Assessing the Capability of Large Language Models for Domain-Specific Ontology Generation

Anna Sofia Lippolis[1,2,*], Mohammad Javad Saeedizade[3,*], Robin Keskisärkkä[3], Aldo Gangemi[1,2], Eva Blomqvist[3] and Andrea Giovanni Nuzzolese[2]

[1]University of Bologna, Italy
[2]ISTC-CNR, Italy
[3]Linköping University, Sweden

Abstract

Large Language Models (LLMs) have shown significant potential for ontology engineering. However, it is still unclear to what extent they are applicable to the task of domain-specific ontology generation. In this study, we explore the application of LLMs for automated ontology generation and evaluate their performance across different domains. Specifically, we investigate the generalizability of two state-of-the-art LLMsDeepSeek and o1-preview, both equipped with reasoning capabilitiesby generating ontologies from a set of competency questions (CQs) and related user stories. Our experimental setup comprises six distinct domains carried out in existing ontology engineering projects and a total of 95 curated CQs designed to test the models reasoning for ontology engineering. Our findings show that with both LLMs, the performance of the experiments is remarkably consistent across all domains, indicating that these methods are capable of generalizing ontology generation tasks irrespective of the domain. These results highlight the potential of LLM-based approaches in achieving scalable and domain-agnostic ontology construction and lay the groundwork for further research into enhancing automated reasoning and knowledge representation techniques.

Keywords

ontology engineering, large language models, ontology generation,

1. Introduction

The rapid evolution of large language models (LLMs) has transformed numerous areas of natural language processing as well as knowledge engineering, and ontology engineering is no exception. Ontologiesformal representations of knowledge that enable semantic interoperabilityhave long been essential in domains ranging from healthcare to environmental science. However, traditional ontology engineering, as shown by requirement-based methods such as eXtreme Design [1], remains a labour-intensive, expertise-driven task, often hindering timely and scalable knowledge modelling. Recent studies, such as the ones by Saeedizade and Blomqvist [2] and Lippolis et al. [3, 4] suggest a potential for using LLMs in ontology engineering, where LLMs offer a promising addition by automating key aspects of ontology construction through their language understanding and reasoning capabilities.

Building on these earlier studies that showed the feasibility of LLM-supported ontology generation, this paper extends the investigation by suggesting an ontology generation prompt for reasoning models and examining the generalizability of automated ontology generation across diverse and more specific domains. Neither of these aspects were so far addressed in previous work. In particular, we explore the performance of two state-of-the-art LLMsDeepSeek and OpenAI o1-previewwhich are equipped with reasoning capabilities. By leveraging a dataset

comprising 95 carefully curated competency questions paired with corresponding user stories, from specific knowledge domains, our approach systematically tests whether these models can generate good ontology drafts based on requirements that capture domain-specific needs.

The contributions of our study are the following: First, we present an automated pipeline for ontology generation, including a prompt strategy, that harnesses the advanced reasoning capabilities of the most recent commercial and open-source LLMs to interpret natural language requirements and suggest corresponding ontological modules. Second, we evaluate this approach across six distinct knowledge domains, providing empirical evidence of its consistent performance and domain-agnostic potential[1].

The remainder of this paper is organized as follows. Section 2 surveys the relevant literature in LLM-based knowledge engineering. Section 3 establishes the preliminary definitions and concepts used in the remainder of the paper. In Section 4, we describe our methodology, including dataset creation and details about the prompt. Section 5 shows our experimental setup and evaluation framework. Section 6 presents and discusses our results, Section 7 addresses the discussion, and Section 8 focuses on limitations and potential risks. Finally, Section 9 concludes the paper and outlines directions for future research.

