Prompting aligné sur l'ontologie pour l'extraction sémantique RDF : Une étude de cas sur les maladies oro-dentaires rares

Socrates Waka Onyando*, Ali Ayadi*,
Agnès Bloch-Zupan**, Cédric Wemmert*

*SDC team, ICube UMR 7357, University of Strasbourg,
300 Boulevard Sebastien Brant, 67400 Illkirch-Graffenstaden
{swonyando,ali.ayadi,wemmert,agnes.bloch-zupan}@unistra.fr,
**Faculté de chirurgie dentaire, 8 Rue Sainte Elisabeth - F-67000 Strasbourg

Résumé. Les maladies oro-dentaires rares présentent des défis pour l'extraction de connaissances en raison du manque de données et de la dispersion des informations cliniques dans des textes non structurés. Les ontologies constituent une solution pour organiser ces connaissances, mais leur construction manuelle demeure lente et coûteuse. Ce travail étudie si les LLM à poids ouverts peuvent être guidés par des requêtes alignées sur une ontologie afin d'extraire de manière fiable des triplets RDF sans réglage fin spécifique au domaine. Nous proposons une chaîne intégrant un découpage sémantique, des invites contraintes par l'ontologie et une vérification sémantique via BLEURT et SBERT. Trois familles de modèles et leurs variantes biomédicales sont évaluées selon quatre stratégies de prompting : Zero-Shot, Few-Shot, Chain-of-Thought (CoT) et CoT + Few-Shot. Les résultats montrent que le prompting influence la qualité d'extraction que l'architecture ou le fine-tuning, CoT offrant la meilleure fidélité sémantique.

1 Introduction

Le domaine des maladies rares, et en particulier des affections oro-dentaires rares, occupe une niche particulière et difficile au sein de l'informatique médicale et de l'aide à la décision clinique. En raison de leur faible prévalence individuelle mais de leur impact collectif considérable, la recherche dans ce domaine est limitée par la disponibilité restreinte de données structurées (Taruscio et Gahl, 2024). Les connaissances sur les maladies rares sont très fragmentées et principalement intégrées dans des publications scientifiques non structurées, des rapports de cas cliniques et des déclarations consensuelles d'experts (Liu et al., 2022), contrairement aux maladies courantes pour lesquelles des ensembles de données massifs permettent une modélisation statistique robuste. Cette fragmentation crée un effet de cloisonnement des connaissances qui entrave le développement de modèles computationnels fiables. Il est donc essentiel de transformer ces informations dispersées en représentations sémantiques formelles afin de favoriser l'interopérabilité, de permettre le raisonnement sémantique et de faciliter l'analyse avancée des données, ce qui permettra en fin de compte d'améliorer les soins aux patients.

Le paradigme du web sémantique, fondé sur les ontologies pour définir concepts et relations, offre une solution prometteuse au défi de l'interopérabilité (Antezana et al., 2009; Domingue et al., 2011). Dans ce contexte, le texte non structuré peut être transformé en graphes de connaissances (KG) via le mappage des segments de texte à des entités et des propriétés d'objets (Domingue et al., 2011). Cependant, son déploiement pratique reste limité par un goulot d'étranglement dans l'acquisition des connaissances. Le développement et la maintenance des ontologies sont très laborieux, nécessitant une expertise spécialisée souvent rare et coûteuse (Antezana et al., 2009). De plus, la croissance rapide de la littérature biomédicale dépasse la capacité de la curation manuelle (Antezana et al., 2009), rendant les bases de connaissances rapidement obsolètes.

Les méthodes d'extraction automatisée représentent donc une approche plus adaptée, car elles peuvent mieux suivre le dynamisme des données et des connaissances. Les techniques traditionnelles de Traitement du Langage Naturel (TLN), bien qu'utiles, peinent à capturer la variabilité sémantique, la nuance contextuelle et la complexité syntaxique du texte biomédical (Leaman et al., 2015). De plus, de nombreuses approches TLN reposent sur des pipelines à règles ou des modèles d'apprentissage supervisé nécessitant d'importants corpus annotés (Spasic et Nenadic, 2020), particulièrement rares pour les maladies oro-dentaires rares. L'Intelligence Artificielle (IA) générative, et en particulier les Grands Modèles de Langage (LLM), offre une alternative prometteuse pour automatiser l'extraction de connaissances. Les LLM ont démontré de fortes capacités de compréhension du langage, de raisonnement et de suivi d'instructions complexes qui surpassent de loin les générations précédentes d'outils TLN (Cao et al., 2024). Entraînés sur de vastes corpus de textes diversifiés, ces modèles possèdent une compréhension latente de la structure linguistique et, dans une certaine mesure, des concepts biomédicaux généraux (Cao et al., 2024). Cependant, leur application à des domaines très spécialisés et à fort enjeu, tels que les maladies oro-dentaires rares, demeure un champ de recherche actif. Un défi majeur réside dans l'exploitation de leur puissance générative tout en limitant leur tendance à l'hallucination.

