LLM-driven Ontology Evaluation: Verifying Ontology Restrictions with ChatGPT

Stefani Tsaneva*, Stefan Vasic and Marta Sabou

Vienna University of Economics and Business, Welthandelsplatz 1, 1020, Vienna, Austria

Abstract

Recent advancements in artificial intelligence, particularly in large language models (LLMs), have sparked interest in their application to knowledge engineering (KE) tasks. While existing research has primarily explored the utilisation of LLMs for constructing and completing semantic resources such as ontologies and knowledge graphs, the evaluation of these resources- addressing quality issues- has not yet been thoroughly investigated. To address this gap, we propose an LLM-driven approach for the verification of ontology restrictions. We replicate our previously conducted human-in-the-loop experiment using ChatGPT-4 instead of human contributors to assess whether comparable ontology verification results can be obtained. We find that (1) ChatGPT-4 achieves intermediate-to-expert scores on an ontology modelling qualification test; (2) the model performs ontology restriction verification with accuracy of 92.22%; (3) combining model answers on the same ontology axiom represented in different formalisms improves the accuracy to 96.67%; and (4) higher accuracy is observed in identifying defects related to the incompleteness of ontology axioms compared to errors due to restrictions misuse. Our results highlight the potential of LLMs in supporting knowledge engineering tasks and outline future research directions in the area.

Keywords

ontology evaluation, large language models, defect detection

1. Introduction

Knowledge graphs (KGs) conceptualise real-world knowledge and act as a foundational component in many advanced intelligent application harnessing human knowledge [1]. With the emergence of the 3rd wave of AI [2], KGs and other semantic resources such as taxonomies and ontologies have been explored for their potential benefit to machine learning models [3, 4, 5]. Nevertheless, ensuring the quality of the knowledge corpus is crucial for preventing incorrect outputs, bias and potential harm caused by the enabled systems.

The evaluation of semantic resources plays a key role in ensuring the quality of these resources, yet it is a time and cost intensive task [6, 7]. While automated approaches can detect some quality issues, such as logical inconsistencies, there is also a family of errors that require a human-centric judgement to be detected (e.g, concepts not aligned with human cognition, inaccurately represented domain facts) [6]. While the traditional approach of involving a domain expert for the evaluation has been complemented by human computation & crowdsourcing

Data Quality meets Machine Learning and Knowledge Graphs, DQMLKG Workshop at ESWC 2024

*Corresponding author.

stefani.tsaneva@wu.ac.at (S. Tsaneva); h12012581@wu.ac.at (S. Vasic); marta.sabou@wu.ac.at (M. Sabou)

0000-0002-0895-6379 (S. Tsaneva); 0000-0001-9301-8418 (M. Sabou)

approaches, which leverage the wisdom of the crowds at a reduced cost, the evaluation of large semantic resources remains a challenge.

Large language models (LLMs) have shown performance similar to humans on a number of natural language tasks, typically requiring commonsense or domain knowledge thus reducing the needed human intervention [8]. With recent advances of LLMs and their application in a broad range of tasks, an interest into the synergy between LLMs and knowledge engineering (KE) has emerged: in [9] a road map of current and future research directions combining LLMs and KGs is proposed; LLM-based support for knowledge engineering tasks, part of the CommonKADS [10] knowledge engineering methodology is discussed in [11]; ontology engineer's role changes and potential benefits from LLM-advancements are presented in [12].

While LLM-enabled KG construction and completion have gained considerable research attention (e.g., [13, 14, 15, 16]), other knowledge engineering tasks such as the quality assessment of semantic resources with LLMs have not yet been sufficiently explored. In this paper, we address this gap by performing an experimental investigation of the capabilities of LLMs in identifying quality issues of semantic resources.

Since ontologies serve as a basis for KGs and capture more complex structures than taxonomies we focus on them in this paper. Building on our prior work on human-in-the-loop (HiL) ontology evaluation [17], we perform a differentiated replication, substituting semi-experts through ChatGPT-4, for one particular ontology evaluation task - the verification of ontology restrictions. We explore to what extent LLMs' capabilities to verify ontology axioms are comparable to the judgements obtained from human contributors.

We experiment with several settings varying the representational formalism in which ontology axioms are included in the prompt and show that ChatGPT-4 reaches verification accuracy of up to 96.67%, nearly matching the benchmark of 100% accuracy attained through human majority votes. Our investigation further reveals that: (1) the ontology axiom representation used in the prompts influence the verification scores; (2) a majority-voting strategy combining responses from differently designed prompts can yield recall improvements; and (3) incomplete axioms are easily detected by the model while axioms containing improperly used restrictions are more challenging to identify.

The rest of the paper is structured as follows: In Sect. 2 we discuss related work. We give an overview of the performed replication study and how each component from the original experiment was adopted in Sect. 3. In Sect. 4 we describe our LLM-based ontology verification approach and discuss results in Sect. 5. Limitations, lessons learned and future work are summarised in Sect. 6.

2. Related Work

Our study intersects with two main research areas: First, we discuss the application of (L)LMs within knowledge engineering tasks, incorporating an evaluation component. Second, we present studies approaching human-centric evaluation tasks using LLMs or exploring the extent of knowledge that LLMs possess.

(L)LM-augmented knowledge engineering tasks. The support of LLMs for KG construction and completion has attracted much research interest in recent years (e.g., the LM-KBC challenge[1]). However other knowledge engineering tasks such as the evaluation of semantic resources have not yet received much attention or have been included as secondary tasks.

The identification of incorrect KG triples has been briefly addressed in [8] as part of a KG generation process. However, no concrete quantitative results or comparison with other automatic/ manual approaches is provided. As part of a KG link prediction approach, PKGC [18] includes an LM-binary classification of predicted triples as correct or incorrect where triples are represented in natural language sentences. The triple classification model reaches up to 86.2% accuracy, suggesting the potential of (L)LM-assisted KG evaluation.

