Taming Hallucinations: A Semantic Matching Evaluation Framework for LLM-Generated Ontologies

Nadeen Fathallah[1], Steffen Staab[1,2] and Alsayed Algergawy[3,4]

[1]Analytic Computing, Institute for Artificial Intelligence, University of Stuttgart, Stuttgart, Germany
[2]University of Southampton, Southhampton, UK
[3]Data and Knowledge Engineering, University of Passau, Passau, Germany
[4]Institute for Informatics, Friedrich-Schiller-University Jena, Jena, Germany

Abstract

Ontology learning using Large Language Models (LLMs) has shown promise, yet remains challenged by hallucinations—spurious or inaccurate concepts and relationships that undermine domain validity. This issue is particularly critical in highly specialized fields such as life sciences, where ontology accuracy directly impacts knowledge representation and decision-making. In this work, we introduce an automated evaluation framework that systematically assesses the quality of LLM-generated ontologies by comparing their concepts and relationship triples against expert-curated domain ontologies. Our approach leverages transformer-based semantic similarity methods to detect inconsistencies, ensuring that generated ontologies align with real-world knowledge. We evaluate our framework using six LLM-generated ontologies, validating them against three reference ontologies with increasing domain specificity. Results demonstrate that our framework significantly enhances ontology reliability, reducing hallucinations while maintaining high semantic alignment with expert knowledge. This work establishes a scalable, automated approach for validating LLM-generated ontologies, paving the way for their broader adoption in complex, knowledge-intensive domains.

Keywords

Large Language Models, Life Science Domain, NeOn-GPT, Ontology Learning, Ontology Matching.

1. Introduction

Ontologies provide structured frameworks for representing domain knowledge, enabling interoperability, reasoning, and information organization. Large Language Models (LLMs) have shown promise in tasks like ontology generation and ontology population [1, 2, 3, 4, 5, 6, 7, 8]. However, one major challenge is the tendency of LLMs to produce hallucinations—instances, where they generate concepts or relationships that either do not exist or are irrelevant to the domain [9], citelavrinovics2025knowledge. This issue can lead to significant errors in fields like life sciences, where ontologies support decision-making and knowledge representation. The tendency of LLMs to hallucinate is particularly pronounced when tasked to model highly specialized domains like ecology and biology, as the lack of domain-specific training data increases the likelihood of generating inaccurate or irrelevant concepts. Although manual validation of LLM-generated ontologies by domain experts is effective, it is resource-intensive and does not scale. This work addresses the need for an automated framework to evaluate LLM-generated ontologies against domain knowledge, ultimately reducing the manual verification efforts required by domain experts.

Our evaluation framework is based on semantic ontology matching [10], identifying correspondences between concepts and relationships across ontologies. A pressing question emerges in this context: How well do LLM-generated ontology concepts and relationships align with real-world domain-specific knowledge? To address this question, we leverage six LLM-generated ontologies as a case study; those ontologies were generated in our previous work [11] using our enhanced NeOn-GPT pipeline [12] for ontology learning proposed in [11]. We validate concepts and triples generated by LLM against

ESWC 2025
nadeen.fathallah@ki.uni-stuttgart.de (N. Fathallah); steffen.staab@ki.uni-stuttgart.de (S. Staab); alsayed.algergawy@uni-passau.de (A. Algergawy)
0000-0001-7921-034X (N. Fathallah); 0000-0002-0780-4154 (S. Staab); 0000-0002-8550-4720 (A. Algergawy)


three domain-specific ontologies recommended by experts using our automated evaluation framework. These ontologies increase in relevance to the domain, allowing us to assess whether LLM-generated knowledge aligns with established domain knowledge rather than being generic.

Our results demonstrate that LLM-generated ontologies exhibit increasing domain alignment, supporting their use as automated ontology generation and population tools in highly specialized domains, with concepts and triples aligning more strongly as the reference ontology becomes increasingly domain-specific. Furthermore, our findings show that our automated evaluation framework effectively captures these alignments, improving the accuracy of LLM-generated ontologies while significantly reducing the manual efforts required for validation by domain experts.

The paper is structured as follows: Section 2 reviews related work, Section 3 outlines our methodology, Section 4 presents results, Section 5 discusses findings, and Section 6 concludes with future directions.

2. Related Work

Recent work shows that LLMs hold considerable promise for knowledge engineering tasks [13, 14, 15, 16, 17], particularly in the realm of ontology creation [1, 2, 3, 4, 5]. Several recent approaches employ structured prompting to facilitate ontology creation tasks. Notable works such as OntoChat [5], Ontogenia [18] and our own NeOn-GPT [12] illustrate the promising capabilities of LLMs in generating ontologies. These works identify challenges with ontology generation using LLMs, such as syntax errors, logical inconsistencies, and common modeling pitfalls, as well as hallucinations, where LLMs generate incorrect or irrelevant ontology elements to the domain due to sparse domain-specific training data. Unlike other methods, our NeOn-GPT framework is designed to internally address syntax and logical consistency issues and common pitfalls. It integrates detection tools such as RDFLib for syntax checking, reasoners such as Pellet and HermiT to verify logical consistency and a pitfall detection tool. Error messages from these tools, which describe the problems encountered, are used to prompt the LLM to automatically fix these issues. However, while these mechanisms effectively handle syntactical errors, logical inconsistency, and common pitfalls (e.g., wrong inverse relations, cycles in the class hierarchy), reducing hallucinations remains a significant challenge. This motivates our current work, where we propose an automatic evaluation framework aimed at mitigating hallucinations and reducing the manual effort required to validate LLM-generated ontologies.

