Évaluation des LLM pour la correction d'ontologies environnementales

Davide Di Pierro[1], Danai Symeonidou[2], Lylia Abrouk[3]

[1] Université de Montpellier, France
[2] MISTEA, INRAE & Institut Agro, France
[3] Université Bourgogne Europe, France

davide.di-pierro@umontpellier.fr

Résumé

La construction d'ontologies représente une des tâches les plus pertinentes par rapport à l'ingénierie des connaissances aujourd'hui. L'essor des LLM a récemment amené la communauté à introduire des changements sur les pistes de modélisation des connaissances, alors que peu a été fait sur l'évaluation ou la correction assistée par les LLM. Bien que l'évaluation « dynamique » dans des contextes réels puisse être traitée à travers les tâches de l'apprentissage automatique, notamment la prédiction de lien, dans l'évaluation « statique » l'effort humain reste considérable. Dans cet article, nous présentons la capacité des LLM à identifier des axiomes qui résolvent des problèmes de modélisation et leur capacité à identifier des erreurs. Nous faisons référence principalement aux métriques OOPS pour comparer les résultats. L'expérimentation inclut des ontologies de domaines similaires : biologie, environnement et médecine. Non seulement ils représentent des domaines d'intérêt, mais aussi nous envisageons que les grands modèles de langage disposeraient d'un contexte suffisant pour traiter ces sujets-là. Dans le cadre de ces thématiques, nous avons implémenté l'ontologie OntoPFAS, le sujet principal de l'expérimentation, et nous présentons ici une comparaison avec des ontologies de domaine comparables. Ce travail s'inscrit dans le cadre du projet interdisciplinaire DAE (Détection d'Anomalies Environnementales).

Mots-clés

Ontologie, LLM, évaluation.

Abstract

The construction of ontologies is one of the most relevant tasks in knowledge engineering today. The rise of LLMs has recently led the community to introduce changes in knowledge modelling approaches, while little has been done on LLM-assisted evaluation or correction. Although 'dynamic' evaluation in real-world contexts can be addressed through machine learning tasks, such as link prediction, 'static' evaluation still requires considerable human effort. In this article, we present the ability of LLMs to identify axioms that solve modelling problems and their ability to identify critical issues themselves. We refer primarily to OOPS metrics to compare results. The experiment includes ontologies from similar domains : biology, environment, and medicine. Not only do they represent areas of interest, but we also consider that large language models would have sufficient context to handle these topics. Within the framework of these themes, we have implemented the OntoPFAS ontology, the main subject of the experiment, and we present here a comparison with ontologies from comparable domains. This work is part of the interdisciplinary DAE (Détection d'Anomalies Environnementales) project.

Keywords

Ontology, LLM, evaluation.

1 Introduction

L'ingénierie des ontologies représente encore aujourd'hui un défi scientifique majeur dans les thématiques de recherche en modélisation et intelligence artificielle [1]. En dépit de sa longue histoire, elle continue de stimuler la communauté grâce aux nouvelles technologies et besoins qui apparaissent régulièrement. Une « ontologie » constitue une représentation formelle des éléments qui composent un domaine spécifique [16]. Il est pratique de concevoir ces éléments du domaine comme des nœuds (concepts indépendants) et des arêtes (connexions entre concepts). Ensuite, nous pouvons définir un « graphe de connaissances » comme un graphe de données destiné à accumuler et à transmettre des connaissances sur le monde réel, dont les nœuds représentent des entités d'intérêt et les arêtes des relations entre ces entités [9]. Dans quelques contextes, les deux termes sont utilisés de manière interchangeable. Grâce à l'expressivité des langages formels utilisés pour décrire les ontologies, il est possible d'introduire un raisonnement logique permettant d'inférer de nouvelles connaissances ou de détecter des incohérences dans les informations existantes. Du fait de sa conception humaine beaucoup d'efforts ont été faits pour rechercher comment évaluer les résultats du processus de construction d'ontologies. D'abord, la qualité d'une ontologie est définie par sa capacité à être partagée ou utilisée par ceux qui cherchent les informations sur le domaine représenté. De plus, l'évolution de l'apprentissage automatique a amené une suite d'algorithmes permet-

