

# Benchmarking LLM-based Ontology Conceptualization: A Proposal

Youssra Rebboud<sup>1,\*</sup>, Pasquale Lisena<sup>1</sup>, Lionel Tailhardat<sup>1,2</sup> and Raphael Troncy<sup>1</sup>

<sup>1</sup>EURECOM, Sophia Antipolis, France

<sup>2</sup>Orange, Paris, France

## Abstract

This study presents a benchmark proposal designed to enhance knowledge engineering tasks through the use of large language models (LLMs). As LLMs become increasingly pivotal in knowledge extraction and modeling, it is crucial to evaluate and improve their performance. Building on prior work aiming at reverse generating competency questions (CQs) from existing ontologies, we introduce a benchmark focused on specific knowledge modeling tasks including ontology documentation, ontology generation, and query generation. In addition, we propose a baseline evaluation framework that applies various techniques, such as semantic comparison, ontology evaluation criteria, and structural comparison, using both existing ground truth datasets and newly proposed ontologies with corresponding CQs and documentation. This rigorous evaluation aims to provide a deeper understanding of LLM capabilities and contribute to their optimization in knowledge engineering applications.

## Keywords

Benchmark Proposal, Knowledge Engineering, Knowledge Representation, Large Language Models

## 1. Introduction

The knowledge engineering and semantic web communities are increasingly experimenting with Large Language Models (LLMs) to build ontologies and knowledge graphs. Key tasks being explored include: creating views on heterogeneous data lakes [1], RDF triples and SPARQL query generation [2], named entity recognition and relation extraction [3], RML mappings creation [4] or schema and ontology matching [5, 6, 7]. Hence, we observe that the various stages of the knowledge engineering process are revisited in the era of LLMs (e.g. LOT [8]). However, their systematic usage need to be further assessed as the results greatly vary depending on the underlying LLM being used and other factors.

In previous work [9], we have evaluated six LLMs using zero- and few-shot approaches with three prompting strategies, inputting either classes alone, classes with properties, or a schema summary.<sup>1</sup> These configurations were tested across five ontologies to assess the LLMs' ability to reverse generate Competency Questions (CQs). These ontologies were precisely selected because expert made competency questions having lead to their conceptualization were made available. We observed that while providing competency question examples generally improved performance for this task, in some cases, adding more detailed information from certain ontologies unexpectedly reduced LLM effectiveness. This highlights the need to further investigate the characteristics of the ontologies that impact the accuracy of LLM responses and vice versa.

In this paper, we propose to develop a benchmark to systematically compare the performance of LLMs for knowledge engineering tasks, specifically focusing on the stages of specification, conceptualization, and validation of an ontology. The core of our proposal is to leverage ontologies that have been published alongside a set of CQs and have been evaluated through the corresponding authoring tests expressed in SPARQL. The goal of the LLM will typically be to understand user intents expressed in natural language

ISWC 2024 Special Session on Harmonising Generative AI and Semantic Web Technologies, November 13, 2024, Baltimore, Maryland

\*Corresponding author.

✉ youssra.rebboud@eurecom.fr (Y. Rebboud); pasquale.lisena@eurecom.fr (P. Lisena); lionel.tailhardat@orange.com (L. Tailhardat); raphael.troncy@eurecom.fr (R. Troncy)

ORCID 0000-0003-3507-5646 (Y. Rebboud); 0000-0003-3094-5585 (P. Lisena); 0000-0001-5887-899X (L. Tailhardat); 0000-0003-0457-1436 (R. Troncy)



© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

<sup>1</sup>The code is available at <https://github.com/D2KLab/llm4ke>

regarding a given domain of discourse and to produce the axiomatization of that domain in an ontology. Typical knowledge engineering tasks in the scope of the proposal are: 1) conceptualize an ontology from CQs; 2) generate CQs from an ontology (e.g. for evaluating its completeness or discovering new usages); 3) produce the documentation of an ontology; 4) implement queries corresponding to CQs; 5) verbalize knowledge graph excerpts using instance data and an ontology.

Beyond providing a comparison of the performance of LLMs to perform these tasks, we hypothesize that the implementation of this benchmark will yield the necessary insights to understand to what extent does the language bias of a LLM [10] influence its interpretation of the axiomatic structure of a domain of discourse represented in an ontology, notably by exploring the role of competency question formulation and the influence of object properties defined in ontologies, including their names, descriptions, and associated logic.