Recent literature underscores the necessity of rigorous evaluation frameworks for systematically assessing semantic accuracy and detecting LLM-induced errors [19, 20]. Agrawal et al. [21] survey knowledge-augmented LLM methods and find that incorporating Knowledge Graphs (KGs) has shown promising results in mitigating hallucinations and improving reasoning accuracy. Essentially, KG acts as a real-world grounding by retrieving relevant facts from a KG and feeding them into the prompt (or fine-tuning the LLM on domain facts); the model is less likely to hallucinate because it can access authoritative information. Lavrinovics et al. [22] categorize various hallucination types, illustrating their negative impacts on the reliability and trustworthiness of ontology outputs. However, despite these advances, hallucinations, i.e., incorrect or irrelevant ontology elements, remain challenging, especially in specialized domains with sparse training data. Our current work proposes a novel automated framework that explicitly targets this gap. By integrating embedding and transformer-based ontology matching methods with automated semantic validation against expert-curated domain ontologies, our method significantly reduces manual verification efforts, systematically identifies hallucinations, and ensures enhanced semantic accuracy and domain relevance.

In the broader context of ontology matching, traditional lexical and heuristic methods such as PROMPT [23] and COMA [24] struggle with capturing deep semantic relationships. Recent advancements leverage embedding-based methods and LLMs to improve alignment accuracy and robustness against structural mismatches. Embedding-based models like BERTMap [25] fine-tune BERT on ontology text and integrate logical refinement to enforce consistency. Other unsupervised approaches, such as TEXTO [26], PropMatch [27], and [28], utilize transformer embeddings combined with additional structural representations. These methods enhance recall by identifying semantic equivalents beyond

string similarity. The LLMs4OM framework [29] systematically evaluates LLMs in ontology matching, employing retrieval-augmented generation (RAG) to combine semantic retrieval with LLM-based classification. It explores multiple retrieval models (e.g., sentence-BERT, OpenAI's text-embedding-ada) and evaluates LLMs across 20 datasets, demonstrating competitive performance against traditional systems like LogMap and AML. These studies highlight the promise of transformer-based models in identifying semantic similarities and capturing deep semantic relationships and contextual nuances. Consequently, we adopt a transformer-based methodology as the cornerstone of our evaluation framework.

3. Methodology

In this study, we introduce an automated evaluation framework designed to assess the reliability of LLM-generated ontologies by systematically comparing their concepts and relationships with established domain knowledge. Ontology concepts and their relationships (triples) serve as the foundational elements of structured knowledge representation, defining entities and their connections. However, LLM-generated ontologies occasionally generate hallucinated concepts and triples. To address this, we propose evaluating LLM-generated ontologies at both the concept level (to assess entity correctness) and the triple level (to validate relational integrity). By matching these elements against expert-curated ontologies, we ensure that generated knowledge aligns with established domain standards rather than being artificially constructed. The framework leverages semantic ontology matching techniques, including sentence embeddings and similarity-based alignment, to quantify the degree of conceptual and relational consistency between LLM-generated ontologies and expert-curated reference ontologies.

To apply and validate our framework, we utilize six LLM-generated ontologies that were previously developed in [11] using our enhanced NeOn-GPT pipeline [12, 11] for ontology learning with GPT-4o [30]. These ontologies focus on different aspects of the AquaDiva research domain, which investigates microbial ecology, biogeochemical cycles, and environmental processes in subsurface ecosystems:

- **AquaDiva Ontology (Version 1):** Represents concepts in groundwater ecosystems, including aquifers, microbial communities, and biogeochemical processes, but with limited structural depth.
- **AquaDiva Ontology (Version 2):** Expands the AquaDiva domain representation by incorporating a deeper class hierarchy and more object properties, improving relational depth between entities.
- **AquaDiva Ontology (Version 3):** Merges previous AquaDiva ontology versions 1 and 2.
- **Habitat Ontology:** A module of the AquaDiva ontology that captures knowledge about different habitat types within groundwater ecosystems, including their environmental conditions, microbial populations, and ecological interactions.
- **Role Ontology:** A module of the AquaDiva ontology that models the functional roles of biological, chemical, and environmental agents in groundwater systems, defining their interactions and contributions to ecosystem dynamics.
- **Carbon & Nitrogen Cycling Ontology:** A module of the AquaDiva ontology that represents biochemical processes related to carbon and nitrogen cycles in groundwater, including fixation, transformation, and exchange between environmental compartments.

These ontologies serve as test cases for our framework, allowing us to assess how well LLM-generated knowledge aligns with domain-specific standards. To validate the accuracy and domain relevance of these ontologies, we compare their concepts and triples against three expert-recommended reference ontologies:

- **OBOE-SBC (Santa Barbara Coastal Observation Ontology)** [31]: Describes environmental observations specific to the Santa Barbara Coastal Long Term Ecological Research Project. It defines site-specific measurement protocols, data collection methods, and observational contexts relevant to coastal ecosystems.

