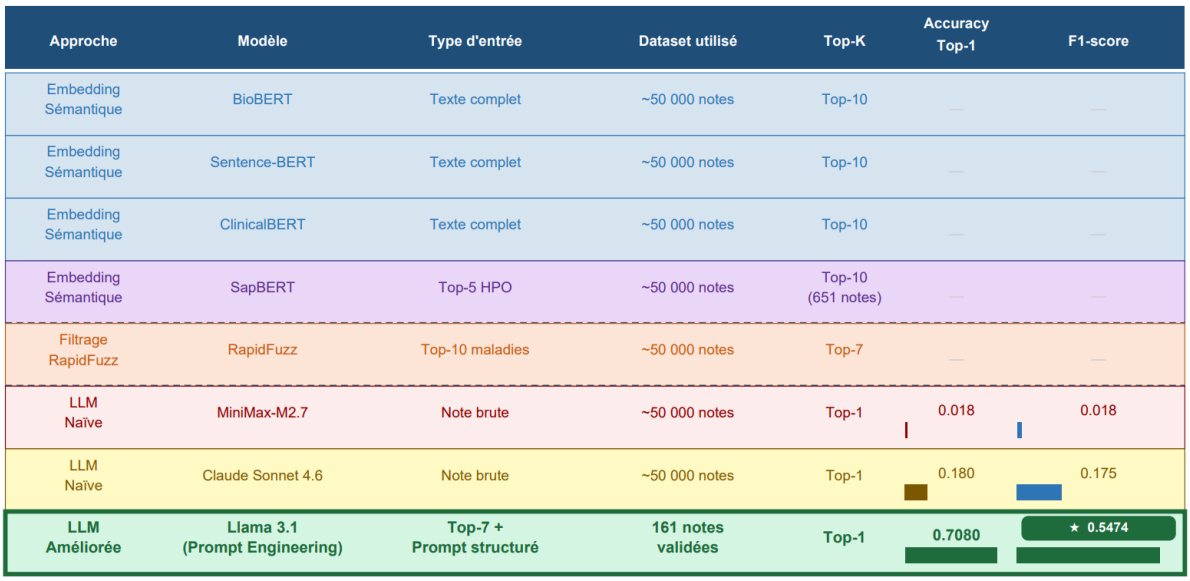


Fig. 3.1 : Comparaison des performances des modèles d'embeddings et des modèles LLM pour le diagnostic automatique des maladies rares

3.6 Discussion générale

Cette section présente une synthèse globale des résultats obtenus ainsi que les principales limites du pipeline développé. Bien que l'approche améliorée avec Meta Llama 3.1 ait atteint une précision de 70 %, plusieurs contraintes restent présentes dans le domaine des maladies rares.

La rareté et le déséquilibre des données limitent les performances des modèles d'embedding et des modèles de langage. De plus, la variabilité et le bruit des notes cliniques rendent le raisonnement diagnostique plus difficile. Le manque de ressources médicales annotées et de ressources computationnelles constitue également une limite pour l'entraînement et l'amélioration des modèles.

Comme perspectives futures, il serait intéressant d'intégrer des techniques de RAG (*Retrieval-Augmented Generation*) ainsi que des ontologies médicales afin d'améliorer l'accès aux connaissances médicales et la précision du diagnostic. L'utilisation du fine-tuning sur des datasets biomédicaux plus larges et mieux annotés pourrait également améliorer les capacités de raisonnement clinique des modèles.

3.7 Conclusion

Ce chapitre a présenté l'évaluation complète du pipeline proposé. Les résultats obtenus montrent une progression significative des performances à chaque étape, culminant avec une accuracy de 70 % pour l'approche améliorée par prompt engineering avec Llama 3.1. Ces résultats valident la pertinence de l'approche hybride combinant embeddings biomédicaux, filtrage textuel et LLM pour le diagnostic automatique des maladies rares.

Conclusion Générale

Ce travail a proposé une approche hybride et multi-étapes pour le diagnostic automatique des maladies rares à partir de notes cliniques textuelles. En combinant des modèles d'embeddings biomédicaux (BioBERT, ClinicalBERT, SentenceBERT, SapBERT), un filtrage textuel par RapidFuzz et un raisonnement diagnostique par grands modèles de langage, le pipeline développé permet d'identifier la maladie rare la plus probable parmi un ensemble de candidats. L'approche améliorée par prompt engineering avec Llama 3.1 a atteint une accuracy de 70 %, démontrant l'apport significatif du guidage structuré des LLM dans des tâches médicales complexes. Comme perspectives, l'intégration de techniques RAG et le fine-tuning sur des datasets biomédicaux plus larges constituent des pistes prometteuses pour améliorer davantage les performances du système.

Bibliographie

[1] Emily Alsentzer, John R. Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew B. A. McDermott. Publicly available clinical bert embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, 2019. https://aclanthology.org/W19-1909/.

[2] Anthropic. Claude sonnet 4.6 – model documentation, 2025. https://www.anthropic.com/claude.

[3] Ehsan Asnaashari. Les domaines de l'intelligence artificielle, 2025. https://doi.org/10.6084/m9.figshare.29954543.

[4] Max Bachmann. Rapidfuzz documentation, 2021. https://rapidfuzz.github.io/RapidFuzz/.

[5] Meriem Bahi. Deep neural network architecture, 2026. https://www.researchgate.net/figure/Deep-Neural-Network-architecture_fig1_330120030.

[6] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 2020. https://arxiv.org/abs/2005.14165.