## 2. Towards a More Comprehensive Benchmark

In this section, we further detail the elements of this benchmark: the use cases and tasks to be evaluated (Section 2.1), the curated datasets (Section 2.2), and the evaluation metrics for improving LLM performance (Section 2.3).

### 2.1. Benchmark Tasks

**Conceptualize an ontology.** This task aims to conceptualize an ontology given a set of competency questions and a domain scope. A variant of it can start from a partial ontology, and the task of the LLM is to complete the ontology by adding the missing classes and properties. In this case, we will evaluate the LLM’s performance in accurately completing an ontology based solely on the upper-level structure. This will help understanding the depth to which the LLM can reach with and without CQs that could assist in clarifying user intent.

**Generate Competency Questions.** This task focuses on generating competency questions given an ontology or specific parts of it, similar to the approach outlined in [9]. The performance of LLMs will be evaluated according to the prompts being used and the nature of the information provided, such as a general description of an ontology, taxonomical branches, or even the entire ontology. The impact of each type of input will be analyzed in order to assess the generalization capabilities of the model and its limitations related to the domain, the structure or the size of the ontology. Additionally, this task aims to identify best practices for prompt formatting to enhance communication between human language and LLMs.

**Produce the ontology documentation.** In this task, the LLM should produce a human-readable documentation of the ontology, emphasizing on its main classes and properties. This can be further expanded in generating useful API calls, e.g. following the SPARQL Transformers approach [11]. This can be achieved by either inputting the entire ontology or using a chain-of-thought (CoT) [12] approach, in case of complex ontologies.

**Implement queries corresponding to CQs.** This task will assess the LLM’s ability to generate autonomously relevant queries given the ontology structure and the user intent expressed with competency questions.

**Verbalize Knowledge Graph Excerpts.** The task involves generating human-readable interpretations of a dataset, using an ontology as a guide for the LLM to structure the information. The goal is to go beyond the verbalization of all possible triples (subject, verb, object) and to generate paragraphs that summarize the graphs.

## 2.2. Datasets

We propose to consider the following criteria to include ontologies in the benchmark: the availability of competency questions that have lead to their conceptualization, comprehensive documentation of the ontology, FAIR-related scores attached to the ontology, SPARQL queries demonstrating the usage of the ontology and/or SHACL [13] shapes constraining its usage. Furthermore, we advocate for a selection that covers diverse domain of discourses. From a conceptualization point of view, the selected ontologies should exhibit different structure (e.g. relatively flat ontologies such as [schema.org](https://schema.org/)<sup>2</sup> versus nested models such as FRBR [14]) and make use of well-known ontology design patterns (e.g. event-based modeling like CIDOC-CRM<sup>3</sup>). Although not all ontologies provide the requisite information for every criteria, they should still be suitable for evaluation in specific tasks.

**Domain of Discourse.** The benchmark should cover as many domains as possible to be able to draw the line between the performance of LLMs and a given domain. Examples include: general purpose ontologies, IT and sensors, creative industries and medias, cultural heritage and museums, healthcare and medicine, biology and life sciences, education and e-learning, e-commerce and retail, finance and banking, and legal sectors.

**Candidate ontologies.** The following ontologies fulfill the selection criteria discussed above:

- **DOREMUS** [15]: related to **music** and **cultural heritage** domains, the ontology comes with a documentation, competency questions, SPARQL queries and APIs and a large knowledge graph.
- **Polifonia** [16]: related to **music** and **cultural heritage**, the ontology enables to capture musical and historical knowledge. In addition to CQs and a comprehensive documentation, the authors provides a set of queries that we can leverage as ground truth and a knowledge graph.
- **DemCare** [17]: related to the **medical** domain, specifically tailored for dementia care and monitoring. Provided with CQs, a dataset, and well-structured documentation.
- **Odeuropa** [18]: related to **sensory experiences** and **cultural heritage**, focusing on olfactory experiences in historical contexts. Provided with CQs, documentation, queries and a dataset.
- **NORIA-O** [19]: related to the **IT** domain, and designed for network monitoring and performing **anomaly detection**. Provided with CQs, a documentation, queries, and a knowledge graph.
- **FIBO** [20]: related to the **financial** domain, this ontology is composed of numerous very specialized terms and comes with a set of SPARQL queries aimed at testing its compliance.<sup>4</sup>

We plan to incorporate additional ontologies from other domains using well-known catalogs such as Linked Open Vocabulary (LOV) [21] and the Industry Portal [22]. We can also rely on [23] which provides a dataset of competency questions for different ontologies and domains (e.g. the African wildlife ontology (AWO) [24], the Software Ontology (SWO) [25], the Generic Ontology of Datatypes (OntoDT) [26]) together with their translation into SPARQL queries.