- **ENVO (Environmental Ontology)** [32]: Provides a controlled vocabulary for describing environmental entities, including ecosystems, environmental processes, and qualities.
- **CHEBI (Chemical Entities of Biological Interest)** [33]: Provides a structured classification of chemical compounds of biological relevance.

3.1. Ontology Concept and Triple Extraction

Ontology concepts and their relationships (triples) serve as the foundational elements of structured knowledge representation; concepts establish what exists, while triples describe how these concepts relate. Ensuring that these elements are human-readable is essential for effective matching, interpretation, and validation, as transformer-based models rely on textual semantics to compute similarity. Human-readable labels preserve the natural language structure, allowing models to capture meaningful relationships rather than treating ontology elements as arbitrary tokens. Without readable labels, embeddings may fail to reflect the actual meaning of concepts and triples, leading to misalignment. For example, an identifier like `OBO:0003742` provides no semantic value, whereas `Microbial Biomass` enables a model to contextualize the concept within biological and ecological domains, improving similarity computation and alignment accuracy.

Concept labels alone can be ambiguous, making it difficult to determine meaning without additional context. A term like "cell" can refer to a biological unit or a prison room, which highlights the need to extract both human-readable labels and definitions. Definitions provide critical contextual disambiguation, improving alignment accuracy with expert-curated ontologies. For instance, an LLM-generated concept labeled “Soil” without a definition may be difficult to align with the ENVO ontology’s “Agricultural Soil,” which is explicitly defined as "Soil which is part of an ecosystem used for agricultural activities." By extracting both labels and definitions, our framework ensures better semantic comparison. Triples capture relationships between concepts and form the backbone of structured ontologies. Extracting human-readable labels for subjects, predicates, and objects ensures meaningful comparison. For example, a raw triple like `(OBO:0003742) - [obo:RO_0002234] → (OBO:0000270)` is transformed into `(Microbial Biomass) - [is affected by] → (Dissolved Organic Carbon)`, making the relationship clear.

Concept Extraction: We extract ontology concepts by parsing class labels and their definitions from the ontology structure. Each extracted concept (class label) is accompanied by its associated comment, which serves as its definition. The definition is critical in our concept matching pipeline, as it aids in concept disambiguation and standardizing terminological variations. Concept disambiguation ensures that identical terms with different meanings are correctly classified (e.g., cell as a biological unit vs. cell as a prison room). Standardizing terminological variations helps align different representations of the same concept (e.g., CO_2 vs. carbon dioxide).

We developed an automated extraction pipeline to extract concepts and their definitions from LLM ontologies represented in Turtle (TTL) format. The pipeline applies regular expressions to identify ontology classes (`owl:Class`) and extract their corresponding labels and descriptions (`rdfs:comment`). The extracted concepts and their definitions are stored in a structured dictionary; each class is paired with its corresponding description or labeled as an empty string if missing. During our analysis, we observed that LLM-generated ontologies sometimes contain duplicate classes with identical labels, leading to redundant entries. To prevent inflating the results, we implemented a filtering step to remove these duplicates before storing the processed data in JSON format for further analysis. For example, in the Carbon & Nitrogen Cycling Ontology, we extracted the concept: `"Forest Ecosystem": "An ecosystem dominated by trees and other vegetation, playing a key role in carbon and nitrogen cycling."`

Similarly, a pipeline was developed to extract concepts and their definitions from reference ontologies using the BioPortal API. Our pipeline retrieves ontology classes from OBOE-SBC, ENVO, and ChEBI repositories by making iterative API requests. The data extraction process involves querying the API, parsing JSON responses to extract concept labels and definitions, and handling pagination to ensure the retrieval of all available entries. To ensure meaningful semantic content in the extracted

concepts, we excluded blank nodes (BNodes), as they often lack clear labels or definitions. Additionally, we removed UUID-like alphanumeric strings using regular expression filtering, as these randomly generated identifiers do not contribute to the ontology's conceptual structure. Concepts containing ORCID IDs, ontology prefixes (e.g., "foodon:01234" or "chebi:12345"), or database-specific notations were also filtered out or replaced with more readable terms. For example, instead of retaining "chebi:15377", we used API-based label retrieval to replace it with its human-readable name, "Water." This ensured that the extracted concepts remained interpretable and useful for semantic matching. The extracted data is stored in structured JSON files for further analysis.

Triple Extraction We extract ontology SPO triples (Subject-Predicate-Object relationships) to evaluate LLM-generated ontologies; triples capture semantic relationships between ontology entities and are essential for ontology reasoning.

