

# Benchmarking Ontology Validation Capabilities of LLMs

Stefani Tsaneva\*, Guntur Budi Herwanto and Marta Sabou

*Institute of Data, Process and Knowledge Management, Vienna University of Economics and Business, Austria*

## Abstract

With the advent of Generative AI, numerous approaches exploring large language models (LLMs) have been proposed for addressing a number of Knowledge Engineering (KE) tasks. Yet, the status of this research field is rather preliminary and there is, for now, no systematic and comprehensive understanding on how LLMs perform on selected knowledge engineering tasks (e.g., what is their *expertise level* in understanding ontology modeling concepts). Such insights would be crucial for researchers working in this field to support with selecting the most suitable LLMs during experiment design. This situation is exacerbated by the rapid expansion in the number of available LLMs. We therefore see the need for methodologies and tools that allow (comparatively) assessing LLM capabilities. To address this need, we propose the creation of an assessment test benchmark for evaluating the LLM knowledge engineering skills. We present ongoing work and preliminary results on assessing the expertise of LLMs in terms of a concrete KE task, namely ontology validation. Our experiments highlight the superiority of proprietary models on this task, particularly GPT-4o and Claude-Sonnet-3.5, over open source models. Lastly, we identify the need of a community-driven comparative LLM assessment platform that facilitates resource sharing and experience exchange, while protecting the integrity and privacy of the envisioned benchmark. We share (i) the current version of the qualification tests and (ii) its implementation for assessing LLM capabilities for ontology validation.

## Keywords

knowledge engineering, ontology validation, LLMs, expertise evaluation, assessment tests

## 1. Context and Research Need

*Context.* Advances in Generative AI, and specifically large language models (LLMs), offer many opportunities for enhancing Knowledge Engineering (KE) activities [1, 2, 3]. Numerous LLM-based solutions have already been successfully implemented for KE tasks. For instance, the construction and completion of knowledge graphs (KGs) have gained considerable research attention [4, 5, 6, 7]. Several approaches have also been proposed supporting the evaluation of semantic resources (i.e., KGs, ontologies, etc.): ontology requirements compliance has been approached through LLM-powered competency questions validation [8] and coverage testing [9]; KG triple validation with LLMs has been explored in [10, 11]; and ontology modeling error detection and correction through LLMs have been proposed in [12] and [13].

As research on the use of LLMs in KE is intensifying, it is essential to investigate whether LLM-produced results can be replicated with other models or are heavily dependent on the used LLM. Thus, it is important to test each approach with several LLMs, with different characteristics. Yet, LLMs are being published at a staggering rate. In this situation, it is increasingly challenging to decide which LLM to choose for which task, and to understand the “expertise” level of each LLM relevant for a given task. Having an understanding of which LLM is capable of which KE task and to what level, would enable adaptations of LLM-based solutions across different use cases with various data privacy needs and available budget. Moreover, it would allow the construction of more complex workflows, involving several LLMs, each responsible for a particular KE task.

*Research Need.* We therefore see a research need for the (comparative) assessment of LLM capabilities in terms of performing a variety of KE tasks. This requires the community to develop a collection of qualification tests as an instrument to assess the expertise level of LLMs in various KE tasks. For

---

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

\*Corresponding author.

✉ stefani.tsaneva@wu.ac.at (S. Tsaneva); guntur.budi.herwanto@wu.ac.at (G. B. Herwanto); marta.sabou@wu.ac.at (M. Sabou)

ORCID 0000-0002-0895-6379 (S. Tsaneva); 0000-0003-0250-6884 (G. B. Herwanto); 0000-0001-9301-8418 (M. Sabou)



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

example, in an ontology validation scenario, basic understanding of modeling constructs, or validation of ontology restriction correctness would be required. Having such assessment tests for LLMs is a pre-requisite for a straightforward LLM selection process for experimentation.

*Related Work.* The need to understand the strengths and limitations of numerous LLMs in concrete tasks has become apparent across domains. For instance, a systematic LLM evaluation framework is designed in the chemistry domain [14]. The framework relies on a number of diverse chemistry tasks selected to assess the capabilities of the LLM in terms of reasoning, understanding, and explaining within the domain. The deductive, inductive and abductive reasoning skills of LLMs are systematically assessed through a collection of evaluation methods across different dimensions (answer correctness, explanation redundancy, etc.) in [15]. Similarly, LLMs have been assessed in terms of their psychological profile by recording and analysing their answers to standard psychometric inventories [16].

## 2. Proposed Approach: a Collection of Assessment Tests for LLM Ontology Validation Skills

We take inspiration from research in the human-in-the-loop (HiL) area, where qualification test have frequently been applied to select the participants with the required skills to solve a particular knowledge validation task. Typically, a set of questions will be created to assess contributors' background knowledge and only those who score above a certain threshold will qualify to work on the available tasks.

