LLM-driven Ontology Evaluation: Verifying Ontology Restrictions with ChatGPT

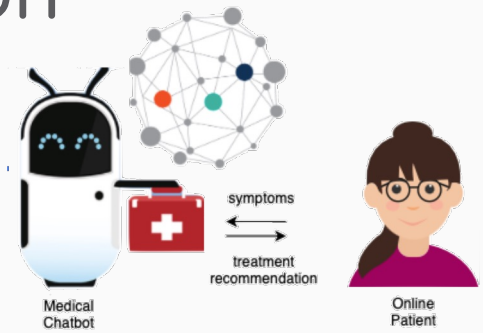
Stefani Tsaneva, Stefan Vasic and Marta Sabou

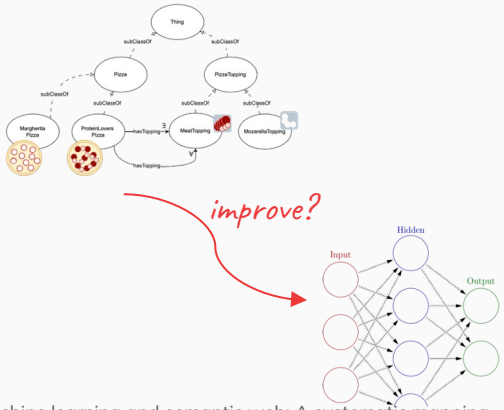
HOnEst
FWF Austrian Science Fund
WU VIENNA
ESWC24
DQMLKG

Importance of Ontology Evaluation

Google, 30.11.2022

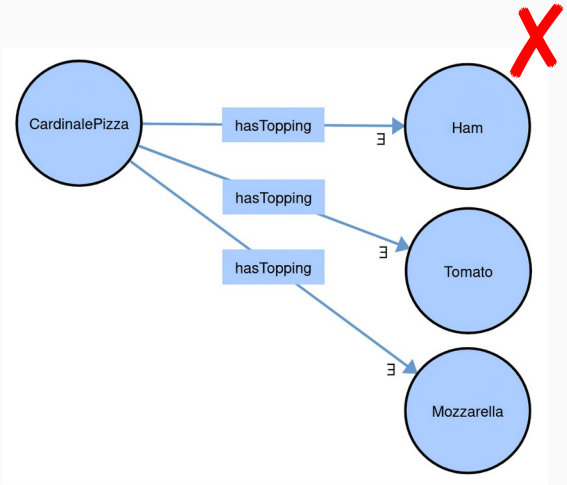
- Ontologies and other semantic resources are rarely perfect
- The quality of the ontology can result in erroneous system outputs
- Ontologies are used as input to machine learning algorithms aiming to improve their performance [1, 2, 3]



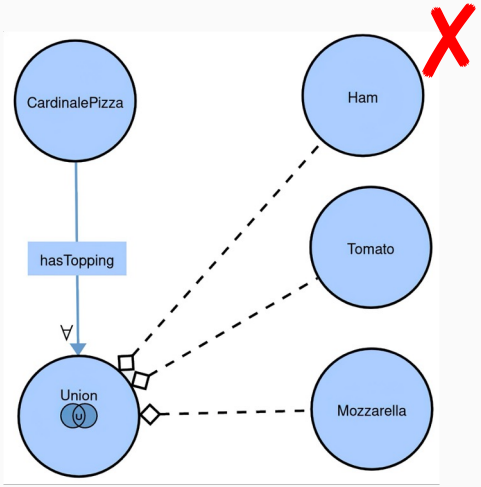


[1] A. Breit, L. Waltersdorfer, F. J. Ekaputra, M. Sabou, A. Ekelhart, A. Iana, H. Paulheim, J. Portisch, A. Revenko, A. t. Teije, F. van Harmelen, Combining machine learning and semantic web: A systematic mapping study, ACM Comput. Surv. (2023).
[2] M.Kulmanov,F.Z.Smaili,X.Gao,R.Hoehndorf,Semantic similarity and machine learning with ontologies, Briefings in Bioinformatics 22 (2020).
[3] 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.

