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 domain knowledge (i.e. expert-curated domain ontologies). Our approach leverages transformer-based semantic similarity methods to detect hallucinations, 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. 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]. 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 [6, 7]. 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 and relationships. 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. Importantly, our evaluation framework is not limited solely to mitigating hallucinations in LLM-generated ontologies; it can also be adapted for other knowledge engineering tasks performed by LLMs. For instance, the framework can validate LLM-generated knowledge graphs, ensure accuracy in semantic annotations, and verify consistency in automated taxonomy creation. To achieve these goals, our proposed evaluation framework is based on semantic ontology matching, identifying correspondences between concepts and relationships across ontologies [8].

Third International Workshop on Semantic Technologies and Deep Learning Models for Scientific, Technical and Legal Data SemTech4STLD, ESWC '25, June 1st or 2nd, 2025, Portoroz, Slovenia
*Corresponding author.
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)

A pressing question emerges in this context: How well can LLMs model domain-specific concepts and relationships that align with real-world domain knowledge? To address this question, we leverage six LLM-generated ontologies as a case study; those ontologies were generated in our previous work [9] using our enhanced NeOn-GPT pipeline for ontology learning proposed in [10, 9]. We validate concepts and relationship triples generated by LLM against three domain-specific ontologies recommended by domain 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 relationship triples aligning more strongly as the reference ontology becomes more domain-specific. Furthermore, our findings show that our automated evaluation framework effectively captures these alignments 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 [11, 12, 13, 14, 15], 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], OntoGenix [16], Ontogenia [17] and our own NeOn-GPT [10] illustrate the promising capabilities of LLMs in generating ontologies. These works identify challenges with ontology generation using LLMs, such as syntax errors, logical inconsistencies, common modeling pitfalls, and 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 address syntax and logical consistency issues and common pitfalls internally. 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, prompt the LLM to fix these issues automatically. 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 to mitigate hallucinations and reduce 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 [18, 19]. Lavrinovics et al. [7] categorize various hallucination types, illustrating their negative impacts on the reliability and trustworthiness of ontology outputs. Agrawal et al. [20] survey knowledge-augmented LLM methods, showing how incorporating Knowledge Graphs (KGs) can mitigate hallucinations by grounding model outputs in validated, domain-specific knowledge. However, despite these advances, hallucinations remain challenging, i.e., incorrect or irrelevant ontology elements, especially in highly specialized domains with sparse training data. Our current work proposes a novel automated framework that targets this gap. To determine whether a generated ontology contains hallucinations—such as non-existent or irrelevant concepts and triples—we compare it against expert-curated, domain-specific ontologies. This process, known as ontology matching, aims to identify correspondences between semantically related concepts across different ontologies [21]. Our framework performs ontology matching by leveraging transformer-based embedding techniques to semantically align concepts and triples between the generated and reference ontologies. The percentage of matched elements serves as an indicator of semantic accuracy and domain relevance, helping to flag potential hallucinations in the generated output.

In the broader context of ontology matching, lexical and heuristic methods such as PROMPT [22]

and COMA [23] have been widely used for their ability to identify matches based on name or string similarity. However, these methods often fail to capture conceptual equivalence when similar terms are expressed differently. Recent approaches use embedding-based models and LLMs to identify semantically equivalent concepts based on meaning and context rather than wording alone. Embedding-based models like BERTMap [24] fine-tune BERT on ontology texts and apply logic-based constraints—such as disjointness and hierarchy rules—to ensure consistent alignment of equivalent concepts. Unsupervised methods like TEXTO [25], PropMatch [26], and [27] enhance matching by combining transformer embeddings with structural features, such as class hierarchies and property relationships, allowing them to identify concept equivalence beyond string similarity. The LLMs4OM framework [28] 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 [29] and AML [30]. These studies show 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 [1] designed to assess the reliability of LLM-generated ontologies by systematically comparing their concepts and relationships with established domain knowledge (i.e. expert-curated domain ontologies). We evaluate LLM-generated ontologies at both the concept level (to assess entity correctness) and the relationship-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- sentence embeddings and similarity-based alignment, to quantify the degree of conceptual and relational consistency between LLM-generated ontologies and expert-curated reference ontologies, an overview of our proposed framework is shown in Figure 1. We utilize six LLM-generated ontologies that were previously developed in [9] using our enhanced NeOn-GPT pipeline [10, 9] for ontology learning with GPT-4o [31] as a case study. These ontologies represent different aspects of the AquaDiva[2] research domain [32, 33], 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.
- **Role Ontology:** A module of the AquaDiva ontology that models the functional roles of biological, chemical, and environmental agents in groundwater systems.
- **Carbon & Nitrogen Cycling Ontology:** A module of the AquaDiva ontology that represents biochemical processes related to carbon and nitrogen cycles in groundwater.