Compte tenu de ces défis, cet article propose, met en œuvre et évalue un cadre basé sur les LLM et adapté aux besoins spécifiques des pathologies oro-dentaires rares. Nous émettons l'hypothèse que les LLM facilement accessibles et à poids ouvert, lorsqu'ils sont guidés par des stratégies de prompting structurées qui imposent des contraintes ontologiques, peuvent extraire avec précision des informations à partir de textes scientifiques et alimenter une ontologie dentaire spécialisée sans nécessiter de réglages fins gourmands en ressources informatiques ou en données. Dans notre formulation, le remplissage de l'ontologie est considéré comme une tâche de traduction contrainte, dans laquelle les entrées en langage naturel sont converties en syntaxe triple RDF. Pour tester cette hypothèse, nous procédons à une évaluation comparative de trois familles de modèles de premier plan (Mistral, Qwen et Llama), en analysant leurs performances dans le cadre de différentes stratégies de prompting : Zero-Shot, Few-Shot, Chain-of-Thought (CoT) et un paramètre hybride CoT+Few-Shot.

Nos contributions sont doubles. Premièrement, nous adaptons un cadre ontologique polyvalent au domaine des maladies bucco-dentaires rares, en définissant un schéma rigoureux permettant de saisir les assertions phénotypiques et génotypiques. Deuxièmement, nous présentons une comparaison empirique approfondie de plusieurs modèles, en fournissant des mesures de performance détaillées qui quantifient leur efficacité dans le traitement du langage relatif aux maladies bucco-dentaires rares.

2 Travaux connexes

Cette section propose une revue exhaustive de la littérature, retraçant l'évolution des techniques de population ontologique et examinant l'application spécifique des modèles linguistiques à grande échelle (LLM) dans le contexte des soins de santé.

2.1 Population ontologique et extraction de connaissances

Le peuplement ontologique, formellement défini comme le processus d'insertion d'instances de concepts et de relations dans une ontologie existante (enrichissant l'ABox tout en préservant la TBox), fait l'objet de recherches intensives depuis des décennies (Ayadi et al., 2019; Cao et al., 2024; Sahbi et al., 2025; Wu et al., 2024). La trajectoire de ce domaine reflète l'évolution plus large de l'IA, qui est passée de systèmes symboliques rigides à l'apprentissage statistique et, plus récemment, à des paradigmes génératifs.

Aux débuts, les systèmes s'appuyaient sur des approches basées sur des règles (Ayadi et al., 2019; Du et al., 2024). Ces méthodes utilisaient des patrons lexico-syntaxiques, des statistiques de cooccurrence et la programmation logique inductive pour identifier les termes et les relations candidats (Du et al., 2024). Bien que ces approches sémantiques offraient une grande précision lorsque le texte respectait des formats attendus, elles nécessitaient une supervision humaine étendue pour définir les règles d'extraction et manquaient de flexibilité face à l'immense variabilité linguistique de la littérature scientifique (Ayadi et al., 2019). Dans le contexte des maladies rares, où les descriptions phénotypiques sont très idiosyncratiques, les systèmes basés sur des règles échouent souvent à capturer l'ensemble du tableau clinique.

Avec l'augmentation de la puissance de calcul, le domaine s'est orienté vers des approches d'apprentissage automatique et d'apprentissage profond. Ces méthodes ont défini le remplissage de l'ontologie comme une tâche de classification. Des modèles supervisés tels que les machines à vecteurs de support (SVM), puis les réseaux neuronaux récurrents (RNN) et les réseaux à mémoire à court et long terme (LSTM), ont été entraînés à classer des séquences de tokens en tant qu'entités ou relations spécifiques (Ayadi et al., 2019; Ivanisenko et al., 2024). Cependant, ces modèles discriminatifs se sont heurtés à un obstacle de taille. Ils nécessitaient de grands corpus d'entraînement annotés pour apprendre des caractéristiques efficaces (Du et al., 2024). Dans des domaines spécialisés tels que les maladies oro-dentaires rares, de tels corpus n'existent tout simplement pas, et leur création nécessite des efforts d'experts dont le coût est prohibitif. De plus, ces modèles avaient souvent du mal à classer les instances présentant des propriétés « floues », c'est-à-dire des concepts qui ne correspondaient pas parfaitement aux catégories d'entraînement prédéfinies (Jia et al., 2019). Cela nécessite la présence d'un expert pour une validation continue, ce qui augmente encore les coûts de main-d'œuvre.

L'état actuel de la technique s'est orienté vers l'IA générative et les LLM (Mihindukulasooriya et al., 2023; Wu et al., 2024). Le benchmark Text2KGBench met en évidence ce changement de paradigme, en proposant par exemple que les LLM puissent être utilisés pour l'extraction de relations guidée par des ontologies de manière Zero-Shot ou Few-Shot (Mihindukulasooriya et al., 2023). Cette approche traite le remplissage d'ontologies non pas comme une tâche de classification, mais comme une tâche de traduction générative qui traduit le langage naturel en triplets structurés. Notre travail s'appuient sur cette base, en appliquant le paradigme de l'extraction générative à un domaine où le coût de l'erreur est élevé et où la complexité de l'ontologie est importante.