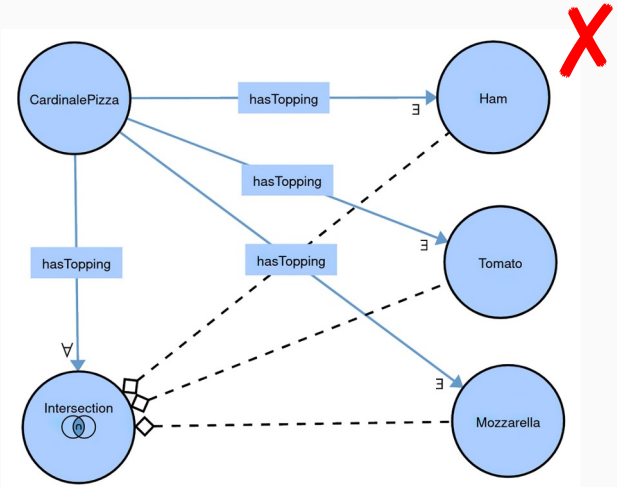
Human-Centric Ontology Defects



Missing closure axiom

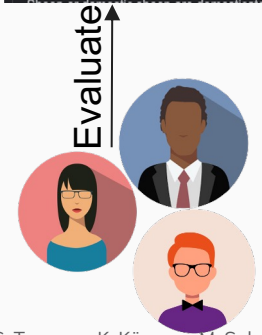


Trivially satisfiable universal restriction



Confusion between logical and linguistic *and*

Human-in-the-Loop (HiL) Ontology Evaluation



- High planning efforts of evaluation campaigns [4]
- Limited availability of experts
- High costs for high-quality crowd work
- Do not scale well [5]

LLM-in-the-Loop vs Human-in-the-Loop

- Comparable expertise on human-centric tasks for quality inspection (of text) [6]
- Comparable results with post-graduate students on qualification tests from various domains [6, 7]
- Comparable reproducibility of results [6]
- Time efficiency

[4] S. Tsaneva, K. Käsznar, M. Sabou, Human-centric ontology evaluation: Process and tool support, in: Knowledge Engineering and Knowledge Management, Springer International Publishing, Cham, 2022, pp. 182–197.
[5] H. Paulheim, Knowledge graph refinement: A survey of approaches and evaluation methods, Semantic Web J. 8 (2017) 489–508.
[6] C.-H.Chiang,H.-y.Lee,Can large language models be an alternative human evaluations?, arXiv preprint arXiv:2305.01937 (2023).
[7] 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).

Are LLMs capable of detecting quality issues in ontologies?

Our **Approach** : Differentiated Experiment **Replication** of a HiL Evaluation Campaign [8]

- Pizza Ontology - 30 axioms (15 correct, 15 incorrect)
- Assessment based on a predefined defect taxonomy
- Using **ChatGPT-4** in place of human participants (students)

Goals:

1. Assess ChatGPT4's **expertise on ontology modelling** (Self Assessment & Qualification Test)
2. Assess ChatGPT4's **capability to detect misuse of ontology restrictions** (owl:allValuesFrom, owl:someValuesFrom)

\+ Assess ChatGPT4's **capability based on the representation of the ontology** axiom

[8] S. Tsaneva, M. Sabou, Enhancing human-in-the-loop ontology curation results through task design, J. Data and Information Quality (2023).

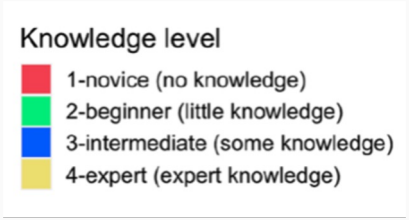
1. ChatGPT4's expertise on ontology modelling

Self-Assessment

Test Available in Zenodo

Different areas tested:

- English Language
- Formal Logics
- General Modelling Skills
- Ontology Modelling Skills



How would you rate your knowledge in ontology modeling?

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.

1. ChatGPT4's expertise on ontology modelling

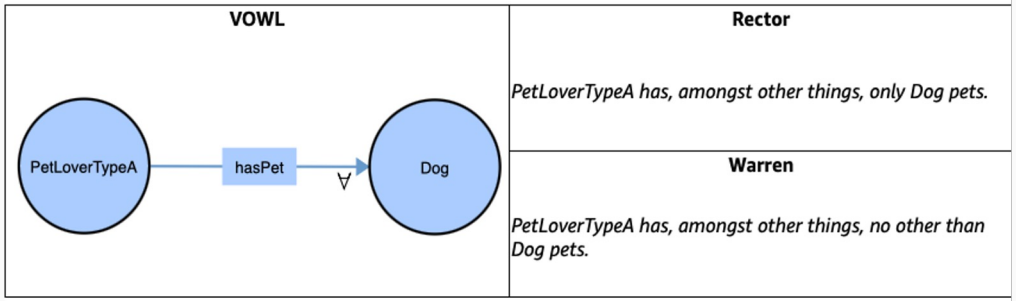
Qualification Test

- Focus on ontology restrictions
- 11 questions with increasing difficulty

Test Available in Zenodo

Turtle used for the replication instead instead of VOWL

Consider the model, represented in 3 equivalent formalisms (VOWL | Rector | Warren) and answer question 3 below.



