Enhancing Knowledge Engineering with LLMs

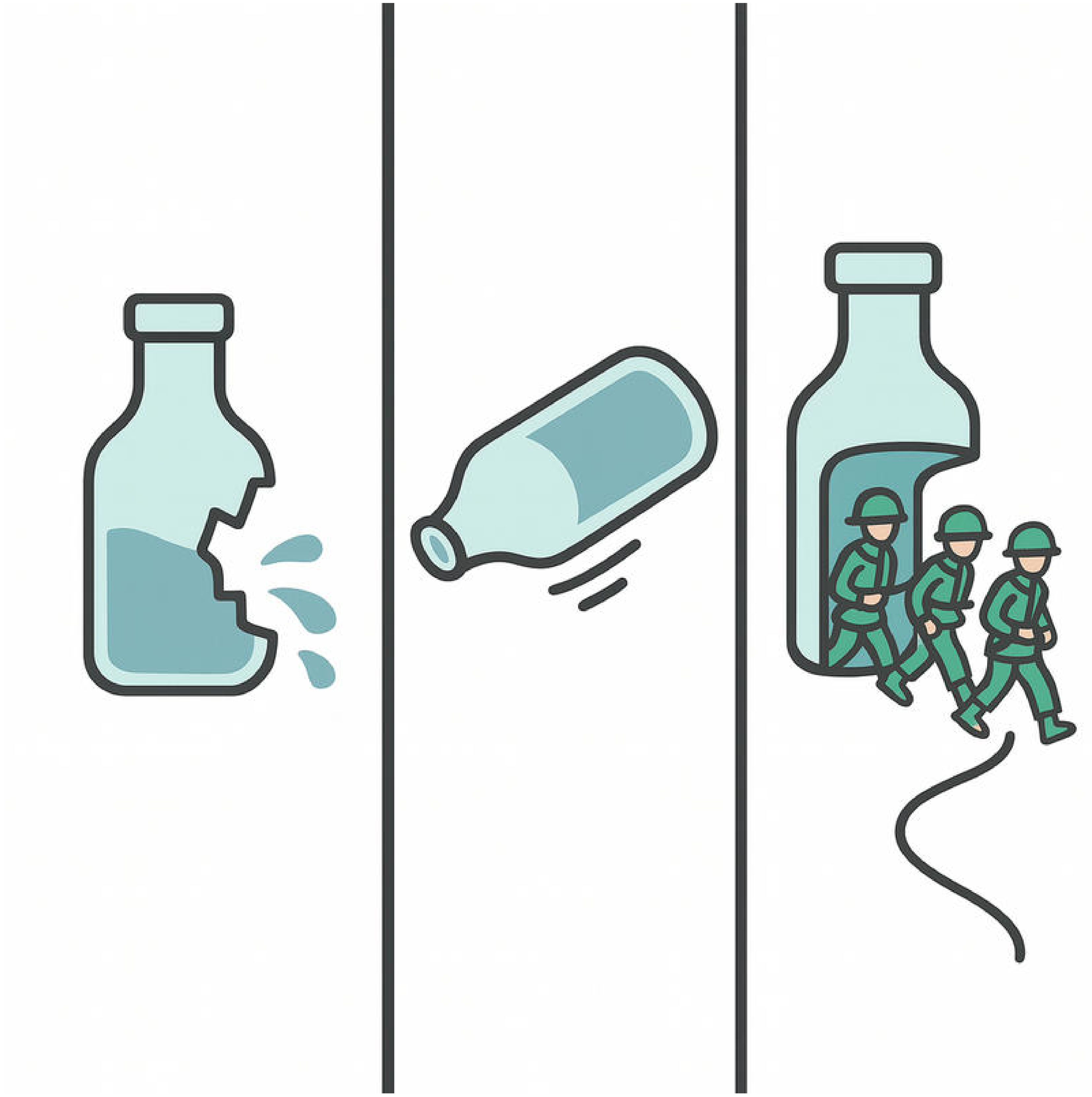
Anna Sofia Lippolis

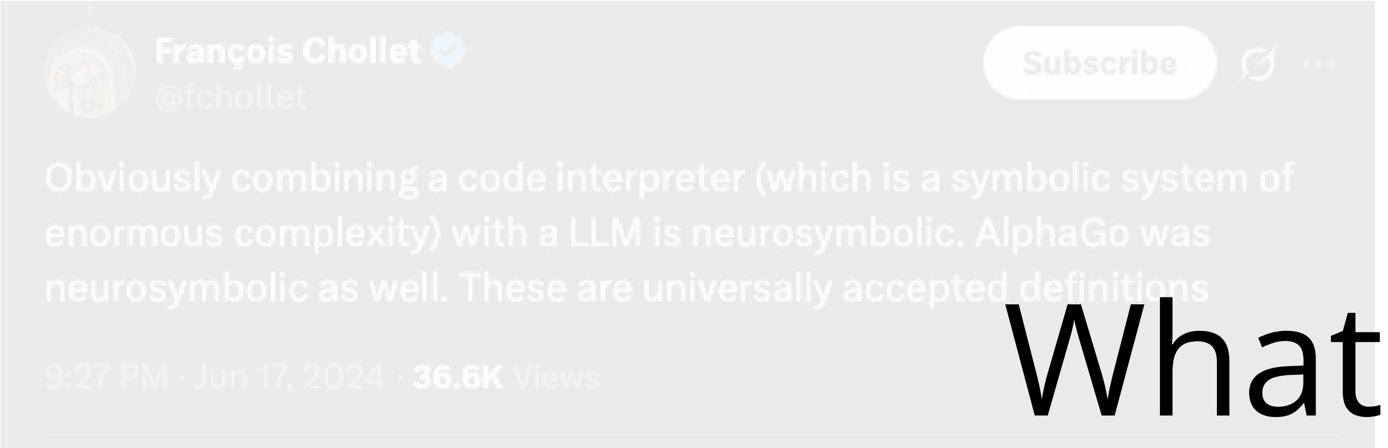
AIRA Seminar

11.12.2025

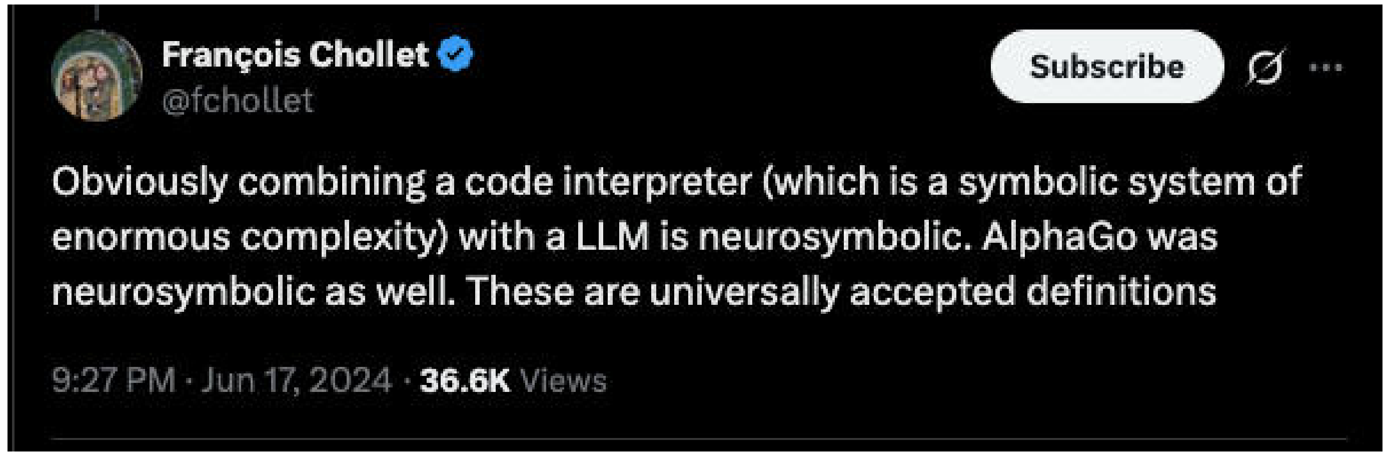
The case for (explicit) knowledge

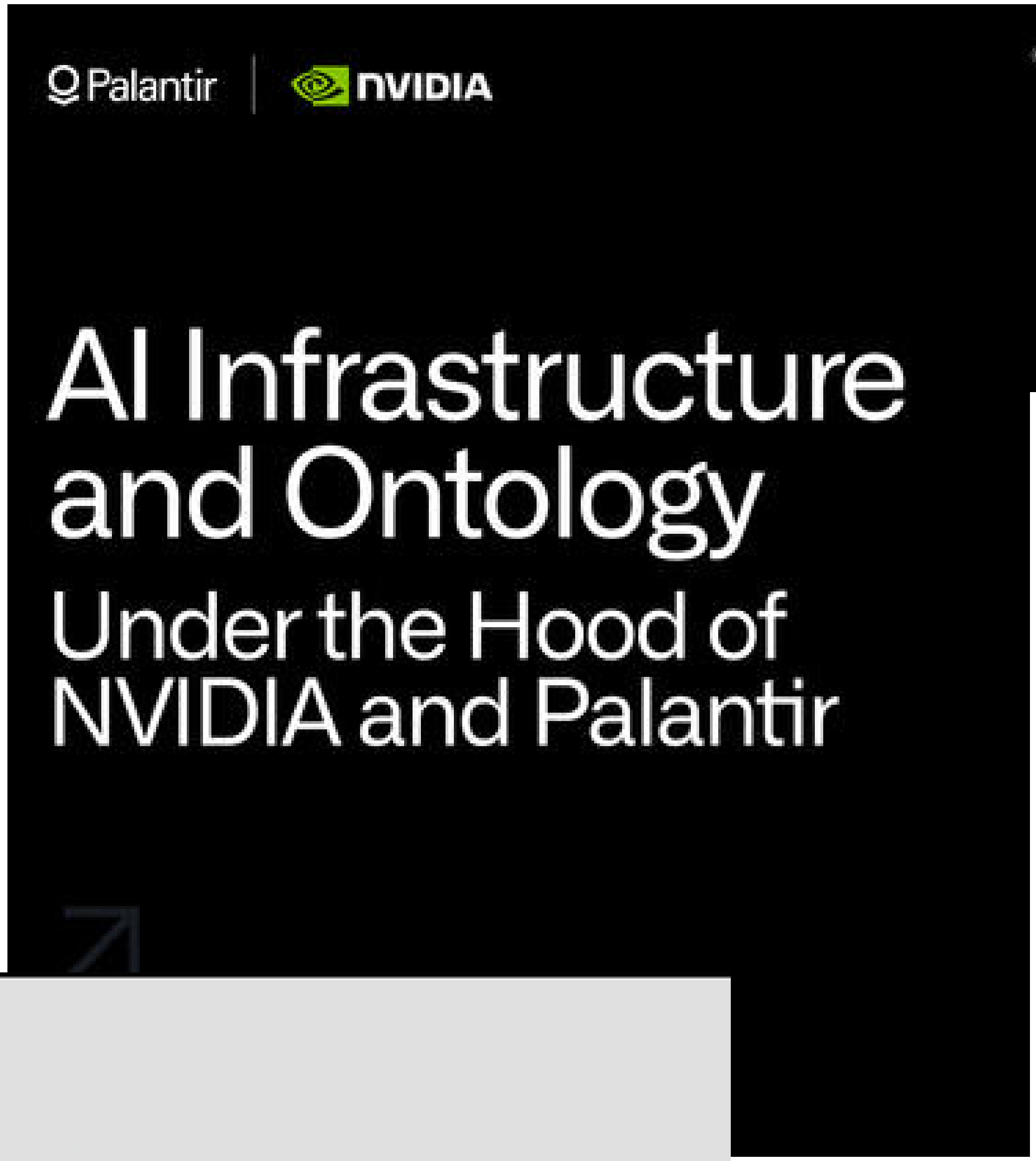
- Much of what we know is abstract and general
- LLMs are deeply unreliable when it comes to knowledge like this
- Current systems have no real concept of a bottle, a soldier, etc.
- **Without explicit, manipulable knowledge, our models will remain *uninterpretable* and *unreliable***





What can be a **solution**?





The Future is Neuro-Symbolic: Where has it been, and where is it going?*

Vaishak Belle   Gary Marcus

Abstract

This report explores the evolution and current state of neuro-symbolic artificial intelligence, an approach that integrates neural network capabilities with symbolic reasoning. We trace the historical context from early AI aspirations to modern implementations and successes, highlighting key paradigms, and other logical and semantical considerations. We argue against the "scaling is all you need" hypothesis, and point to persistent challenges in reliable symbolic reasoning with deep and large models. We conclude by suggesting that despite numerous implementation choices and the "broad church" nature of neuro-symbolic AI, these approaches offer the most promising path towards AI systems that combine pattern recognition with robust reasoning, particularly for applications requiring structured knowledge, explainability, and trustworthiness.

Introduction

In its original inception, the field of artificial intelligence (AI) was deeply concerned with how human cognition could be automated on a computer (Turing 1950). John McCarthy, for example, argued for the development of so-called *common-sense programs* that could reason and problem solve, using the mathematical apparatus of logic for formalising the application domain. The idea was arguably radical at that time (although Leibniz had already wondered about thinking as an algebraic process (Levesque 2012)).

Despite continuing extensions of logical formalisms to deal with actions, plans and agents (Levesque 1996; Kelly and Pearce 2008), three representational aspects severely challenge this program: (a) the lack of *complete* knowledge (that is, an exhaustive formalisation) in almost all applications, (b) the difficulty in asserting that logical assertions are *categorically* true, and (c) the need to leverage sensorimotor data and relevant statistical patterns from that data. Sensorimotor data, including visual and auditory information, moreover, is sampled from high-dimensional spaces, which lead to the problem of *scalability* with purely logical frameworks. For example, a 16×16 black and white image would require, say, a data structure with 216×16 propositions to capture the pixel values, and for a dataset of 100,000 images would require as many instances of such data structures.

To deal with possibly false facts, and uncertainty in general, *probabilistic logics* emerged (Bacchus 1990; Halpern 2003), that allowed for a mixture of probabilistic and logical assertions. However, because they inherited the expressive power of (often first-order) logic but then further extended it for probabilistic knowledge, they too suffered from scalability issues. And they largely glossed over the point of where the probabilities came from. Presumably, these need to be drawn from data, but integrating the learning of statistical information and ensuring its consistency with prior logical knowledge is a non-trivial matter (Valiant 1999).