2.2 Les LLM dans les domaines médical et des maladies rares

L'intégration des LLM à l'informatique médicale a connu une croissance, portée par leur capacité à traiter et synthétiser du texte complexe. Des études récentes ont démontré que des modèles comme Med-PaLM et des variantes open-source peuvent obtenir des notes de passage aux examens de licence médicale et effectuer des résumés de haute qualité de notes cliniques (Naemi et Sahafi, 2025; Singhal et al., 2025). Cependant, le déploiement de modèles généralistes en milieu clinique est semé d'embûches, en raison de leur fiabilité et de l'hallucination (Ivanisenko et al., 2024). Les modèles génératifs sont des systèmes probabilistes, conçus pour produire du texte plausible plutôt que des vérités vérifiables (Ivanisenko et al., 2024). Dans le domaine médical, cela peut engendrer des faits hallucinés, comme lier un gène à une maladie sur la base d'une cooccurrence faible, ou mal interpréter une négation (Wu et al., 2024).

Dans le contexte des maladies rares, la rareté des données exacerbe ces défis. Ces conditions étant peu fréquentes dans les corpus d'entraînement des LLM, les représentations internes des modèles sont éparses et donc plus sujettes aux erreurs. Des travaux en text mining biomédical ont commencé à résoudre ce problème en intégrant des contraintes symboliques à l'extraction. Le framework RELATE montre que la requête alignée sur une ontologie améliore l'extraction de relations à partir de abstracts biomédicaux en assurant que les prédicats respectent un vocabulaire contrôlé (Olasunkanmi et al., 2025). De même, les approches hybrides pour le phénotypage des maladies rares, qui combinent dictionnaires dérivés d'ontologie et extraction par LLM, montrent que les modèles peuvent être orientés vers des sorties plus précises guidées par des terminologies de domaine (Wu et al., 2024). Ces résultats indiquent que les contraintes symboliques servent de garde-fous, ancrant les prédictions du modèle dans des concepts bien définis plutôt que dans des associations statistiques non contraintes. Notre travail adopte ce principe dans le domaine textuel, tirant parti de l'ancrage ontologique pour promouvoir une extraction plus sûre et fiable dans le contexte des maladies oro-dentaires rares.

3 Méthodologie

Notre méthodologie simule un flux de travail à ressources limitées pour la création d'une base de connaissances à partir de corpus scientifiques actualisés, sans nécessiter d'entraîner un modèle de fondation. Formellement, étant donné un texte non structuré S et une ontologie cible O, le système doit extraire un ensemble de triplets RDF $T = \{(s, p, o) \mid s, o \in Entities,\ p \in Relations\}$ qui représente fidèlement l'information clinique de S tout en adhérant aux contraintes de schéma de O. Le pipeline complet est résumé dans la Fig. 1.

3.1 Construction ontologique : la TBox

La TBox est un élément central de notre cadre d'extraction. Nous utilisons une ontologie spécifique au domaine conçue par Elgohary et al. (2025) pour coder formellement les maladies oro-dentaires rares. Elle a été conçue en étroite collaboration avec des experts cliniques afin de garantir que l'ontologie capture les concepts nécessaires dans le domaine. Il convient de noter que cette ontologie sert non seulement de base de données, mais aussi de contrainte de schéma qui guide le processus de génération du LLM.

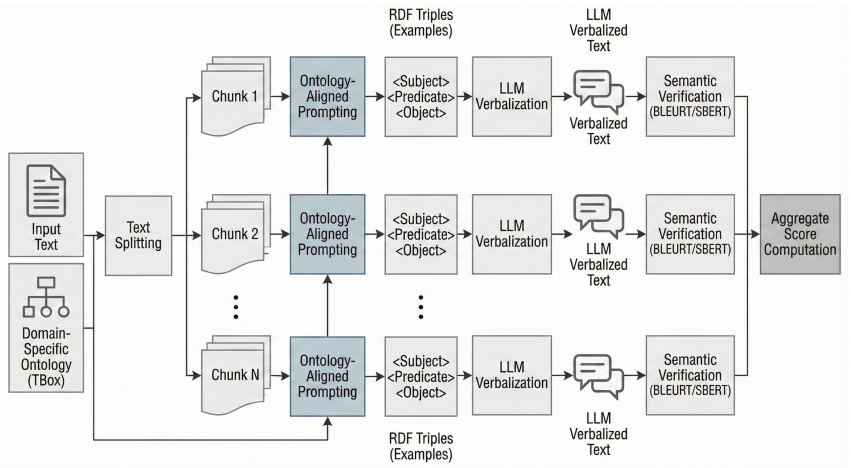


FIG. 1 – *Pipeline d'extraction d'ontologie. Le texte est divisé en fragments, chacun envoyé à une requête LLM ; les triplets de sortie sont agrégés puis verbalisés. Le texte verbalisé est comparé au texte source via BLEURT/SBERT ; les triplets validés forment le graphe RDF final.*