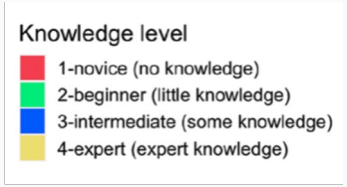
3. Select the statement that describes instances of PetLoverTypeA correctly.

- Instances of PetLoverTypeA must have a Dog pet and cannot have other types of pets.
- Instances of PetLoverTypeA might not have a Dog pet and cannot have other types of pets.
- Instances of PetLoverTypeA must have a Dog pet and can also have other types of pets.
- Instances of PetLoverTypeA might not have a Dog pet and can also have other types of pets.

1. ChatGPT4's expertise on ontology modelling
Qualification Test

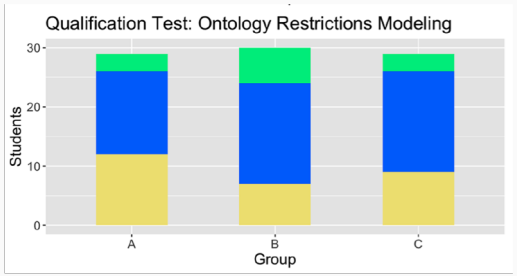
ChatGPT4 is an intermediate/expert in ontology modelling:

- Intermediate: when a single ontology representation is shown
- Expert: when several ontology representations are shown



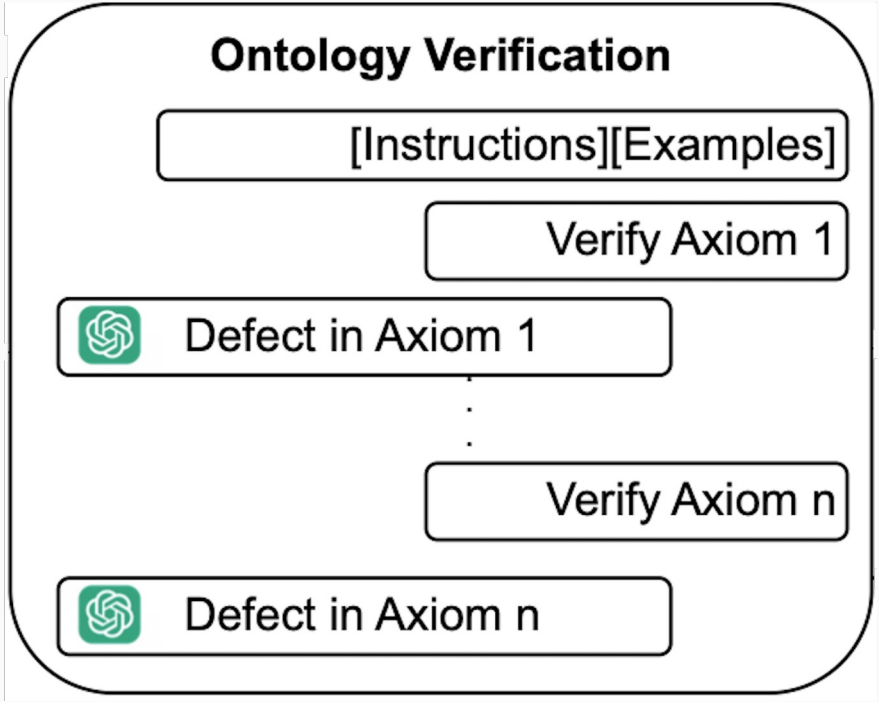
HiL Results

(A) **Rector** Verbalisation with keywords **some & only**	(B) **Warren** Verbalisation with keywords **at least one & no other than**	(C) **Turtle** **owl:allValuesFrom, owl:someValuesFrom**	**(A) + (B) + (C)** within the same prompt	**Majority** (A_prompt, B_prompt, C_prompt)
intermediate	intermediate	intermediate	**expert**	**expert**



VOWL (∀ & ∃) with **GPT-4o** : intermediate

2. ChatGPT4's capability to detect misuse of ontology restrictions



Human Evaluation Task

See Instructions

5 instructions on the correct usage of ontology restrictions

Please make sure you are familiar with the rules and examples provided in the **Instructions** before answering the question.