In our earlier work, we developed a qualification test for assessing the expertise level of students in understanding basic ontology representations with focus on the meaning of *ontology restrictions* [17]. To the best of our knowledge, this is the only assessment test for knowledge engineering skills, which is publicly available in full. Indeed, authors of other similar resources opted to only present selected example questions from these to prevent participant bias (e.g., [18]). Moreover, unlike similar tests, our qualification test classifies examinees according to their expertise level (*novice*, *beginner*, *intermediate*, or *expert*) rather than using a binary classification of *qualified/unqualified*, providing a more comprehensive understanding of their skill levels.

We propose adopting HiL qualification testing approaches for LLM skills assessment. Particularly, we focus on assessing ontology validation capabilities in terms of the detection of modelling errors. As a starting point, we utilise the qualification test from [17]. The test should be extended and diversified to also assess other skill aspects relevant for ontology defect detection, such as understanding of disjointness axioms or the correct usage of intersections and unions. Creating a common collection of KE assessment tests allows researchers to evaluate available LLMs and select those that best fit their research need. While we focus on ontology validation skills, a similar approach can be followed for assessing other knowledge engineering skills as well.

## 3. Ongoing Work and First Results

We briefly describe the current version of the qualification test and subsequently the evaluation of a number of LLMs when they were administered this test.

### 3.1. Qualification test for LLM assessment in evaluating ontology restrictions


We start from a qualification test on ontology restrictions modeling initially developed in [17]. The test classifies the examinee as a *novice*, *beginner*, *intermediate*, or *expert* based on the achieved scores across questions with varying levels difficulty. It consists of 11 questions grouped in three categories:

- 4 beginner-level questions, which test the ability to identify ontology components (i.e., classes, relations and restrictions) in graphical and textual representations of an ontology.
- 3 intermediate level questions, assessing the understanding of the implications of ontology axioms containing the universal and existential restrictions.

- 4 expert questions, examining the ability to reason with ontology models, as well as compare and relate these ontology models to each other.

The expertise classification takes into account both the number of correctly answered questions from a specific category and the correct answers overall, as explained in detail in [17].

An example question, requiring intermediate modelling skills, is shown in Figure 1. To answer the question correctly, one needs to (1) identify the usage of a universal restriction; (2) understand that the universal restriction implies that instances of *PetLoverTypeA* cannot have pets that are not dogs; (3) be aware that the universal restriction, does not imply that instances of *PerLoverTypeA* must have a dog.

<p>Consider the model, represented in 3 equivalent formalisms (VOWL   Rector   Warren) and select the statement (A-D) that describes instances of <i>PetLoverTypeA</i> correctly.</p>	
<p><b>VOWL</b></p> 	<p><b>Rector</b></p> <p>PetLoverTypeA has, amongst other things, only Dog pets.</p>
	<p><b>Warren</b></p> <p>PetLoverTypeA has, amongst other things, no other than Dog pets.</p>
<p>A. Instances of <i>PetLoverTypeA</i> must have a Dog pet and cannot have other types of pets.          B. Instances of <i>PetLoverTypeA</i> might not have a Dog pet and cannot have other types of pets.          C. Instances of <i>PetLoverTypeA</i> must have a Dog pet and can also have other types of pets.          D. Instances of <i>PetLoverTypeA</i> might not have a Dog pet and can also have other types of pets.</p>	

**Figure 1:** Example question from the qualification test presented in [17].

The complete test<sup>1</sup> is designed using three ontology axiom representation modalities: the graphical representation VOWL, and two natural language formalisms proposed by Rector [19] and Warren [20].

### 3.2. Comparative assessment of LLM expertise in evaluating ontology restrictions

In previous work [12], we tested ChatGPT-4’s capabilities with the qualification test described above and concluded expertise levels comparable to an *expert*. Subsequently, we run the test on a variety of LLMs with different characteristics. We consider both proprietary and open-source foundational LLMs. While proprietary models, such as ChatGPT-4, have demonstrated strong performance [12], they bring high costs and reliance on cloud services. In contrast, open models offer the opportunity for local installation, ensuring data privacy and the potential for fine-tuning specific tasks. Thus, we evaluate models across three size categories: small (Llama3-8b<sup>2</sup>, Gemma-7b<sup>3</sup>), medium (Llama3-70b, Mixtral 8x22b<sup>4</sup>, Qwen2-72b<sup>5</sup>), and large (DeepSeek model<sup>6</sup>: deepseek-coder-v2 at 236b and deepseek-v2 at 236b). We compare the performance of these open-source models against the state-of-the-art proprietary LLM GPT-4o<sup>7</sup> and Claude-3.5-Sonnet<sup>8</sup>. We report the results of each LLM on the qualification test in Table 1.