## 2.3. Evaluation Metrics and Process

In this section, we outline the process for evaluating and improving LLM performance using a factorial experiment design based on the following factors:

- 1) *Prompting strategies*: presence or absence of partial knowledge of the competency questions, taxonomy, and documentation in the LLM’s context, depending on the specific task similar to an ablation study.

---

<sup>2</sup><https://schema.org/>

<sup>3</sup><https://cidoc-crm.org/>

<sup>4</sup><https://shorturl.at/rwAPc>

- 2) *Data instance*: presence or absence of instances from a knowledge graph structured by a given ontology to guide the LLM.

The evaluation process will be iterative, refining the integration of the knowledge graph with the LLM at each step and assessing performance improvements along the way. Multiple iterations will be conducted and the results will be analyzed using statistical methods to quantify progress. To facilitate comparison between different models or methods, we propose to rely on a CI/CD-enabled pipeline based on the tools developed in [9], with performance results tracked using a leaderboard.

Table 1 summarizes the evaluation techniques intended for each of the benchmark tasks (Section 2.1).

Task	Evaluation Techniques
<b>Conceptualize an ontology</b>	Ontology Evaluation Criteria, Logical Consistency.
<b>Generate Competency Questions</b>	Semantic Similarity.
<b>Produce the ontology documentation</b>	Semantic Similarity, between the generated documentation and the existing definition of classes and properties).
<b>Implement queries corresponding to CQs</b>	Structure Comparison.
<b>Verbalize Knowledge Graph Excerpts</b>	Fluency and Coherence.

**Table 1**  
Benchmark Tasks and Evaluation Techniques

- **Semantic Similarity.** This is typically implemented as a cosine similarity between vectors embedding a ground truth sentence and a generated response from a LLM. SentenceBERT [27] is generally used for evaluating the CQ generation and ontology documentation tasks.
- **Ontology Evaluation Criteria.** Using an existing ontology as the gold standard, we can assess the *accuracy*, *completeness*, and *conciseness* of the generated ontology [28]. This serves the tasks of ontology generation and ontology enrichment. However, *adaptability*, *clarity*, and *computational efficiency* are not addressed in this research, as they depend on the ground truth ontology.
- **Logical Consistency.** This enables us to validate the semantic formalization of an ontology, typically using tools such as the Hermit reasoner [29].
- **Structure Comparison.** When evaluating the generation of queries, this measure will compare the structure of the generated query with a ground truth query. We can leverage the RTED algorithm, which calculates the Tree Edit Distance (TED) [30] for this purpose.
- **Fluency and Coherence.** When verbalizing and summarizing knowledge graph excerpts (instance data guided by an ontology), this metric will assess the fluency (grammatical correctness) and adequacy (referring to the accurate integration of triples [31]) of the generated text.

### 3. Conclusion and Future Work

In this work, we propose a comprehensive benchmark for knowledge engineering tasks utilizing large language models (LLMs), specifically focusing on the knowledge conceptualization aspect. Building on our previous research [9], which explored the use of LLMs for reverse generating competency questions (CQs) from existing ontologies, we aim to expand the benchmark to encompass additional knowledge engineering tasks, including ontology and query generation, ontology documentation and enrichment, as well as knowledge graph verbalization. It is important to note that the set of tasks we propose is not finite, and there are significant opportunities for extending this benchmark to accommodate evolving challenges and new developments in the field.

Furthermore, we advocate for the inclusion of a broader range of ontologies, extending beyond our initial focus on the cultural heritage, education, network operations, and medical domains. By incorporating ontologies from other sectors, we aim to establish a robust foundation for comparing LLM

performance across diverse domains of discourse. This expansion would enhance our understanding of LLM capabilities and facilitate their fine-tuning within the context of knowledge engineering. Additionally, we propose a baseline evaluation framework for the various tasks, which includes semantic comparisons, ontology evaluation criteria, logical consistency, and structure comparison. We detail the evaluation technique for each task. Moreover, the diverse range of proposed datasets will provide valuable insights into how LLM performance correlates with the size, complexity, and domain of the ontology.