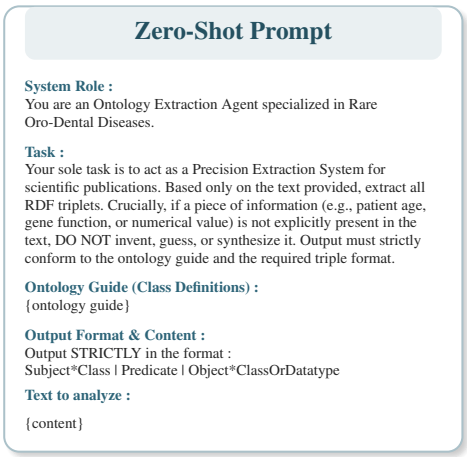
3.2 Corpus de données et prétraitement

Notre corpus expérimental est constitué de fragments textuels issus de la littérature scientifique anglaise publiquement disponible. Pour préparer ce corpus à l'extraction, un pipeline de prétraitement en deux étapes a été mis en œuvre : filtrage du contenu et segmentation par phrases (chunking). Le filtrage a permis de supprimer le bruit structurel du texte (informations de contact, références, etc.). Le texte nettoyé a ensuite été segmenté en unités sémantiquement cohérentes en utilisant une stratégie de segmentation par phrases, assurant le respect de la structure linguistique plutôt que des limites arbitraires de caractères. Pour préserver la continuité contextuelle entre les limites des fragments, chaque nouveau fragment commençait par un chevauchement des trois phrases précédentes. Cette fenêtre de chevauchement prévenait la perte de sens pouvant survenir si les arguments chevauchaient des fragments consécutifs.

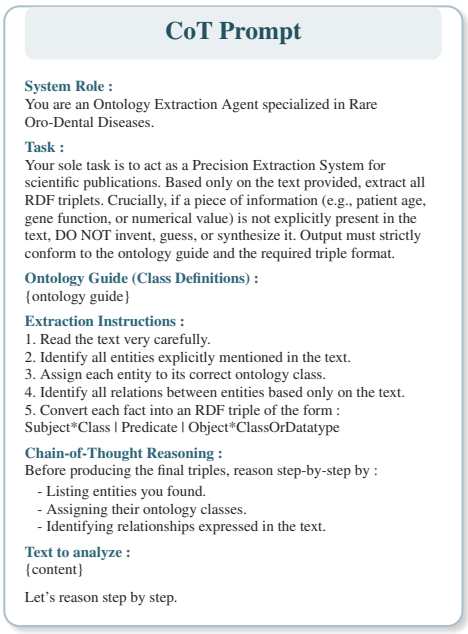
3.3 Invite alignée sur l'ontologie

Pour extraire les informations relationnelles structurées de chaque fragment, nous avons utilisé un framework de requête alignée sur l'ontologie, conçu pour guider le LLM vers des triplets cohérents avec notre ontologie. Chaque requête demandait au modèle d'identifier entités et relations dans le texte, les fournissant au format standard ⟨ Sujet, Prédicat, Objet⟩, en utilisant les classes et les types de relations ontologiques comme référence. Nous avons évalué plusieurs stratégies de requête pour déterminer leur impact sur la qualité d'extraction :

1. Requête Zero-Shot : Le modèle recevait uniquement la description de la tâche et la spécification du format de triplet (Fig. 2a).
2. Requête Few-Shot : Deux exemples de texte similaire étaient fournis pour illustrer le mappage correct des phrases en triplets alignés sur l'ontologie.



(a) Requête Zero-Shot utilisée pour l'extraction de triplets alignés sur une ontologie.



(b) Requête de type "Chain-of-Thought" (CoT) illustrant le raisonnement par étapes.

FIG. 2 – *Stratégies de requêtes pour l'extraction alignée sur l'ontologie.*

3. Requête CoT : Les requêtes incluaient des signaux de raisonnement par étapes (Fig. 2b), encourageant un raisonnement intermédiaire structuré.
4. Requête CoT + Few-Shot : Les requêtes incluaient à la fois des triplets exemplaires et des instructions de raisonnement.

3.4 Vérification Sémantique

Pour garantir que les triplets alignés sur l'ontologie représentent le sens des segments de texte originaux, nous avons mis en œuvre une technique de vérification sémantique combinant la verbalisation graphe-vers-texte avec la notation de similarité par référence. Pour chaque fragment et ses triplets, nous générons d'abord une reconstruction en langage naturel du graphe par verbalisation graphe-vers-texte. En utilisant les triplets extraits, nous construisons une requête few-shot de verbalisation envoyée à un modèle GPT-4.1 configuré avec une faible température pour des sorties cohérentes et concises. Le texte synthétique est la meilleure tentative de reformuler le sens encodé, permettant une comparaison directe entre la reconstruction dérivée du graphe et le fragment original. Pour quantifier cet alignement, nous appliquons deux métriques complémentaires. BLEURT (métrique fine-tunée basée sur référence) est utilisée pour détecter les déviations sémantiques subtiles (omissions, hallucinations). Parallèlement, nous calculons les embeddings SBERT pour les deux textes et mesurons leur similarité cosinus. Cela fournit