tant d'estimer la plus-value d'un graphe de connaissances.
L'évaluation « dynamique » reflète la capacité d'une ontologie d'améliorer la prestation sur des tâches précises. Autrement dit, la « sémantique » ajoutée permet d'améliorer, par exemple, la prédiction de liens manquants ou la classification. Par contre, l'évaluation « statique » prend en compte la façon dont le modèle a été formalisé. Même si la modélisation aide pour certains objectifs, des erreurs sont généralement présentes. Encore aujourd'hui, le travail le plus connu sur les erreurs de construction est *OntOlogy Pitfall Scanner (OOPS)* [14] où une liste des pièges (*pitfalls*) de conceptualisations a été publiée. La liste inclut plus de quarante erreurs communes. OOPS est, en même temps, un outil en ligne [1] pour la détection de ces erreurs, bien que tous ne soient pas détectés par le logiciel.
Grâce à l'énorme capacité des *Large Language Models (LLM)* [13] à comprendre les cas d'usage en forme écrite, ils peuvent être utilisés par les experts pour automatiser partiellement quelques étapes de la construction. Les travaux suivant cette piste ont démontré que l'introduction est généralement favorable, alors qu'une approche définie qui les inclut n'existe pas à présent. En fait, l'absence d'une méthodologie partagée menace la reproducibilité et la confiance de l'aboutissement.
Dans le cadre du projet DAE, nous avons développé une première version de l'ontologie *OntoPFAS* [2] comprenant différentes évaluations statiques, OOPS compris. Un des objectifs, donc, est représenté par la possibilité d'examiner l'« opinion » des LLM sur notre modélisation. OntoPFAS est une ontologie dédiée à la représentation des molécules PFAS [8] dans l'environnement. Il s'agit d'une ressource qui rassemble plusieurs domaines d'intérêt comme la chimie, la géographie, l'exposition aux substances et leurs effets. C'est la raison pour laquelle nous avons collecté des ontologies reconnues dans la communauté sur des thématiques inhérentes. Plusieurs modèles de langages ont été testés sur cinq ontologies qui traitent à peu près des concepts similaires ou en relation entre elles.
La suite de cet article est organisée de la manière suivante : dans la section 2, nous présentons l'état de l'art sur les méthodes et les cas d'usage des LLM dans le domaine de la construction d'ontologies. La section 3 introduit les ontologies de notre expérimentation. La section 4 décrit notre approche d'évaluation, ainsi que les étapes de prétraitement nécessaires. La section 5 montre les résultats des LLM face à plusieurs tâches d'amélioration demandées et la perception générale soutenue par les résultats. Enfin, la section 6 résume les résultats, explore les extensions de ce travail et conclut l'article.