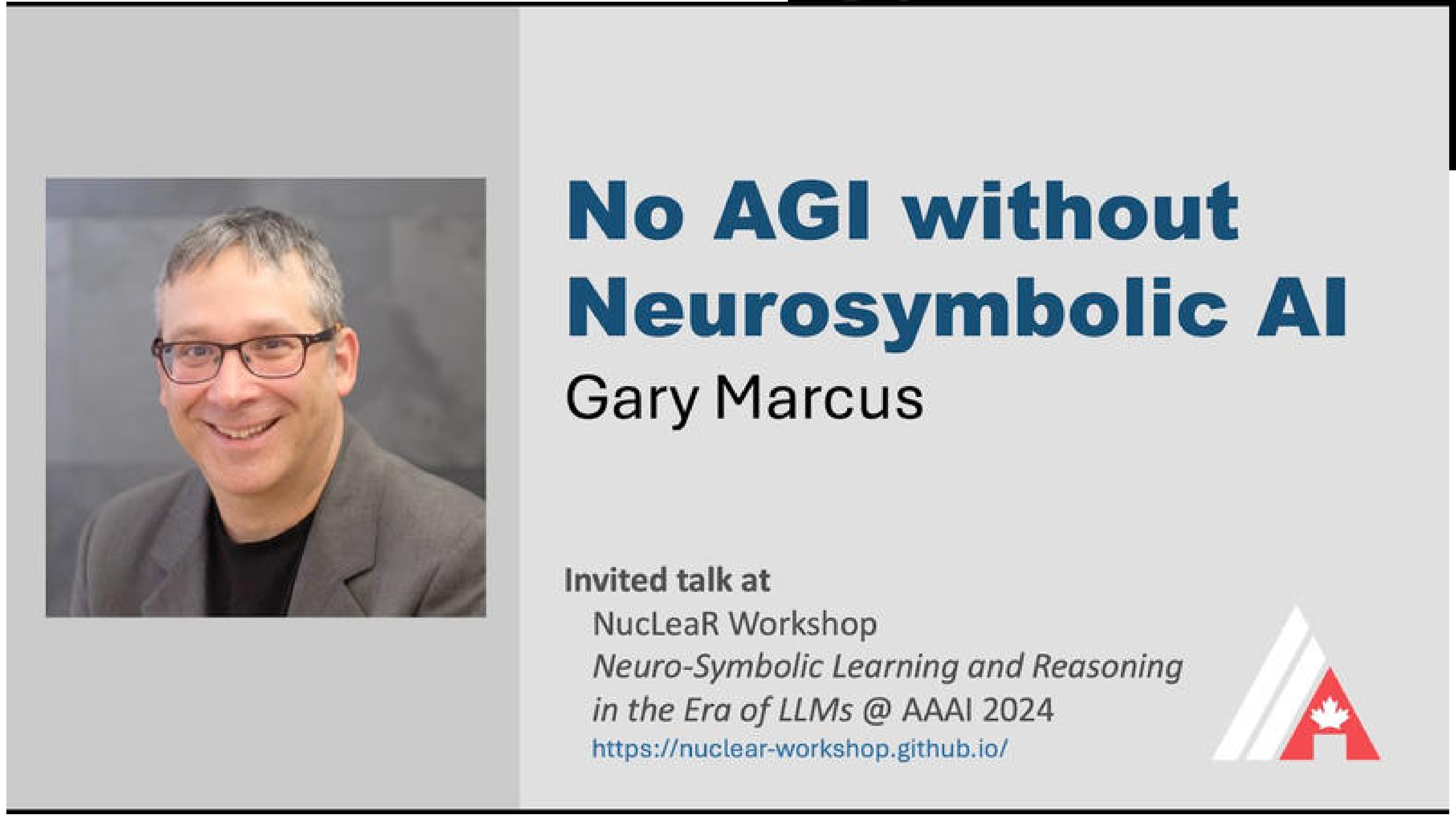
When limited to graphs, Bayesian (and later causal) networks offered a reasonable compromise between expert knowledge together with probabilistic and statistical information (Pearl 1998), and were amenable to certain types of learning from data (Koller and Friedman 2009). Building on this success, the field of statistical relational learning (SRL) aimed to unify logical and probabilistic frameworks by controlling the expressiveness (Raedt et al. 2016), such as by investigating relational extensions to Bayesian networks (Koller and Pfeffer 1997; Getoor et al. 2001). Be that as it may, a purely expert-driven paradigm for dealing with high-dimensional data is unlikely to succeed – unless they outsource data-heavy computation to neural networks.

From Neural to Neuro-Symbolic

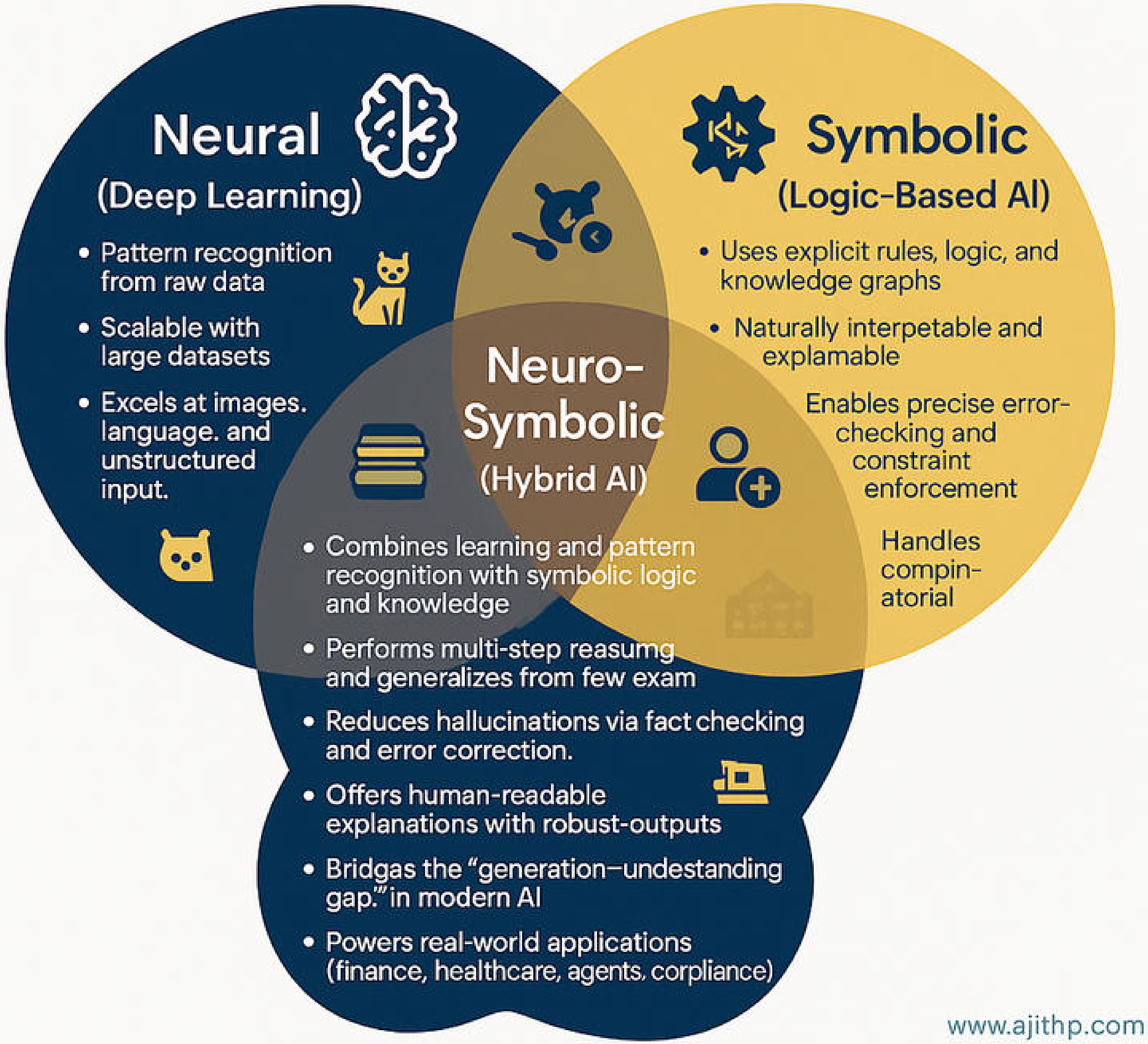
In the areas of vision and speech, in the last two decades, neural network-based learning began demonstrated outstanding performance in pattern recognition across computer vision (Krizhevsky, Sutskever, and Hinton 2012), natural language processing, and recommendation systems. This is owing to a cascade of improvements. Classically, neural networks were built consisting of input, hidden, and output layers with limited depth, but then "deep" learning architectures introduced multiple hidden layers, likely enabling the learning of deeper hierarchical representations (although these internal constructions are largely opaque to humans). There were also:

1. **Architectural improvements:** The development of convolutional neural networks (CNNs) for computer vision, recurrent neural networks (RNNs) (Goodfellow et al.

*Contact author: Vaishak Belle, University of Edinburgh, UK; email: vaishak@ed.ac.uk.
Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.



AI Paradigms: Neural, Symbolic, and Hybrid (Neuro-Symbolic)



Knowledge engineering is the disciplined process of designing, structuring, and maintaining explicit, machine-readable representations of domain knowledge to enable reliable reasoning and decision-making in AI systems.

- How LLMs can automate knowledge engineering
- How knowledge engineering can improve LLMs

Research Questions (RQs)

Evaluation

of ontologies

Which LLMs?

Which prompting techniques?

Which evaluation metrics?

Which benefits and weaknesses?

How to guarantee the reproducibility of the results?

Generation

of ontologies

Reproducibility

of the results

Use case: implicit knowledge (**metaphor**)

- How LLMs can automate knowledge engineering
- How knowledge engineering can improve LLMs