We developed an automated extraction pipeline to extract triples from LLM-generated ontologies represented in Turtle (TTL) format. These triples represent hierarchical structures, entity relationships, and property constraints. The extracted triples include (a) Class Hierarchies (`subClassOf` and `is a` relationships), (b) Object Properties (links between ontology concepts), and (c) Data Properties (attributes associated with ontology entities). The extraction process begins with domain and range identification to determine property domain and range constraints, specifying the types of entities a property can connect. Using this structured information, we then proceed to construct SPO triples; for instance, if the property `"is consumed by"` is defined with `"Trace Gas"` as its domain and `"Microbial Community"` as its range, the extracted triple would be: `(Trace Gas) -[is consumed by]-> (Microbial Community)`. The final set of triples is stored in structured CSV files for further ontology matching. Additionally, the pipeline identifies hierarchical relationships, extracting `subClassOf` relationships that define taxonomic structures within the ontology: `(Methane Production) -[subClassOf]-> (Carbon Cycling Process)`. We also capture `"is a"` (`rdf:type`) relationships, which categorize entities into specific classes, such as `(North Sea) -[is a]-> (Marine Ecosystem)`. The examples presented above were extracted from the Carbon and Nitrogen Cycling Ontology.

Similarly, a pipeline was developed to extract triples from reference ontologies using the BioPortal API. Our pipeline retrieves ontology triples from the OBOE-SBC, ENVO, and ChEBI repositories by making iterative API requests. The data extraction process involves querying the API and handling the same types of triple (a) Class Hierarchies (`subClassOf` and `is a` relationships), (b) Object Properties (links between ontology concepts), and (c) Data Properties (attributes associated with ontology entities). We applied the same filtering mechanisms as in concept extraction to ensure semantic relevance.

3.2. Ontology Concept and Triple Matching

Concept Matching: We match ontology concepts by comparing class labels and their definitions across LLM-generated ontologies and reference ontologies. To achieve this, we employ a concatenation-based embedding strategy, where the concept name and its definition are merged into a single text representation before generating an embedding. Specifically, each concept is formatted as: `"concept tokenizer.sep_token definition"`. This approach allows the model to simultaneously process the concept and its associated definition, ensuring that contextual meaning is preserved when computing similarities. Instead of treating the concept and the definition separately, this method generates a single vector representation that captures the semantics of the label and the descriptive information. Our concept matching pipeline utilizes `all-MiniLM-L6-v2` [34], a pre-trained sentence transformer model, to generate fixed-size vector embeddings for both LLM-generated and reference ontology concepts. We selected `all-MiniLM-L6-v2` as our embedding model due to its lightweight architecture, efficiency, and strong performance in semantic similarity tasks. This model generates 384-dimensional sentence embeddings, effectively capturing the semantic meaning of the text while maintaining a compact size of only 22MB. Its efficiency makes it suited for handling large-scale ontology matching without requiring extensive computational resources. Furthermore, `all-MiniLM-L6-v2` has demonstrated strong semantic search, clustering, and sentence similarity performance [35, 36], making it particularly effective for concept and triple matching in ontology alignment.

Embeddings are then compared using cosine similarity, a mathematical measure that calculates the angle between two vectors in a high-dimensional space [37]. Unlike Euclidean distance, which measures absolute differences, cosine similarity evaluates how directionally similar two vectors are, making it suited for semantic comparisons. A score of 1 indicates identical meanings, while 0 suggests no similarity. Concepts that exceed a similarity threshold (e.g., 0.50) are retained as valid matches, while concepts that fail to find a meaningful match are flagged as hallucinations for domain experts to verify. For example, in the Carbon & Nitrogen Cycling Ontology, the concept: `"Soil": ""` was matched with the concept `"agricultural soil": "Soil which is part of an ecosystem used for agricultural activities."` in ENVO ontology with a similarity score of 0.74. Similarly, concepts such as `"Trace Gas Consumption"` that lacked strong matches were flagged as hallucinations and excluded. The final output consists of three main components: (a) Accepted Matches - LLM-generated concepts successfully aligned with reference ontology concepts; (b) Hallucinated Concepts - Concepts with no meaningful match, indicating potential LLM errors that need manual verification by domain experts; and (c) Match Confidence Statistics - A breakdown of how many LLM concepts were validated and their match distribution across reference ontologies.

Triple Matching: We match ontology triples by comparing Subject-Predicate-Object (SPO) relationships across LLM-generated ontologies and reference ontologies. To achieve this, we employ a sentence-based embedding strategy, where each SPO triple is converted into a natural language sentence representation before generating an embedding. For example, `(TraceGas) - [is consumed by] -> (MicrobialCommunity)` is transformed to `"TraceGas is consumed by MicrobialCommunity"`. This approach ensures that the semantic relationships within triples are preserved, allowing the model to process them holistically rather than as disjointed components. Our triple matching pipeline utilizes the same model `all-MiniLM-L6-v2`. These embeddings are then compared using cosine similarity as well. Triples that exceed a similarity threshold (e.g., 0.50) are retained as valid matches, while triples that fail to find a meaningful match are flagged as hallucinations for domain experts to verify. For example, the triple `"Karst Groundwater is a Water"` extracted from Carbon & Nitrogen Cycling ontology was matched with the following triple from the ENVO ontology: `"fresh water subcass of water"` with a similarity score of 0.58 and `"TraceGas is consumed by MicrobialCommunity"` was matched with `"methane has role bacterial metabolite"` from CHEBI ontology with similarity score 0.52.