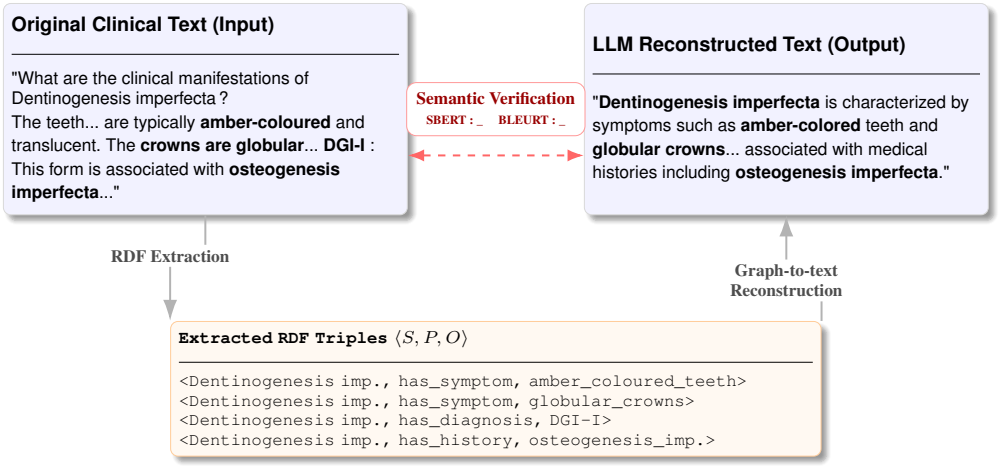


FIG. 3 – *Exemple de vérification sémantique pour les triplets alignés sur une ontologie. La figure présente le fragment de texte original, les triplets RDF extraits au format ⟨ Sujet, Prédicat, Objet⟩ et la reconstruction du graphe en texte produite par le LLM. Cette configuration permet d'évaluer la fidélité sémantique à l'aide des métriques BLEURT et SBERT.*

une évaluation sémantique au niveau de l'embedding moins sensible à la variation lexicale. Un exemple de ce processus (segment original, triplets, verbalisation) est fourni dans la Fig. 3.

4 Expériences

4.1 Jeu de données

Nous avons utilisé un ensemble de documents sur les maladies bucco-dentaires rares provenant de sources accessibles au public. Il s'agit notamment des fiches d'information Genosmile sur l'amélogenèse imparfaite, la dentinogenèse imparfaite, la dysplasie dentinaire, l'oligodontie, la perte précoce des dents, le syndrome de l'incisive centrale maxillaire médiane solitaire (SMMCI) (GenoSmile, 2015) et des articles de la revue Orphanet (Crawford et al., 2007; Reibel et al., 2009). Ces fichiers PDF contiennent les définitions cliniques, l'étiologie, les gènes et les symptômes de maladies oro-dentaires rares spécifiques.

4.2 Architectures de LLM et Stratégies de Prompting

Notre configuration expérimentale est organisée selon une architecture LLM à deux niveaux, comprenant des LLM de base ajustés par instruction et des variantes adaptées au domaine, fine-tunées sur des corpus biomédicaux. Nous employons trois LLM récents open-source suivant les instructions, qui équilibrent capacité et efficacité de calcul : Mistral-7B-Instruct-v0.3 (7B paramètres), Qwen2.5-7B-Instruct (7B paramètres), Meta-Llama-3.1-8B-Instruct (8B paramètres). Nous incluons des variantes médicales fine-tunées de chaque ar-

chitecture de base : FlowerTune-Mistral-7B-Medical-PEFT et FlowerTune-LLaMA-3.1-8B-Medical-LoRA (fine-tunés sur les jeux de données PubMedQA, MedMCQA et MedQA), et Qwen-UMLS-7B-Instruct (fine-tuné sur le jeu de données UMLS. Cette structure permet d'isoler l'impact de l'adaptation au domaine sur l'extraction d'informations alignées sur l'ontologie.

Pour simuler un environnement à ressources limitées sans clusters haut de gamme, tous les modèles ont été quantifiés à 4 bits (NF4). Cela permet aux modèles de fonctionner sur un seul GPU 12 Go VRAM. Pour les LLM, la température a été réglée sur 0.2. Ceci visait à minimiser l'aléatoire et à encourager des sorties déterministes, évitant les sorties créatives qui compromettraient la fiabilité requise pour l'extraction.

Pour évaluer l'influence des styles de prompting sur l'extraction de triplets alignés sur l'ontologie, nous comparons les quatre stratégies de requête sur tous les LLM. Chaque stratégie est appliquée uniformément sur les modèles de base et ajustés médicalement pour permettre la comparaison de leur sensibilité au style de requête. Toutes les requêtes sont exécutées avec des paramètres identiques, et la performance est mesurée sur les mêmes segments d'entrée fragmentés. Ceci garantit une évaluation contrôlée des interactions prompt-modèle.

4.3 Évaluation

Les scores BLEURT et SBERT par bloc ont été calculés pour chaque configuration de modèle. Les moyennes et les écarts-types de ces scores sont indiqués en raison de l'hétérogénéité naturelle entre les différents blocs et textes biomédicaux. Le score BLEURT est la note humaine prédite ($\hat{y}$) dérivée d'une métrique apprise basée sur le modèle BERT (Sellam et al., 2020). Il est calculé en appliquant une couche linéaire à la représentation du jeton spécial `[CLS]` ($\tilde{v}_{[\text{CLS}]}$) dérivé du modèle BERT :

$$\hat{y} = f(x, \tilde{x}) = W\tilde{v}_{[\text{CLS}]} + b$$

où x est la phrase de référence, $\tilde{x}$ est la phrase prédite, W est la matrice de poids et b est le vecteur de biais (Sellam et al., 2020). Le score SBERT ($\text{Score}_{\text{SBERT}}$) est une mesure de la similarité sémantique textuelle calculée à l'aide de la similarité cosinus entre deux plongements de phrases de taille fixe (Reimers et Gurevych, 2019). Étant donné deux plongements de phrases dérivés, u et v, le score SBERT est calculé à l'aide de la fonction de similarité cosinus :

$$\text{Score}_{\text{SBERT}}(u, v) = \cos(u, v) = \frac{u \cdot v}{\|u\|\|v\|}$$

où $\|\cdot\|$ désigne la norme euclidienne (Reimers et Gurevych, 2019).