Notably, no models fell into the lowest (novice) expertise category, indicating that all tested LLMs possess at least a basic level of ontology understanding competence. In fact, the expertise level tends to increase with the size of the model. The smallest models (Llama3-8b, Gemma-7b) showed *beginner*-expertise, while medium-sized LLMs performed slightly better and received *beginner* (Mixtral-8x7b) to *intermediate* qualifications (Llama3-70b, Qwen2-72b). The open-source large models (DeepSeek-

<sup>1</sup>The qualification test is included withing the Zenodo resource available at <https://zenodo.org/records/7643357>

<sup>2</sup><https://huggingface.co/meta-llama>

<sup>3</sup><https://huggingface.co/google>

<sup>4</sup><https://huggingface.co/mistralai>

<sup>5</sup><https://huggingface.co/Qwen/Qwen2-72B>

<sup>6</sup><https://deepseek.com>

<sup>7</sup><https://openai.com>

<sup>8</sup><https://claude.ai>

**Table 1**

Ontology Restriction Understanding Skills of LLMs according to the qualification test developed in [17]

Model	Size Category	Parameters	Assessed Expertise
Gemma-7b	Small Models	7B	Beginner
Llama3-8b		8B	Beginner
Mixtral-8x7b	Medium Models	47B (8x7B)	Beginner
Llama3-70b		70B	Intermediate
Qwen2-72b		72B	Intermediate
DeepSeek-V2	Large Models	236B	Intermediate
DeepSeek-Coder-V2		236B	Intermediate
Claude-Sonnet-3.5		Not disclosed	Expert
GPT-4o		Not disclosed	Expert

V2, DeepSeek-Coder-V2) also achieved an *intermediate* score, while Claude-Sonnet-3.5 and GPT-4o answered the most qualification questions correctly, showcasing *expert* skills.

The experiment was carried out on June 20th, 2024. We share the translation of the qualification test into a format suitable for LLMs<sup>9</sup>.

## 4. Outlook

First results of assessing LLM ontology validation skills indicate the power and *superiority of proprietary models*, in particular GPT-4o and Claude-Sonnet-3.5, compared to smaller open source models. Nevertheless, small and medium size models should not be ignored since they offer other benefits such as processing of privacy-sensitive tasks on local installations. Further investigations are needed whether providing additional instructions and context would improve the achieved results and whether these smaller LLMs could achieve comparable or better results to other knowledge validation tasks.

In the future, we plan to *design and evaluate additional qualification tests* focusing on complementary ontology validation aspects to allow for a comprehensive LLM profiling according to their capabilities. We would further like to motivate fellow researchers to publish previously created ontology engineering qualification tests, utilised in human-in-the-loop experiments or other assessment settings, to allow for the reuse of these valuable resources.

Motivated by recent work, assessing LLM capabilities with regards to their understanding of SPARQL and Turtle over time [21], we believe the collection of assessment tests would support the re-evaluation of LLMs upon new releases to allow for a *historic analysis* of selected capabilities.

Moreover, we see a need for a *community-driven platform*, where researchers can share their LLM assessment tests and contribute to LLM capability profiling on previously conducted skill assessments. This approach would not only promote the reuse of resources and facilitate experience sharing, but it would also support a more sustainable approach for advancing LLM assessment by preventing that similar tests are performed by several researchers in parallel, with high computational costs.

A crucial aspect to consider when making LLM assessment tests accessible to the community is the potential risk that these tests could be included in future model training datasets. Therefore, a community-wide strategy is needed to mitigate this risk and avoid evaluation biases. A community-driven assessment platform, could allow this by requiring authorised access to view tests, while providing functionality for “*blind*” LLM assessments.

## Acknowledgments

This work was funded by the Austrian Science Fund (FWF) Bilateral AI (Grant Nr. 10.55776/COE12) and HOnEst (V 745) projects.