4. Results

To evaluate the alignment of LLM-generated ontologies with domain-specific reference ontologies, our evaluation framework for concept and triple matching follows a stepwise methodology. Each LLM-generated ontology was matched against three reference ontologies ranked by domain experts in ascending order of relevance to the AquaDiva ontology domain. The matching process proceeded in the following stages: (1) Matching with the least relevant reference ontology (OBOE-SBC), (2) Matching with the combination of the least and second least relevant reference ontologies (ENVO + OBOE-SBC), and (3) Matching with all three reference ontologies together (ENVO + OBOE-SBC + CHEBI). This hierarchical approach allowed us to assess how the gradual incorporation of more relevant ontologies influenced the percentage of matched concepts and triples, providing insights into the semantic coverage and relevance of LLM-generated ontologies. Our code base is publicly available for research and development purposes, accessible at: https://github.com/NadeenAhmad/TamingHallucinations.

4.1. Concept Matching Results

The percentage of matched concepts across different reference ontology combinations is summarized in Table 1. The results show that matching only with OBOE-SBC resulted in relatively low concept match percentages across all ontologies (e.g., 46.27% for AquaDiva (Version1) and 36.94% for Carbon and Nitrogen Cycling). Adding ENVO significantly increased the percentage of matched concepts to almost double the first stage. Incorporating all three reference ontologies (ENVO + OBOE-SBC + CHEBI)

led to marginal improvements beyond the second stage, with all ontologies exceeding 90% alignment. The Carbon & Nitrogen Cycling ontology achieved the highest match percentages, likely due to their alignment with the CHEBI ontology, which classifies biologically relevant chemical compounds. Since this ontology focuses on biochemical processes related to carbon and nitrogen cycles, its terminology closely matches ChEBI's structured vocabulary.

Ontology	% of Matched Concepts with OBOE-SBC	% of Matched Concepts with ENVO + OBOE-SBC	% of Matched Concepts with ENVO + OBOE-SBC + CHEBI
AquaDiva (Version1)	46.27%	91.04%	94.03%
AquaDiva (Version2)	43.36%	91.15%	91.15%
AquaDiva (Version3)	41.72%	88.34%	90.18%
Habitat	32.53%	89.16%	90.36%
Role	39.31%	94.02%	94.02%
Carbon and Nitrogen Cycling	36.94%	94.90%	97.45%

Table 1
Concept Matching Results

4.2. Triple Matching Results

The percentage of matched triples across different reference ontology combinations is summarized in Table 2. The results show that matching only with OBOE-SBC resulted in significantly lower match percentages for triples compared to concepts (e.g., 15.98% for AquaDiva (Version1) and 13.29% for Carbon and Nitrogen Cycling). Adding ENVO led to a substantial improvement in triple alignment, with match percentages increasing by more than 40 percentage points in all cases. Including all three reference ontologies (ENVO + OBOE-SBC + CHEBI) further improved the match percentages, though the gain was less pronounced compared to the second stage.

Ontology	% of Matched Triples with OBOE-SBC	% of Matched Triples with ENVO + OBOE-SBC	% of Matched Triples with ENVO + OBOE-SBC + CHEBI
AquaDiva (Version1)	15.98%	55.03%	63.91%
AquaDiva (Version2)	12.90%	44.35%	56.45%
AquaDiva (Version3)	15.52%	52.41%	62.07%
Habitat	20.00%	63.33%	71.90%
Role	26.57%	66.18%	76.33%
Carbon and Nitrogen Cycling	13.29%	67.48%	74.83%

Table 2
Triple Matching Results

5. Discussion

Despite the high concept alignment observed in our matching process, some LLM-generated concepts and triples remained unmatched, highlighting semantic consistency and structured representation challenges. The incremental ontology matching approach revealed that as more relevant ontologies were included, the match rate increased significantly, especially for concepts. However, triples continued to show a lower match rate, emphasizing the difficulty of aligning LLM-generated relationships with structured ontological knowledge.

A review of the unmatched concepts suggests that many terms were either highly specialized (highly relevant to the AquaDiva ontology domain, therefore not found in any reference ontology) or overly generic to align with reference ontologies. Highly specialized concepts such as "Hainich Critical Zone" from the AquaDiva (Version 3) ontology, which is a valid scientific term but is absent from standard reference ontologies because it is highly relevant to the AquaDiva ontology domain only. This indicates that LLMs can generate scientifically relevant terms requiring expert manual verification. The high concept matching rates indicate that LLMs are effective at generating entity-level knowledge, likely due to their ability to synthesize widely occurring terms from large training corpora. Overly generic concepts generated by LLMs are not part of formalized ontological vocabularies, even if they are conceptually meaningful in domain contexts such as "Extreme Weather Event" in the AquaDiva (Version 1) ontology.

Unlike concept matching, triple matching showed lower alignment rates. Similar to highly specialized unmatch concepts, some triples remained unmatched because they were highly relevant to the AquaDiva ontology domain only, such as the triple: `(TriassicLimestone) - [is a] -> (GeologicalFormation)` from the AquaDiva (Version 3) ontology. Many unmatched triples lacked clear hierarchical or property constraints, making them difficult to align. For example, the unmatched triple `(reflects changes in) - [is a] -> (ObjectProperty)`, (reflects changes in) suggests a causal relationship, but standard ontologies often use more rigid property constraints, such as "has Process" or "affects". The absence of standardized predicates in LLM-generated ontologies makes direct alignment with structured ontologies challenging. Unlike traditional ontology engineering methods that rely on formal logic and domain expertise, LLMs rely on statistical correlations and vector-based search methods rather than deductive reasoning. As a result, LLMs struggle to generate subject-relation-object triples that conform to well-defined ontological structures. This explains why concept alignment is significantly higher than triple alignment—while LLMs can extract and generate entity-level knowledge effectively, they lack the ability to formalize structured semantic relationships.