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

[1] Our code base is publicly available for research and development purposes, accessible at: https://github.com/NadeenAhmad/TamingHallucinations

[2] https://www.aquadiva.uni-jena.de/

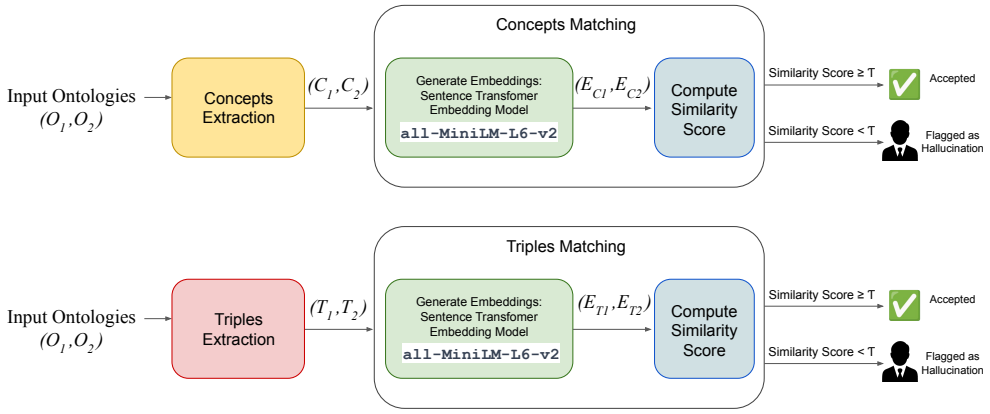

Figure 1: Overview of our automated evaluation framework for flagging potential hallucinations in LLM-generated ontologies. Given two input ontologies, LLM-generated ontology O_1 and expert-curated reference ontology O_2, the framework extracts **set of concepts** C_1 from O_1 and **set of concepts** C_2 from O_2, as well as **set of triples** T_1 from O_1 and **set of triples** T_2 from O_2. Each extracted concept and triple is transformed into an embedding representation using the `all-MiniLM-L6-v2` sentence transformer model, yielding embeddings (E_{C1}, E_{C2}) for concepts and (E_{T1}, E_{T2}) for triples. In the matching process, we use Cosine similarity to compute the similarity score of each concept embedding from E_{C1} against **all concepts** in E_{C2}, and each triple embedding from E_{T1} against **all triples** in E_{T2} to determine semantic alignment. If the similarity score meets or exceeds a predefined threshold τ, the concept or triple is **accepted**. Otherwise, it is **flagged as a hallucination**, requiring expert validation.

these ontologies, we compare their concepts and triples against three expert-recommended reference ontologies:

- **OBOE-SBC (Santa Barbara Coastal Observation Ontology)** [34]: Describes environmental observations specific to the Santa Barbara Coastal Long Term Ecological Research Project.
- **ENVO (Environmental Ontology)** [35]: Provides a controlled vocabulary for describing environmental entities, including ecosystems, environmental processes, and qualities.
- **CHEBI (Chemical Entities of Biological Interest)** [36]: Provides a structured classification of chemical compounds of biological relevance.

3.1. Ontology Concept and Triple Extraction

Our evaluation framework starts by extracting concepts and triples from LLM-generated and expert-curated reference ontologies.

Concept Extraction: We use transformer-based models that rely on textual semantics to compute embeddings, so extracting human-readable labels for ontology concepts is crucial for accurate matching, interpretation, and validation. These labels retain natural language structure, helping models understand relationships instead of treating concepts as meaningless identifiers. Without readable labels, embeddings 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 its label: `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 without context. For example, "cell" refers to a biological unit or a prison room, highlighting the need to extract labels and definitions. Definitions provide crucial disambiguation, improving alignment with expert-curated ontologies. For instance, an LLM-generated concept labeled "Microbial Activity" without a definition may be difficult to align with the ENVO ontology's "Microbial Biogeochemical Process," defined in the ENVO ontology as "A process mediated by microbial activity influencing the transformation of chemical compounds in an ecosystem." Extracting both labels and definitions ensures a more accurate semantic comparison.

Our framework extracts ontology concepts by parsing class labels and their associated definitions. Definitions are key in concept matching by resolving ambiguities and standardizing variations. Disambiguation ensures that terms with multiple meanings are classified correctly (e.g., "cell" as a biological unit vs. a prison room). At the same time, standardization aligns 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 definitions (`rdfs:comment`). The extracted concepts and their definitions are stored in a structured dictionary; each class is paired with its corresponding definition or labeled as an empty string if missing. Our analysis 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 replaced with readable terms by retrieving the class label from their corresponding data sources. 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.