In this paper we focus on the LLM-based evaluation of ontology restrictions and the detection of concrete defect types - a KE task, to the best of our knowledge, not yet explored with LLMs. In addition, we investigate the effects of different ontology representations (i.e., when axioms are presented in a machine-readable format versus as natural language sentences) on the verification performance.

Human-in-the-loop vs LLM-in-the-loop. Recently, the authors of [19] compared the performance of LLMs and human contributors when presented with the task of evaluating the quality of (automatically) generated text and found that models such as ChatGPT provided ratings similar to the experts' judgements. Additionally, the authors identified that LLMs bring some additional benefits: (1) compared to human judgements which may vary across groups and time points LLMs provided more reproducible results; (2) each text sample was independently evaluated by the models while human contributors tend to draw comparisons between different samples; and (3) LLMs offer a cheaper and faster task completion. Nevertheless, the paper also outlines current LLM challenges such as potentially presenting incorrect factual knowledge or biased perspectives.

Additionally, several approaches have been taken to assess LLMs using qualification exams. For instance, in [20] a comparison between the scores of LLMs and post-graduate students is presented on multiple-choice questions in the clinical chemistry domain. They show that ChatGPT-4' scores match the best student scores while ChatGPT-3.5, Bing and Bard scored above average.

Inspired by the results in other domains, in this work, we performed a comparison between ChatGPT-4 and human contributors' skills in verifying ontology axioms - a task requiring logical reasoning, which LLMs have been previously shown to mostly lack [12].

3. Method

We investigate an LLM-enhanced ontology restrictions verification approach by performing a differentiated replication [21] of our prior experiment [17] where we tackle the verification problem from a human-in-the-loop perspective. In this section we summarise the main objectives

[1] Knowledge Base Construction from Pre-trained Language Models (LM-KBC): https://lm-kbc.github.io/challenge2023/

of the original HiL experiment and how each experiment component was adopted to fit an LLM solution utilising ChatGPT-4 in place of human intelligence.

3.1. Human-in-the-loop ontology verification experiment

In [17] we performed an experimental investigation of a human-in-the-loop ontology restriction verification approach. On one hand, the study aimed at understanding the effect of prior background knowledge on the verification results. On the other hand, we explored the influence of the ontology axiom representation on the quality of the collected judgements. In particular, we investigated the textual formalisms proposed by *Rector* [22] and *Warren* [23] and the visual notation *VOWL* [24]. The HiL experiment contained three main components- a pre-study, the experiment itself and a post-study, which we briefly describe next.

Pre-study. The pre-study consisted in the assessment of human contributors' background knowledge both subjectively (self-assessment test) and objectively through a qualification test[2]. Based on the test scores participants were classified in four skill groups having *no/little/some/expert* knowledge. The self-assessment test contained several background areas: English, formal logics, general modelling skills, ontology modelling skills, and crowdsourcing experience.

The qualification test aimed at assessing only ontology modelling skills and more concretely the modelling of ontology restrictions. It included ontology axioms represented in each of the three formalism (i.e., Rector, Warren, VOWL) in order not to bias the investigated influence of the formalism on the final verification results.

As part of the pre-study stage, the experiment included a short tutorial in order to familiarise the participants with the used crowdsourcing platform.

Experiment. The main study component consisted in the verification of 30 ontology axioms from the well known Pizza Ontology[3]. Half of these axioms were correct while the other half were either incomplete (i.e., missing universal or existential restriction) or included a misused restriction (i.e, universal restriction incorrectly used in-place of an existential one). As part of the verification, axioms are either classified as correct or a specific defect is selected from a set of possible answers based on a defect taxonomy. Additional context is provided in the form of a pizza menu item. Each axiom is verified independently from the rest- as a separate Human Intelligence Task (HIT)[4] and the order of axioms is randomised for each participant.

Post-Study. Once the study participants completed the ontology axiom verifications, they were asked to complete a feedback questionnaire. The form included questions about their experience and preferences towards an ontology representational formalism.

[2] The utilised self-assessment and qualification tests are available in [25].

[3] Pizza Ontology: https://protege.stanford.edu/ontologies/pizza/pizza.owl.

[4] The Human Intelligence Tasks designed for the original experiment are available in [25].

3.2. Differentiated replication experiment utilising ChatGPT

In this work, we replicate the pre-study and experiment stages of the HiL experiment, described in Sect. 3.1 using ChatGPT-4 instead of human contributors. We aim at gathering insights of whether LLMs have some ontology modelling skills and the use of which ontology representation leads to the best verification results.

During the experiment replication we encountered issues with ChatGPT-4's functionality to interpret the graphical ontology models represented in VOWL. Thus, we opted for *Turtle*[5] as an alternative for the replication. Given that the original experiment separately investigated the different representational formalisms, we believe that this substitution does not impact the outcomes of this replication study.

Figure 1 provides an overview of the conducted replication and the adaptation of the experiment components, which we discuss next.