5 Résultats et Discussion

5.1 Résultats

Le Tableau 1 présente la performance des six LLM à travers les quatre stratégies de prompting. Il est évident que, toutes stratégies confondues, le CoT a généralement donné les scores de similarité sémantique les plus élevés. Par exemple, le CoT a produit le score BLEURT le plus élevé (33.21%) ainsi que le score SBERT le plus élevé (74.95%) pour les modèles Meta-Llama-3.1-8B-Instruct et Qwen-UMLS-7B-Instruct, respectivement. Il a également généré les

meilleurs scores sur les deux métriques, à l'exception de Qwen2.5-7B-Instruct (SBERT), où le score maximal a été obtenu avec la configuration CoT + Few-Shot. Fait intéressant, la combinaison CoT avec des exemples Few-Shot n'a pas produit de gains additifs, à l'exception du cas susmentionné de Qwen2.5-7B-Instruct (SBERT).

Model	BLEURT				SBERT			
	Zero-Shot	Few-Shot	CoT	CoT+Few-Shot	Zero-Shot	Few-Shot	CoT	CoT+Few-Shot
Mistral-7B-Instruct-v0.3	31.48 ± 3.7	30.70 ± 4.2	31.63 ± 3.0	31.62 ± 3.3	68.98 ± 13.0	67.27 ± 14.6	72.44 ± 9.8	69.11 ± 12.3
Qwen2.5-7B-Instruct	32.14 ± 4.2	32.15 ± 4.0	32.39 ± 5.0	32.15 ± 3.4	67.42 ± 14.9	71.27 ± 15.0	71.97 ± 12.6	73.18 ± 10.7
Meta-Llama-3.1-8B-Instruct	30.32 ± 2.6	30.11 ± 3.2	$\mathbf{33.21 \pm 3.5}$	31.62 ± 3.2	50.57 ± 17.2	58.12 ± 14.4	73.28 ± 16.1	69.11 ± 12.2
FlowerTune-Mistral-7B-Medical-PEFT	31.13 ± 3.8	30.10 ± 4.0	32.91 ± 4.9	31.55 ± 3.3	67.23 ± 14.9	63.53 ± 14.7	70.43 ± 13.9	64.32 ± 14.5
Qwen-UMLS-7B-Instruct	32.10 ± 3.6	31.74 ± 4.6	32.79 ± 4.5	31.86 ± 3.8	69.48 ± 15.8	65.55 ± 18.5	$\mathbf{74.95 \pm 11.9}$	74.54 ± 10.0
FlowerTune-LLaMA-3.1-8B-Medical-LoRA	28.48 ± 2.0	29.24 ± 2.6	32.41 ± 3.4	30.55 ± 3.4	53.14 ± 18.5	62.29 ± 12.8	72.18 ± 17.7	64.04 ± 12.0

TAB. 1 – *Résultats d'évaluation (moyenne $\pm$ écart-type) des scores BLEURT et SBERT pour chaque modèle et requête.*

La variance de performance différait significativement entre les deux métriques. Les scores BLEURT étaient relativement stables sur tous les modèles et stratégies de prompting, allant principalement de 28.48% à 33.21%. En revanche, les scores SBERT montraient plus de variabilité, les écarts types allant de 9.8% à 18.5% avec les valeurs les plus élevées pour tous les modèles en configuration Zero-Shot ou Few-Shot. Les trois LLM de base ajustés par instruction produisaient généralement des scores BLEURT et SBERT supérieurs à leurs homologues fine-tunés. Le score BLEURT le plus élevé du Tab. 1 a été obtenu par Meta-Llama-3.1-8B-Instruct. Néanmoins, Qwen-UMLS-7B-Instruct (fine-tuné) a surpassé le modèle de base en configuration CoT, atteignant le score SBERT global le plus élevé de 74.95%.

5.2 Discussion

Nos résultats expérimentaux révèlent trois principaux enseignements concernant la performance des LLM pour l'extraction de triplets RDF de maladies oro-dentaires rares.

Dominance de la stratégie de requête sur l'architecture Dans l'ensemble, nos résultats suggèrent que, pour cette tâche, la stratégie de prompting a un impact plus important sur la performance que les différences architecturales (dans une gamme de paramètres fixe) ou le pré-entraînement spécifique au domaine. En particulier, les gains de performance obtenus en changeant de stratégie de prompting—surtout en passant de Zero-Shot à Chain-of-Thought (CoT)—étaient supérieurs à ceux associés aux changements d'architecture ou à la spécialisation biomédicale. Par exemple, Meta-Llama-3.1-8B-Instruct a montré une amélioration d'environ 45% en passant du Zero-Shot au prompting CoT. Cette observation est cohérente avec les résultats de Wei et al. (2022), qui montrent que le prompting CoT peut débloquer des capacités de raisonnement latentes sous-utilisées avec les stratégies standards.