The extracted set C_1 from the LLM-generated ontology consists of concepts paired with their corresponding definitions. The set from the expert-curated reference ontology is denoted as C_2 (as shown in Figure 1).

Triple Extraction: We extract subject–predicate–object (SPO) triples from the ontology and obtain their human-readable forms by extracting labels for each entity in the triple. For example, the triple `(OBO:0003742) - [obo:RO_0002234] → (OBO:0000270)` is transformed into `(Microbial Biomass) - [is affected by] → (Dissolved Organic Carbon)`.

We developed an automated extraction pipeline to extract triples from LLM-generated ontologies represented in Turtle (TTL) format. 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, and (c) Data Properties. We applied the same filtering mechanisms as in concept extraction to ensure semantic relevance. The extracted set of triples from the LLM-generated ontology is T_1, while the expert-curated reference ontology set is T_2 (see Figure 1).

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. Each concept is formatted as: `"concept tokenizer.sep_token definition"`. This approach allows the model to process the concept and its associated definition simultaneously. Including a separator token explicitly signals the boundary between the concept label and its definition, helping the embedding model distinguish and appropriately weigh the semantic contributions of each component. Thus, instead of merging labels and definitions into one potentially ambiguous sentence, our concatenation approach ensures accurate contextualization and improved embedding quality. Our concept matching pipeline utilizes `all-MiniLM-L6-v2` [37], 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 22MB. Its efficiency suits it 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 [38, 39]. In this process, the sets of concepts C_1 and C_2 are transformed into the sets of embeddings E_{C1} and E_{C2}, respectively.

Embeddings are then compared using cosine similarity, a mathematical measure that calculates the angle between two vectors in a high-dimensional space [40]. 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. In this process, each concept and definition embedding from E_{C1} is compared against all concepts in E_{C2} using cosine similarity (see Figure 1). Concepts that exceed a similarity threshold (τ) (e.g., 0.50) are retained as valid matches. Concepts that fail to find a meaningful match are flagged as hallucinations for domain experts to verify. For example, in the AquaDiva Ontology (Version 2), the concept: `"GroundwaterPrecipitation": "The process where water precipitates, either through chemical means within groundwater systems or as a part of the hydrological cycle impacting groundwater recharge."` was matched with the concept `"PrecipitationWaterSample": "PrecipicationWater falls from the atmosphere to earth, as rain or snow. Also, see the process called Precipitation."` in OBOE-SBC ontology with a similarity score of 0.66. Concepts such as `"Trace Gas Consumption"` that lacked strong matches were flagged as hallucinations for experts to review. 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. We employ a sentence-based embedding strategy to achieve this, 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, transforming the sets of triples T_1 and T_2 into the sets of embeddings E_{T1} and E_{T2}, respectively. These embeddings are then compared using cosine similarity as well. Each triple embedding from E_{T1} is compared against all triples in E_{T2} using cosine similarity (see Figure 1). Triples that exceed a similarity threshold (τ) (e.g., 0.50) are retained as valid matches. In contrast, 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: "freshwater subclass 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, each LLM-generated ontology was matched against three reference ontologies ranked by domain experts (i.e. ecologists) 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 (OBOE-SBC + ENVO), and (3) Matching with all three reference ontologies together (OBOE-SBC + ENVO + CHEBI). This stepwise approach reveals whether alignment improves with more domain-specific references—indicating higher semantic relevance in the LLM-generated ontologies.

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 rate of matched concepts, almost doubling 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.

LLM-Generated Ontology via NeOn-GPT	% of Matched Concepts with OBOE-SBC	% of Matched Concepts with (OBOE-SBC +ENVO)	% of Matched Concepts with (OBOE-SBC +ENVO +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
Percentage of matched concepts between LLM-generated ontologies and reference ontologies at three stages: using OBOE-SBC alone, OBOE-SBC combined with ENVO, and all three (OBOE-SBC + ENVO + CHEBI).

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 (OBOE-SBC + ENVO + CHEBI) further improved the match percentages, though the gain was less pronounced than in the second stage.

LLM-Generated Ontology via NeOn-GPT	% of Matched Triples with OBOE-SBC	% of Matched Triples with (OBOE-SBC +ENVO)	% of Matched Triples with (OBOE-SBC +ENVO +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
Percentage of matched triples between LLM-generated ontologies and reference ontologies at three stages: using OBOE-SBC alone, OBOE-SBC combined with ENVO, and all three (OBOE-SBC + ENVO + CHEBI).