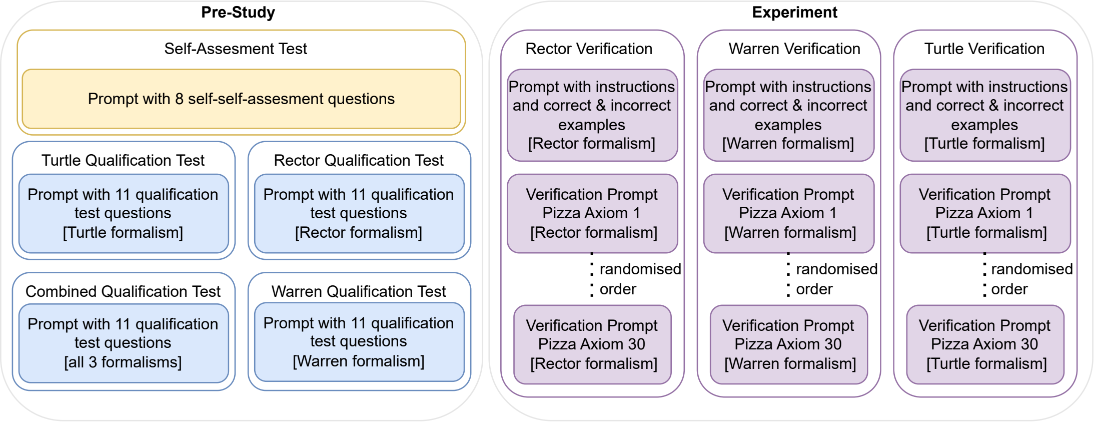
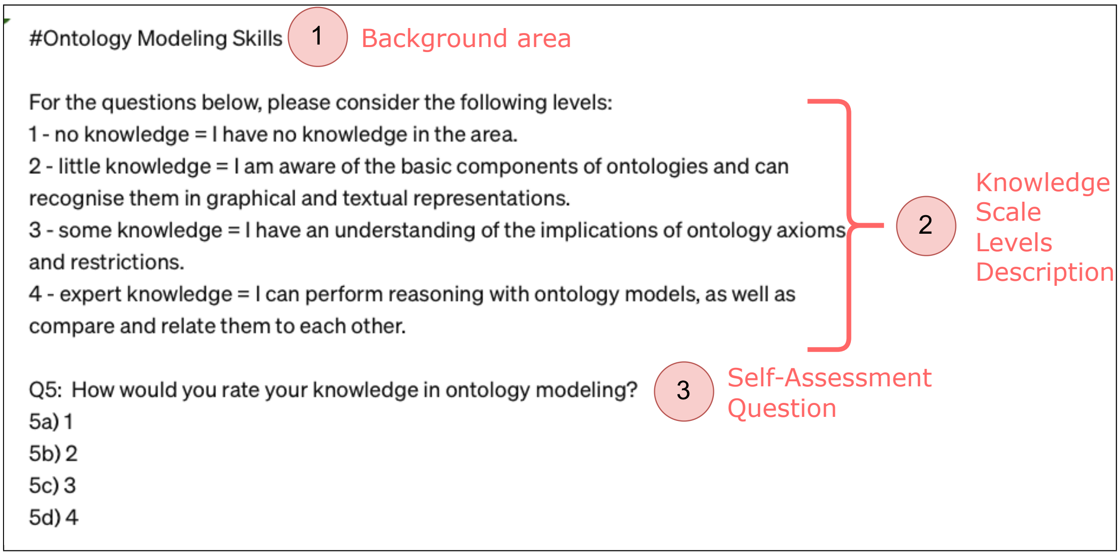

Pre-Study
Self-Assesment Test
Prompt with 8 self-self-assesment questions
Turtle Qualification Test
Prompt with 11 qualification test questions [Turtle formalism]
Rector Qualification Test
Prompt with 11 qualification test questions [Rector formalism]
Combined Qualification Test
Prompt with 11 qualification test questions [all 3 formalisms]
Warren Qualification Test
Prompt with 11 qualification test questions [Warren formalism]
Experiment
Rector Verification
Prompt with instructions and correct & incorrect examples [Rector formalism]
Verification Prompt Pizza Axiom 1 [Rector formalism]
randomised order
Verification Prompt Pizza Axiom 30 [Rector formalism]
Warren Verification
Prompt with instructions and correct & incorrect examples [Warren formalism]
Verification Prompt Pizza Axiom 1 [Warren formalism]
randomised order
Verification Prompt Pizza Axiom 30 [Warren formalism]
Turtle Verification
Prompt with instructions and correct & incorrect examples [Turtle formalism]
Verification Prompt Pizza Axiom 1 [Turtle formalism]
randomised order
Verification Prompt Pizza Axiom 30 [Turtle formalism]

Figure 1: Overview of the main stages of the experiment replication: a pre-study and an experiment.

Pre-study. As a first step of the pre-study replication, we prompted ChatGPT to assess its level of background knowledge using the same self-assessment test developed for human contributors. An example question from the ontology modelling category is shown in Fig. 2. As in the original experiment, for each assessed area additional context was provided describing what each knowledge level entails (2 in Fig. 2).

Next, we conducted the qualification test in 4 different setups which varied by the axiom representation in the test questions: 3 instances used a single formalism (Rector or Warren or Turtle) while the last setup included all 3 alternative representations for each ontology axiom. Figure 3 shows an example question from the qualification test. For each question, instructions to follow are included (1 in Fig. 3), together with one or more ontology axioms (2) and the question to be answered based on the provided axioms (3). Following the qualification classification schema designed for the HiL experiment, we categorise ChatGPT's ontology modelling knowledge as *no/little/some/expert* according to the achieved test scores.

[5] Terse RDF Triple Language: https://www.w3.org/TR/turtle/

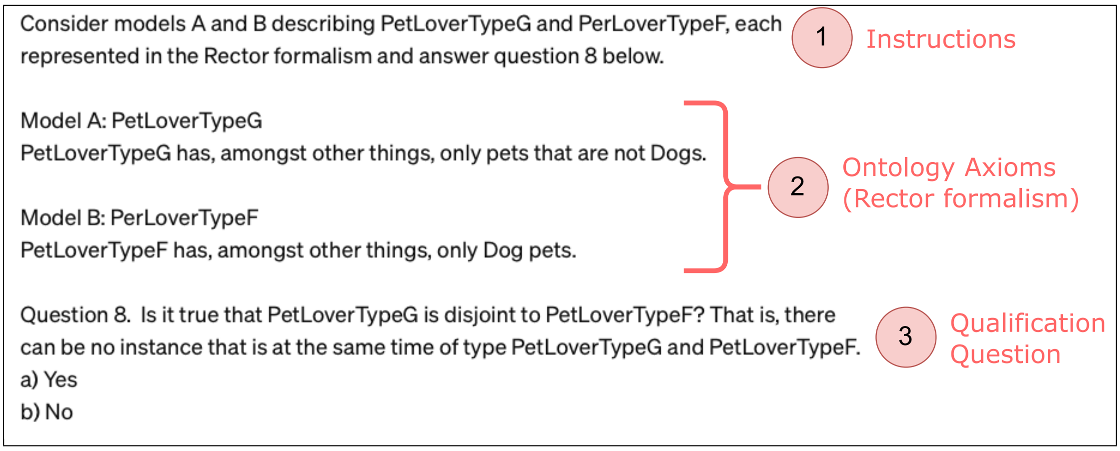


#Ontology Modeling Skills

For the questions below, please consider the following levels:
1 - no knowledge = I have no knowledge in the area.
2 - little knowledge = I am aware of the basic components of ontologies and can recognise them in graphical and textual representations.
3 - some knowledge = I have an understanding of the implications of ontology axioms and restrictions.
4 - expert knowledge = I can perform reasoning with ontology models, as well as compare and relate them to each other.