2 Travaux antérieurs

Méthodes pour l'évaluation statique d'ontologies L'évaluation statique d'ontologies nous permet de vérifier les bonnes pratiques dans le processus de modélisation. Typiquement, plusieurs aspects sont pris en compte, notamment la présence de métadonnées, l'absence d'éléments isolés et l'exhaustivité des concepts, ainsi qu'une convention de nommage cohérente. OOPS [3] est un service (disponible aussi en ligne) qui arrive à détecter la majorité des erreurs de modélisation citées dans le papier original. Parmi les *pitfalls*, nous listons ici les plus fréquents par rapport à notre expérience : *Creating unconnected ontology elements*, *missing annotations*, *missing domain or range in properties*, *inverse relationships not explicitly declared*, *using different naming conventions in the ontology*.
Bien que les métriques OOPS soient les plus connues et presque obligatoires pour l'analyse statique d'ontologie, il y a également d'autres évaluations. Parallèlement à OOPS, *FAIR OOPS (FOOPS)* [6] cherche les métadonnées manquantes ayant un impact sur les aspects *FAIR (Findable, Accessible, Interoperable, Reusable)*. Ces propriétés sont de plus en plus attendues dans les nouveaux logiciels mis à disposition. FOOPS est aussi un tool en ligne [4] qui vérifie la présence et la cohérence des triplets qui garantissent une correcte réutilisation de la ressource. Pour illustrer, les propriétés suivantes s'accordent aux aspects respectifs : *Ontology has a persistent URL* pour Findable, *Ontology uses an open protocol* pour Accessible, *Ontology imports or reuses well established vocabularies* pour Interoperable, *Ontology documentation : all terms have labels* pour Reusable.
D'autres évaluations sont encore possibles. Vrandečić et al. [19] ont proposé des caractéristiques qui définissent la qualité d'une ontologie selon plusieurs aspects - avec une forte intersection avec FAIR -, mais il n'existe pas de méthode de détection automatique de ces défauts. De plus, il existe des méthodes nécessitant une interaction humaine, comme la validation par requêtes SPARQL [4] de *questions de compétence* [2] ou à travers des interfaces et des cas d'usage [20, 17].
LLM dans la construction des ontologies La capacité des LLM de s'adapter à différents domaines, vocabulaires et exigences a amené les ingénieurs des connaissances à s'en fier et à les introduire partiellement dans les tâches de construction ou d'évaluation d'ontologies. Une utilisation très commune consiste en la vérification des axiomes d'ontologies proposée par Tsaneva et al. [18]. Les axiomes peuvent être traduits en langage naturel et donnés comme support pour le *prompt*. Rebboud et al. [15] ont évalué les grands modèles de langage par rapport aux principales tâches de la construction d'ontologie : conceptualisation, formalisation des questions de compétence, écriture des requêtes SPARQL et production de la documentation. Cependant, le travail ne se concentre pas sur la validation, mais plutôt sur l'écriture des requêtes. Lippolis et al. [10] ont en détail évalué les principaux modèles de langage, principalement de manière semi-automatique. L'évaluation de qualité automatique se concentrait aussi sur l'écriture des requêtes SPARQL, sans considérer l'opportunité de corriger des axiomes. Enfin, des travaux comme [7] montrent la possibilité d'entraîner les LLM à travers le *fine-tuning* pour

1. https://oops.linkeddata.es/
2. https://github.com/davide1797/OntoPFAS
3. https://oops.linkeddata.es/index.jsp
4. https://foops.linkeddata.es/FAIR_validator.html

répondre aux exigences spécifiques dans certains domaines. Globalement, de plus en plus les LLM entrent dans le processus de construction des ontologies. À présent, une méthodologie partagée n'existe pas encore, mais les premières tentatives sont promettantes. Ensuite, les évaluations dans l'état de l'art se déroulent sur l'écriture des requêtes et négligent encore les aspects statiques évidenciés par OOPS. Voilà pourquoi ce travail peut aider à comprendre dans quelle mesure les LLM sont capables d'introduire des axiomes pertinents tout en gardant la sémantique de ce qu'ils représentent. La spécificité des pièges OOPS nous garantit une évaluation précise sur quels types de changements ils prévoient.

3 Sources de données

Afin de comparer les résultats de notre ontologie, nous avons pris en compte des ressources dans des domaines similaires. Nous fournissons ici une description détaillée de notre ontologie ainsi que de petites descriptions pour d'autres bases de connaissances connues.

OntoPFAS [3] est une conceptualisation du domaine PFAS, concernant ses aspects d'exposition, de mesure et de propagation. Combinant des données chimiques, géographiques et issues de capteurs, il vise à représenter une plateforme pour les praticiens de différents domaines désireux de contribuer avec des données, des modèles ou des applications d'intelligence artificielle. Avec le développement parallèle de l'ensemble de données PFAS Data Hub[5], nous avons jeté les bases d'une synergie positive entre la disponibilité des données et les solutions à disposition. De nombreux concepts sont tirés d'ontologies existantes : *ChEBI*, *ExO*, *ENVO* et *SOSA*, *Schema.org*, *PROV-O* et *DCAT*.
Cette ontologie représente principalement ces trois aspects différents des préoccupations liées aux PFAS : la chimie, les mesures et l'exposition. L'architecture graphique donne un aperçu des éléments fournis. Elle est disponible sur Github[6] ainsi que sur *AgroPortal*[7].