5. Discussion

The high concept matching rates indicate that LLMs are effective at generating widely accepted entity-level knowledge, likely due to their ability to synthesize common terms from large-scale training corpora. The incremental ontology matching approach revealed that as more relevant ontologies were included, the match rate increased significantly, especially for concepts. Despite the high concept alignment observed in our matching process, some LLM-generated concepts remained unmatched, highlighting some semantic consistency and structured representation challenges. A manual review of the unmatched concepts suggests that many terms were either highly specialized (i.e., highly relevant to the AquaDiva ontology domain but not represented in reference ontologies) or overly generic to align with structured reference vocabularies. Highly specialized concepts such as `"Hainich Critical Zone"` from the AquaDiva (Version 3) ontology represent valid scientific terms that reflect domain-specific knowledge. Their absence from reference ontologies does not imply inaccuracy but rather illustrates the potential of LLMs to surface novel or underrepresented entities relevant to the target domain. On the other hand, generic terms like `"Extreme Weather Event"` in the AquaDiva (Version 1) ontology are meaningful but often not formalized in structured vocabularies.

This is where human-in-the-loop validation becomes essential, enabling domain experts to assess such unmatched concepts' correctness, relevance, and potential value, as shown in Figure 1. For this reason, our approach complements—rather than replaces—expert validation, helping reduce the manual effort required. It flags unmatched concepts and triples for expert review, acknowledging that they may represent legitimate and valuable domain knowledge that falls outside the scope of existing ontologies.

Unlike concept matching, triple matching showed lower alignment rates. Similar to highly specialized unmatched 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) sug-

gests 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 struggle to formalize structured semantic relationships.

6. Conclusion and Future work

In this study, we proposed an evaluation framework for assessing LLM-generated ontologies by matching their concepts and triples against domain-specific reference ontologies, aiming to reduce the manual verification efforts required from domain experts. The results demonstrate that while LLMs excel at generating domain-relevant concepts, their performance declines when it comes to producing structured relationships, as reflected in the lower triple alignment rates. Our stepwise ontology matching strategy further confirmed that the relevance of the reference ontology significantly influences the alignment quality, with higher alignment percentages achieved when using more domain-specific ontologies. Future work should also investigate the potential of leveraging LLMs as the domain expert in this pipeline inspired by previous works that use LLM-as-a-judge [41, 42]. In our previous work, we evaluated LLM-generated ontologies for syntactic correctness, logical consistency, common modeling pitfalls, and structural properties [10, 9], this work extends this evaluation to the semantic level, assessing the alignment of concepts and relationship triples with expert-curated reference ontologies. In future work, we plan to assess the practical utility of these ontologies through task-based evaluations, such as their ability to support competency questions and other real-world applications, providing a deeper understanding of their functional value. Beyond the current use case, we aim to use this framework to evaluate and compare different ontology learning pipelines and LLMs. Additionally, we plan to adapt the framework to support other knowledge engineering tasks, such as validating LLM-generated knowledge graphs, semantic annotations, or taxonomy construction, helping to ensure consistency and domain relevance across a wider range of automated knowledge modeling scenarios. Finally, the results of this semantic evaluation suggest that future ontology generation may benefit from models with improved contextual understanding; thus, we intend to explore the potential of Large Context Models (LCMs) to improve hierarchical structuring in LLM-generated ontologies [43].

Acknowledgement

The authors thank **Mr. Yihang Zhao** (Department of Informatics, King's College London) for kindly presenting this paper at *SemTech4STLD @ ESWC 2025* on their behalf.

Declaration on Generative AI

During the preparation of this work, the author(s) used GPT-4 and Grammarly for Grammar and spelling checks. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

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] 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.

[7] 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.

[8] 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.

[9] 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.

[10] N. Fathallah, A. Das, S. D. Giorgis, A. Poltronieri, P. Haase, L. Kovriguina, Neon-gpt: A large language model-powered pipeline for ontology learning, 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. 36–50. URL: https://doi.org/10.1007/978-3-031-78952-6_4. doi:10.1007/978-3-031-78952-6_4.

[11] V. K. Kommineni, B. König-Ries, S. Samuel, Towards the automation of knowledge graph construction using large language models 3874 (2024) 19–34. URL: https://ceur-ws.org/Vol-3874/paper2.pdf.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] M. Val-Calvo, M. E. Aranguren, J. M. Martínez-Hernández, G. Almagro-Hernández, P. Deshmukh, J. A. Bernabé-Díaz, P. Espinoza-Arias, J. L. Sánchez-Fernández, J. Mueller, J. T. Fernández-Breis, Ontogenix: Leveraging large language models for enhanced ontology engineering from datasets, Inf. Process. Manag. 62 (2025) 104042. URL: https://doi.org/10.1016/j.ipm.2024.104042. doi:10.1016/J.IPM.2024.104042.