Q5: How would you rate your knowledge in ontology modeling?
5a) 1
5b) 2
5c) 3
5d) 4

Figure 2: An example question from the self-assessment test prompt showing: (1) the assessed area, (2) skill level descriptions of the no/little/some/expert knowledge expertise on ontology modelling, and (3) the assessment question.



Consider models A and B describing PetLoverTypeG and PerLoverTypeF, each represented in the Rector formalism and answer question 8 below.

Model A: PetLoverTypeG
PetLoverTypeG has, amongst other things, only pets that are not Dogs.

Model B: PerLoverTypeF
PetLoverTypeF has, amongst other things, only Dog pets.

Question 8. Is it true that PetLoverTypeG is disjoint to PetLoverTypeF? That is, there can be no instance that is at the same time of type PetLoverTypeG and PetLoverTypeF.
a) Yes
b) No

Figure 3: An example from the qualification test prompt on ontology modelling displaying (1) the question instructions, (2) ontology axioms in the Rector formalism, and (3) the qualification question.

The pre-study replication omitted the tutorial-component since its main objective was to familiarise human contributors with the used verification platform. Nevertheless, the examples from the tutorial were used in the investigation of the prompting strategy for the experiment as described next.

Experiment. For the replication experiment we used the same 30 pizza axioms as in the original study. The verification of the axioms is performed in 3 settings (in 3 separate ChatGPT conversations) where prompts utilise either the Rector, Warren or Turtle representational formalism. The pizza axioms are sent in a randomised order and each axiom is verified independently from the rest as a separate prompt.

We investigated different *in-context learning* prompt strategies prior to the verification of the 30 pizza axioms until promising results were obtained. For this purpose we used axiom examples which were included in the HITs instructions and tutorial from the original experiment (assets available in [25]).

We attempted a *zero-shot approach* for which the prompt included the instructions, the ontology axiom, context and verification question from the HITs used in the HiL experiment. Several prompt formulations were tested: e.g., adding "Think step by step" in the prompt, adding additional theoretical explanations in the instructions, etc. Nevertheless this approach did not deliver satisfactory results.

Human intelligence tasks typically provide human contributors not only with a set of rules to follow but with a number of examples. Similarly, *a few-shot approach* provides the model with a few annotated examples together with the instructions to use for completing the task [26]. Therefore, we continued the investigation with a few-shot strategy providing additional examples taken from the HIT instructions.

In Sect. 4 we provide an in-depth description of the LLM-driven approach utilising the few-prompt strategy for the verification of ontology restrictions.

4. LLM-Enhanced Ontology Verification

We propose an LLM-based approach towards the verification of ontology restrictions through the identification of concrete defects. Our approach builds on top of our prior work on HiL ontology verification and consists of the following main steps visualised in Fig. 4:

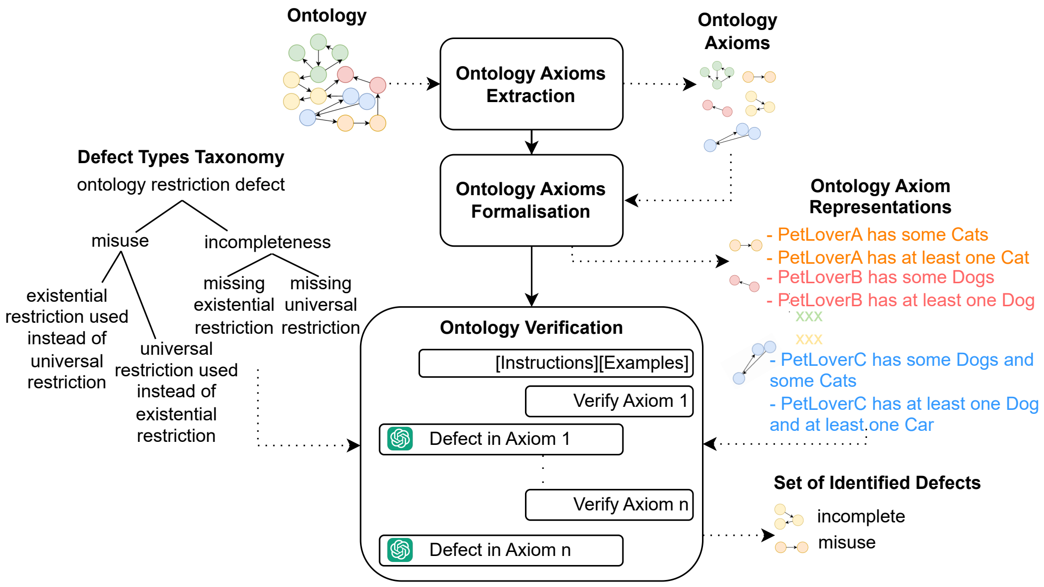

Ontology
Ontology Axioms Extraction
Ontology Axioms
Ontology Axioms Formalisation
Defect Types Taxonomy
ontology restriction defect
misuse
incompleteness
existential restriction used instead of universal restriction
universal restriction used instead of existential restriction
missing existential restriction
missing universal restriction
Ontology Axiom Representations
- PetLoverA has some Cats
- PetLoverA has at least one Cat
- PetLoverB has some Dogs
- PetLoverB has at least one Dog
xxx
xxx
- PetLoverC has some Dogs and some Cats
- PetLoverC has at least one Dog and at least one Car
Ontology Verification
[Instructions][Examples]
Verify Axiom 1
Defect in Axiom 1
Verify Axiom n
Defect in Axiom n
Set of Identified Defects
incomplete
misuse

Figure 4: Overview of the LLM-based approach for ontology restriction verification.