OntoPFAS dégradée Une version volontairement altérée de cette ontologie, ci-après appelée OntoPFAS `OntoPFAS*`, a été modélisée[8] dans laquelle presque tous les pièges implémentés[9] dans OOPS sont présents sauf le *P10*, car il est déjà présent dans les autres ontologies. L'objectif est double : (i) comprendre comment la capacité des LLM change selon le type d'erreur, et (ii) comprendre si le nombre des pièges impacte la capacité ou amène à des solutions incohérentes entre les diverses erreurs.

Exposure Ontology (ExO) [11] vise à combler le fossé entre la science de l'exposition et les disciplines liées à la santé environnementale en facilitant la centralisation et l'intégration des données d'exposition afin d'éclairer la compréhension de la santé environnementale.

Green-AI Ontology [5] Il s'agit de la première tentative de modéliser la consommation d'énergie des modèles d'intelligence artificielle. La base des connaissances inclut les métriques et les instruments pour mesurer la consommation selon plusieurs aspects. La réutilisation des concepts existants est encouragée par la quantité de concepts déjà disponibles dans les domaines de l'énergie, de l'environnement et de l'informatique.

Agri-Food Experiment Ontology (AFEO) [12] est un réseau d'ontologie développé sur deux ressources existantes : l'Ontology for Agriculture Experiment (OAE) et l'Ontology for Food Processing Experiment (OFPE). Elle contient 136 concepts qui couvrent différentes pratiques viticoles, ainsi que produits et opérations liés à la production du vin.

4 Approche

Dans cette section, nous décrivons les différentes étapes pour comparer les LLM par rapport à l'évaluation des ontologies. L'évaluation basée sur OOPS compte combien d'axiomes suggérés par le LLM apportent des améliorations sur les pièges OOPS. Nous signalons aussi si les axiomes introduisent de nouvelles erreurs. Ensuite, pour l'évaluation non supervisée, nous vérifions si les axiomes sont sémantiquement corrects et n'introduisent pas de pièges OOPS.

1. Inspection OOPS Avant de demander aux LLM, nous utilisons le service OOPS comme source initiale pour comprendre les améliorations possibles. Le résultat de l'analyse fera partie du prompt à fournir aux LLM. L'analyse contient une liste d'erreurs avec les concepts associés.

2. Prétraitement Pour cette expérimentation, nous testons plusieurs modèles de langage gratuits. Vu qu'ils sont limités par rapport à la taille du prompt, nous adaptons les bases des connaissances de façon qu'elles puissent être traitées, tout en gardant les informations nécessaires à la correction des axiomes. En tenant compte de tous les possibles pièges OOPS, nous avons choisi un sous-ensemble de prédicats qui permette aux LLM de comprendre la signification des concepts, ainsi que les informations sémantiques des prédicats et les hiérarchies. Finalement, nous avons conçu une liste minimale de prédicats qui n'apportent pas de "faux positifs" sur les pièges et qui aident les LLM à connaître le contexte des éléments mentionnés. Les axiomes suivantes sont gardés : `rdf:type`, `rdfs:label`, `skos:definition`, `owl:inverseOf`, `rdfs:range`, `rdfs:domain`, `rdfs:subClassOf`, `rdfs:subPropertyOf`.

3. Prompts Après avoir vérifié que la longueur des bases de connaissances n'excède pas la taille maximale des LLM pris en compte, nous enrichissons notre prompt avec la réponse détaillée par OOPS et un prompt assez générique mais clair sur les objectifs. Nous rappelons les deux tâches : (i) introduire des axiomes pour résoudre les pièges OOPS, et (ii) modifier ou ajouter des axiomes sur des problèmes de modélisation que lui-même a détectés. Nous proposons les prompts suivants :

— `You are an ontology engineering expert.`

5. https://pdh.cnrs.fr/en/datasets/
6. https://github.com/davide1797/OntoPFAS
7. https://agroportal.lirmm.fr/ontologies/ONTOPFAS
8. https://github.com/davide1797/Bio-ontologies/blob/main/ontopfas_pitfalls.rdf
9. https://oops.linkeddata.es/webservice.html

Ontologie	# axiomes av.	# axiomes ap.
OntoPFAS	1 616	616
OntoPFAS*	1 701	650
ExO	2 355	493
Green AI	581	234
Agri-Food	521	197

TABLE 1 – Nombre de prédicats avant et après prétraitement

```
Based on the ontology and the OOPS
pitfalls of this prompt, if possible,
find concrete axioms solving them, if
they do not introduce new ontological
problems.
```

— `You are an ontology engineering expert. Based on the ontology, find possible ontological pitfalls and how you solve them, with concrete axioms if possible.`

Ces deux prompts ont été le résultat de plusieurs expérimentations. Finalement, la locution “if possible” permet de ne pas s’efforcer de trouver des axiomes résolutifs.