[7] CEA. L'essentiel sur l'intelligence artificielle, 2026. https://www.cea.fr/comprendre/Pages/nouvelles-technologies/essentiel-sur-intelligence-artificielle.aspx.

[8] William W. Cohen, Pradeep Ravikumar, and Stephen E. Fienberg. A comparison of string distance metrics for name-matching tasks. In *IJCAI Workshop on Information Integration on the Web*, 2003. https://www.cs.cmu.edu/~wcohen/postscript/ijcai-ws-2003.pdf.

[9] James Cox. A simple explanation of supervised vs unsupervised vs reinforcement machine learning, 2026. https://medium.com/@cox.366/a-simple-explanation-of-supervised-vs-unsupervised-vs-reinforcement-machine-learning-228ecf78cb45.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert : Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv :1810.04805*, 2018. https://arxiv.org/abs/1810.04805.

[11] Haichao Chen et al. Rarearena : a comprehensive benchmark dataset unveiling the potential of large language models in rare disease diagnosis. *The Lancet Digital Health*, 8(2) :100953, 2026.

[12] European Journal of Human Genetics. Reducing diagnostic odyssey in rare diseases : Approaches and challenges. *European Journal of Human Genetics*, 2024. https://www.nature.com/articles/s41431-024-01604-z.

[13] Carol Friedman, Thomas C. Rindflesch, and Milton Corn. Natural language processing : State of the art and prospects for significant progress. *Journal of Biomedical Informatics*, 46(5) :765–773, 2013. https://pubmed.ncbi.nlm.nih.gov/23810857/.

[14] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. https://www.deeplearningbook.org/.

[15] Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. Clinicalbert : Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv :1904.05342*, 2019. https://arxiv.org/abs/1904.05342.

[16] Sebastian Köhler, Michael Gargano, Nicolas Matentzoglu, Leigh C. Carmody, David Lewis-Smith, Nicole A. Vasilevsky, et al. The human phenotype ontology in 2021. *Nucleic Acids Research*, 49(D1) :D1207–D1217, 2021. https://hpo.jax.org.

[17] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert : A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4) :1234–1240, 2020. https://academic.oup.com/bioinformatics/article/36/4/1234/5566506.

[18] Vladimir I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8) :707–710, 1966.

[19] Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. Self-alignment pretraining for biomedical entity representations (sapbert). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2021. https://github.com/cambridgeltl/sapbert.

[20] Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heinz, and Dan Roth. Recent advances in natural language processing via large pre-trained language models : A survey. *arXiv preprint arXiv :2111.01243*, 2021. https://arxiv.org/abs/2111.01243.

[21] MiniMaxAI. Minimax-m2: A large-scale mixture-of-experts language model, 2025. https://openrouter.ai/minimaxai/minimax-m2.7.

[22] Tom M. Mitchell. *Machine Learning*. McGraw-Hill, 1997. https://www.cs.cmu.edu/~tom/mlbook.html.

[23] Nature Medicine. Large language models in medicine : Opportunities and clinical applications. *Nature Medicine*, 2023. https://www.nature.com/articles/s41591-023-02448-8.

[24] Aurélie Névéol, Hercules Dalianis, Sumithra Velupillai, Guergana Savova, and Pierre Zweigenbaum. Clinical natural language processing in languages other than english : Opportunities and challenges. *Journal of Biomedical Semantics*, 9(12), 2018. https://link.springer.com/article/10.1186/s13326-018-0179-8.

[25] Ziad Obermeyer and Ezekiel J. Emanuel. Predicting the future — big data, machine learning, and clinical medicine. *The New England Journal of Medicine*, 375(13) :1216–1219, 2016. https://www.nejm.org/doi/full/10.1056/NEJMp1606181.

[26] Ollama. Ollama : Run large language models locally, 2024. https://ollama.com.

[27] Orphanet. About rare diseases, 2026. https://www.orpha.net/en/other-information/about-rare-diseases.

[28] Orphanet. Rare disease research, 2026. https://www.orpha.net.

[29] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn : Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. https://jmlr.org/papers/v12/pedregosa11a.html.

[30] RareArena. Rarearena : A large-scale benchmark for rare disease diagnosis. *The Lancet Digital Health*, 2026. https://github.com/zhao-zy15/RareArena.

[31] Nils Reimers and Iryna Gurevych. Sentence-bert : Sentence embeddings using siamese bert-networks. *Proceedings of EMNLP-IJCNLP*, 2019. https://aclanthology.org/D19-1410/.

[32] Laria Reynolds and Kyle McDonell. Prompt programming for large language models : Beyond the few-shot paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–7, 2021. https://arxiv.org/abs/2102.07350.

[33] Rajat Tripathi. Vector embeddings, 2026. https://www.pinecone.io/learn/vector-embeddings/.

[34] Drashtti Vasant, Laetitia Chanas, James Malone, Marc Hanauer, Annie Olry, Simon Jupp, Peter N. Robinson, Helen Parkinson, and Ana Rath. Orphanet rare disease ontology (ordo) : Linking rare diseases, epidemiology and gene information. *Scientific Data*, 2022. https://www.orphadata.com/ordo/.

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv :1706.03762*, 2017. https://arxiv.org/abs/1706.03762.

[36] Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C. Schmidt. A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv :2302.11382*, 2023. https://arxiv.org/abs/2302.11382.

[37] World Health Organization. Rare diseases – questions and answers, 2024. https://www.who.int/news-room/questions-and-answers/item/rare-diseases.