6. Conclusion and Future work

In this study, we proposed an evaluation framework for assessing LLM-generated ontologies, and our framework matches the LLM-generated concept and triple against domain-specific reference ontologies while reducing manual verification efforts required for LLM-generated ontology verification by domain experts. The results demonstrate that while LLMs excel at generating domain-relevant concepts, their performance deteriorates when generating structured relationships, leading to lower triple alignment rates than concept matching. By incorporating a stepwise ontology matching strategy, our evaluation framework shows that LLM-generated concepts and triples are highly relevant to the target domain, rather than being generic entities, as evidenced by the increase in alignment percentages with more relevant reference ontologies.

While our current evaluation framework effectively assesses LLM-generated ontologies, future work should extend its capabilities to provide a deeper analysis of concept and relationship alignment. One key direction is to study the impact of concept context on semantic matching by evaluating different configurations: using only labels, combining labels with definitions, and incorporating synonyms. Additionally, exploring different embedding models beyond sentence-transformers and experimenting with alternative similarity measures could refine ontology alignment accuracy and improve hallucination detection. These enhancements will provide a more comprehensive assessment of LLM-generated ontologies and optimize their validation against expert-curated knowledge. Beyond evaluation improvements, future research should investigate the potential of Large Context Models (LCMs) [38] for ontology generation. Unlike standard LLMs, LCMs can retain long-range dependencies, which could enhance hierarchical ontology structuring and consistency over extended contexts. Evaluating their performance in generating coherent, logically structured ontologies will provide insights into their suitability for knowledge-intensive domains. Further, LLMs can be leveraged for ontology population and enrichment, suggesting new entity classes, refining definitions, and proposing synonyms to reduce

manual effort in ontology expansion. They could also support KG population by identifying missing entities and suggesting relationships. While LLMs demonstrate strong concept generation capabilities, their structured relationship formation remains less precise. Future work should explore hybrid approaches that integrate LLMs with rule-based reasoning and ontology validation methods to enhance the quality and consistency of generated triples. These refinements will contribute to more reliable and scalable ontology learning frameworks for specialized domains.

References

[1] P. Mateiu, A. Groza, Ontology engineering with large language models, in: 25th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing, SYNASC 2023, Nancy, France, September 11-14, 2023, IEEE, 2023, pp. 226–229. URL: https://doi.org/10.1109/SYNASC61333.2023.00038. doi:10.1109/SYNASC61333.2023.00038.

[2] H. B. Giglou, J. D'Souza, S. Auer, Llms4ol: Large language models for ontology learning, in: The Semantic Web - ISWC 2023 - 22nd International Semantic Web Conference, Athens, Greece, November 6-10, 2023, Proceedings, Part I, volume 14265 of *Lecture Notes in Computer Science*, Springer, 2023, pp. 408–427. URL: https://doi.org/10.1007/978-3-031-47240-4_22. doi:10.1007/978-3-031-47240-4_22.

[3] V. K. Kommineni, B. König-Ries, S. Samuel, From human experts to machines: An LLM supported approach to ontology and knowledge graph construction, CoRR abs/2403.08345 (2024). URL: https://doi.org/10.48550/arXiv.2403.08345. doi:10.48550/ARXIV.2403.08345. arXiv:2403.08345.

[4] M. J. Saeedizade, E. Blomqvist, Navigating ontology development with large language models, in: The Semantic Web - 21st International Conference, ESWC 2024, Hersonissos, Crete, Greece, May 26-30, 2024, Proceedings, Part I, volume 14664 of *Lecture Notes in Computer Science*, Springer, 2024, pp. 143–161. URL: https://doi.org/10.1007/978-3-031-60626-7_8. doi:10.1007/978-3-031-60626-7_8.

[5] B. Zhang, V. A. Carriero, K. Schreiberhuber, S. Tsaneva, L. S. González, J. Kim, J. de Berardinis, Ontochat: a framework for conversational ontology engineering using language models, CoRR abs/2403.05921 (2024). URL: https://doi.org/10.48550/arXiv.2403.05921. doi:10.48550/ARXIV.2403.05921. arXiv:2403.05921.

[6] Y. Hu, S. Ghosh, T. Nguyen, S. Razniewski, GPTKB: building very large knowledge bases from language models, CoRR abs/2411.04920 (2024). URL: https://doi.org/10.48550/arXiv.2411.04920. doi:10.48550/ARXIV.2411.04920. arXiv:2411.04920.