4. Évaluation supervisée basée sur OOPS Pour la première partie de l’évaluation, nous nous basons sur l’impact des axiomes sur les pièges OOPS ; autrement dit, si l’ajout des axiomes permet d’éliminer le(s) piège(s) détecté(s). En fait, nous comptons le piège comme “résolu” si la suggestion est en même temps réaliste dans le contexte. Par exemple, si OOPS détecte qu’une propriété manque de “domain”, il ne suffit pas que l’LLM ajoute un axiome pour ça, mais il faut qu’il ait du sens dans la modélisation.

5. Évaluation non supervisée Pour l’évaluation non supervisée, deux métriques sont calculées : (i) la capacité d’identifier des erreurs de modélisation réelles et (ii) la capacité de les résoudre. Globalement, deux scores nous donnent la mesure de leur compétence sur la tâche.

Parmi les modèles accessibles gratuitement, nous avons choisi entre les plus connus et en fonction de la facilitation de la réutilisation. C’est pourquoi nous nous sommes basés sur l’interface *OpenRouter*[10], de façon que les expérimentations soient facilement réplicables et aucun modèle ne soit avantagé. Ensuite, la condition d’égalité entre les contraintes (temps de réponse, taille des données, etc.) est garantie par l’interface commune. Finalement, les modèles suivants ont été testés : `(M)istral-7b-instruct`, `(L)LAMA-3-8b-instruct`, `(G)emma-3-4b-it`, et `(D)eepseek-r1t-chimera:free`.

5 Évaluation et discussion

Évaluation numérique D’abord, nous montrerons la capacité des LLM à résoudre les différents pitfalls qui peuvent être retrouvés dans OOPS. Pour cette évaluation, il fallait créer une base de connaissances comme `OntoPFAS*` qui inclut toutes les erreurs non présentes dans les autres. Même si une seule occurrence de quelques erreurs n’est pas statistiquement robuste, elle nous donne un aperçu de la capacité des modèles de la traiter. Nous pouvons ainsi mesurer, pour chaque modèle, le pourcentage des fois où ils proposent des axiomes réellement résolutifs. Table 2 expose les pourcentages (les meilleurs sont en gras) et le support comme cardinalité totale du pitfall dans les cinq ontologies.

Pitfall	`Mistral`	`Llama`	`Gemma`	`DeepSeek`	Supp.
P02	**100** %	0 %	0 %	**100** %	1
P03	0 %	0 %	0 %	**100** %	1
P04	50 %	20 %	20 %	**80** %	5
P05	**100** %	0 %	0 %	0 %	1
P06	**100 %**	0 %	0 %	**100** %	1
P07	**100** %	0 %	0 %	**100** %	1
P08	**100** %	25 %	20 %	**100** %	4
P10	50 %	25 %	20 %	**75** %	2
P11	75 %	20 %	20 %	**100** %	4
P12	**100** %	0 %	0 %	0 %	1
P13	75 %	20 %	25 %	**100** %	4
P19	**100** %	0 %	0 %	**100** %	1
P20	**100** %	0 %	0 %	**100** %	1
P21	**50** %	0 %	0 %	20 %	1
P22	50 %	20 %	20 %	**75** %	4
P24	**100** %	0 %	0 %	**100** %	1
P25	**100** %	0 %	0 %	**100** %	1
P26	**100** %	0 %	**100** %	**100** %	1
P27	**100** %	0 %	0 %	**100** %	1
P28	**100** %	0 %	0 %	**100** %	1
P29	**100** %	0 %	0 %	**100** %	1
P41	0 %	0 %	0 %	0 %	3

TABLE 2 – Évaluation des modèles pour chaque pitfall.