The problems of knowledge engineering

Manual and time-consuming

Error prone

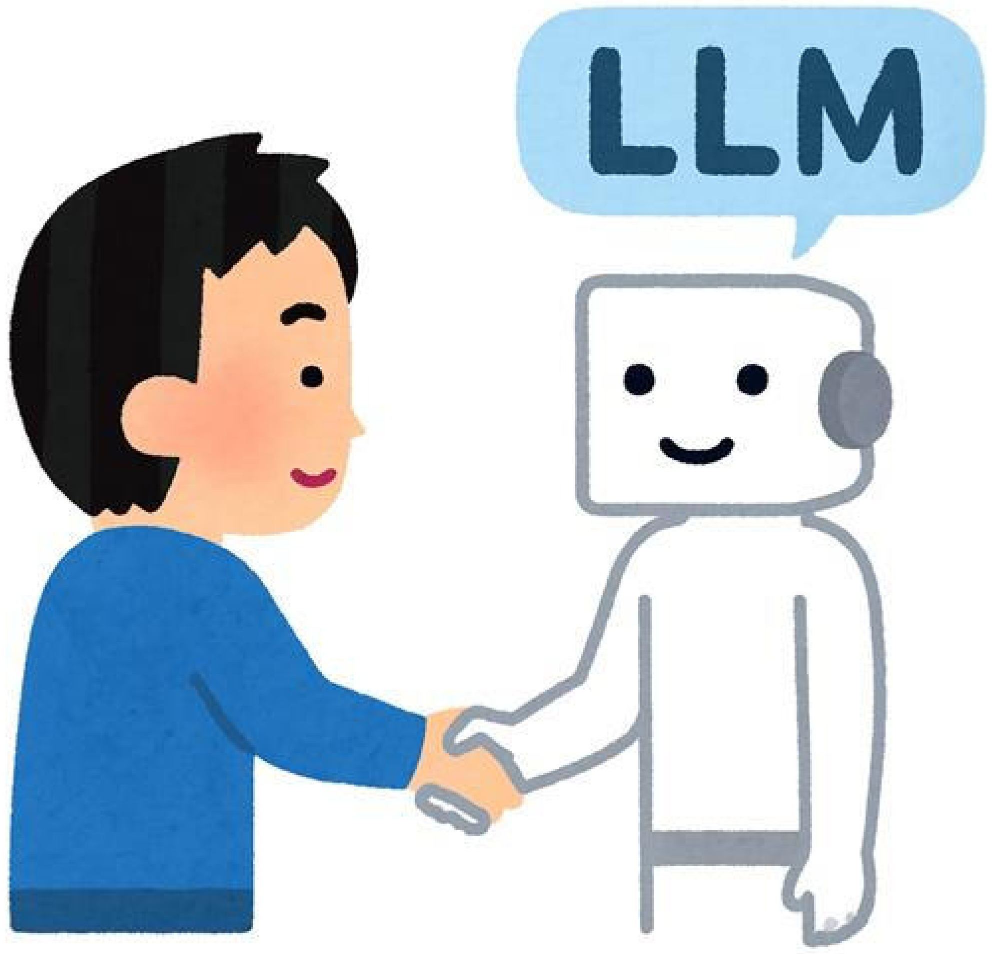
Challenging for novice

ontology engineers

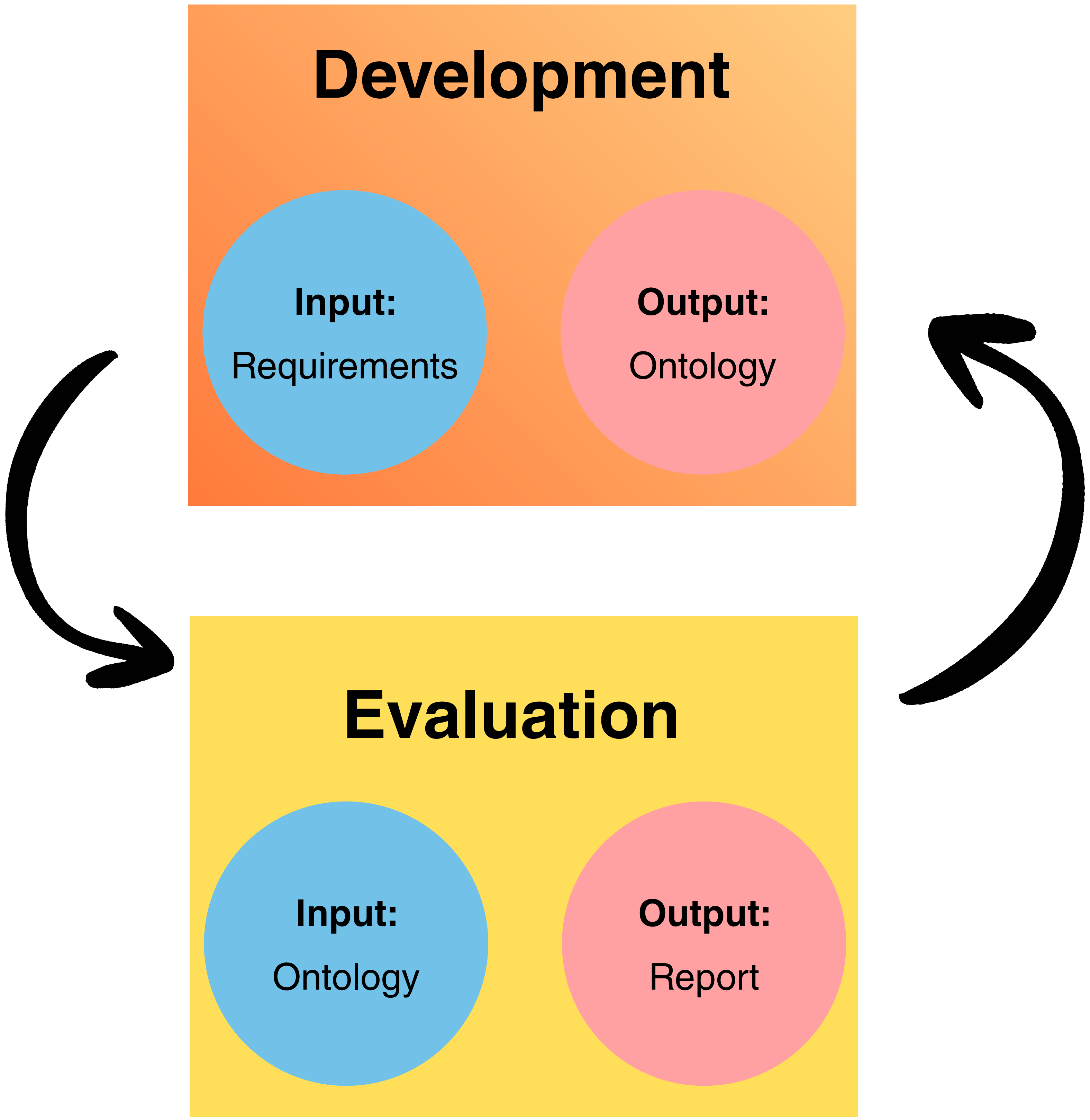
Ontology Engineering Assistant

w/o copilot

w/ copilot



Development and Evaluation



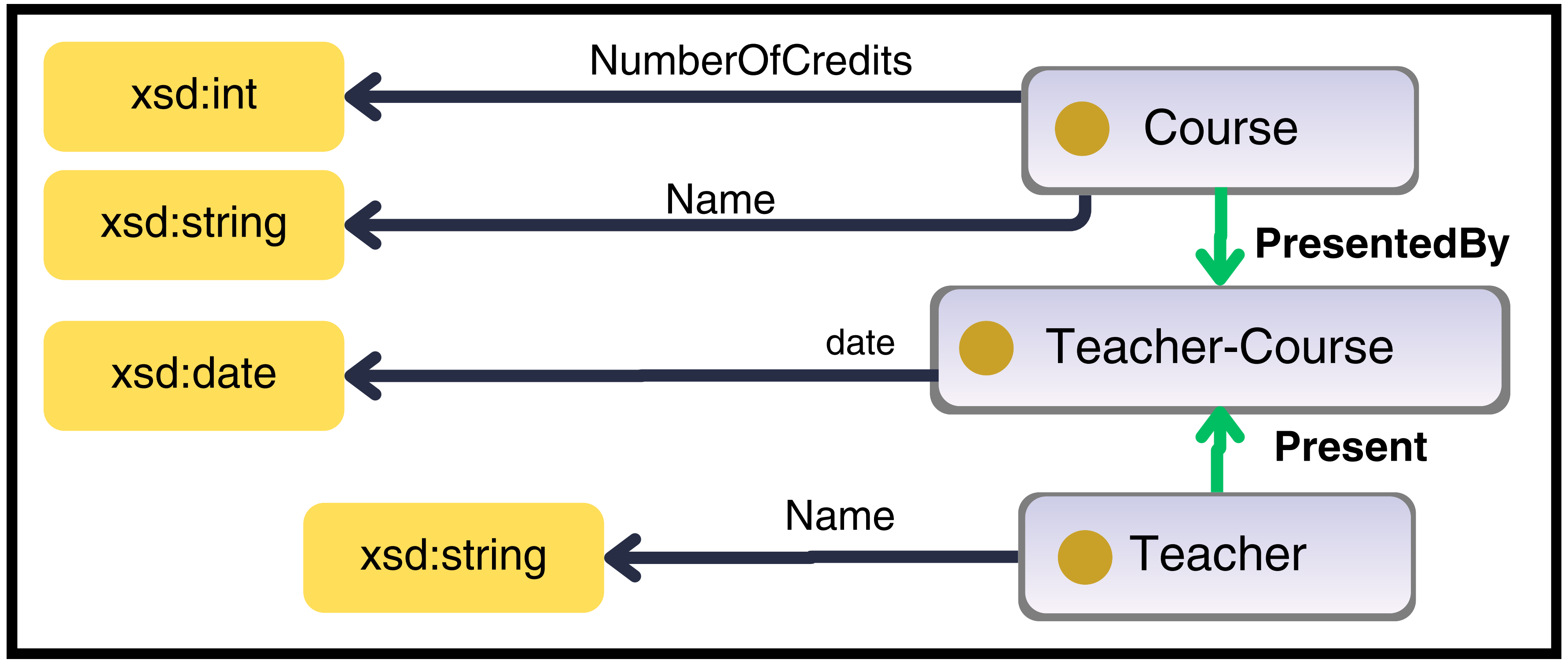
Requirements (Competency questions):

University setting:

1-Which teacher teaches a specific course?

2-Total number of credits a teacher presents this semester?

3-Give me a list of students in a course?



Report: Competency Questions 1 and 2 are modelled in the ontology but the third one is missing

Can LLMs generate ontologies?

Methodology

Prompting techniques

1- Memoryless CQbyCQ

2- Ontogenia

Dataset

- 100 CQs
- 29 user stories
- Gold standard ontologies

Ontology Generation Methods

1- Independent Ontology Generation

2- Incremental Ontology Generation

Evaluation

- OntOlogy Pitfall Scanner! (OOPS)
- CQ verification
- Expert evaluation

Results: OOPS!

Critical Pitfalls by OOPS!

Prompting Techniques

LLMs

		MemorylessCQbyCQ			Ontogenia		
		GPT-4	Llama*	o1	GPT-4	Llama*	o1
P05	Wrong inverse relationships	1	25	0	2	7	5
P06	Cycles in a class hierarchy	0	2	0	5	0	11
P19	Multiple domains or ranges	23	32	1	4	15	0
P29	Wrong transitive relationship	0	0	1	0	0	0
P37	Ontology not available	2	0	0	0	0	0
P39	Ambiguous namespace	2	0	1	1	0	0

Takeaways:

Llama (3.1-405 B-instruct-bf16)→ high number of pitfalls.

o1-preview → yields the least number of pitfalls

Functional evaluation through CQ verification

A competency question is modelled by an ontology if you could write a SPARQL query to extract the answer (Blomqvist et al., 2012)

Story:
The context is about university courses given by different teachers.

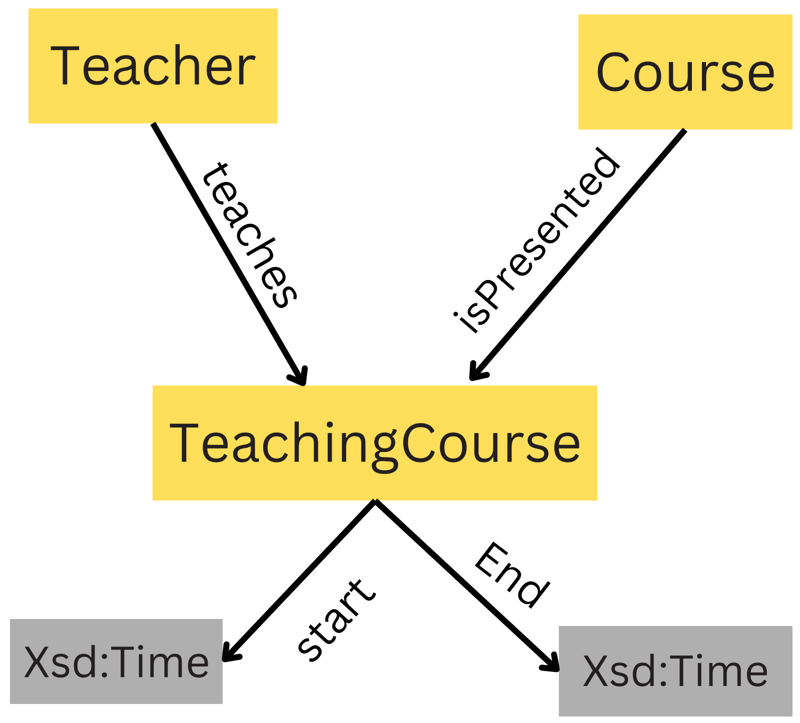
Competency Question:
What courses did Eva teach in this period?

Functional evaluation through CQ verification

A competency question is modelled by an ontology if you could write a SPARQL query to extract the answer (Blomqvist et al., 2012)

Story:
The context is about university courses given by different teachers.

Competency Question:
What courses did Eva teach in this period?

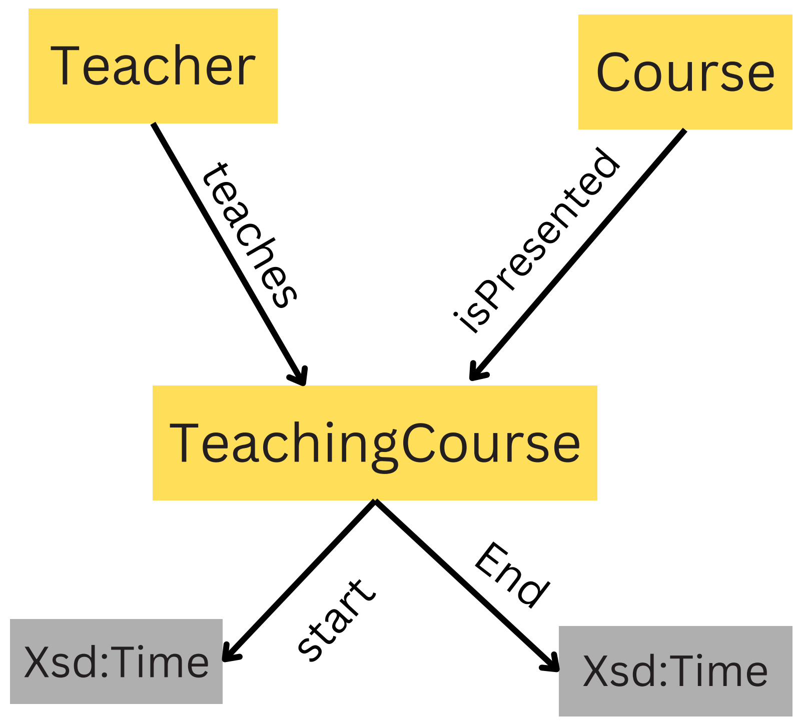


Functional evaluation through CQ verification

A competency question is modelled by an ontology if you could write a SPARQL query to extract the answer (Blomqvist et al., 2012)

Story:
The context is about university courses given by different teachers.

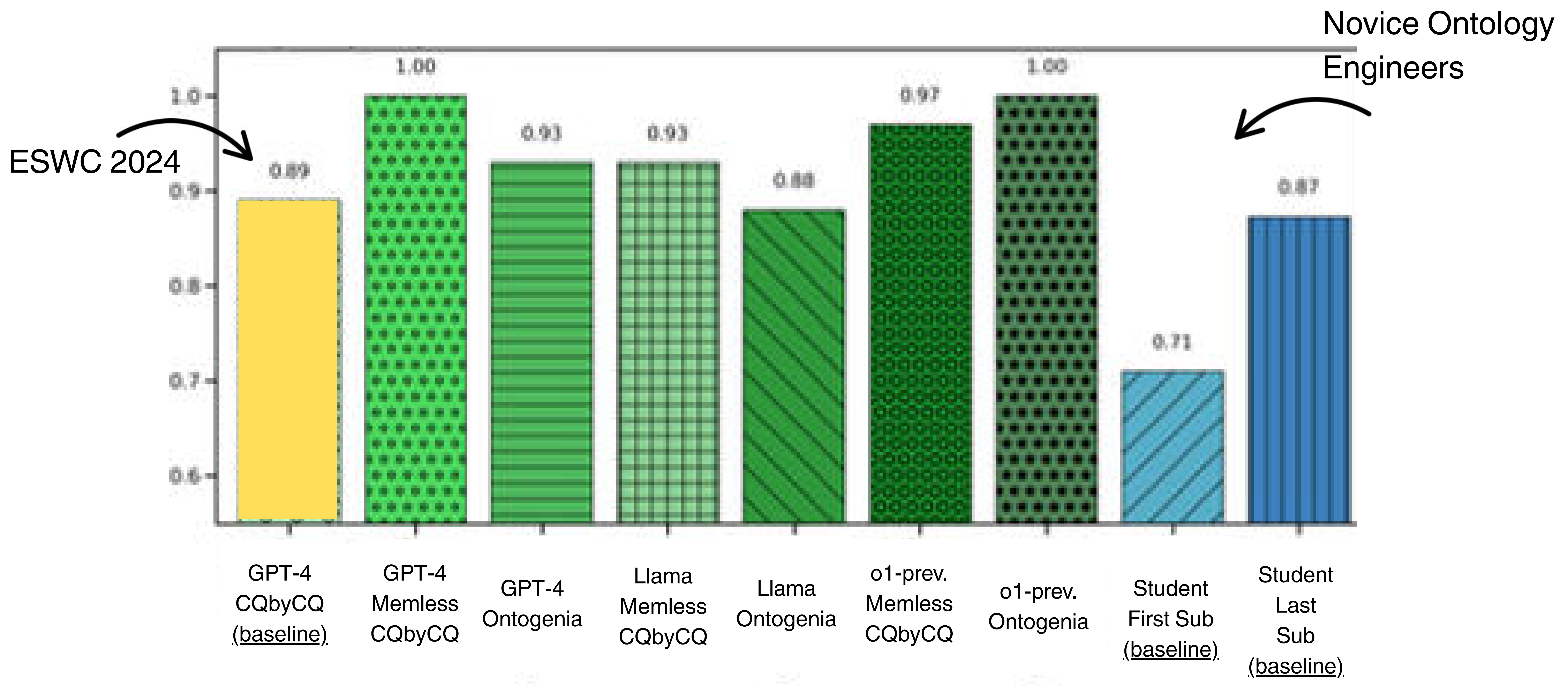
Competency Question:
What courses did Eva teach in this period?




```
SELECT ?course
WHERE {
  ?course ex:isPresented ?tc .
  ?teacher ex:teaches ?tc .
  ?tc ex:start ?start ;
      ex:end ?end .


}
```

Results: CQ verification



Proportion of modelled CQs

Results: Experts Evaluation

1- No label and comments on some classes and properties.

2- Loops in the taxonomy.

3- Multiple domains and ranges for some properties.

4- Unnecessary classes.

5- Duplicate classes or properties.

6- Not all CQs are properly modelled.

7- Axioms are poorly defined for classes or no axioms at all.

8- Bad taxonomy or missing taxonomy.

Detectable by **OOPS!**

Detectable by **CQ verification**

Lippolis, Anna Sofia, Mohammad Javad Saeedizade, Robin Keskisärkkä, Sara Zuppiroli, Miguel Ceriani, Aldo Gangemi, Eva Blomqvist, and Andrea Giovanni Nuzzolese. "Ontology generation using large language models." In *European Semantic Web Conference*, pp. 321-341. Cham: Springer Nature Switzerland, 2025.