Efficacité du réglage fin de domaine des modèles De manière contre-intuitive, nos expériences indiquent que le fine-tuning médical n'améliore pas systématiquement la qualité d'extraction des triplets RDF. Bien que Qwen-UMLS-7B ait constamment surpassé son homologue de base, atteignant le score SBERT global le plus élevé, d'autres variantes ajustées médicalement ont sous-performé leurs modèles de base. Cette tendance suggère que le fine-tuning pourrait imposer un « coût d'alignement » qui peut nuire à la performance sur des tâches structurées comme l'extraction RDF. En même temps, la nature des données de fine-tuning semble jouer un rôle clé. Qwen-UMLS-7B a été fine-tuné sur UMLS, une ressource structurée autour de relations sémantiques et de définitions de concepts médicaux, ce qui est

plus proche de la tâche d'extraction RDF que les jeux de données généraux utilisés pour les autres modèles.

Variance et fiabilité Une autre observation critique est la forte variance observée lors des expériences, notamment dans les scores SBERT, où les écarts types varient de 9.8 à 18.5. Bien que les scores moyens indiquent un niveau de compétence raisonnable, cette variabilité révèle que les modèles sont sujets à une dérive sémantique substantielle échantillon par échantillon. Une telle instabilité représente un défi majeur pour le déploiement pratique dans des environnements cliniques ou biomédicaux, où la robustesse et la fiabilité sont des exigences essentielles.

6 Conclusion

Dans ce travail, nous avons étudié si les LLM à poids ouverts peuvent être guidés pour extraire des triplets RDF alignés sur une ontologie dans le domaine spécialisé des maladies orodentaires rares. Nous avons proposé un pipeline d'extraction combinant une requête contrainte par ontologie, des instructions multi-stratégies et une vérification sémantique avec BLEURT et SBERT. Nos résultats montrent qu'une connaissance scientifique substantielle peut être convertie en représentations structurées sans entraînement supervisé, atténuant ainsi le goulot d'étranglement de l'acquisition de connaissances pour le développement d'ontologies biomédicales.

Sur six LLM et quatre stratégies, la conception de la requête a eu une influence plus grande sur la qualité de l'extraction que l'architecture ou le fine-tuning biomédical. La requête "Chain-of-Thought" (CoT) a constamment amélioré la fidélité sémantique, dépassant parfois les gains des modèles adaptés au domaine. Alors que certaines variantes ajustées (Qwen-UMLS-7B notamment) ont surpassé leurs modèles de base, d'autres ont sous-performé, suggérant que le fine-tuning de type questions-réponses peut nuire à l'extraction structurée. Enfin, la forte variance observée dans la similarité SBERT indique que les modèles, bien que performants en moyenne, restent susceptibles à la dérive sémantique.

Remerciements

Ce travail a été soutenu conjointement par l'Institut Thématique Interdisciplinaire HealthTech, dans le cadre du programme ITI 2021-2028 de l'Université de Strasbourg, du CNRS et de l'Inserm, via IdEx Unistra (ANR-10-IDEX-0002) et SFRI (projet STRAT'US, ANR-20-SFRI-0012) dans le cadre de France 2030, et par un financement gouvernemental géré par l'Agence Nationale de la Recherche au titre de France 2030 via le Pôle IA ENACT (ANR-23-IACL-0004). La contribution des données sur les maladies rares a été soutenue par, notamment : le projet e-GenoDENT (Fonds d'Intervention Régionale – FIR, ARS Grand Est, 2019–2022) ; l'atelier de co-conception en e-santé de la Fondation Maladies Rares (2019) ; l'AMI Économie numérique Grand Est – i-Dent (2020–2021) ; les dispositifs Impulsion Recherche Filière TETECOU (2020, 2022) ; Bpifrance (dans le cadre de la stratégie d'accélération du numérique en santé de la DNS et du SGPI, Health Data Hub, stratégie « France 2030 », 2021–2023) ; la Fondation Force DIAGNODENT (2023–2026) ; le projet ANR 3DBioDENT (ANR-23-CE17-0048-01, 2023–2026) ; ainsi que la MIG F04 pour les CRMR, DGOS, Ministère français de la Santé et de la Prévention.

Références

Antezana, E., M. Kuiper, et V. Mironov (2009). Biological knowledge management : the emerging role of the semantic web technologies. *Briefings in Bioinformatics 10*(4), 392.

Ayadi, A., A. Samet, F. d. B. de Beuvron, et C. Zanni-Merk (2019). Ontology population with deep learning-based NLP : a case study on the Biomolecular Network Ontology. *Procedia Computer Science 159*, 572–581.

Cao, L., J. Sun, A. Cross, et al. (2024). An automatic and end-to-end system for rare disease knowledge graph construction based on ontology-enhanced large language models : Development study. *JMIR Medical Informatics 12*(1), e60665.