Ensuite, pour chaque base de connaissances, nous attestons les deux tâches : (i) la capacité de résoudre les pitfalls et (ii) la capacité d’identifier et résoudre de nouvelles erreurs conceptuelles. Pour la tâche (i) nous montrons quels sont les pitfalls résolus pour chaque modèle et chaque ontologie, et combien d’hallucinations il y avait dans la réponse ; c’est-à-dire, si le modèle propose des solutions à des problèmes qui ne faisaient pas partie du résultat de l’inspection OOPS (étape 1 de la section 4). Le comptage des hallucinations est fait pour nombre d’erreurs non existantes plutôt que pour nombre d’axiomes vu qu’une erreur pourrait avoir besoin de plusieurs. Dans la table 3, la correspondance entre les pitfalls et les modèles est montrée. Finalement, pour la tâche (ii), nous avons demandé aux LLM de détecter des anomalies dans la modélisation, ainsi que de proposer des axiomes qui l’améliorent. Pour chaque modèle et ontologie, nous comptons le nombre de pitfalls suggérés et, parmi eux, combien sont *objectifs* ; c’est-à-dire, qui ne dépendent pas de l’interprétation de l’expert du domaine. La figure 1 montre, avec des barres, la quantité de corrections proposées, en divisant les ontologies par couleur. Par contre, la quantité des corrections vraies est démontrée par des lignes obliques. Nous rappelons que les corrections qui ne sont pas *objectivement* vraies ne sont pas fausses, mais qu’elles nécessitent une évaluation et une acceptation supplémentaires par les experts ou les créateurs. Nous montrons les taux de précision obtenue dans la table 4.

10. https://openrouter.ai/

Ontologie	Issue(s)	M	L	G	D
OntoPFAS	P04				✓
	P11	✓		✓	✓
	P22	✓	✓	✓	✓
	P41				
	Hallucination Rate	0 %	20 %	33 %	0 %
OntoPFAS*	2, 4, 6, 7, 13, 19 - 21, 24 - 29	✓			✓
	P03				✓
	P05 - P12	✓			
	P08, P11, P22, P26	✓		✓	✓
	P41				
	Hallucination Rate	0 %	0 %	0 %	0 %
Green AI	P04, P08, P10, P13, P34	✓	✓	✓	✓
	P41				
	Hallucination Rate	0 %	0 %	0 %	0 %
ExO	P04, P08	✓			✓
	P11, P13, P22		✓		✓
	Hallucination Rate	0 %	20 %	0 %	0 %
Agri-Food	P04, P10, P21, P22				
	P08	✓			
	P11	✓	✓		✓
	P13	✓			✓
	Hallucination Rate	60 %	0 %	100 %	25 %

TABLE 3 – Table sommaire de la résolution des pitfalls.

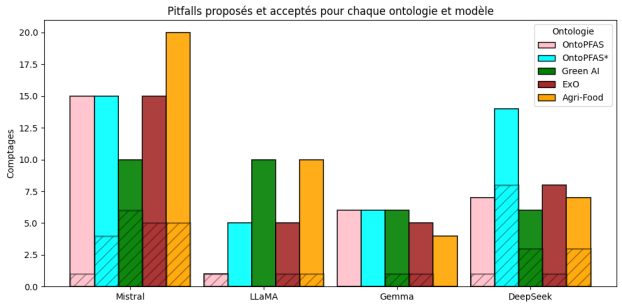

FIGURE 1 – Graphe des pitfalls rajoutés par les LLM en fonction des ontologies

Ontologie	Modèle	Précision
OntoPFAS	Mistral	0.06
	Llama	**1.00**
	Gemma	0.00
	DeepSeek	0.14
OntoPFAS*	Mistral	0.13
	Llama	0.00
	Gemma	0.00
	DeepSeek	**0.53**
Green AI	Mistral	**0.60**
	Llama	0.00
	Gemma	0.17
	DeepSeek	0.50
ExO	Mistral	**0.33**
	Llama	0.20
	Gemma	0.20
	DeepSeek	0.13
Agri-Food	Mistral	0.20
	Llama	0.10
	Gemma	0.00
	DeepSeek	**0.50**

TABLE 4 – Précision par rapport à la suggestion des pitfalls.