Read the paper

GitHub repository

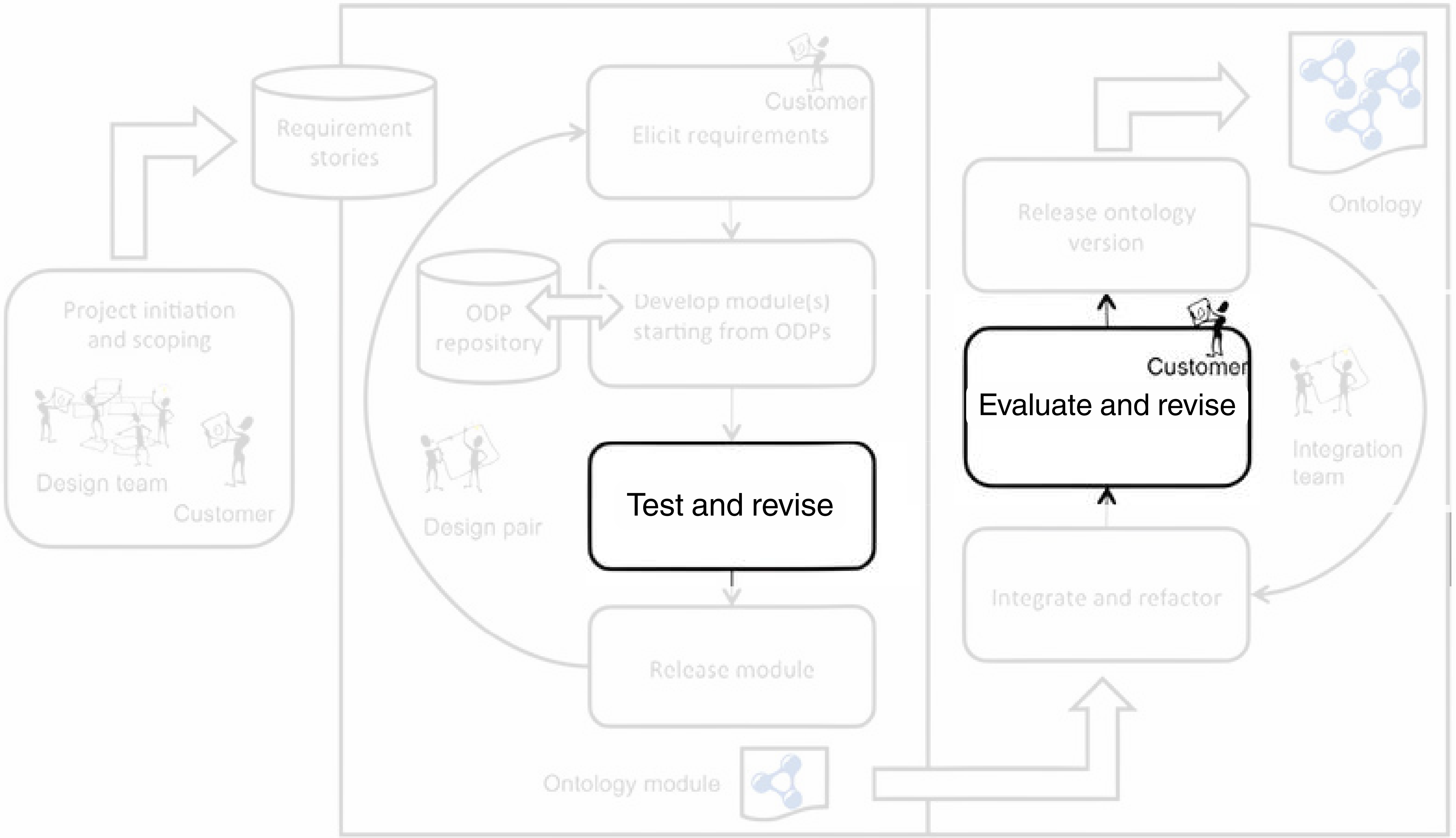
LINKÖPINGS UNIVERSITET

ALMA MATER STUDIORUM A.D. 1088 UNIVERSITÀ DI BOLOGNA

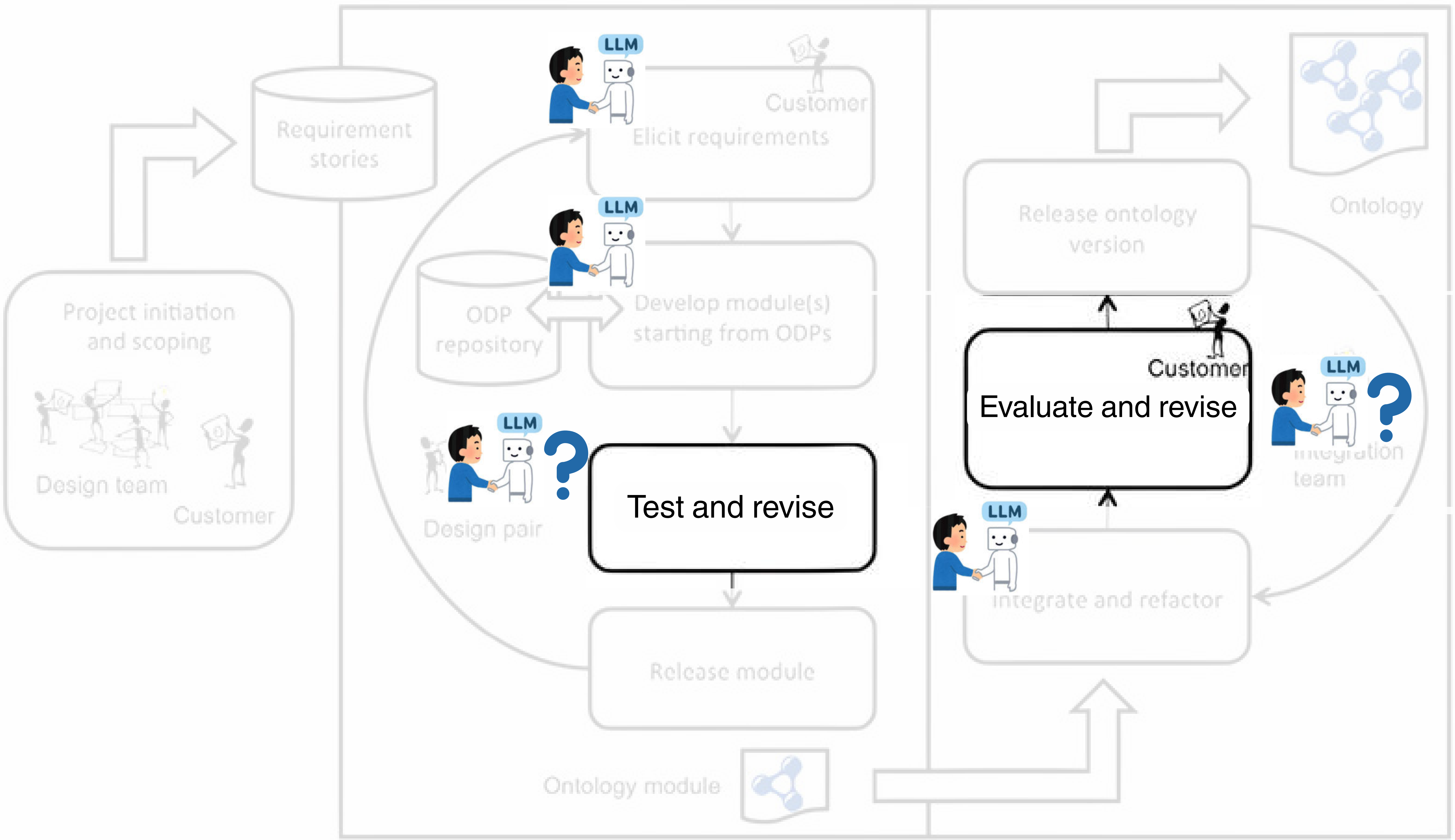
Italian National Research Council

Can LLMs evaluate ontologies?

The problem of evaluation



LLMs in the loop?

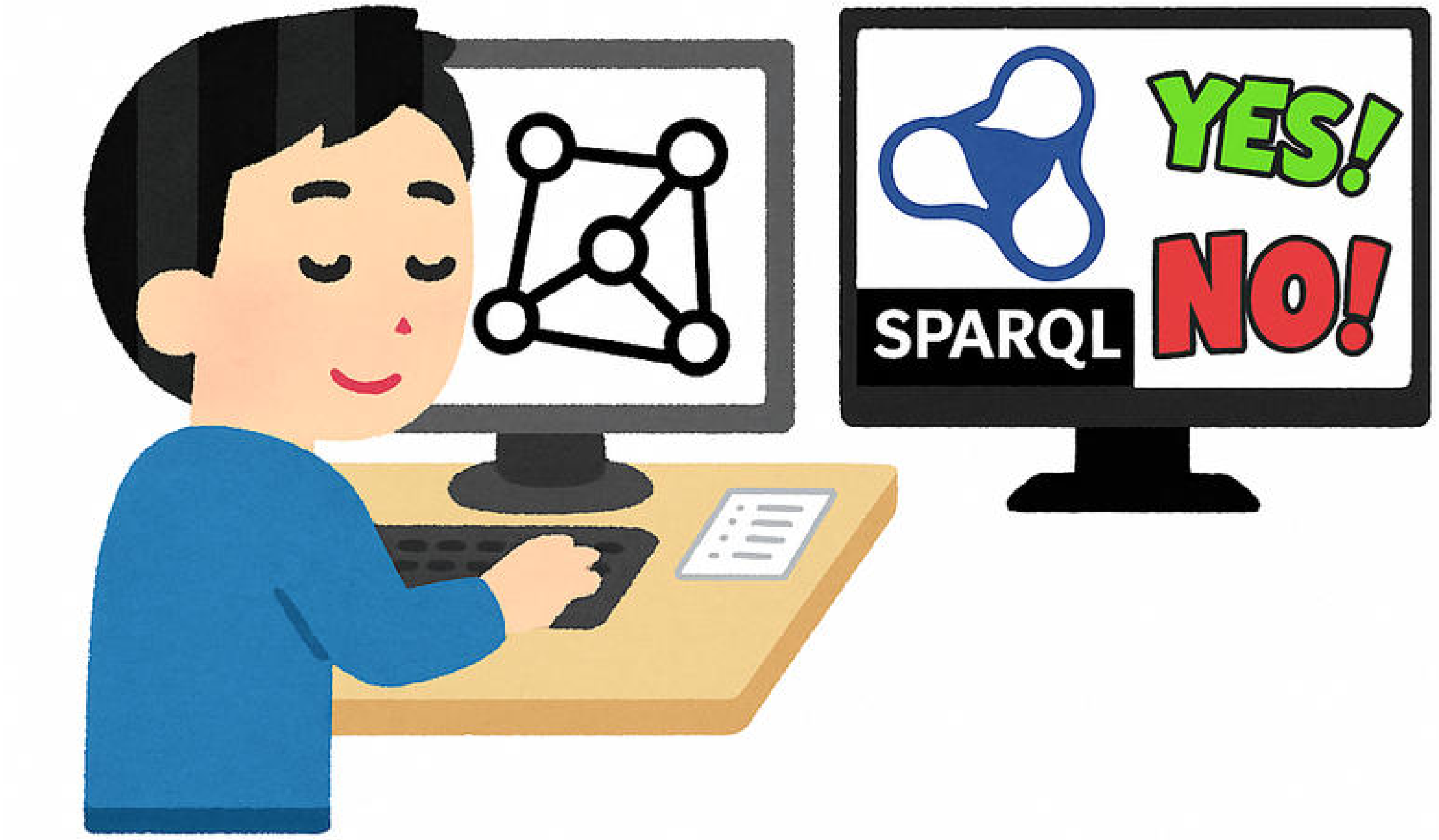


Functional evaluation through CQ verification

The user wants to know which competency questions are properly modelled in the ontology.

Unassisted mode

Assisted mode



Research questions

RQ 1 To what extent can LLMs **evaluate** ontologies using CQ verification?

RQ 2 To what extent can LLMs **assist** ontology engineers in evaluating ontologies through CQ verification, and what are the benefits and drawbacks of a hybrid approach combining LLM suggestions with expert validation compared to traditional human-only methods?

Contributions

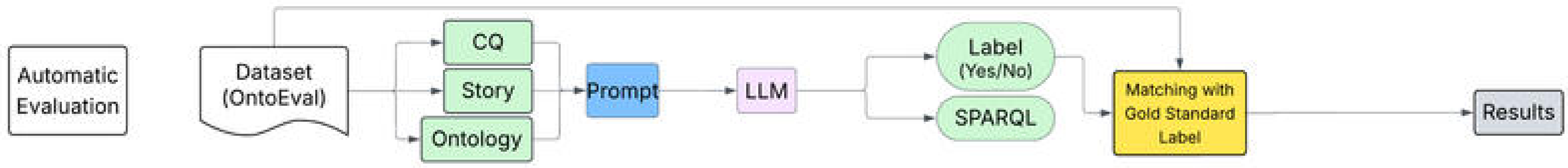
- New, dedicated dataset for ontology evaluation
- Automatic and semi-automatic CQ verification

The OntoEval Datasets

Metric	OntoEval	OntoEval-small
Total CQs	1,393	20
Modelled CQs	1,204	10
Difficulty: simple	725	0
Difficulty: complex	135	20
Domains	33	6

RQ 1

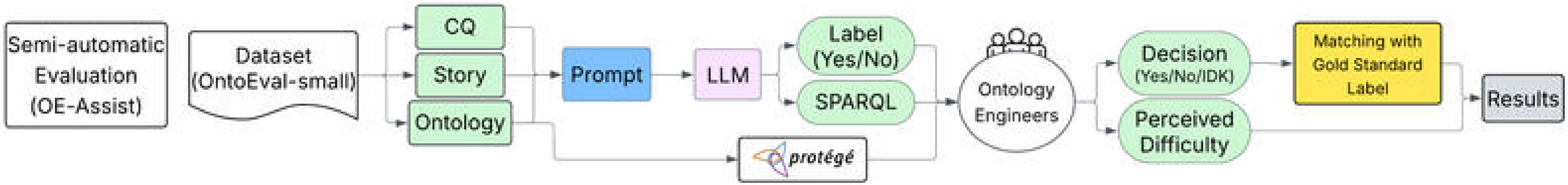
Automatic evaluation



- Models: o1-preview, o3-mini, GPT-4o-0513
- Metric: macro-F1 on the full set and accuracy on the balanced set

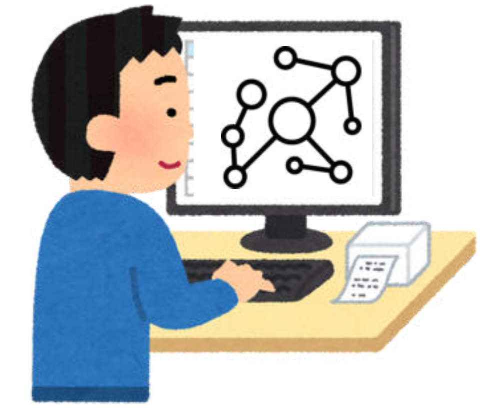
RQ 2

Semi-automatic evaluation

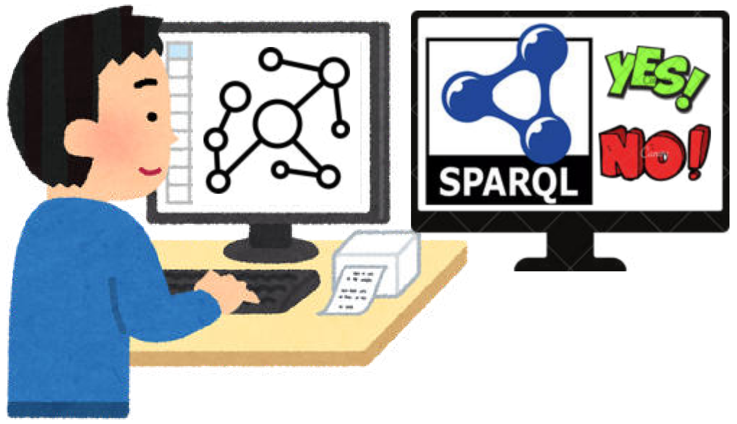


- 19 users either expert and non-expert in ontology engineering
- 20 CQ verification tasks, 10 with LLM suggestions, 10 with not
- Answers were *Yes*, *No*, or *I don't know*, followed by a 1–5 difficulty rating
- The order of the tasks was randomized every time
- A survey was given to obtain more information from the users