[7] E. Motta, A. A. Salatino, F. Osborne, L. Pompianu, A. Pisu, D. R. Recupero, D. Riboni, Leveraging language models for generating ontologies of research topics, in: S. Tiwari, N. Mihindukulasooriya, F. Osborne, D. Kontokostas, J. D'Souza, M. Kejriwal, M. A. Pellegrino, A. Rula, J. E. L. Gayo, M. Cochez, M. Alam (Eds.), Joint proceedings of the 3rd International workshop on knowledge graph generation from text (TEXT2KG) and Data Quality meets Machine Learning and Knowledge Graphs (DQMLKG) co-located with the Extended Semantic Web Conference (ESWC 2024), Hersonissos, Greece, May 26-30, 2024, volume 3747 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2024, p. 11. URL: https://ceur-ws.org/Vol-3747/text2kg_paper6.pdf.

[8] M. Funk, S. Hosemann, J. C. Jung, C. Lutz, Towards ontology construction with language models, in: S. Razniewski, J. Kalo, S. Singhania, J. Z. Pan (Eds.), Joint proceedings of the 1st workshop on Knowledge Base Construction from Pre-Trained Language Models (KBC-LM) and the 2nd challenge on Language Models for Knowledge Base Construction (LM-KBC) co-located with the 22nd International Semantic Web Conference (ISWC 2023), Athens, Greece, November 6, 2023, volume 3577 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023. URL: https://ceur-ws.org/Vol-3577/paper16.pdf.

[9] J. Yao, K. Ning, Z. Liu, M. Ning, L. Yuan, LLM lies: Hallucinations are not bugs, but features as adversarial examples, CoRR abs/2310.01469 (2023). URL: https://doi.org/10.48550/arXiv.2310.01469. doi:10.48550/ARXIV.2310.01469. arXiv:2310.01469.

[10] Y. R. Jean-Mary, E. P. Shironoshita, M. R. Kabuka, Ontology matching with semantic verification, J. Web Semant. 7 (2009) 235–251. URL: https://doi.org/10.1016/j.websem.2009.04.001. doi:10.1016/J.WEBSEM.2009.04.001.

[11] N. Fathallah, S. Staab, A. Algergawy, LLMs4Life: Large Language Models for Ontology Learning in Life Sciences, in: Proceedings of the ELMKE Workshop on Evaluation of Language Models in Knowledge Engineering, EKAW-24 (24th International Conference on Knowledge Engineering and Knowledge Management), 2024. URL: https://arxiv.org/abs/2412.02035. arXiv:2412.02035.

[12] N. Fathallah, A. Das, S. De Giorgis, A. Poltronieri, P. Haase, L. Kovriguina, Neon-gpt: A large language model-powered pipeline for ontology learning, in: The Extended Semantic Web Conference, 2024.

[13] V. K. Kommineni, B. König-Ries, S. Samuel, Towards the automation of knowledge graph construction using large language models (2024).

[14] B. P. Allen, L. Stork, P. Groth, Knowledge engineering using large language models, TGDK 1 (2023) 3:1–3:19. URL: https://doi.org/10.4230/TGDK.1.1.3. doi:10.4230/TGDK.1.1.3.

[15] T. Xu, Y. Gu, M. Xue, R. Gu, B. Li, X. Gu, Knowledge graph construction for heart failure using large language models with prompt engineering, Frontiers Comput. Neurosci. 18 (2024). URL: https://doi.org/10.3389/fncom.2024.1389475. doi:10.3389/FNCOM.2024.1389475.

[16] R. Alharbi, U. Ahmed, D. Dobriy, W. Łajewska, L. Menotti, M. J. Saeedizade, M. Dumontier, Exploring the role of generative ai in constructing knowledge graphs for drug indications with medical context, Proceedings http://ceur-ws. org ISSN 1613 (2023) 0073.

[17] B. Zhang, I. Reklos, N. Jain, A. Meroño-Peñuela, E. Simperl, Using large language models for knowledge engineering (LLMKE): A case study on wikidata, in: S. Razniewski, J. Kalo, S. Singhania, J. Z. Pan (Eds.), Joint proceedings of the 1st workshop on Knowledge Base Construction from Pre-Trained Language Models (KBC-LM) and the 2nd challenge on Language Models for Knowledge Base Construction (LM-KBC) co-located with the 22nd International Semantic Web Conference (ISWC 2023), Athens, Greece, November 6, 2023, volume 3577 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023. URL: https://ceur-ws.org/Vol-3577/paper8.pdf.

[18] A. S. Lippolis, M. Ceriani, S. Zuppiroli, A. G. Nuzzolese, Ontogenia: Ontology generation with metacognitive prompting in large language models, in: A. Meroño-Peñuela, Ó. Corcho, P. Groth, E. Simperl, V. Tamma, A. G. Nuzzolese, M. Poveda-Villalón, M. Sabou, V. Presutti, I. Celino, A. Revenko, J. Raad, B. Sartini, P. Lisena (Eds.), The Semantic Web: ESWC 2024 Satellite Events - Hersonissos, Crete, Greece, May 26-30, 2024, Proceedings, Part I, volume 15344 of *Lecture Notes in Computer Science*, Springer, 2024, pp. 259–265. URL: https://doi.org/10.1007/978-3-031-78952-6_38. doi:10.1007/978-3-031-78952-6_38.