## Acknowledgements

This work has been partially supported by the French National Research Agency (ANR) within the kFLOW project (Grant n°ANR-21-CE23-0028).

## References

- [1] S. Arora, B. Yang, S. Eyuboglu, A. Narayan, A. Hojel, I. Trummer, C. Ré, Language Models Enable Simple Systems for Generating Structured Views of Heterogeneous Data Lakes, *VLDB Endowment* 17 (2023) 92–105. doi:10.14778/3626292.3626294.
- [2] J. Frey, L.-P. Meyer, N. Arndt, F. Brei, K. Bulert, Benchmarking the Abilities of Large Language Models for RDF Knowledge Graph Creation and Comprehension: How Well Do LLMs Speak Turtle?, in: *Workshop on Deep Learning for Knowledge Graphs (DL4KG)*, 2023. URL: <https://ceur-ws.org/Vol-3559/paper-3.pdf>.
- [3] J. Wang, Y. Chang, Z. Li, N. An, Q. Ma, L. Hei, H. Luo, Y. Lu, F. Ren, TechGPT-2.0: A large language model project to solve the task of knowledge graph construction, 2024. arXiv:2401.04507.
- [4] M. Hofer, J. Frey, E. Rahm, Towards self-configuring Knowledge Graph Construction Pipelines using LLMs – A Case Study with RML, in: *5<sup>th</sup> International Workshop on Knowledge Graph Construction*, 2024. URL: <https://ceur-ws.org/Vol-3718/paper6.pdf>.
- [5] S. Hertling, H. Paulheim, OLaLa: Ontology Matching with Large Language Models, in: *12<sup>th</sup> Knowledge Capture Conference (KCAP)*, Association for Computing Machinery, 2023. doi:10.1145/3587259.3627571.
- [6] H. B. Giglou, J. D’Souza, F. Engel, S. Auer, LLMs4OM: Matching Ontologies with Large Language Models, in: *21<sup>st</sup> Extended Semantic Web Conference (ESWC)*, Special Track on Large Language Models for Knowledge Engineering, 2024.
- [7] B. P. Allen, P. T. Groth, Evaluating Class Membership Relations in Knowledge Graphs using Large Language Models, in: *21<sup>st</sup> Extended Semantic Web Conference (ESWC)*, Special Track on Large Language Models for Knowledge Engineering, 2024.
- [8] M. Poveda-Villalón, A. Fernández-Izquierdo, M. Fernández-López, R. García-Castro, LOT: An industrial oriented ontology engineering framework. *Engineering Applications of Artificial Intelligence*, Engineering Applications of Artificial Intelligence 111 (2022). doi:10.1016/j.engappai.2022.104755.
- [9] Y. Rebboud, L. Tailhardat, P. Lisena, R. Troncy, Can LLMs generate competency questions?, in: *21<sup>st</sup> Extended Semantic Web Conference (ESWC)*, Special Track on Large Language Models for Knowledge Engineering, 2024.
- [10] Queenie Luo, Michael J. Puett, Michael D. Smith, A Perspectival Mirror of the Elephant, *Communications of the ACM* 67 (2024) 98–105. doi:10.1145/3670241.
- [11] P. Lisena, A. Meroño-Peñuela, T. Kuhn, R. Troncy, Easy Web API Development with SPARQL Transformer, in: *Proceedings of the 18<sup>th</sup> International Semantic Web Conference (ISWC)*, Semantic Web Science Association (SWSA), Auckland, New Zealand, 2019. doi:10.1007/978-3-030-30796-7\_28.
- [12] J. Wei, X. Wang, D. Schuurmans, M. Bosma, E. H. Hsin Chi, F. Xia, Q. Le, D. Zhou, Chain of thought prompting elicits reasoning in large language models, *ArXiv abs/2201.11903* (2022).



