Évaluation des LLM pour la correction d'ontologies environnementales

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

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

davide.di-pierro@umontpellier.fr

Résumé

L'évaluation statique d'ontologies reste une tâche coûteuse, car les outils existants détectent certains défauts de modélisation sans nécessairement proposer de corrections directement exploitables. Dans cet article, nous étudions la capacité de plusieurs grands modèles de langage (LLM) à identifier et corriger des pitfalls OOPS dans des ontologies environnementales. L'expérimentation porte sur cinq ontologies, dont OntoPFAS, dédiée à la représentation des PFAS dans l'environnement, et une version volontairement dégradée permettant de contrôler différents types d'erreurs. Les résultats montrent que certains modèles, en particulier DeepSeek et Mistral, proposent des corrections pertinentes pour plusieurs pièges de modélisation, mais que la validation experte reste indispensable, notamment pour détecter les hallucinations et préserver la cohérence sémantique du domaine.

Mots-clés

Ontologie, LLM, évaluation, OOPS, environnement.

1 Introduction

L'ingénierie des ontologies constitue un enjeu important pour la représentation formelle des connaissances, notamment dans des domaines interdisciplinaires comme l'environnement, la santé et l'exposition aux polluants. Cependant, la construction d'une ontologie de qualité reste difficile : elle nécessite non seulement de représenter correctement les concepts du domaine, mais aussi de respecter des principes de modélisation permettant la réutilisation, le raisonnement et l'interopérabilité. L'évaluation statique vise précisément à vérifier ces aspects de modélisation indépendamment d'une tâche applicative particulière.

Parmi les approches existantes, OOPS (*OntOlogy Pitfall Scanner*) [4] est l'un des outils les plus utilisés pour détecter des erreurs fréquentes, telles que l'absence d'annotations, des éléments non connectés, des domaines manquants, ou encore des relations inverses non déclarées. Néanmoins, la détection d'un problème ne suffit pas : la correction reste généralement manuelle et demande une expertise à la fois ontologique et métier.

Les grands modèles de langage (LLM) sont aujourd'hui de plus en plus mobilisés pour assister différentes étapes de l'ingénierie ontologique, comme la formulation de questions de compétence, la génération de requêtes SPARQL ou la documentation [5, 3]. En revanche, leur capacité à corriger automatiquement des erreurs de modélisation reste encore peu étudiée. Notre contribution est donc d'évaluer plusieurs LLM sur deux tâches : (i) proposer des axiomes permettant de corriger des *pitfalls* OOPS déjà détectés ; (ii) identifier problèmes de modélisation et suggérer des corrections. Ce travail s'inscrit dans le cadre du projet interdisciplinaire DAE (Détection d'Anomalies Environnementales) et prend comme cas central l'ontologie *OntoPFAS* [1].

Ontologie	Domaine	Rôle dans l'étude
OntoPFAS	PFAS, exposition, environnement	Ontologie principale
OntoPFAS*	Version altérée d'OntoPFAS	Évaluation contrôlée des *pitfalls*
ExO	Exposition et santé environnementale	Comparaison domaine proche
Green AI	Énergie et IA	Comparaison environnementale
AFEO	Agroalimentaire et expérimentation	Comparaison domaine appliqué

TABLE 1 – Ontologies utilisées dans l'expérimentation

2 Méthodologie

L'expérimentation porte sur cinq ontologies issues de domaines proches : *OntoPFAS*, une version volontairement dégradée *OntoPFAS**, Exposure Ontology (ExO), Green-AI Ontology et Agri-Food Experiment Ontology (AFEO). Le tableau 1 résume leur rôle dans l'étude.

Pour chaque ontologie, nous appliquons d'abord OOPS afin d'obtenir une liste de *pitfalls*. Les ontologies sont ensuite prétraitées pour respecter les limites de taille des prompts tout en conservant les informations nécessaires à la correction : types RDF, labels, définitions, hiérarchies de classes et propriétés, domaines, codomaines et relations inverses. Quatre modèles gratuits accessibles via OpenRouter sont évalués : `Mistral-7b-instruct`, `Llama-3-8b-instruct`, `Gemma-3-4b-it` et `Deepseek-r1t-chimera:free`.

Deux prompts sont utilisés : Le premier fournit au modèle l'ontologie et les *pitfalls* détectés par OOPS, puis lui de-

Ontologie	Mistral	Llama
OntoPFAS	P11,P22/0%	P22/20%
OntoPFAS*	maj./ 0%	–/0%
Green AI	P04,P08,P10,P13,P34/0%	mêmes/0%
ExO	P04, P08/0%	P11,P13,P22/20%
Agri-Food	P08,P11,P13/60%	P11/0%

TABLE 2 – Synthèse pour Mistral et Llama.