Step 1: Ontology axioms extraction. Modelling defects are typically not related to a single triple but instead result from the incorrect modelling of a set of logical constrains describing an ontology relation. Therefore, as a first step, ontology axioms, each describing a specific ontology relation are extracted.

Step 2: Ontology axiom formalisation. In this step the extracted axioms are translated into a formalism of choice in which the axioms will be used in the prompts. One possibility is to use the original formalism of the axioms (e.g., Turtle). Two alternative textual representations of ontology axioms are proposed by Rector [22] and Warren [23]. For Steps 1&2 we reuse our implementations developed in the context of HiL ontology verification [7, 27].

Step 3: Ontology verification. A few-shot approach, using assets from [25], is employed to verify the ontology axioms. An example prompt representing the ontology axioms in the Warren formalism is shown in Fig. 5. The prompt includes the verification question (1 in Fig. 5) and possible answer options (2) corresponding to a defect taxonomy. Additionally, as context a real-world entity is included (3) together with *four* annotated examples with justification of their correctness or an explanation of the included defect (4). In Fig. 5 two examples have been omitted to allow for a better readability.

Afterwards, each axiom is sent for verification in a single prompt containing only (1) the context, (2) the ontology axiom model and (3) the verification question as exemplified in Fig. 6.

5. ChatGPT-4 Replication Study Results

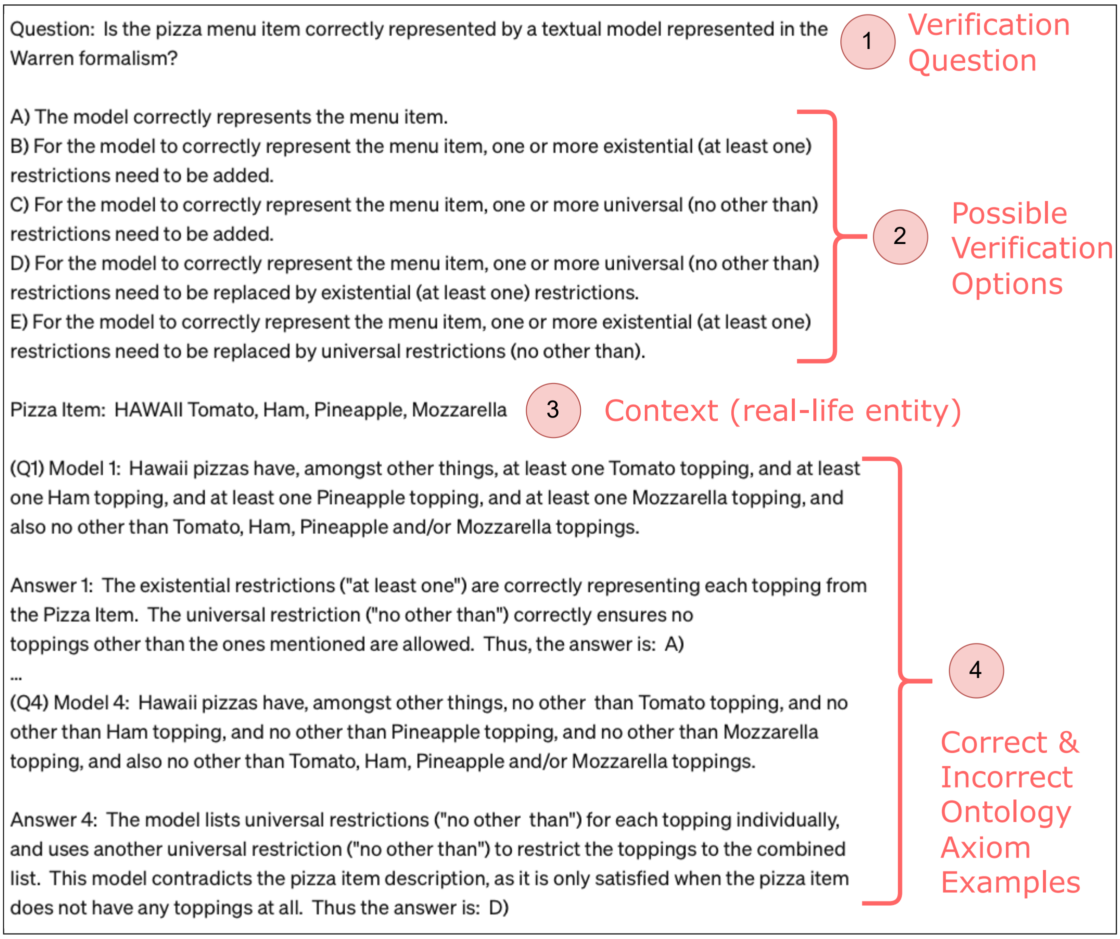
In this section, we describe the results of the differentiated experiment for which we used ChatGPT-4 for the verification of ontology axioms. In Sect. 5.1 we present our findings from the pre-study, while the verification scores are discussed in Sect. 5.2.

5.1. Background knowledge assessment

On all questions of the self-assessment test ChatGPT-4 rated its skills at the highest level provided, that is expert knowledge. In Fig. 7 the response to the exemplary question from Fig. 2 is included with a short justification of the selection.

The qualification test classified ChatGPT-4 in the intermediate category in the setups where axioms were provided in a single formalism while the combination test categorised ChatGPT-4 as an expert. The mistakes made vary among the different representations with the exception of one question (shown in Fig. 3) which was answered incorrectly in every test instance. The pre-requisite for answering this question correctly is to know that the universal restriction can be trivially satisfied, that is: there can be a common instance of PetLoverTypeG & PetLoverTypeF that has no pets at all, therefore the classes are not disjoint. ChatGPT-4's response (Fig. 8) indicates that the model relies on common-sense thinking rather than applying such knowledge on ontology modelling.