2. Related Work

Ontology generation based on requirements using large language models has been recently studied. In Benson et al.[5], the authors present a study on producing outputs that are consistent with BFO through the chat interface of ChatGPT (with GPT-4). However, this work was based on a few examples and was solely focused on LLMs with non-reasoning capabilities, noting reasoning models needed a new, ad-hoc analysis. In Saeedizade and Blomqvist [2], LLM-generated ontologies from a set of competency questions from a semantic web course were compared to student performance in that course. The domain of those ontology stories was limited to ontology stories designed for that specific course. The study by Lippolis et al. [3] on ontology generation was based on the African Wildlife Ontology, a gold standard outlined in Potoniec et al. [6]. Additionally, Doumanas et al. [7] fine-tuned LLMs to generate ontologies by considering the domain of the ontology. However, the comparison of its performance between different domains is not mentioned in the study. In the same way, the work by Lippolis et al. [4] addresses LLM-based ontology generation on a dataset with 100 CQs of different domains, but doesn't address the LLMs' performance on the involved domains and focus on three ontology homework from a semantic web course, the one presented in [2]. In this study, it is concluded that OpenAI's o1-preview model is the best performing one with respect to non-reasoning models. For what concerns studies about domain-specific applications with LLMs, Fathallah [8] focuses on ontology generation in life sciences, Huang et al. [9] in drinking water distribution network and Xu et al. [10] evaluate retrieval augmented generation for E-commerce and [11] on drug indication, assessing the capability of LLMs in generating knowledge graphs from text using predefined ontologies. Val-Calvo et al. propose a pipeline to automate ontology development and knowledge graph generation with GPT-4 from CSV files from the e-commerce domain to streamline the engineering process in enterprise settings [12]. From these studies, it is possible to conclude there is no systematic evaluation of one method across different ontology domains of varying complexity, especially with different LLMs with reasoning capabilities.

3. Preliminaries

In this section, we present and define terminology that will be used throughout this paper, building on the one already outlined in Lippolis et al. [4].

---

[1]Data, code and other material for this study is available at this link: https://github.com/dersuchendee/Domain-OntoGen

Ontology Generation: We define ontology generation as the process of creating formal representations of knowledge by identifying the necessary classes, properties, and axioms to capture domain-specific concepts and their relationships. This process can be performed manually or supported by automated techniques, such as LLMs, to draft ontologies from natural language inputs in the form of requirements.

Modelling Competency Question: Following the definitions outlined in previous work [4], a CQ is considered modelled in an ontology if and only if a SPARQL query exists to extract the answer of the CQ from the ontology irrespective of the quality of the ontology modelling or adherence to optimal modelling practices.

Minor issue: Similar to [4], for a pair ontology and a competency question, if the ontology includes all necessary elements (named classes or properties) except for only one object property or only one data property, and adding this single element would make the CQ modelled, this is considered a minor issue in modelling the CQ.

Reasoning Capability of LLMs: The reasoning capability of LLMs refers to their ability to generate logically structured and coherent responses by recognizing and using patterns learned from large amounts of text. This definition includes simulating various forms of reasoningsuch as deducing, generalizing from examples, and drawing analogiesoften by processing information step-by-step (a process sometimes called chain-of-thought, which is natively employed in models like o1-preview). Importantly, LLM reasoning is based on statistical patterns rather than true human understanding or deliberate logic.

4. Methodology

In this section, we outline the methodology employed in this study. We begin by describing the dataset creation process and its statistical properties, followed by the explanation of the prompting technique used in this work. Finally, we present the pipeline for ontology generation.

4.1. Dataset Creation

To evaluate ontologies generated by LLMs with reasoning capabilities, we constructed a dataset comprising 95 competency questions (CQs) with 12 ontology stories across 6 distinct domains. Each dataset entry includes: (i) a CQ, (ii) a user story (providing the contextual background), (iii) the binary label of the difficulty level of the CQ (easy or hard), and (iv) the domain of the ontology. Table 1 shows statistics for the dataset.

4.1.1. Dataset composition

The CQs were extracted from four sources: Onto-DESIDE [2] (domain: Circular Economy), the Polifonia Project (domains: Music and Events)[13], the WHOW Project (domain: Water and Health) [14] and AquaDiva (domains: Microbe Habitat and Carbon and Nitrogen Cycling)[15]. In particular, these derive from carefully curated large-scale projects, with the fourth one being a semi-automatically annotated dataset.