[17] 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.

[18] 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.

[19] 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.

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

[21] J. Euzenat, P. Shvaiko, et al., Ontology matching, volume 18, Springer, 2007.

[22] N. F. Noy, M. A. Musen, PROMPT: algorithm and tool for automated ontology merging and alignment, in: H. A. Kautz, B. W. Porter (Eds.), Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on on Innovative Applications of Artificial Intelligence, July 30 - August 3, 2000, Austin, Texas, USA, AAAI Press / The MIT Press, 2000, pp. 450–455. URL: http://www.aaai.org/Library/AAAI/2000/aaai00-069.php.

[23] 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.

[24] 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.

[25] Y. Peng, M. Alam, T. Bonald, Ontology matching using textual class descriptions, in: P. Shvaiko, J. Euzenat, E. Jiménez-Ruiz, O. Hassanzadeh, C. Trojahn (Eds.), Proceedings of the 18th International Workshop on Ontology Matching co-located with the 22nd International Semantic Web Conference (ISWC 2023), Athens, Greece, November 7, 2023, volume 3591 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023, pp. 67–72. URL: https://ceur-ws.org/Vol-3591/om2023_STpaper2.pdf.

[26] G. Sousa, R. Lima, C. Trojahn, Combining word and sentence embeddings with alignment extension for property matching., in: OM@ ISWC, 2023, pp. 91–96.

[27] G. Sousa, R. Lima, C. Trojahn, Complex ontology matching with large language model embeddings, arXiv preprint arXiv:2502.13619 (2025).

[28] H. B. Giglou, J. D'Souza, F. Engel, S. Auer, Llms4om: Matching ontologies with large language models, arXiv preprint arXiv:2404.10317 (2024).

[29] E. Jiménez-Ruiz, B. C. Grau, Logmap: Logic-based and scalable ontology matching, in: L. Aroyo, C. Welty, H. Alani, J. Taylor, A. Bernstein, L. Kagal, N. F. Noy, E. Blomqvist (Eds.), The Semantic Web - ISWC 2011 - 10th International Semantic Web Conference, Bonn, Germany, October 23-27, 2011, Proceedings, Part I, volume 7031 of *Lecture Notes in Computer Science*, Springer, 2011, pp. 273–288.

URL: https://doi.org/10.1007/978-3-642-25073-6_18. doi:10.1007/978-3-642-25073-6_18.
[30] D. Faria, C. Pesquita, E. Santos, M. Palmonari, I. F. Cruz, F. M. Couto, The agreementmakerlight ontology matching system, in: On the Move to Meaningful Internet Systems: OTM 2013 Conferences: Confederated International Conferences: CoopIS, DOA-Trusted Cloud, and ODBASE 2013, Graz, Austria, September 9-13, 2013. Proceedings, Springer, 2013, pp. 527–541.
[31] OpenAI, Hello gpt-4o, https://openai.com/index/hello-gpt-4o/, 2024. Accessed: 2024-05-18.
[32] A. Algergawy, H. Hamed, B. König-Ries, Towards scientific data synthesis using deep learning and semantic web, in: The Semantic Web: ESWC 2021 Satellite Events: Virtual Event, June 6–10, 2021, Revised Selected Papers 18, Springer, 2021, pp. 54–59.
[33] A. Algergawy, H. Hamed, S. Thiel, B. König-Ries, Towards semantic annotation for scientific datasets, in: The Semantic Web: ESWC 2024 Satellite Events: May 26–30, 2024, ???? URL: https://api.semanticscholar.org/CorpusID:269758799.
[34] 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.
[35] 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.
[36] 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.
[37] 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.
[38] 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.
[39] 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.
[40] 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.
[41] J. Gu, X. Jiang, Z. Shi, H. Tan, X. Zhai, C. Xu, W. Li, Y. Shen, S. Ma, H. Liu, Y. Wang, J. Guo, A survey on llm-as-a-judge, CoRR abs/2411.15594 (2024). URL: https://doi.org/10.48550/arXiv.2411.15594. doi:10.48550/ARXIV.2411.15594. arXiv:2411.15594.
[42] J. Gu, X. Jiang, Z. Shi, H. Tan, X. Zhai, C. Xu, W. Li, Y. Shen, S. Ma, H. Liu, et al., A survey on llm-as-a-judge, arXiv preprint arXiv:2411.15594 (2024).
[43] 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.