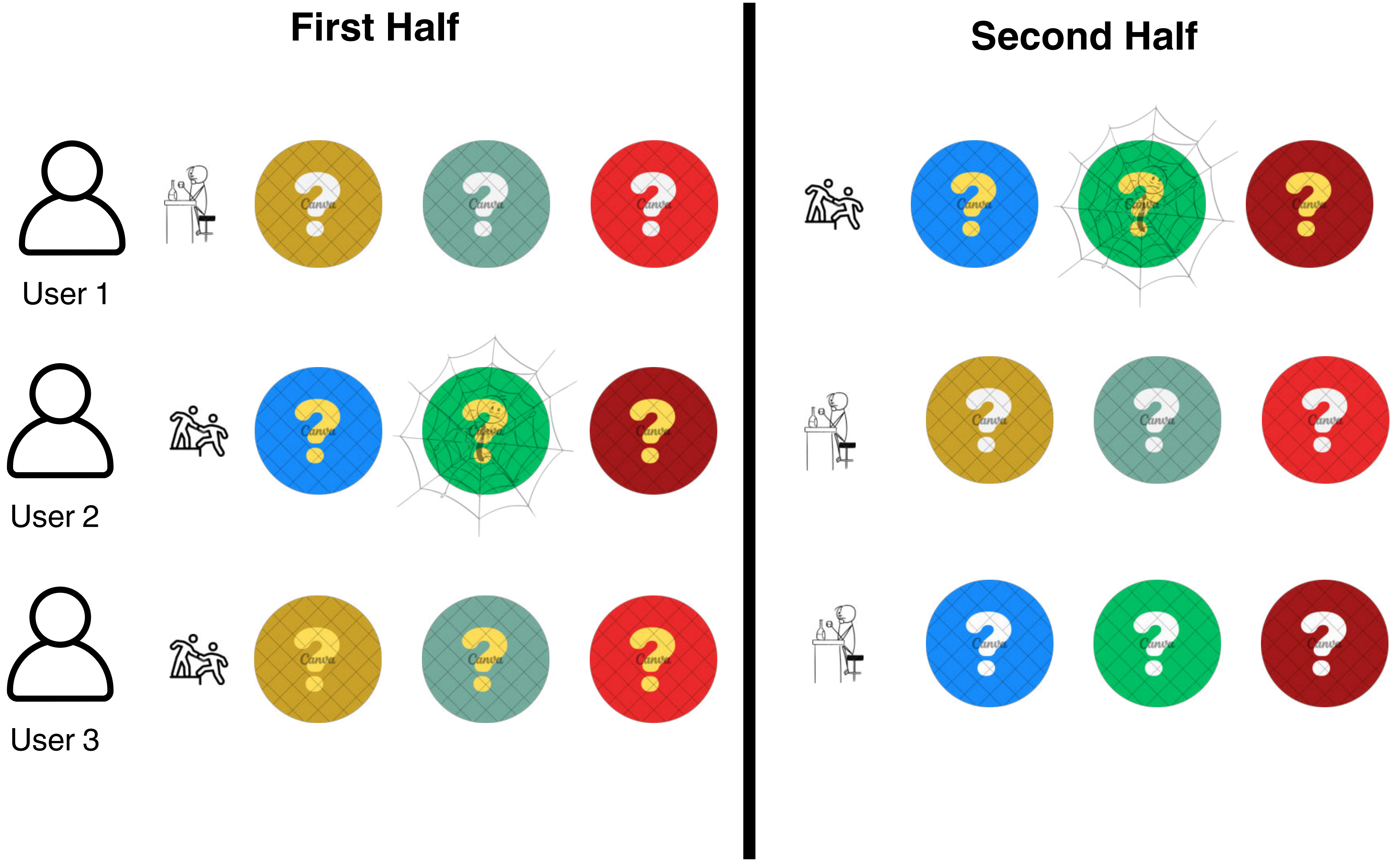
Ontology Evaluation

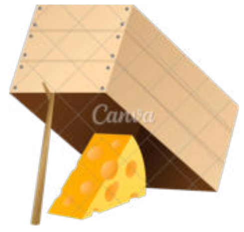


Unassisted mode



Assisted mode





Not all assisted answers were correct. We wanted to know the effect of wrong answers on people's decisions.

Semi-automatic evaluation: setup (assisted)

CQ: When is the level of a chemical substance recorded in a water body?

Scenario: Context is about monitoring water quality by recording chemical substance levels to investigate potential links between contaminated water sources and adverse public health outcomes.

LLM suggested answer: Yes

```
SELECT ?chemicalSubstance ?waterBody ?date
WHERE {
  ?clr a :ChemicalLevelRecording .
  ?clr :hasRecordingDate ?date .
  ?clr :recordsChemical ?chemicalSubstance .
  ?clr :recordedIn ?waterBody .
}
```

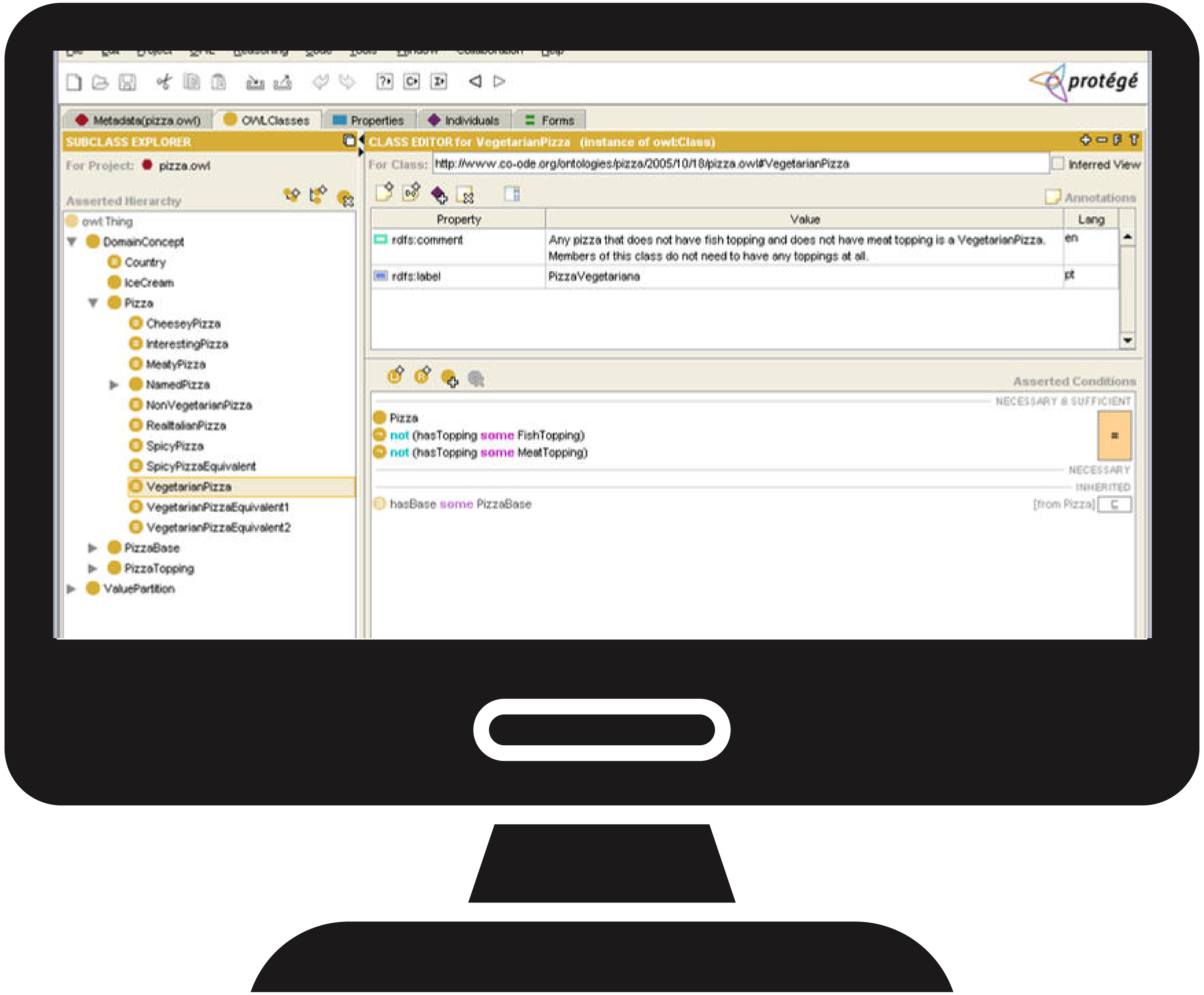
Is the CQ modelled?

Yes	No	I don't know

How difficult was the CQ for you to solve?

(1: Very easy to 5: Very hard)

1	2	3	4	5



Semi-automatic evaluation: setup (unassisted)

CQ: When is the level of a chemical substance recorded in a water body?

Scenario: Context is about monitoring water quality by recording chemical substance levels to investigate potential links between contaminated water sources and adverse public health outcomes.

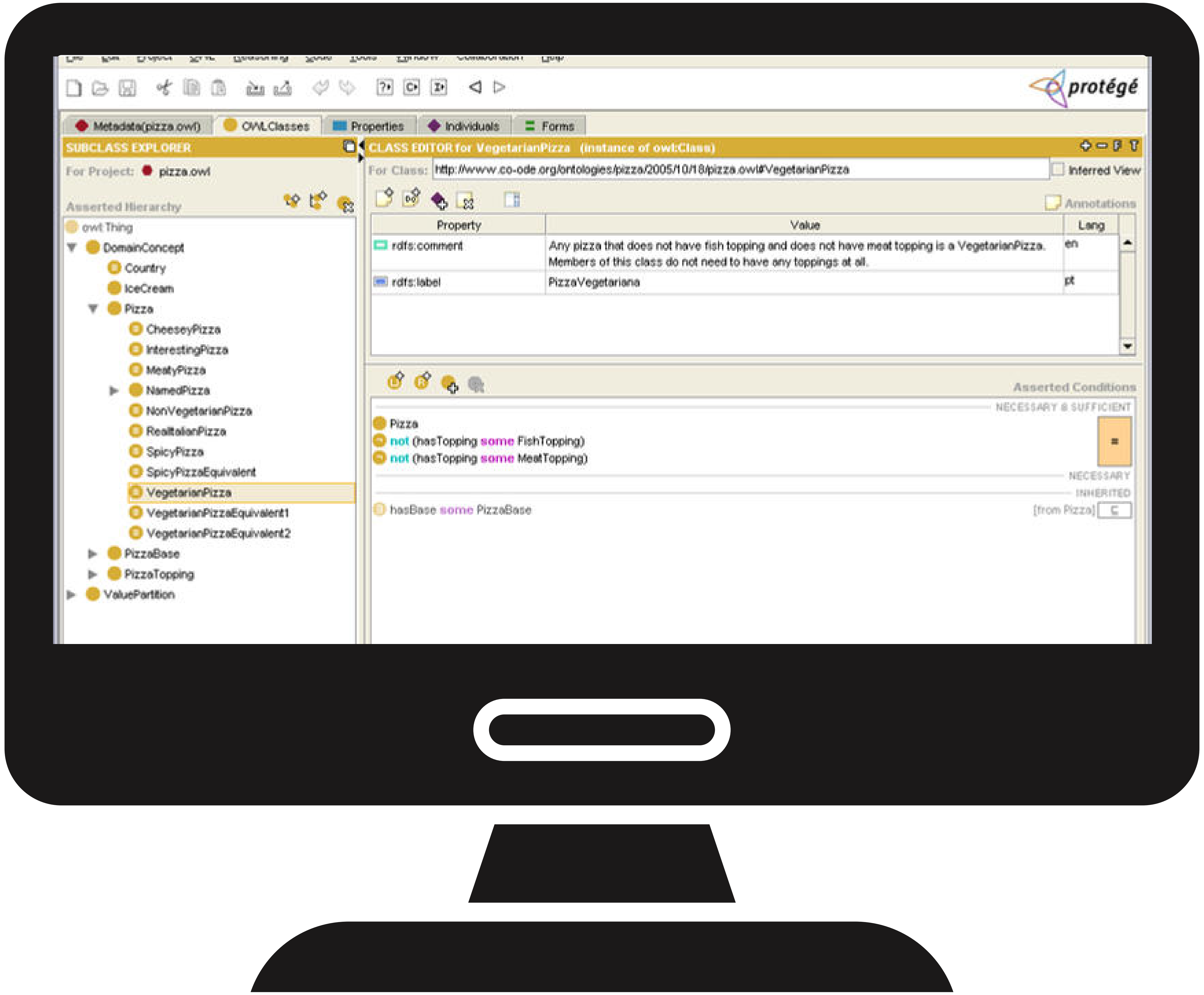
Is the CQ modelled?

Yes	No	I don't know

How difficult was the CQ for you to solve?

(1: Very easy to 5: Very hard)

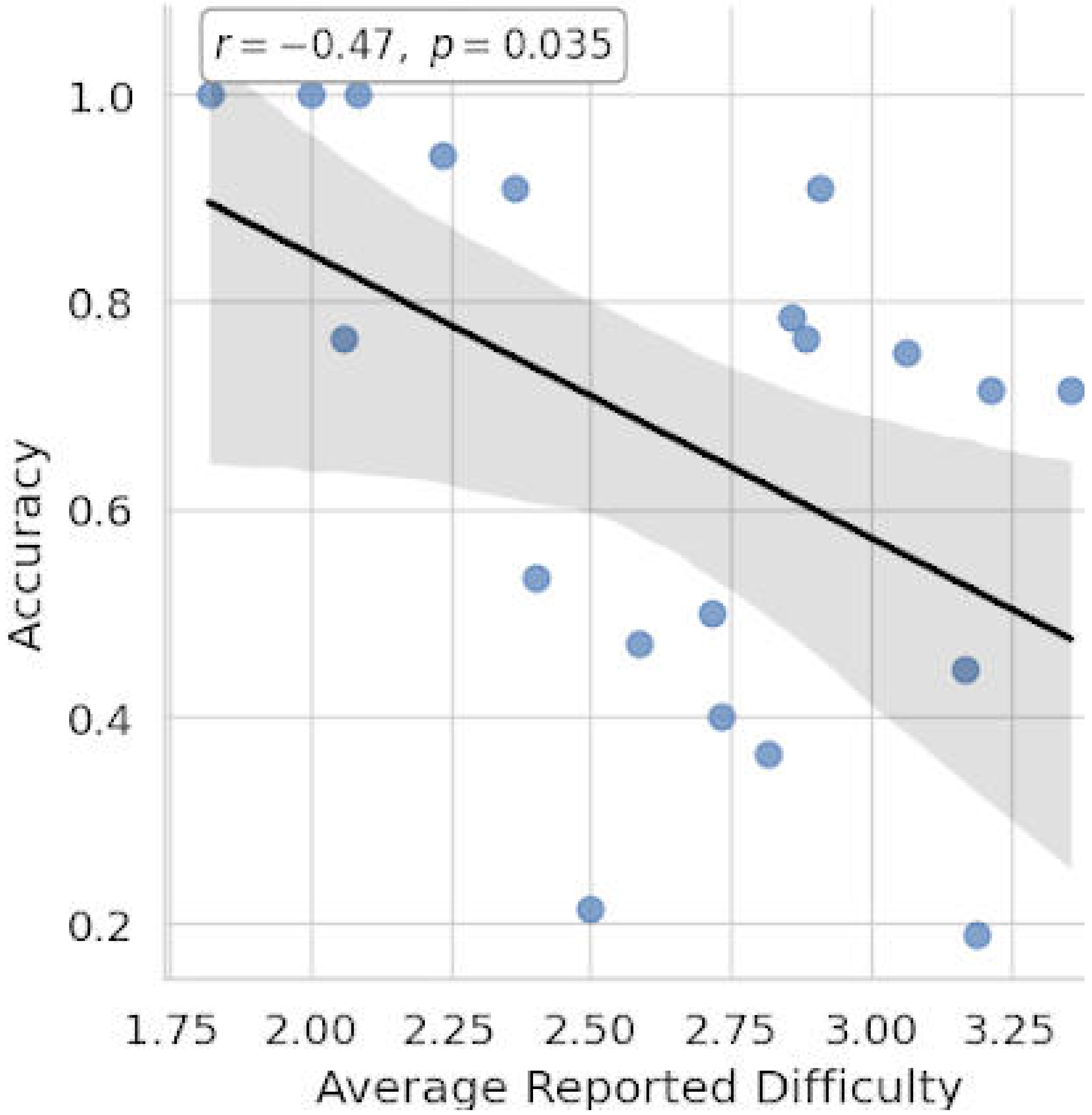
1	2	3	4	5



Automatic evaluation: results

Model	Macro-F_1	Accuracy
Dataset	OntoEval	OntoEval-small
Random baseline	0.43	0.50
GPT-4o-0513	0.48 ± 0	0.55 ± 0
o3-mini	0.58 ± 0.01	0.72 ± 0.02
o1-preview	**0.66**	**0.75** ± 0.05