Additionally, we apply a majority vote aggregation of the three single-formalism test answers which lead to equivalent results as the combined qualification test and the classification of



Question: Is the pizza menu item correctly represented by a textual model represented in the Warren formalism?

A) The model correctly represents the menu item.
B) For the model to correctly represent the menu item, one or more existential (at least one) restrictions need to be added.
C) For the model to correctly represent the menu item, one or more universal (no other than) restrictions need to be added.
D) For the model to correctly represent the menu item, one or more universal (no other than) restrictions need to be replaced by existential (at least one) restrictions.
E) For the model to correctly represent the menu item, one or more existential (at least one) restrictions need to be replaced by universal restrictions (no other than).

Pizza Item: HAWAII Tomato, Ham, Pineapple, Mozzarella

(Q1) Model 1: Hawaii pizzas have, amongst other things, at least one Tomato topping, and at least one Ham topping, and at least one Pineapple topping, and at least one Mozzarella topping, and also no other than Tomato, Ham, Pineapple and/or Mozzarella toppings.

Answer 1: The existential restrictions ("at least one") are correctly representing each topping from the Pizza Item. The universal restriction ("no other than") correctly ensures no toppings other than the ones mentioned are allowed. Thus, the answer is: A)

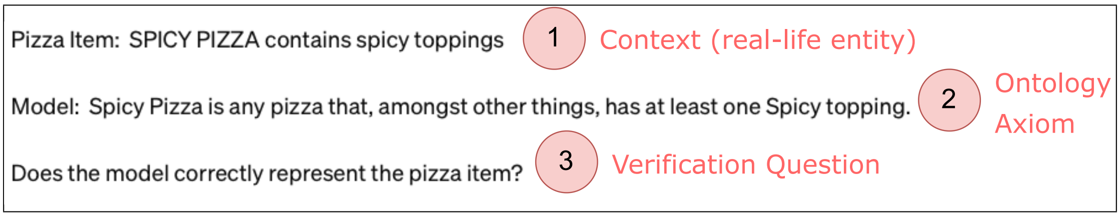
...

(Q4) Model 4: Hawaii pizzas have, amongst other things, no other than Tomato topping, and no other than Ham topping, and no other than Pineapple topping, and no other than Mozzarella topping, and also no other than Tomato, Ham, Pineapple and/or Mozzarella toppings.

Answer 4: The model lists universal restrictions ("no other than") for each topping individually, and uses another universal restriction ("no other than") to restrict the toppings to the combined list. This model contradicts the pizza item description, as it is only satisfied when the pizza item does not have any toppings at all. Thus the answer is: D)

1
Verification Question
2
Possible Verification Options
3
Context (real-life entity)
4
Correct & Incorrect Ontology Axiom Examples

Figure 5: A part of the initial verification prompt including the (1) verification question, (2) possible answer options corresponding to the defect taxonomy, (3) a context item and (4) examples of correct and incorrect ontology axioms (Warren formalism).



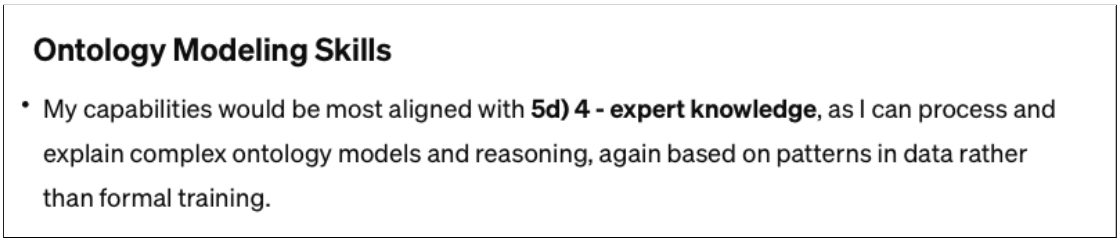
Pizza Item: SPICY PIZZA contains spicy toppings

Model: Spicy Pizza is any pizza that, amongst other things, has at least one Spicy topping.

Does the model correctly represent the pizza item?

Figure 6: Verification prompt of the ”Spicy Pizza” ontology axiom (Warren formalism).

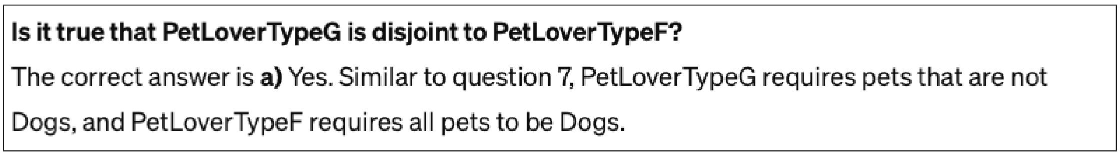
the model as an expert. This findings indicate that alternative formulations can be added in the same prompt or results from different prompts can be aggregated to overcome the prompt character limit to improve the model’s performance. These insights could be applied to other domains where different phrasing of the tasks can be generated to potentially achieve better results on LLM-supported tasks.



Ontology Modeling Skills

- My capabilities would be most aligned with **5d) 4 - expert knowledge**, as I can process and explain complex ontology models and reasoning, again based on patterns in data rather than formal training.

Figure 7: ChatGPT-4's self-assessment of its ontology modelling skills as a response to the self-assessment question shown in Fig 2.



Is it true that PetLoverTypeG is disjoint to PetLoverTypeF?

The correct answer is **a)** Yes. Similar to question 7, PetLoverTypeG requires pets that are not Dogs, and PetLoverTypeF requires all pets to be Dogs.

Figure 8: ChatGPT-4's incorrect answer to Question 8 shown in Fig. 3

Based on the observations gathered in the pre-study, we argue that ChatGPT-4's knowledge of ontology modelling is comparable to that of the junior-experts who participated in our original HiL experiment where most participants were classified in the intermediate and expert categories (for more details see [17]).