Pizza Menu	Model
 **ROSA** Ⓥ Gorgonzola, Mozzarella, Tomato	Rosa pizzas have, amongst other things, some Tomato topping, and some Mozzarella topping, and some Gorgonzola topping, and also only Gorgonzola, Mozzarella, and/or Tomato toppings.

1 context entity (EVORA) in a representational format of choice

2 ontology restriction axiom (ORA) in a representational format of choice

Does the model represent the pizza menu item correctly ?

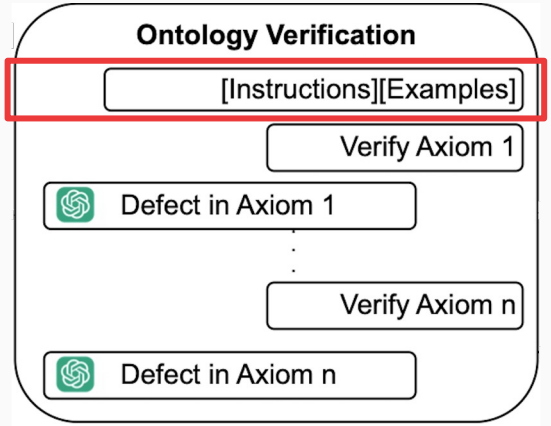
- The model correctly represents the menu item.
- For the model to correctly represent the menu item, one or more existential (some) restrictions need to be added.
- For the model to correctly represent the menu item, one or more universal (only) restrictions need to be added.
- For the model to correctly represent the menu item, one or more universal (only) restrictions need to be replaced by existential (some) restrictions.
- For the model to correctly represent the menu item, one or more existential (some) restrictions need to be replaced by universal restrictions (only).

3 verification options corresponding to a defect taxonomy

Comment (optional)

In case you have any remarks please add them here

ChatGPT Evaluation Task



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

1 Verification Question

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).

2 Possible Verification Options

Pizza Item: HAWAII Tomato, Ham, Pineapple, Mozzarella

3 Context (real-life entity)

(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)

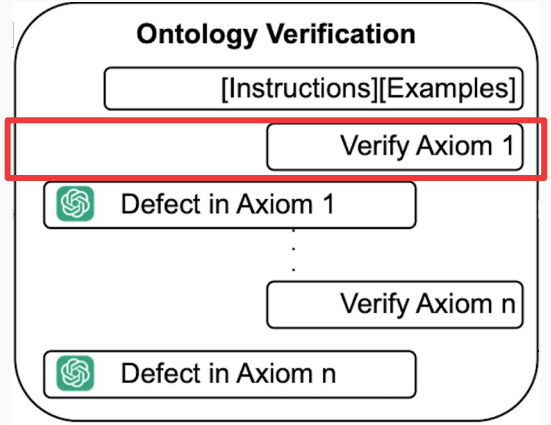
...

(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)

4 Correct & Incorrect Ontology Axiom Examples

ChatGPT Evaluation Task



Pizza Item: SPICY PIZZA contains spicy toppings (1) Context (real-life entity)

Model: Spicy Pizza is any pizza that, amongst other things, has at least one Spicy topping. (2) Ontology Axiom

Does the model correctly represent the pizza item? (3) Verification Question

2. ChatGPT4's capability to detect misuse of ontology restrictions

- Up to 96.67% accuracy of axiom evaluations
- Defect-based scores: incompleteness defects (100%accuracy misuse defects (73 %)
- Combination of different verbalizations can improve the recall.

	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%**		

Main Findings & Future Work

- ChatGPT's **expertise** on ontology modeling is equivalent to **intermediates/experts**.
- **Capability**. LLMs, and in particular ChatGPT achieve high accuracy (96.67%) in verifying ontology restrictions.
- **Verbalisation**. The concrete language used played a role in the achieved performance.
- **HiL inspiration**. There are many similarities between human intelligence tasks and LLM prompts.

- Experiments on **more complex / real ontologies & knowledge graphs**
- Experiments on **various quality issues**
- Experiments on **various verbalisations**
- Experiments with **different LLMs**
- Experiments on **hybrid workflows** combining both Human-in-the-Loop and LLM-in-the-Loop

Thank you!

LLM-driven Ontology Evaluation: Verifying Ontology Restrictions with ChatGPT

(These) Slides

Full Paper

Original Human-in-the-Loop Experiment

Full Paper

Zenodo Resources