Ontologie	Gemma	DeepSeek
OntoPFAS	P11,P22/33%	P04,P11,P22/0%
OntoPFAS*	P08,P11,P22,P26/0%	maj./0%
Green AI	mêmes/0%	mêmes/0%
ExO	–/0%	P04,P08,P11,P13,P22/0%
Agri-Food	–/100%	P11,P13/25%

TABLE 3 – Synthèse pour Gemma et DeepSeek.

mande de proposer des axiomes correctifs. Le second demande au modèle d'identifier lui-même d'éventuels problèmes de modélisation et de proposer des corrections. Les réponses sont évaluées manuellement selon trois critères : (i) le *pitfall* est effectivement corrigé ; (ii) l'axiome proposé est sémantiquement cohérent avec le domaine ; (iii) la réponse n'introduit pas de nouvelle erreur ou d'hallucination. Nous utilisons donc des comptages et des pourcentages de corrections validées, plutôt qu'une métrique de rappel, car nous ne disposons pas d'un ensemble exhaustif d'instances positives et négatives à classifier.

3 Résultats et discussion

Les tableaux 2 et 3 synthétisent les résultats supervisés par OOPS. Il indique, pour chaque ontologie, les modèles qui corrigent au moins une partie des *pitfalls* détectés, ainsi que le taux d'hallucination observé. Les résultats montrent que DeepSeek et Mistral sont globalement les modèles les plus efficaces pour proposer des axiomes correctifs. Llama et Gemma produisent moins de corrections valides, même s'ils hallucinent moins dans certains cas.

L'analyse par type de *pitfall* montre que certaines erreurs sont plus faciles à corriger. Par exemple, P11 (absence de domaine ou de codomaine pour une propriété) est souvent résolu. En revanche, P41 (absence de licence déclarée) n'est corrigé par aucun modèle. Ce résultat suggère que les LLM ne réagissent pas uniquement à la simplicité syntaxique d'une correction, mais aussi à la manière dont le problème est contextualisé dans le prompt.

La discussion qualitative met également en évidence le rôle de la clarté des concepts. Agri-Food produit davantage d'hallucinations, probablement parce que certaines distinctions conceptuelles sont moins explicites et moins bien documentées. À l'inverse, les deux versions d'OntoPFAS sont mieux corrigées lorsque les annotations et définitions disponibles permettent au modèle de relier les axiomes proposés au sens métier. Nous n'observons pas de lien direct avec l'expressivité OWL, identique dans les expériences, mais plutôt avec la qualité descriptive des entités et des relations.

Un exemple d'hallucination concerne la correction de relations inverses : un modèle peut proposer une relation syntaxiquement plausible mais linguistiquement ou sémantiquement incorrecte, comme une relation inverse construite à partir d'un nom non attesté ou ambigu. Ainsi, même lorsque la structure OWL paraît correcte, la correction peut rester invalide du point de vue du domaine. Ces résultats confirment que les LLM peuvent assister l'évaluation d'ontologies, mais ne peuvent pas encore remplacer une validation experte.

4 Conclusion

Nous avons évalué plusieurs LLM pour la correction d'ontologies environnementales à partir des *pitfalls* OOPS. Les résultats montrent que les modèles peuvent proposer des axiomes utiles, en particulier pour certains problèmes localisés comme les domaines, codomaines ou relations inverses. DeepSeek et Mistral se distinguent globalement, tandis que Llama et Gemma restent moins fiables pour cette tâche. Toutefois, les hallucinations et les erreurs sémantiques montrent que ces outils doivent être considérés comme une aide à l'expertise plutôt que comme des correcteurs automatiques autonomes. Les perspectives concernent l'élargissement de l'évaluation à d'autres modèles, l'étude de stratégies de *prompt engineering* ou de *fine-tuning*, ainsi que l'intégration d'autres critères de qualité comme FOOPS [2] ou OntoMetrics.

Remerciements

Ce travail a bénéficié d'une aide : de l'Etat gérée par l'Agence Nationale de la Recherche au titre du programme d'investissements d'avenir France 2030 portant la référence « ANR-21-EXES-0005 » ; de la Région Occitanie ; et de l'Institut ExposUM de l'Université de Montpellier.

Références

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.