4.1.2. CQ classification

To ensure balanced and comparable evaluations, each CQ was classified into one of two categories, "Easy" and "Hard", based on the complexity required for its formal representation. An LLM, o1-preview, was used to assign these labels by estimating the difficulty level of classes and properties needed to model the CQ. Specifically, if a CQ required at most 2 classes and 1 property (either a data or an object property), it was classified as "Easy"; otherwise, it was classified as "Hard". The labels were then checked manually by an ontology engineer. This

---

[2]https://ontodeside.eu

labelling strategy was implemented to maintain a consistent distribution of CQ difficulties across all domains.

Table 1
Domain Statistics for Competency Questions.

Domain	Total CQs	Easy	Hard
Circular Economy	16	5	11
Music (Bells, Organs, Tunes)	16	10	6
Events	18	8	10
Microbe Habitat	15	8	7
Carbon and Nitrogen Cycling	15	4	11
Water and Health	15	7	8
Total	95	42	53

4.1.3. Manual annotations

The labelling of the dataset for the verification of CQs has been carried out manually by two ontology engineers. Following the error classification outlined in previous work [4], the output ontologies were considered to be modelled, not modelled, or not modelled due to a minor error.

4.2. Prompting Techniques

A prompt serves as an instruction for LLMs to generate responses to a given input. The prompt employed in this study follows a few-shot prompting technique, with some examples, building on and distinguishing it from previous approaches which mainly employed a decomposed prompting technique [3, 4, 2]. This choice is motivated by the observation that reasoning models have been shown in Lippolis et al. [4] to perform more effectively with minimal instructions rather than being explicitly guided to think step by step. After trials with and without a few-shot example, the performance of the few-shot has proven more accurate than the other. The prompt is structured into three key components:

- Instruction: A description of the task (ontology generation) and its objective (ontology must model the input CQ).
- Example: A Turtle syntax sample ontology model corresponding to a CQ and its associated ontology story, illustrating an ideal response.
- Input: A designated section where a new CQ and ontology story are inserted, prompting the LLM to generate an ontology model accordingly.

4.3. Ontology Generation

We used the Independent Ontology Generation method described in [4]. In this approach, each CQ and its associated ontology story are provided to an LLM through a prompt to generate the corresponding ontology. This method treats each CQ as an independent unit, enabling a precise evaluation of the ontology generation process on a per-question level. By handling CQs in isolation, the impact of the settings can be assessed.

5. Experiment Setup

In this section, we describe the experimental setup. To run the experiments, we used Azure API to call o1-preview and the original DeepSeek API with default hyperparameters. The prompt

with a fixed example (few-shot) is appended with ontology story and CQ and sent to the API to get the resulting output file. The results are saved in a Turtle file.

LLMs: In the present study, we have selected two large language models with reasoning capabilities: DeepSeek and OpenAI o1-preview. This selection was motivated by the desire to perform a comparative analysis between the current state-of-the-art closed-source and open-source LLMs. The OpenAI o1-preview model was accessed through the Microsoft Azure API, whereas DeepSeek was used according to the specifications detailed in its official documentation[3].

Prompt: The prompt used in this work is available in the supplementary materials on GitHub[4]. Initially, several prompt configurations were evaluated, including versions with and without examples (i.e., few-shot and zero-shot settings). The results demonstrated that prompts incorporating examples consistently outperformed those without, justifying their use. Furthermore, the impact of varying the complexity of the examples was examined in a separate experiment involving multiple ontology generation tasks. In the interest of conciseness, the detailed results of these additional experiments are not reported here.

Hyperparameters: The default hyperparameters for the LLMs, including temperature and penalty, were employed throughout this study.

6. Results

Overall, both the o1-preview and DeepSeek models exhibited high and comparable accuracy. Initially, nine and ten CQs remained unmodeled for the o1-preview and DeepSeek models, respectively, out of a total of 95 CQs. However, when minor modelling issues are excluded, the number of unmodeled CQs decreases to eight for the o1-preview model and five for the DeepSeek model.