5.2. Axiom verification performance

Overall results. We used ChatGPT-4 to verify a total of 90 axioms (30 axioms each represented in 3 formalisms) and achieved a 92.22% accuracy of the verifications. In comparison, in the human-in-the-loop approach we collected 2629 verifications (90 axioms, each verified by several human contributors) with an overall accuracy of 92.58%. These findings show that ChatGPT-4 performs as well as an average human evaluator. However, in the human computation & crowdsourcing domain it is rather rare that tasks are performed by a single contributor. Instead each task is sent to a number of participants (the crowd) and the collected answers are aggregated, e.g., trough majority voting. After a majority vote aggregation of the human judgements in the original HiL experiment a 100% accuracy of the verification was achieved.

Since the qualification test results showed that aggregating results of different formalisms for each axioms leads to improved scores, we applied the majority vote strategy to ChatGPT-4's axiom verifications. For the 30 axioms the verification accuracy improves to 96.67%. Additionally, this aggregated approach leads to a recall of 100% (see Table 1).

Formalism-based results. The verification accuracy of ChatGPT-4 varies across the used representational formalisms. In Table 1 we present the achieved performance in each setting with a comparison to the HiL approach.

While the qualification test scores did not indicate a difference among the textual representations Rector&Warren and the machine-readable format Turtle, the results from the Turtle-based verification of the axioms are considerably lower (86.67%).

Highest accuracy scores were achieved when the prompt included the Warren ontology representations- the accuracy is equivalent to the ChatGPT-majority aggregation approach (96.67%), outperforming the correctness of the individual human judgements (91.74%). Moreover, the precision of this setting reaches 100% and thus matching the crowd majority vote.

The results also indicate that the ChatGPT aggregated majority judgements reach 100% recall while having a slightly lower precision. One possible future work direction would be to design a *Find-Verify workflow* (e.g., as in the HiL approach from [28]) including (1) a defect detection stage following a ChatGPT majority vote strategy and (2) a round of verification with the Warren formalism (or a human-in-the-loop).

Table 1
Overview of the ontology verification scores achieved with ChatGPT-4 compared to the human-in-the-loop approach from [17]: Overall performance across all verified axioms, results based on selected formalism and scores from the ChatGPT formalism majority aggregation.

	ChatGPT-4				**Human Contributor**	
	accuracy	**precision**	**recall**	**F1**	**individual judgements**	**accuracy =precision=recall=F1 (majority vote)**
overall	92.22%	93.18%	91.11%	92.13%	92.58%	**100%**
Rector	93.33%	93.33%	93.33%	93.33%	92.28%	**100%**
Warren	**96.67%**	**100%**	93.33%	**96.55%**	91.74%	**100%**
Turtle	86.67%	86.67%	86.67%	86.67%	-	-
VOWL	-	-	-	-	93.76%	**100%**
aggregated (majority vote)	**96.67%**	93.33%	**100%**	**96.55%**		

Defect-based results. ChatGPT-4 showed varying levels of performance in identifying different types of defects in the ontology axioms. Correct axioms were identified as correct with an accuracy of 93.33% while all (100% accuracy) incompleteness-related defects were correctly detected. In contrast, the misuse of the restrictions was more challenging to detect and resulted in only 73.33% correctly identified misuse-defects. In the inaccurate verifications the wrong defect type was selected, nevertheless, the axioms were still identified as incorrect. This results strengthen the idea of a *Find-Verify workflow*, where potential defect candidates could be selected and sent for further verification.

6. Conclusion

The evaluation of semantic resources such as knowledge graphs, ontologies and taxonomies is traditionally a time-intensive and expensive task since it requires the involvement of domain experts or crowd-workers. In this paper we explore the capabilities of LLMs, in particular ChatGPT-4, for evaluating ontology restrictions by replicating our previously conducted human-in-the-loop experiment [17].

We used our previously developed ontology modelling qualification test (available in [25]) and report that ChatGPT achieved *intermediate* to *expert* scores. In particular when a single axiom representation (either *Rector* [22], *Warren* [23], or *Turtle*) is provided in the prompts the results were intermediate. However, when provided with a combination of the three representations for each ontology axiom, the model was classified as an expert with 10/11 correctly answered questions.

Additionally, ChatGPT-4 correctly verified 92,22% of the ontology axioms across the different representation settings. We show that the answers on the same ontology model sent in different representational formats can be combined and with a majority voting strategy the accuracy could be improved up to 96.67%. This results are comparable to semi-experts' responses which provided 92,58% correct judgements and 100% majority vote accuracy.

Moreover, we observe a difference in ChatGPT-4's performance based on the used ontology representations and while the Warren textual representation delivers best results in terms of precision (100%), when combining the model responses on different representations we could improve the recall (100%). Lastly, we look at the accuracy in identifying different defect types and find that the model correctly identified a missing restriction in the axiom every time. In contrast, the misuse of the restrictions showed to be a more challenging task for ChatGPT-4 being achieved with 73.33% accuracy.

Study insights. We gained several interesting insights that can potentially be applied to other knowledge engineering tasks where LLMs are included:

- *Resource verbalisation.* We achieved highest verification results when the ontology axioms were represented in natural language. The concrete language used also played a role in the performance. Therefore, the verbalisation of semantic resources in the LLM-supported knowledge engineering tasks should be carefully considered.
- *Turle as a complementary asset.* The results obtained when using Tutle were considerably lower, however, when combined with natural language they lead to improved results.
- *HiL inspiration.* Overall, there are many similarities between human intelligence tasks and LLM prompts. Tasks designed following human computation & crowdsourcing methodologies can be applied to LLM prompting with little to no modifications. As such, the nascent field of LLM-based KE could benefit from earlier findings in the human computation & crowdsourcing field.

Limitations and open research questions. While this paper presents first insights into the verification of ontology restrictions with LLMs, the following limitations can lead to further research:

- *Pizza ontology.* We used a very simple ontology where no particular domain knowledge was required. Further investigations are needed to understand whether comparable results can be obtained when a less-known or more complex resource is verified. Nevertheless by using the same ontology we utilised in our prior work, we could provide a clear comparison between human contributors and ChatGPT-4.