Crawford, P. J., M. Aldred, et A. Bloch-Zupan (2007). Amelogenesis imperfecta. *Orphanet journal of rare diseases 2*(1), 17.

Domingue, J., D. Fensel, et J. A. Hendler (2011). *Handbook of semantic web technologies*. Springer Science & Business Media.

Du, R., H. An, K. Wang, et W. Liu (2024). A short review for ontology learning from text : Stride from shallow learning, deep learning to large language models trend. *arXiv preprint arXiv :2404.14991*.

Elgohary, K., A. Ayadi, M. Kawczynski, A. Bloch-Zupan, et C. Wemmert (2025). Ontology-guided prompting for reasoning in multimodal vision-language models : An application to rare dental disease. In *Workshop on Multimodal Knowledge and Language Modeling at IJCAI*.

GenoSmile (2015). D[4]/Phenodent. `https://www.genosmile.eu/en/resources`.

Ivanisenko, T. V., P. S. Demenkov, et V. A. Ivanisenko (2024). An accurate and efficient approach to knowledge extraction from scientific publications using structured ontology models, graph neural networks, and large language models. *International Journal of Molecular Sciences 25*(21), 11811.

Jia, C., X. Liang, et Y. Zhang (2019). Cross-domain ner using cross-domain language modeling. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 2464–2474. Association for Computational Linguistics.

Leaman, R., R. Khare, et Z. Lu (2015). Challenges in clinical natural language processing for automated disorder normalization. *Journal of biomedical informatics 57*, 28–37.

Liu, J., J. S. Barrett, E. T. Leonardi, L. Lee, S. Roychoudhury, Y. Chen, et P. Trifillis (2022). Natural history and real-world data in rare diseases : applications, limitations, and future perspectives. *The Journal of Clinical Pharmacology 62*, S38–S55.

Mihindukulasooriya, N., S. Tiwari, C. F. Enguix, et K. Lata (2023). Text2KGBench : A benchmark for ontology-driven knowledge graph generation from text. In *International semantic web conference*, pp. 247–265. Springer.

Naemi, A. et A. Sahafi (2025). Benchmarking large language models for MIMIC-IV clinical note summarization. *Journal of Healthcare Informatics Research 10*, 1–21.

Olasunkanmi, O., M. Satursky, H. Yi, C. Bizon, H. Lee, et S. Ahalt (2025). RELATE : Relation extraction in biomedical abstracts with LLMs and ontology constraints. *arXiv preprint arXiv :2509.19057*.

Reibel, A., M.-C. Manière, F. Clauss, D. Droz, Y. Alembik, E. Mornet, et A. Bloch-Zupan (2009). Orodental phenotype and genotype findings in all subtypes of hypophosphatasia. *Orphanet Journal of Rare Diseases 4*(1), 6.

Reimers, N. et I. Gurevych (2019). Sentence-BERT : Sentence embeddings using Siamese BERT-networks. In K. Inui, J. Jiang, V. Ng, et X. Wan (Eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 3982–3992. Association for Computational Linguistics.

Sahbi, A., C. Alec, et P. Beust (2025). Semantic vs. LLM-based approach : A case study of KOnPoTe vs. Claude for ontology population from French advertisements. *Data & Knowledge Engineering 156*, 102392.

Sellam, T., D. Das, et A. Parikh (2020). BLEURT : Learning robust metrics for text generation. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 7881–7892.

Singhal, K., T. Tu, J. Gottweis, R. Sayres, E. Wulczyn, M. Amin, L. Hou, K. Clark, S. R. Pfohl, H. Cole-Lewis, et al. (2025). Toward expert-level medical question answering with large language models. *Nature Medicine 31*(3), 943–950.

Spasic, I. et G. Nenadic (2020). Clinical text data in machine learning : systematic review. *JMIR medical informatics 8*(3), e17984.

Taruscio, D. et W. A. Gahl (2024). Rare diseases : challenges and opportunities for research and public health. *Nature Reviews Disease Primers 10*(1), 13.

Wei, J., X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems 35*, 24824–24837.

Wu, J., H. Dong, Z. Li, H. Wang, R. Li, A. Patra, C. Dai, W. Ali, P. Scordis, et H. Wu (2024). A hybrid framework with large language models for rare disease phenotyping. *BMC Medical Informatics and Decision Making 24*(1), 289.

Summary

Rare oro-dental diseases pose significant challenges for knowledge extraction due to scarce structured data and the fragmentation of clinical information across unstructured biomedical texts. Ontologies offer a principled solution for structuring such knowledge, but manual curation remains slow and labor-intensive. This work investigates whether open-weight LLMs can be guided through ontology-aligned prompting to reliably extract RDF triples without domain-specific fine-tuning. We propose a full pipeline combining sentence-level chunking, ontology-constrained prompts, and semantic verification through BLEURT and SBERT. Using three model families and their biomedical variants, we evaluate four prompting strategies: Zero-Shot, Few-Shot, Chain-of-Thought (CoT), and CoT + Few-Shot. Results show that prompting strategy has a stronger impact on extraction quality than model architecture or medical fine-tuning, with CoT yielding the highest semantic fidelity. Our findings demonstrate the potential of prompt-guided LLMs for scalable, ontology-consistent knowledge acquisition in rare disease domains.