Table 2 provides a breakdown of mistakes on both easy and hard CQs by domain. Regarding the o1-preview scores, performance was consistent across all domains, suggesting that the ontology generation method is broadly applicable rather than being limited to a specific domain. With slight differences, although DeepSeek generally demonstrated similar consistency, the scores for the Events domain were noticeably lower, indicating a need for further investigation into the model's performance in that particular domain.

Table 2
Count of errors in modelling for each LLM, divided into Easy and Hard competency questions grouped by ontology domain

Category	o1-preview					DeepSeek				
	Not modelled		Contains a minor issue		Total	Not modelled		Contains a minor issue		Total
	Easy	Hard	Easy	Hard		Easy	Hard	Easy	Hard	
Circular Economy	0	0	0	2	2	0	0	0	0	0
Music	0	1	1	0	2	2	0	0	0	2
Events	0	0	1	1	2	1	2	0	2	5
Microbe Habitat	0	0	0	0	0	0	0	1	0	1
Carbon and Nitrogen Cycling	0	0	0	1	1	0	0	0	1	1
Water and Health	0	0	0	2	2	0	0	1	0	1

7. Discussion

Our results reveal a performance trade-off between the two models. Specifically, while o1-preview tends to commit more "minor" errors, DeepSeek is prone to a higher incidence of incomplete modellings. When we consider cases with "minor" mistakes equivalent to complete modellings, both models demonstrate similar performance across the domains. This finding suggests that differences in error types may not translate into significant disparities in overall outcomes.

---

[3]https://api-docs.deepseek.com/
[4]https://github.com/dersuchendee/Domain-OntoGen

Domain generalizability. Our work extends previous findings by Lippolis et al. [4] by showing that reasoning models working for a specific domain can generalize effectively across other datasets. Therefore, the improvements achieved in earlier studies are not limited to a single domain, affirming the broader applicability of the developed techniques.

Reasoning models architecture. One possible explanation for the similar outputs observed from both DeepSeek and o1-preview is their reasoning capabilities. It is plausible that these models benefit from mutual fine-tuning or shared architectural strengths. The collective evidence points to a significant enhancement in reasoning with respect to models with non-reasoning capabilities, which appears to underlie their comparable performance.

Difficulty levels of the requirements. Furthermore, the analysis indicates that both models handle hard and easy CQs with similar performance. This observation challenges the common assumption that increased question complexity necessarily results in lower scores. Instead, our findings suggest that the models are robust enough to process questions of varying complexity without a marked performance penalty. Nonetheless, this is limited to only one evaluation criterion.

Error analysis. Finally, a closer look at the errors reveals that the majority occur solely within the Events domain. The corresponding CQs in this category are much shorter than those in other domains, but there are multiple user stories that serve as requirements. This pattern could suggest that longer, more descriptive CQs could provide additional context that aids the models in processing and reasoning, leading to improved performance, especially if multiple scenarios are provided. Therefore, the lack of standardization in requirements is also fundamental when dealing with automated techniques.

Overall, these insights highlight the importance of both model architecture and the design and selection of requirements in achieving balanced and generalizable performance.

8. Limitations and Risks

One potential risk associated with this study is dataset leakage, which may impact the quality and reliability of the LLM-generated ontology. Specifically, in certain domains, an LLM may have encountered parts of the dataset during pretraining, while in others, it relies solely on its reasoning capabilities. To the best of our knowledge, the CQs and ontology modules were not publicly available in a single dataset, reducing the likelihood that LLMs were trained on these elements in direct succession as consecutive tokens. Another limitation of this work is the exclusive use of reasoning-based LLMs. These models, while beneficial for ontology generation tasks, are generally resource-intensive and less widely available compared to standard LLMs. Additionally, the scope of this study is restricted to only two LLMs, which may limit the generalizability of the findings.