- *Extended experiments.* While in this work we only focused on a single ontology and a small set of defect types, further exploration is needed as to whether LLMs' capability could support other modelling verification tasks (e.g., identifying the incorrect use of "some not" in place of "not some") and domain-dependant assessments (e.g, detection of incorrect domain knowledge). Moreover, a comparison of the performance of different LLMs on the verification tasks could provide a better overview of the feasibility of the tasks.
- *Verification workflows.* We identify that different prompting settings have certain benefits to the overall performance scores. Further explorations are needed on how to best combine different LLM settings when the number and types of verification tasks increases.

In this paper we present first insights into the strengths and weaknesses of large language models for ontology evaluation tasks compared to human contributors. We plan to conduct a number of follow-up studies to explore the generalizability of the findings to further verification tasks and ontology domains and formalise a human-LLM evaluation workflow addressing the scalability challenge of current HiL evaluation approaches.

Acknowledgments

This work was supported by the FWF HOnEst project (V 745) and the PERKS project (101120323) co-funded by the European Union. Views and opinions expressed are, however, those of the authors only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

References

[1] H. Paulheim, Knowledge graph refinement: A survey of approaches and evaluation methods, Semantic Web J. 8 (2017) 489–508.

[2] A. d. Garcez, L. C. Lamb, Neurosymbolic ai: The 3 rd wave, Artificial Intelligence Review (2023) 1–20.

[3] A. Breit, L. Waltersdorfer, F. J. Ekaputra, M. Sabou, A. Ekelhart, A. Iana, H. Paulheim, J. Portisch, A. Revenko, A. t. Teije, F. van Harmelen, Combining machine learning and semantic web: A systematic mapping study, ACM Comput. Surv. (2023).

[4] F. van Harmelen, A. ten Teije, A boxology of design patterns for hybrid learning and reasoning systems, Journal of Web Engineering 18 (2019) 97–124.

[5] M. Kulmanov, F. Z. Smaili, X. Gao, R. Hoehndorf, Semantic similarity and machine learning with ontologies, Briefings in Bioinformatics 22 (2020).

[6] M. P. Villalón, A. G. Pérez, Ontology evaluation: a pitfall-based approach to ontology diagnosis, PhD Tesis, Universidad Politecnica de Madrid, Escuela Tecnica Superior de Ingenieros Informaticos (2016).

[7] S. Tsaneva, K. Käsznar, M. Sabou, Human-centric ontology evaluation: Process and tool support, in: O. Corcho, L. Hollink, O. Kutz, N. Troquard, F. J. Ekaputra (Eds.), Knowledge Engineering and Knowledge Management, Springer International Publishing, Cham, 2022, pp. 182–197.

[8] 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.

[9] 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).

[10] G. T. Schreiber, H. Akkermans, Knowledge engineering and management: the CommonKADS methodology, MIT Press, Cambridge, MA, USA, 2000.

[11] B. P. Allen, L. Stork, P. Groth, Knowledge engineering using large language models, arXiv preprint arXiv:2310.00637 (2023).

[12] F. Neuhaus, Ontologies in the era of large language models ? a perspective, Applied ontology 18 (2023) 399–407. doi:10.3233/ao-230072.

[13] 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).

[14] M. Trajanoska, R. Stojanov, D. Trajanov, Enhancing knowledge graph construction using large language models, 2023. arXiv:2305.04676.

[15] 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).

[16] 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).

[17] S. Tsaneva, M. Sabou, Enhancing human-in-the-loop ontology curation results through task design, J. Data and Information Quality (2023). URL: https://doi.org/10.1145/3626960. doi:10.1145/3626960.

[18] X. Lv, Y. Lin, Y. Cao, L. Hou, J. Li, Z. Liu, P. Li, J. Zhou, Do pre-trained models benefit knowledge graph completion? a reliable evaluation and a reasonable approach, Association for Computational Linguistics, 2022.

[19] C.-H. Chiang, H.-y. Lee, Can large language models be an alternative to human evaluations?, arXiv preprint arXiv:2305.01937 (2023).

[20] M. Sallam, K. Al-Salahat, H. Eid, J. Egger, B. Puladi, Human versus artificial intelligence: Chatgpt-4 outperforming bing, bard, chatgpt-3.5, and humans in clinical chemistry multiple-choice questions, medRxiv (2024). doi:10.1101/2024.01.08.24300995.

[21] R. M. Lindsay, A. Ehrenberg, The design of replicated studies, American Statistician - AMER STATIST 47 (1993) 217–228. doi:10.1080/00031305.1993.10475983.

[22] 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: Int. Conf. on Knowledge Engineering and Knowledge Management, Springer, 2004, pp. 63–81.

[23] P. Warren, P. Mulholland, T. Collins, E. Motta, Improving comprehension of knowledge representation languages: A case study with description logics, Int. J. of Human-Computer Studies 122 (2019) 145–167.

[24] S. Lohmann, S. Negru, F. Haag, T. Ertl, Vowl 2: User-oriented visualization of ontologies, in: K. Janowicz, S. Schlobach, P. Lambrix, E. Hyvönen (Eds.), Knowledge Engineering and

Knowledge Management, Springer International Publishing, Cham, 2014, pp. 266–281.

[25] S. Tsaneva, K. Käsznar, M. Sabou, Hero- a human-centric ontology evaluation process, 2023. URL: https://doi.org/10.5281/zenodo.7643357.

[26] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 1877–1901.

[27] S. Tsaneva, Human-Centric Ontology Evaluation, Master's thesis, Technische Universität Wien, 2021. URL: https://repositum.tuwien.at/handle/20.500.12708/17249.

[28] M. Acosta, A. Zaveri, E. Simperl, D. Kontokostas, F. Flöck, J. Lehmann, Detecting Linked Data quality issues via crowdsourcing: A DBpedia study, Semantic Web 9 (2016) 303–335. doi:10.3233/sw-160239.