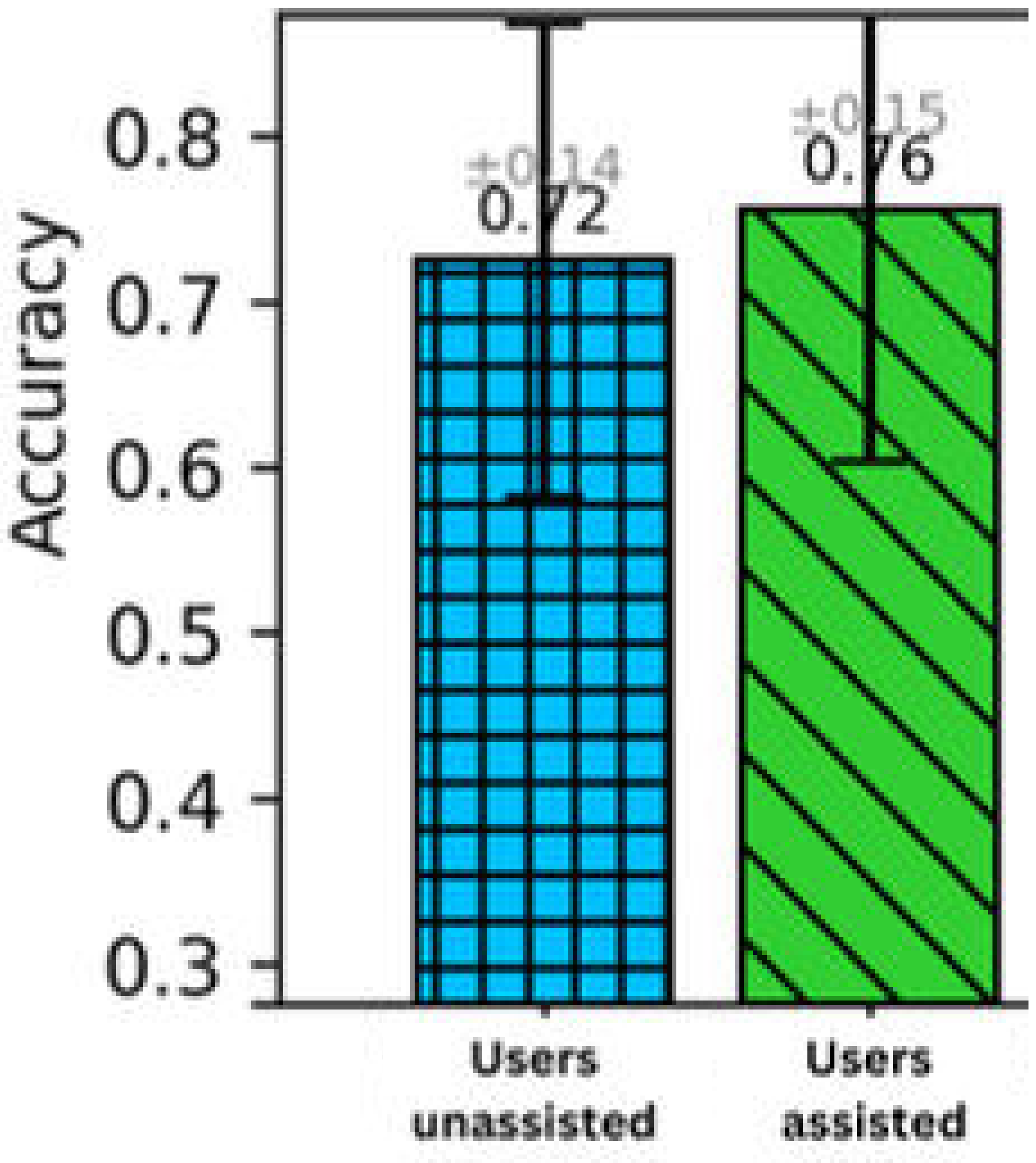
Semi-automatic evaluation: results



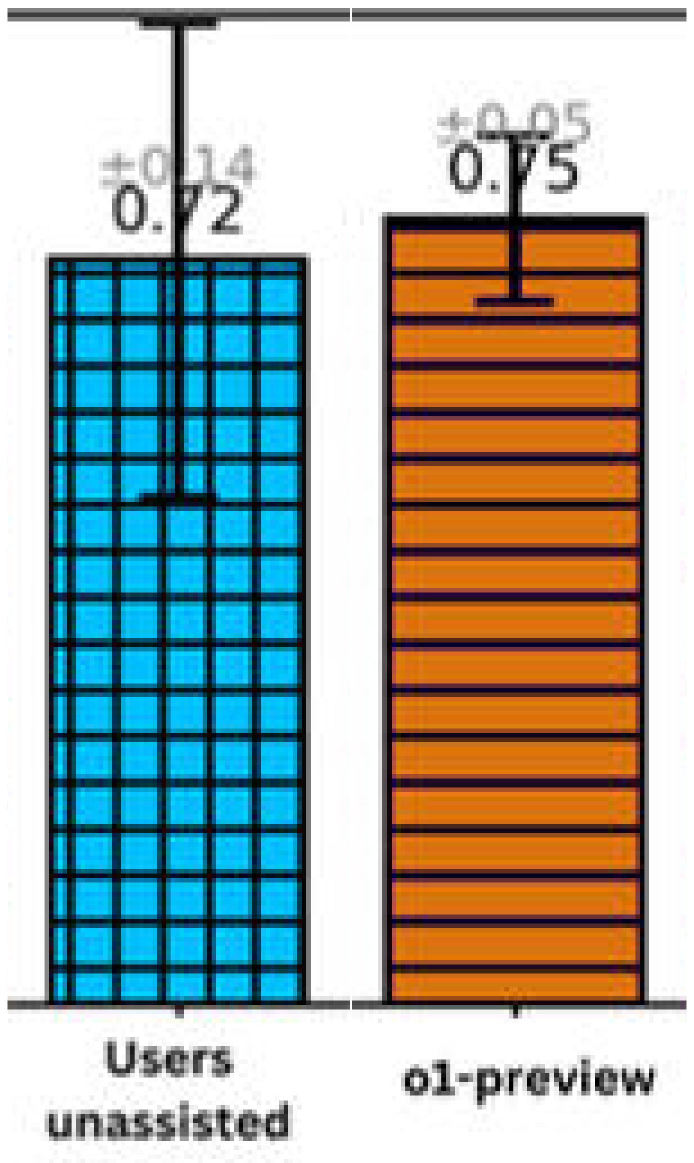
- Perceived difficulty decreased with assistance
- Most participants found the suggestions useful and easy to learn, but about a fifth reported distraction, especially when CQs or labels were unclear

Semi-automatic evaluation: results

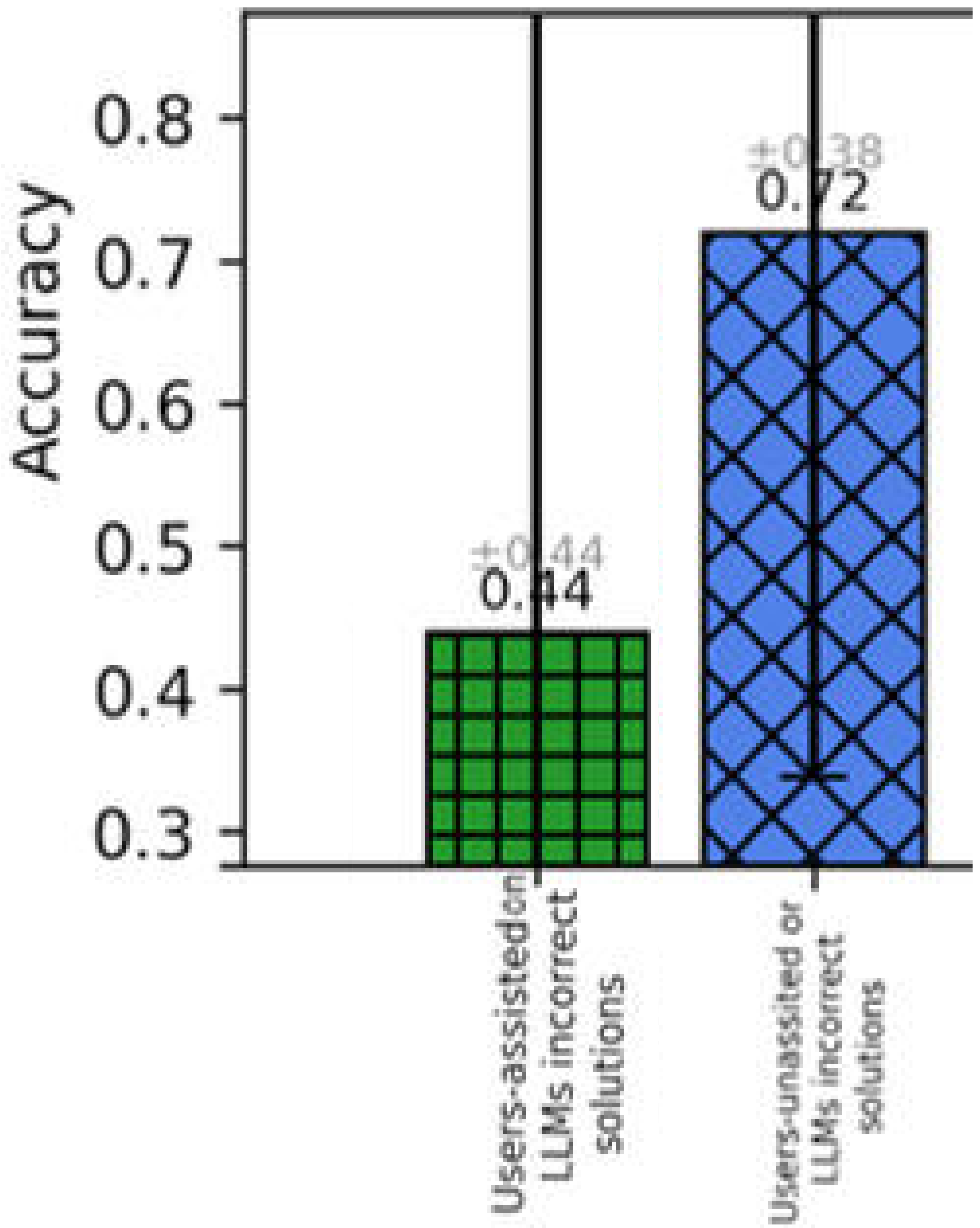
Overall performance did not change significantly.



No significant difference between o1-preview and Ontology Engineers.

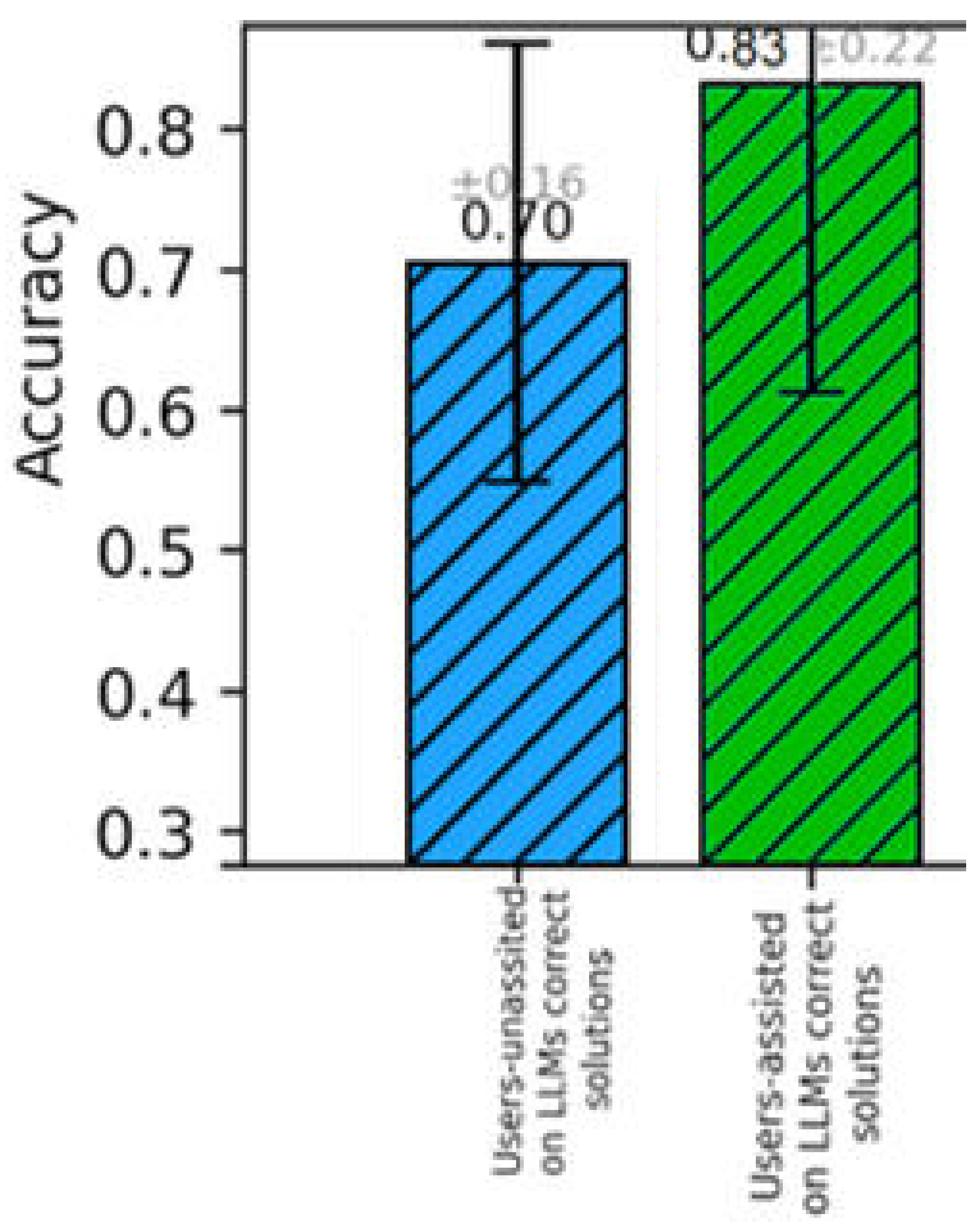


When LLM suggestions are incorrect...



...Accuracy drops.

When LLM suggestions are correct...