<sup>9</sup>The LLM assessment implementation is available at <https://github.com/wu-semsys/llm-ontology-qualification-test>

## References

- [1] S. Pan, L. Luo, Y. Wang, C. Chen, J. Wang, X. Wu, Unifying large language models and knowledge graphs: A roadmap, *IEEE Transactions on Knowledge and Data Engineering* (2024).
- [2] F. Neuhaus, Ontologies in the Era of Large Language Models? a Perspective, *Applied ontology* 18 (2023) 399–407. doi:10.3233/ao-230072.
- [3] B. P. Allen, L. Stork, P. Groth, Knowledge Engineering using Large Language Models, *arXiv preprint arXiv:2310.00637* (2023).
- [4] Y. Zhu, X. Wang, J. Chen, S. Qiao, Y. Ou, Y. Yao, S. Deng, H. Chen, N. Zhang, LLMs for Knowledge Graph Construction and Reasoning: Recent Capabilities and Future Opportunities, *arXiv preprint arXiv:2305.13168* (2023).
- [5] M. Trajanoska, R. Stojanov, D. Trajanov, Enhancing Knowledge Graph Construction Using Large Language Models, 2023. *arXiv:2305.04676*.
- [6] S. Carta, A. Giuliani, L. Piano, A. S. Podda, L. Pompianu, S. G. Tiddia, Iterative zero-shot LLM prompting for knowledge graph construction, *arXiv preprint arXiv:2307.01128* (2023).
- [7] B. Zhang, I. Reklos, N. Jain, A. M. Peñuela, E. Simperl, Using Large Language Models for Knowledge Engineering (LLMKE): A Case Study on Wikidata, *arXiv preprint arXiv:2309.08491* (2023).
- [8] N. Tufek, A. S. S. Thuluva, T. Bandyopadhyay, V. P. Just, M. Sabou, F. J. Ekaputra, A. Hanbury, Validating Semantic Artefacts With Large Language Models, in: *The Semantic Web: ESWC Satellite Events*, 2024.
- [9] 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, in: *The Semantic Web: ESWC Satellite Events*, 2024.
- [10] P. T. G. Bradley P. Allen, Evaluating Class Membership Relations in Knowledge Graphs using Large Language Models, in: *The Semantic Web: ESWC Satellite Events*, 2024.
- [11] H. Khorashadizadeh, N. Mihindukulasooriya, S. Tiwari, J. Groppe, S. Groppe, Exploring In-Context Learning Capabilities of Foundation Models for Generating Knowledge Graphs from Text, 2023. *arXiv:2305.08804*.
- [12] S. Tsaneva, S. Vasic, M. Sabou, LLM-driven Ontology Evaluation: Verifying Ontology Restrictions with ChatGPT, in: *The Semantic Web: ESWC Satellite Events*, 2024.
- [13] N. Fathallah, A. Das, S. De Giorgis, A. Poltronieri, P. Haase, L. Kovriguina, NeOn-GPT: A Large Language Model-Powered Pipeline for Ontology Learning, in: *The Semantic Web: ESWC Satellite Events*, 2024.
- [14] T. Guo, K. Guo, B. Nan, Z. Liang, Z. Guo, N. V. Chawla, O. Wiest, X. Zhang, What can Large Language Models do in chemistry? A comprehensive benchmark on eight tasks, 2023. *arXiv:2305.18365*.
- [15] F. Xu, Q. Lin, J. Han, T. Zhao, J. Liu, E. Cambria, Are Large Language Models Really Good Logical Reasoners? A Comprehensive Evaluation and Beyond, 2024. *arXiv:2306.09841*.
- [16] M. Pellert, C. M. Lechner, C. Wagner, B. Rammstedt, M. Strohmaier, AI Psychometrics: Assessing the Psychological Profiles of Large Language Models Through Psychometric Inventories, *Perspectives on Psychological Science* 19 (2024) 808–826. doi:10.1177/17456916231214460.
- [17] S. Tsaneva, M. Sabou, Enhancing Human-in-the-Loop Ontology Curation Results through Task Design, *J. Data and Information Quality* (2023). doi:10.1145/3626960.
- [18] J. M. Mortensen, M. A. Musen, N. F. Noy, Crowdsourcing the verification of relationships in biomedical ontologies, in: *AMIA Annual symposium proceedings*, volume 2013, American Medical Informatics Association, 2013, pp. 1020–1029.
- [19] A. Rector, N. Drummond, M. Horridge, J. Rogers, H. Knublauch, R. Stevens, H. Wang, C. Wroe, Owl pizzas: Practical experience of teaching owl-dl: Common errors & common patterns, in: *International Conference on Knowledge Engineering and Knowledge Management*, Springer Berlin Heidelberg, Berlin, Heidelberg, 2004, pp. 63–81.
- [20] P. Warren, P. Mulholland, T. Collins, E. Motta, Improving comprehension of knowledge representation languages: A case study with description logics, *International Journal of Human-Computer*

Studies 122 (2019) 145–167.

- [21] J. Frey, L.-P. Meyer, F. Brei, S. Gründer-Fahrer, M. Marti, Assessing the Evolution of LLM capabilities for Knowledge Graph Engineering in 2023, in: The Semantic Web: ESWC Satellite Events, 2024.