[19] F. Petroni, T. Rocktäschel, S. Riedel, P. S. H. Lewis, A. Bakhtin, Y. Wu, A. H. Miller, Language models as knowledge bases?, in: K. Inui, J. Jiang, V. Ng, 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, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019, Association for Computational Linguistics, 2019, pp. 2463–2473. URL: https://doi.org/10.18653/v1/D19-1250. doi:10.18653/V1/D19-1250.

[20] H. Ghanem, C. Cruz, Fine-tuning vs. prompting: evaluating the knowledge graph construction with llms, in: 3rd International Workshop on Knowledge Graph Generation from Text (Text2KG) Co-located with the Extended Semantic Web Conference (ESWC 2024), volume 3747, 2024, p. 7.

[21] G. Agrawal, T. Kumarage, Z. Alghamdi, H. Liu, Can knowledge graphs reduce hallucinations in llms?: A survey, arXiv preprint arXiv:2311.07914 (2023).

[22] E. Lavrinovics, R. Biswas, J. Bjerva, K. Hose, Knowledge graphs, large language models, and hallucinations: An nlp perspective, Journal of Web Semantics 85 (2025) 100844.

[23] N. F. Noy, M. A. Musen, et al., Algorithm and tool for automated ontology merging and alignment, in: Proceedings of the 17th National Conference on Artificial Intelligence (AAAI-00). Available as SMI technical report SMI-2000-0831, volume 115, sn, 2000.

[24] H.-H. Do, E. Rahm, Coma—a system for flexible combination of schema matching approaches, in: VLDB'02: Proceedings of the 28th International Conference on Very Large Databases, Elsevier,

2002, pp. 610–621.
[25] Y. He, J. Chen, D. Antonyrajah, I. Horrocks, Bertmap: a bert-based ontology alignment system, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 36, 2022, pp. 5684–5691.
[26] Y. Peng, M. Alam, T. Bonald, Ontology matching using textual class descriptions, in: International Workshop on Ontology Matching, 2023.
[27] G. Sousa, R. Lima, C. Trojahn, Combining word and sentence embeddings with alignment extension for property matching., in: OM@ ISWC, 2023, pp. 91–96.
[28] G. Sousa, R. Lima, C. Trojahn, Complex ontology matching with large language model embeddings, arXiv preprint arXiv:2502.13619 (2025).
[29] H. B. Giglou, J. D'Souza, F. Engel, S. Auer, Llms4om: Matching ontologies with large language models, arXiv preprint arXiv:2404.10317 (2024).
[30] OpenAI, Hello gpt-4o, https://openai.com/index/hello-gpt-4o/, 2024. Accessed: 2024-05-18.
[31] B. Leinfelder, Santa barbara coastal observation ontology, BioPortal Ontology Repository, 2010. URL: https://bioportal.bioontology.org/ontologies/OBOE-SBC, extensible Observation Ontology for the Santa Barbara Coastal Long Term Ecological Research project (SBC-LTER). OBOE SBC extends core concepts defined in the OBOE suite that are particular to the SBC-LTER project's data collection activities, including specific measurement protocols, sites, etc. This serves as a case study ontology for the Semtools project.
[32] P. L. Buttigieg, N. Morrison, B. Smith, C. Mungall, S. Lewis, Environment ontology (envo), http://obofoundry.org/ontology/envo.html, 2021. Accessed: 2024-09-12.
[33] A. Malik, Chemical entities of biological interest ontology, BioPortal Ontology Repository, 2025. URL: https://bioportal.bioontology.org/ontologies/CHEBI, a structured classification of chemical compounds of biological relevance.
[34] N. Reimers, I. Gurevych, sentence-transformers/all-minilm-l6-v2, Hugging Face Model Hub, 2024. URL: https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2, this model is based on the nreimers/MiniLML6-H384-uncased model and was further fine-tuned using a dataset of 1 billion sentence pairs. The embeddings' length is 384. Accessed on 15 January 2024.
[35] C. Galli, N. Donos, E. Calciolari, Performance of 4 pre-trained sentence transformer models in the semantic query of a systematic review dataset on peri-implantitis, Inf. 15 (2024) 68. URL: https://doi.org/10.3390/info15020068. doi:10.3390/INFO15020068.
[36] E. Vergou, I. Pagouni, M. Nanos, K. L. Kermanidis, Readability classification with wikipedia data and all-minilm embeddings, in: I. Maglogiannis, L. S. Iliadis, A. Papaleonidas, I. P. Chochliouros (Eds.), Artificial Intelligence Applications and Innovations. AIAI 2023 IFIP WG 12.5 International Workshops - MHDW 2023, 5G-PINE 2023, AI-BMG 2023, and VAA-CP-EB 2023, León, Spain, June 14-17, 2023, Proceedings, volume 677 of *IFIP Advances in Information and Communication Technology*, Springer, 2023, pp. 369–380. URL: https://doi.org/10.1007/978-3-031-34171-7_30. doi:10.1007/978-3-031-34171-7_30.
[37] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, in: Y. Bengio, Y. LeCun (Eds.), 1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings, 2013. URL: http://arxiv.org/abs/1301.3781.
[38] H. Ahmad, D. Goel, The future of AI: exploring the potential of large concept models, CoRR abs/2501.05487 (2025). URL: https://doi.org/10.48550/arXiv.2501.05487. doi:10.48550/ARXIV.2501.05487. arXiv:2501.05487.