9. Conclusion

In this study, we investigated the potential of two LLMs, DeepSeek and o1-preview, with reasoning capabilities for automated ontology generation across six diverse domains. Our experiments, conducted on a dataset of 95 competency questions paired with user stories, reveal that both LLMs similarly exhibit robust and consistent performance across all domains and difficulty levels of the CQs. Overall, this work lays a foundation for further advancements in automated ontology engineering, highlighting the potential of LLMs to deliver scalable, domain-agnostic knowledge representation. Future research will build on these results to explore additional LLMs.

Acknowledgments

Disclosure of Interests

The authors have no competing interests to declare that are relevant to the content of this article.

Declaration on Generative AI

During the preparation of this work, the author(s) used GPT-4 and Grammarly in order to: Grammar and spelling check. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publications content.

References

[1] V. Presutti, E. Daga, A. Gangemi, E. Blomqvist, extreme design with content ontology design patterns, in: Proc. Workshop on Ontology Patterns, CEUR-WS, 2009, pp. 83–97.

[2] M. J. Saeedizade, E. Blomqvist, Navigating ontology development with large language models, in: European Semantic Web Conference, Springer, 2024, pp. 143–161.

[3] A. S. Lippolis, M. Ceriani, S. Zuppiroli, A. G. Nuzzolese, Ontogenia: Ontology generation with metacognitive prompting in large language models, in: European Semantic Web Conference, Springer, 2024, pp. 259–265.

[4] A. S. Lippolis, M. J. Saeedizade, R. Keskisärkkä, S. Zuppiroli, M. Ceriani, A. Gangemi, E. Blomqvist, A. G. Nuzzolese, Ontology generation using large language models, arXiv preprint arXiv:2503.05388 (2025).

[5] C. Benson, A. Sculley, A. Liebers, J. Beverley, My ontologist: Evaluating bfo-based ai for definition support, arXiv preprint arXiv:2407.17657 (2024).

[6] J. Potoniec, D. Wiśniewski, A. Ławrynowicz, C. M. Keet, Dataset of ontology competency questions to sparql-owl queries translations, Data in brief 29 (2020) 105098.

[7] D. Doumanas, A. Soularidis, D. Spiliotopoulos, C. Vassilakis, K. Kotis, Fine-tuning large language models for ontology engineering: A comparative analysis of gpt-4 and mistral, Applied Sciences 15 (2025) 2146.

[8] N. Fathallah, S. Staab, A. Algergawy, Llms4life: Large language models for ontology learning in life sciences, arXiv preprint arXiv:2412.02035 (2024).

[9] Y. Huang, E. Karabulut, V. Degeler, Large language model for ontology learning in drinking water distribution network domain (2024).

[10] Z. Xu, Z. Shen, Q. Zhou, P. Ristoski, Knowledge graph-enhanced retrieval augmented generation for e-commerce (2024).

[11] R. Alharbi, U. Ahmed, D. Dobriy, W. ajewska, L. Menotti, M. J. Saeedizade, M. Dumontier, Exploring the role of generative ai in constructing knowledge graphs for drug indications

with medical context, in: 15th International Semantic Web Applications and Tools for Healthcare and Life Sciences (SWAT4HCLS 2024) (to appear), 2024.

[12] M. Val-Calvo, M. Egaña Aranguren, J. Mulero-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, Information Processing & Management 62 (2025) 104042. URL: https://www.sciencedirect.com/science/article/pii/S0306457324004011. doi:https://doi.org/10.1016/j.ipm.2024.104042.

[13] J. de Berardinis, V. A. Carriero, N. Jain, N. Lazzari, A. Meroño-Peñuela, A. Poltronieri, V. Presutti, The polifonia ontology network: Building a semantic backbone for musical heritage, in: International Semantic Web Conference, Springer, 2023, pp. 302–322.

[14] A. S. Lippolis, G. Lodi, A. G. Nuzzolese, The water health open knowledge graph, Scientific Data 12 (2025) 274.

[15] A. Algergawy, H. Hamed, S. Thiel, B. König-Ries, Towards semantic annotation for scientific datasets, in: European Semantic Web Conference, Springer, 2024, pp. 164–167.