- [13] H. Knublauch, D. Kontokostas, Shapes Constraint Language (SHACL), Technical Report, W3C, 2017. URL: <https://www.w3.org/TR/shacl/>.
- [14] IFLA Study Group on the Functional Requirements for Bibliographic Records, Functional Requirements for Bibliographic Records: Final Report, Technical Report, International Federation of Library Associations and Institutions (IFLA), 2009. URL: <https://repository.ifla.org/handle/20.500.14598/811>.
- [15] M. Achichi, P. Lisena, K. Todorov, R. Troncy, J. Delahousse, DOREMUS: A Graph of Linked Musical Works, in: ISWC 2018 - 17th International Semantic Web Conference, volume LNCS of *The Semantic Web – Part II*, Springer, Monterey, CA, United States, 2018, pp. 3–19. doi:10.1007/978-3-030-00668-6\_1.
- [16] 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: *The Semantic Web – ISWC*, 2023. doi:10.1007/978-3-031-47243-5\_17.
- [17] I. Kompatsiaris, Dementia Ambient Care: Multi-Sensing Monitoring for Intelligent Remote Management and Decision Support, <https://demcare.eu/>, 2012.
- [18] P. Lisena, D. Schwabe, M. van Erp, R. Troncy, W. Tullett, I. Leemans, L. Marx, S. C. Ehrich, Capturing the Semantics of Smell: The Odeuropa Data Model for Olfactory Heritage Information, in: *The Semantic Web*, Springer International Publishing, 2022, pp. 387–405.
- [19] L. Tailhardat, Y. Chabot, R. Troncy, NORIA-O: an Ontology for Anomaly Detection and Incident Management in ICT Systems, in: 21<sup>st</sup> International Conference (ESWC), 2024. doi:10.1007/978-3-031-60635-9\_2.
- [20] M. Bennett, The financial industry business ontology: Best practice for big data, *Journal of Banking Regulation* 14 (2013) 255–268.
- [21] I. Stavrakantonakis, A. Fensel, D. Fensel, Linked Open Vocabulary Ranking and Terms Discovery, in: 12th International Conference on Semantic Systems, Association for Computing Machinery, 2016, pp. 1–8. doi:10.1145/2993318.2993338.
- [22] E. Amdouni, A. Sarkar, C. Jonquet, M. H. Karray, IndustryPortal: a Common Repository for FAIR Ontologies in Industry 4.0, in: 22nd International Semantic Web Conference (ISWC), Poster and Demo Tracj, Athens, Greece, 2023. URL: <https://hal.science/hal-04207343>.
- [23] J. Potoniec, D. Wiśniewski, A. Ławrynowicz, C. M. Keet, Dataset of ontology competency questions to SPARQL-OWL queries translations, *Data in Brief* 29 (2020). URL: <https://www.sciencedirect.com/science/article/pii/S2352340919314544>.
- [24] C. M. Keet, The African wildlife ontology tutorial ontologies, *Journal of Biomedical Semantics* 11 (2020). doi:10.1186/s13326-020-00224-y.
- [25] J. Malone, A. Brown, A. L. Lister, et al., The Software Ontology (SWO): a resource for reproducibility in biomedical data analysis, curation and digital preservation, *Journal of Biomedical Semantics* 5 (2014). doi:10.1186/2041-1480-5-25.
- [26] P. Panov, L. N. Soldatova, S. Džeroski, Generic ontology of datatypes, *Information Sciences* 329 (2016) 900–920. doi:10.1016/j.ins.2015.08.006.
- [27] N. Reimers, I. Gurevych, Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, in: *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, 2019.
- [28] J. Raad, C. Cruz, A Survey on Ontology Evaluation Methods, in: 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, Lisbonne, Portugal, 2015. doi:10.5220/0005591001790186.
- [29] R. Shearer, B. Motik, I. Horrocks, HermiT: A Highly-Efficient OWL Reasoner, in: *OWL: Experiences and Directions*, 2008. URL: <https://api.semanticscholar.org/CorpusID:7951194>.
- [30] M. Pawlik, N. Augsten, RTED: a robust algorithm for the tree edit distance, *Vldb Endowment* 5 (2011) 334–345. doi:10.14778/2095686.2095692.
- [31] P. Ke, H. Ji, Y. Ran, X. Cui, L. Wang, L. Song, X. Zhu, M. Huang, JointGT: Graph-Text Joint Representation Learning for Text Generation from Knowledge Graphs, in: *Findings of the Association for Computational Linguistics: ACL-IJCNLP*, Association for Computational Linguistics, 2021, pp. 2526–2538. URL: <https://aclanthology.org/2021.findings-acl.223>.