...Accuracy rises

The Tradeoff

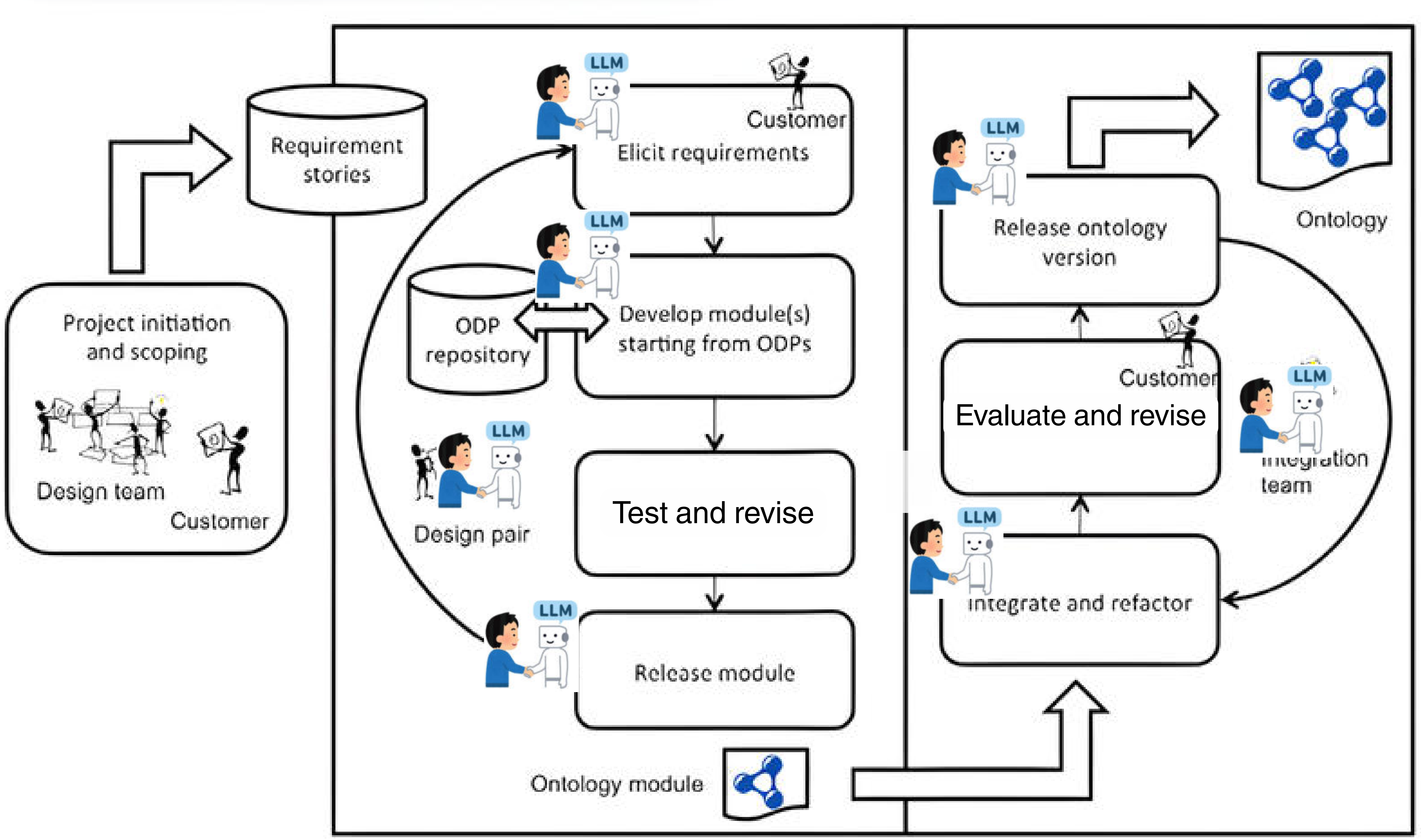
- Efficiency and lower perceived difficulty come with the risk of **over-trust**.
- Two mitigations:
    - pre-check suggested SPARQL against the actual ontology before showing the label
    - show a calibrated confidence or provenance for suggestions

Quality of the requirements matters: domain-specific CQ wording quality and ontology documentation strongly affect outcomes.

The assisted development of ontologies

1- Difficult for both ontology engineers and LLMs--even though for different reasons

2- LLMs still need a human-in-the-loop for accurate results



Lippolis, Anna Sofia, Mohammad Javad Saeedizade, Robin Keskisärkkä, Aldo Gangemi, Eva Blomqvist, and Andrea Giovanni Nuzzolese. "Large Language Models Assisting Ontology Evaluation." In International Semantic Web Conference, pp. 502-520. Cham: Springer Nature Switzerland, 2025.

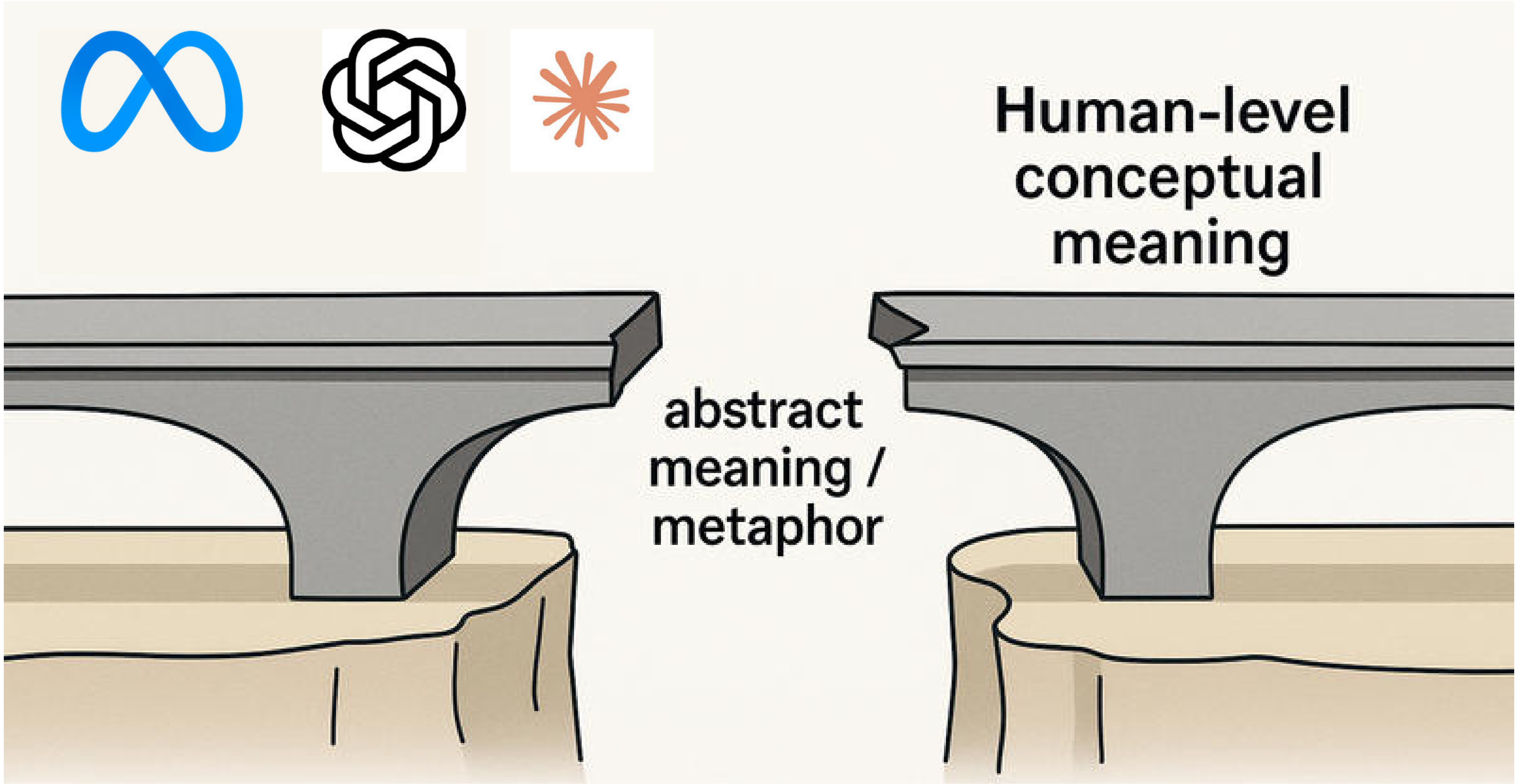
Read the paper

GitHub repository

Italian National Research Council

- How LLMs can automate knowledge engineering
- How knowledge engineering can improve LLMs

The case of metaphor




NHS UNDER SIEGE
The fight to save it in the age of Covid
JOHN LISTER
JACKY DAVIS
PLUS EXPERT CONTRIBUTORS
FOREWORD BY MICHAEL ROSEN
Daniele Cassandro
siamo davvero in guerra
?
rmsten besonders
beiter in Schlachthöfen. Warum trifft
nd? VON MARIA FIEDLER
taz
POLITIK ÖKO
Corona im Idunazentrum Göttingen
Von Prestigeobjekt
In Göttingen häuften sich vor einer W
Fokus der Boulevard-Presse: das Iduna
Bild MENÜ
VIELE BEWOHNER INFIZIERT, MIETE
GST AUS
Aufruhr im Corona-Block in Göttingen
ronavirus
orona-Ausbruch in Neukölln: „Das
st kein Rumänenhaus“
CORRIERE FIORENTINO
LA TOSCANA
corrierefiorentino.it
Caro direttore, dopo aver letto l'intervento del professor Sergio Romagnani sul Corriere Fiorentino di ieri mi farebbe piacere che poteste ospitare anche alcune mie considerazioni in merito a quanto trattato nell'articolo. La Sanità Pubblica non è una scienza esatta, come del resto non lo è la Medicina Clinica.
L'INTERVENTO
IL COVID È UNA GUERRA E PER VINCERLA SERVE OGNI VACCINO POSSIBILE
Ma, ancor più di quest'ultima, ha sempre agito per approssimazioni progressive, spesso trovando ai problemi soluzioni empiriche, i cui meccanismi e ragioni furono compresi molto tempo dopo. Lo stesso Edward Jenner scoprì il primo «vaccino», quello
professor Romagnani. Tanto più che i benefici sono
rilevanti nelle diverse fasce di età. E i dati sul campo non
COVID-19
FROM CHAOS TO CURE
HE BIOLOGY BEHIND
HE FIGHT AGAINST THE
OVEL CORONAVIRUS
ALE YUZUKI, M.A., M.Ed.
Fighting the Virus
How to boost your body's immune response and fight viruses and bacteria naturally
Joseph Veebe


NHS
UNDER SIEGE
The fight to save it in the age of Covid
JOHN LISTER
JACKY DAVIS
PLUS EXPERT CONTRIBUTORS
FOREWORD BY MICHAEL ROSEN
taz
Corona im Idunazentrum Göttingen
Von Prestigeobjekt
VIELE BEWOHNER INFIZIERT, MIETE
Aufruhr im Corona-Block in Göttingen
CORRIERE FIORENTINO
LA TOSCANA
L'INTERVENTO
IL COVID È UNA GUERRA E PER VINCERLA SERVE OGNI VACCINO POSSIBILE
Caro direttore, dopo aver letto l'intervento del professor Sergio Romagnani sul Corriere Fiorentino di ieri mi farebbe piacere che poteste ospitare anche alcune mie considerazioni in merito a quanto trattato nell'articolo. La Sanità pubblica non è una scienza esatta, come del resto non lo è la Medicina Clinica.
Daniele Cassandro
siamo davvero in guerra
?
Corona-Ausbruch in Neukölln: „Das kein Rumänenhaus“
Fighting the Virus
How to boost your body's immune response and fight viruses and bacteria naturally
Joseph Veebe
COVID-19
FROM CHAOS TO CURE
BIOLOGY BEHIND FIGHT AGAINST THE NOVEL CORONAVIRUS

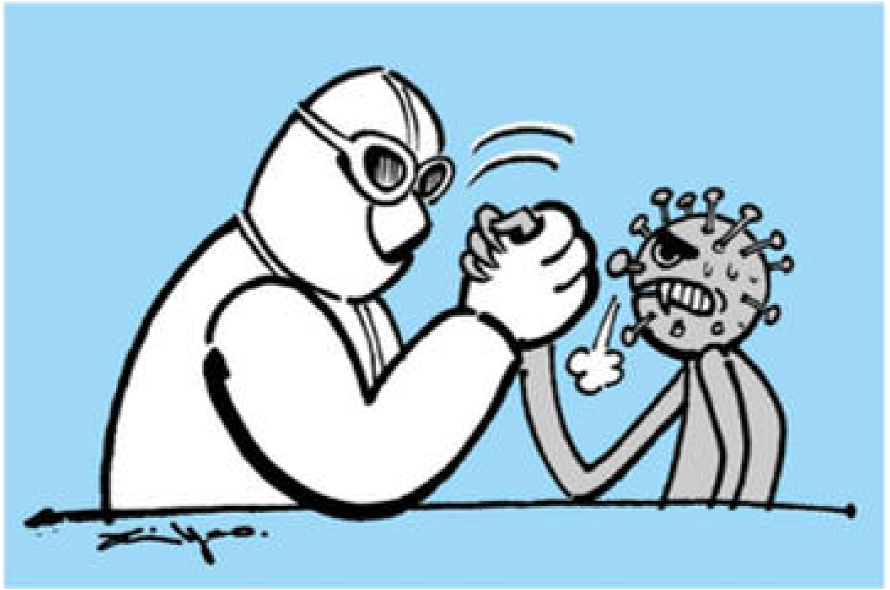


Figure 1: Must-win battle.

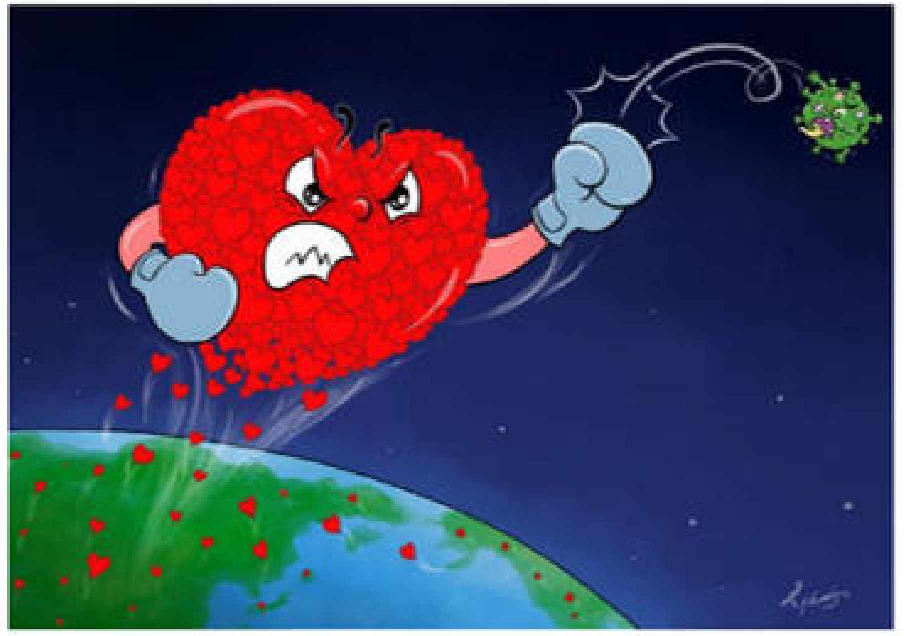
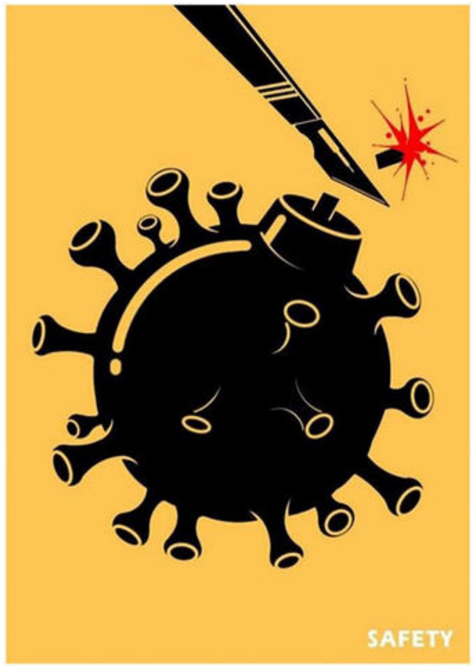


Figure 2: Fighting from the heart.

Figure 3: Unity remains key to virus battle.