Discussion Bien que les LLM aient rarement des résultats parfaitement explicables, nous avons constaté des motifs généraux. Le premier résultat qui émerge est la différence dans le nombre d'hallucinations entre Agri-Food et le reste. La différence n'est pas liée aux pitfalls, vu que les mêmes erreurs se retrouvent dans les autres ontologies. Ce que nous avons noté, c'est la différence de clarté et de spécificité des concepts. En effet, il y a beaucoup de concepts modélisés qui n'ont pas de différence claire, et les axiomes descriptifs sont faibles. Par contre, nous constatons la capacité de Llama de ne pas halluciner dans tous les cas. Au contraire, nous constatons que les deux versions d'OntoPFAS sont plus faciles à corriger, vu la taille des annotations présentes. Elles sont également plus capables de proposer des axiomes supplémentaires pour d'autres erreurs, et leur pourcentage d'axiomes corrects est généralement plus grand, Mistral émergeant parmi les autres. De façon générale, les modèles Mistral et DeepSeek démontrent une efficacité supérieure à celle des autres sur la correction, mais DeepSeek est évidemment le meilleur sur la génération des nouveaux pitfalls. Ces résultats sont conforms aux impressions des utilisateurs des LLM gratuits[11]. En effet, DeepSeek montre une capacité exceptionnelle sur les raisonnement et les calcules, qui évidemment favorise la résolution par règles. En revanche, Llama, bien qu'il ait une bonne capacité d'écrire des textes de nature différente, ne peux pas encore être utilisé pour des tâches concernant les ontologies formelles.

Parmi les pitfall, P11 ("Missing domain or range in properties") est généralement résolu, même à travers une solution évident comme utiliser directement `owl:Thing` comme subject/object. Par contre, nous mettons en évidence que la langue peut être source d'hallucination. Par exemple, DeepSeek essaye de résoudre le pitfall P13 ("Inverse relationships not explictly declared") avec des noms de rélations incorrects; à titre d'exemple : `exo:#is_associated_with> owl:inverseOf exo:#associated_to>` mais "associated to" n'est pas une expression correcte en anglais, et même s'elle existait, la signification serait un synonym de la relation originale. Finalement, on constate que le P41 ("No Licence declared") n'est résolu par aucun modèle, même s'il ne demande qu'un axiome.

6 Conclusion

En guise de conclusion, nous rappelons la contribution de ce travail : l'évaluation de la capacité des LLM (i) à résoudre les pitfalls OOPS et (ii) à proposer et corriger de nouveaux pitfalls de modélisation. Dans le cadre des LLM gratuits, nous avons réduit les bases des connaissances mentionnées tout en conservant les aspects nécessaires à la compréhension de la sémantique. Enfin, les résultats de l'évaluation montrent que certains modèles sont évidemment plus capables que d'autres pour ces tâches et que les descriptions et métadonnées disponibles aident les LLM à

11. https://enricopiovano.com/blog/open-source-llms-guide/

ne pas halluciner et à proposer des solutions cohérentes. Ce travail ouvre plusieurs pistes de recherche, notamment l'amélioration des tâches avec des techniques de *fine-tuning* ou *prompt engineering*, l'évaluation massive des modèles payants et la mesure d'autres métriques de qualité comme FOOPS [12] ou OntoMetrics [13].

Remerciements

Ce travail bénéficie du soutien conjoint de l'Institut EposUM.

Références

[1] Theodore Arabatzis. Towards a historical ontology ? *Studies in History and Philosophy of science*, 34(2) :431–442, 2003.

[2] Camila Bezerra, Fred Freitas, and Filipe Santana. Evaluating ontologies with competency questions. In *2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, volume 3, pages 284–285. IEEE, 2013.

[3] Davide Di Pierro, Lylia Abrouk, Alexis Guyot, Danai Symeonidou, Benjamin Lysaniuk, and Pierre Labadie. Ontopfas : Ontologie des pfas et de leur exposition. In *PFIA*, 2025.

[4] Bob DuCharme. *Learning SPARQL : querying and updating with SPARQL 1.1*. " O'Reilly Media, Inc.", 2013.

[5] Michael Färber and David Lamprecht. The green ai ontology : An ontology for modeling the energy consumption of ai models. In *ISWC (Posters/Demos/Industry)*, 2022.

[6] Daniel Garijo, Oscar Corcho, and María Poveda-Villalón. Foops ! : An ontology pitfall scanner for the fair principles. In *ISWC (Posters/Demos/Industry)*, page 0, 2021.

[7] Rikhiya Ghosh, Hans-Martin von Stockhausen, Martin Schmitt, George Marica Vasile, Sanjeev Kumar Karn, and Oladimeji Farri. Cve-llm : Ontology-assisted automatic vulnerability evaluation using large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 28757–28765, 2025.

[8] Juliane Glüge, Martin Scheringer, Ian T Cousins, Jamie C DeWitt, Gretta Goldenman, Dorte Herzke, Rainer Lohmann, Carla A Ng, Xenia Trier, and Zhanyun Wang. An overview of the uses of per-and polyfluoroalkyl substances (pfas). *Environmental Science : Processes & Impacts*, 22(12) :2345–2373, 2020.

[9] Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia d'Amato, Gerard De Melo, Claudio Gutierrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, et al. Knowledge graphs. *ACM Computing Surveys (Csur)*, 54(4) :1–37, 2021.

[10] Anna Sofia Lippolis, Mohammad Javad Saeedizade, Robin Keskisärkkä, Aldo Gangemi, Eva Blomqvist, and Andrea Giovanni Nuzzolese. Large language models assisting ontology evaluation. In *International Semantic Web Conference*, pages 502–520. Springer, 2025.

[11] Carolyn Mattingly, Thomas McKone, Michael Callahan, Judith Blake, and Elaine Cohen Hubal. Exo : an ontology for exposure science. *Nature Precedings*, pages 1–1, 2011.

[12] Aunur Rofiq Muljarto, Jean-Michel Salmon, Brigitte Charnomordic, Patrice Buche, Anne Tireau, and Pascal Neveu. A generic ontological network for agri-food experiment integration–application to viticulture and winemaking. *Computers and electronics in agriculture*, 140 :433–442, 2017.

[13] Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. A comprehensive overview of large language models. *ACM Transactions on Intelligent Systems and Technology*, 16(5) :1–72, 2025.

[14] María Poveda-Villalón, Asunción Gómez-Pérez, and Mari Carmen Suárez-Figueroa. Oops !(ontology pitfall scanner !) : An on-line tool for ontology evaluation. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 10(2) :7–34, 2014.

[15] Youssra Rebboud, Pasquale Lisena, Lionel Tailhardat, and Raphael Troncy. Benchmarking llm-based ontology conceptualization : A proposal. In *ISWC 2024, 23rd International Semantic Web Conference*, 2024.

[16] Barry Smith. Ontology. In *The furniture of the world*, pages 47–68. Brill, 2012.

[17] Filipi Miranda Soares, Antonio Mauro Saraiva, Luís Ferreira Pires, Debora Pignatari Drucker, Kelly Rosa Braghetto, Luiz Olavo Bonino da Silva Santos, Dilvan de Abreu Moreira, Fernando Elias Corrêa, and Alexandre Cláudio Botazzo Delbem. A novel ux-based approach for ontology evaluation : Applying tree testing to the agricultural product types ontology. *IEEE Access*, 2025.

[18] Stefani Tsaneva, Stefan Vasic, and Marta Sabou. Llm-driven ontology evaluation : Verifying ontology restrictions with chatgpt. *The semantic web : ESWC satellite events*, 2024, 2024.

[19] Denny Vrandečić. Ontology evaluation. In *Handbook on ontologies*, pages 293–313. Springer, 2009.

[20] Shyama I Wilson, Jeevani S Goonetillake, Athula Ginige, and Anusha I Walisadeera. Towards a usable ontology : the identification of quality characteristics for an ontology-driven decision support system. *IEEE Access*, 10 :12889–12912, 2022.

12. https://foops.linkeddata.es/FAIR_validator.html
13. https://ontometrics.informatik.uni-rostock.de/ontologymetrics/