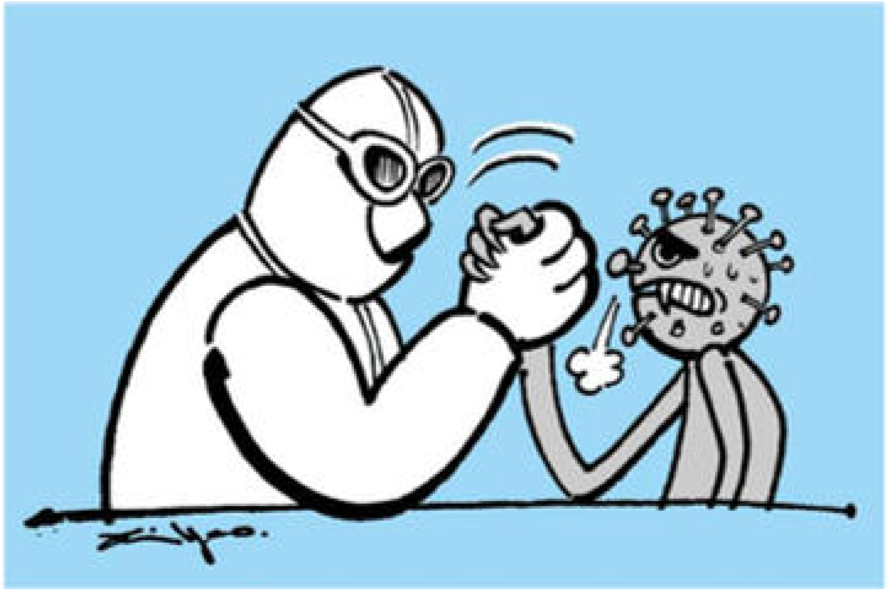


Figure 1: Must-win battle.

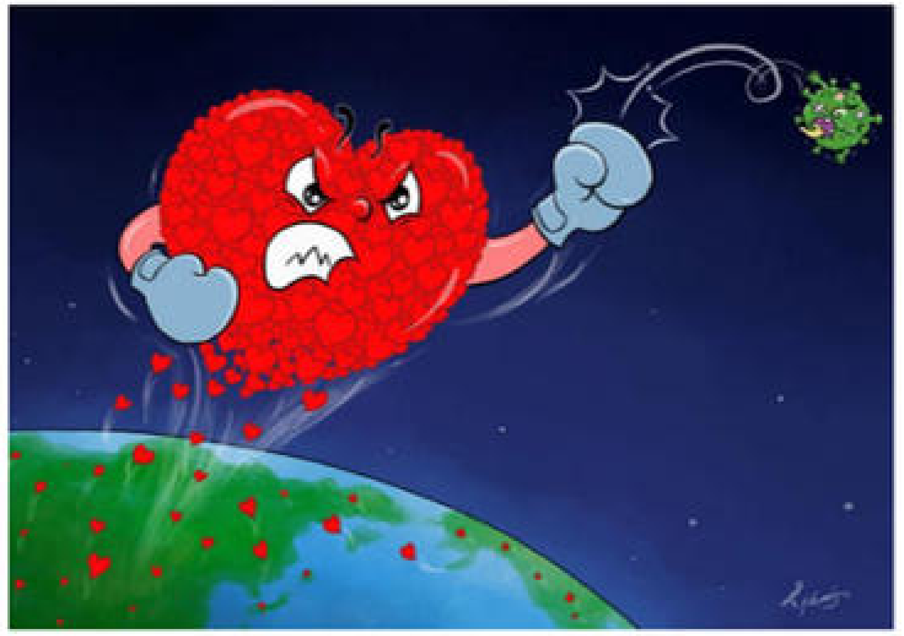
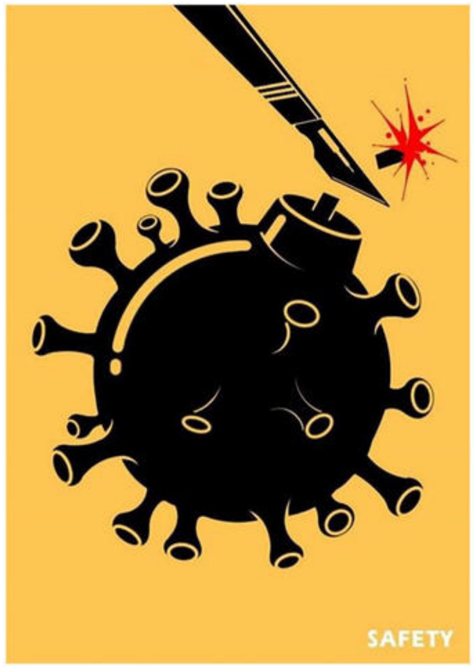


Figure 2: Fighting from the heart.

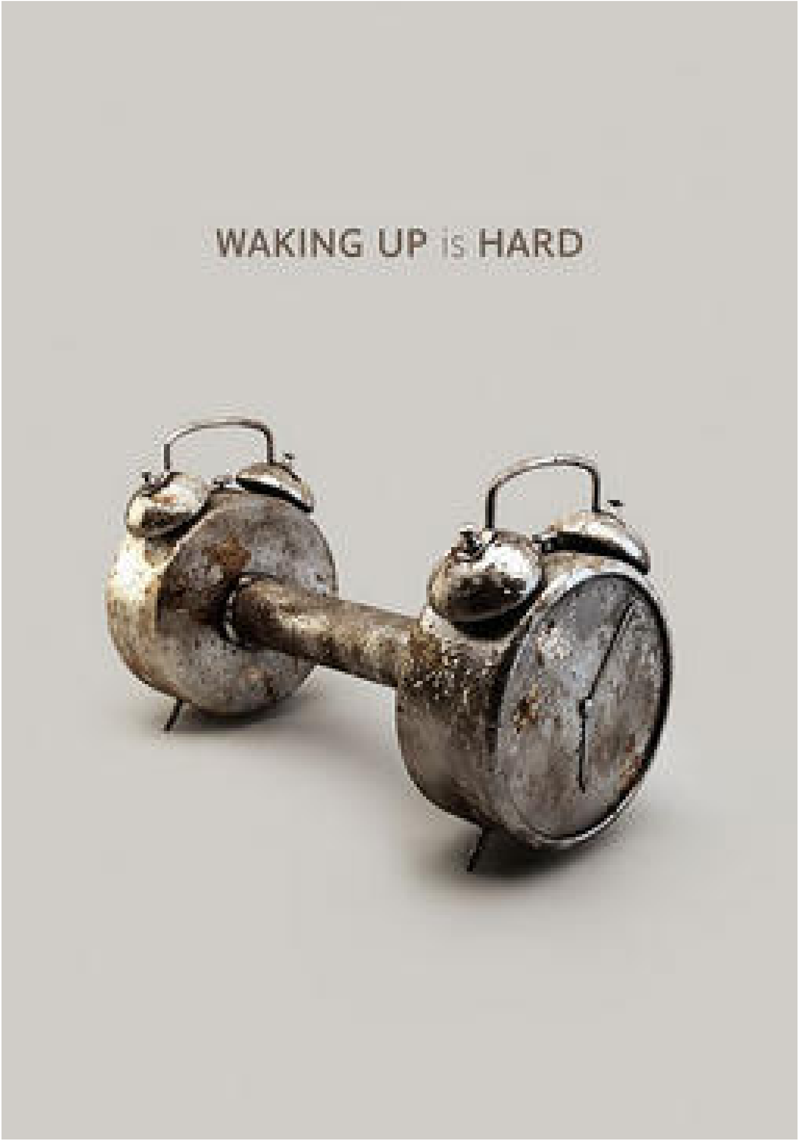
Figure 3: Unity remains key to virus battle



Title: one of them is my alt

Caption: Meme poster appreciates their only two followers and one of them is their alternative account.

Figure 1: A meme and its title. The caption describes what the meme poster was trying to convey.



weapon
law enforcement
victim
agent
place
CRIME
target domain

*The **crime epidemic** is affecting our cities*

source domain

DISEASE
type
source
agent
patient
doctor

What is a metaphor?

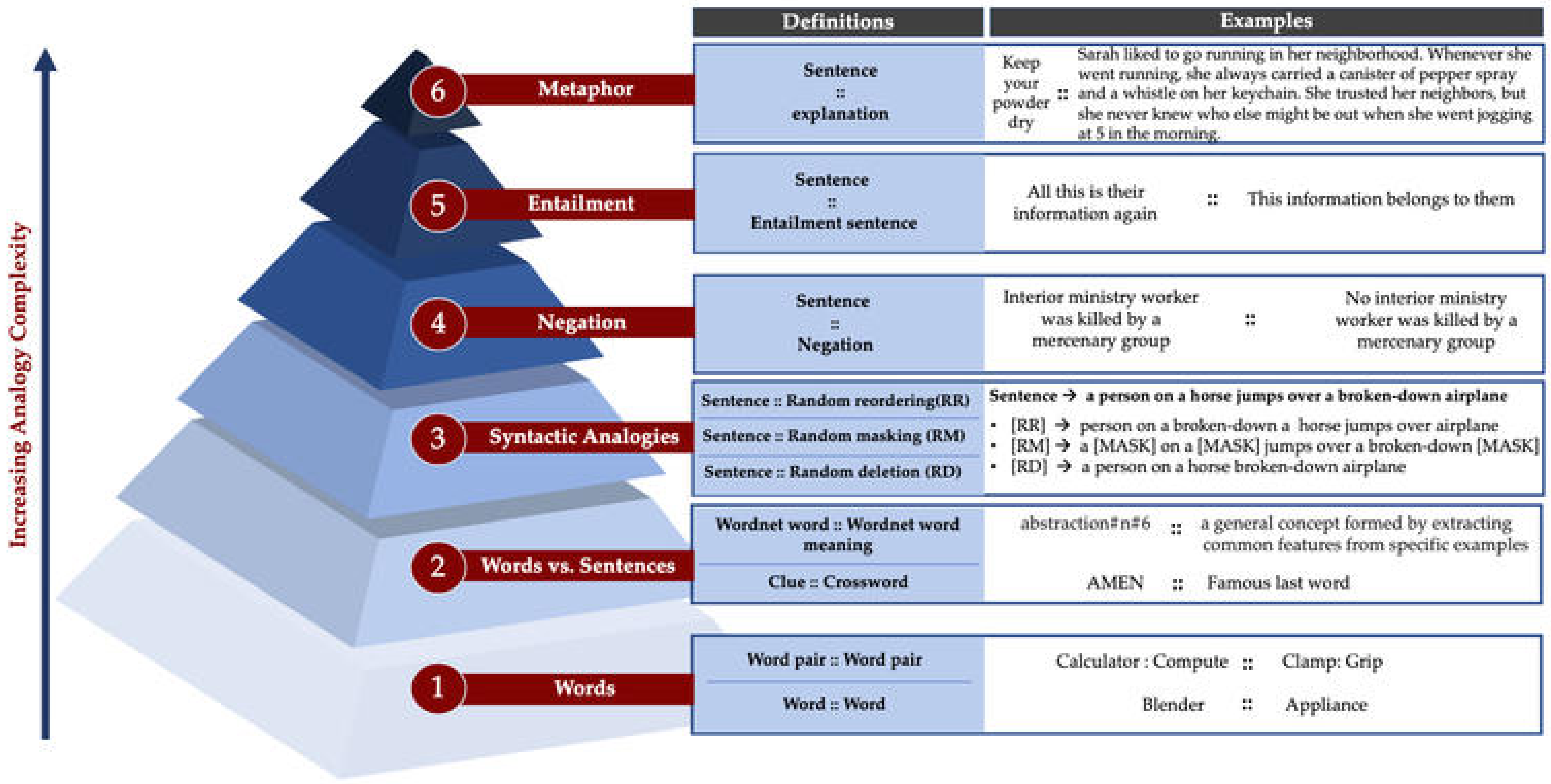
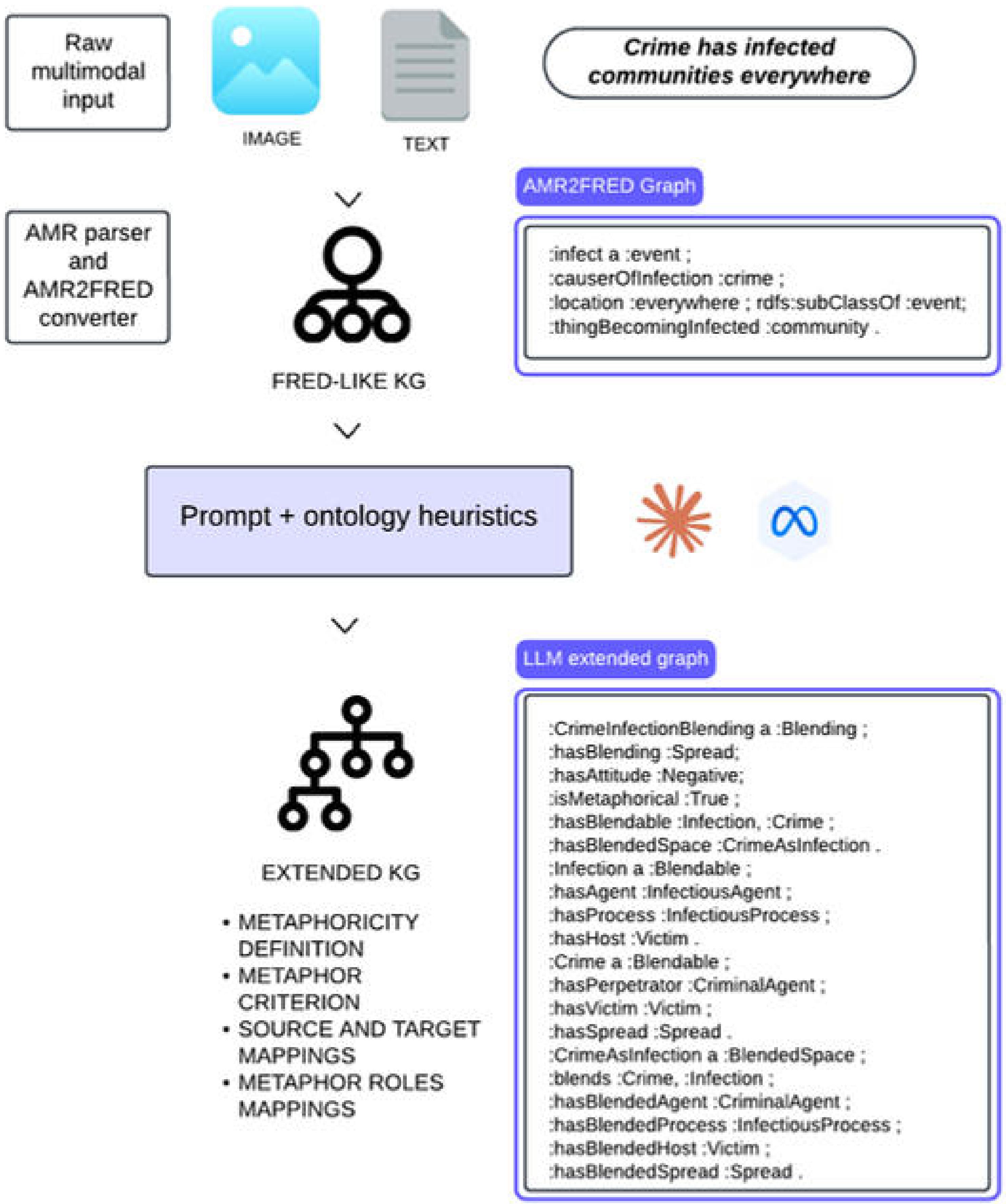


Figure 2: Analogy taxonomy with six levels. The definitions of the analogies at each level and examples for each analogy type from the datasets are indicated.

Wijesiriwardene, Thilini, et al. "ANALOGICAL--A Novel Benchmark for Long Text Analogy Evaluation in Large Language Models." arXiv preprint arXiv:2305.05050 (2023).

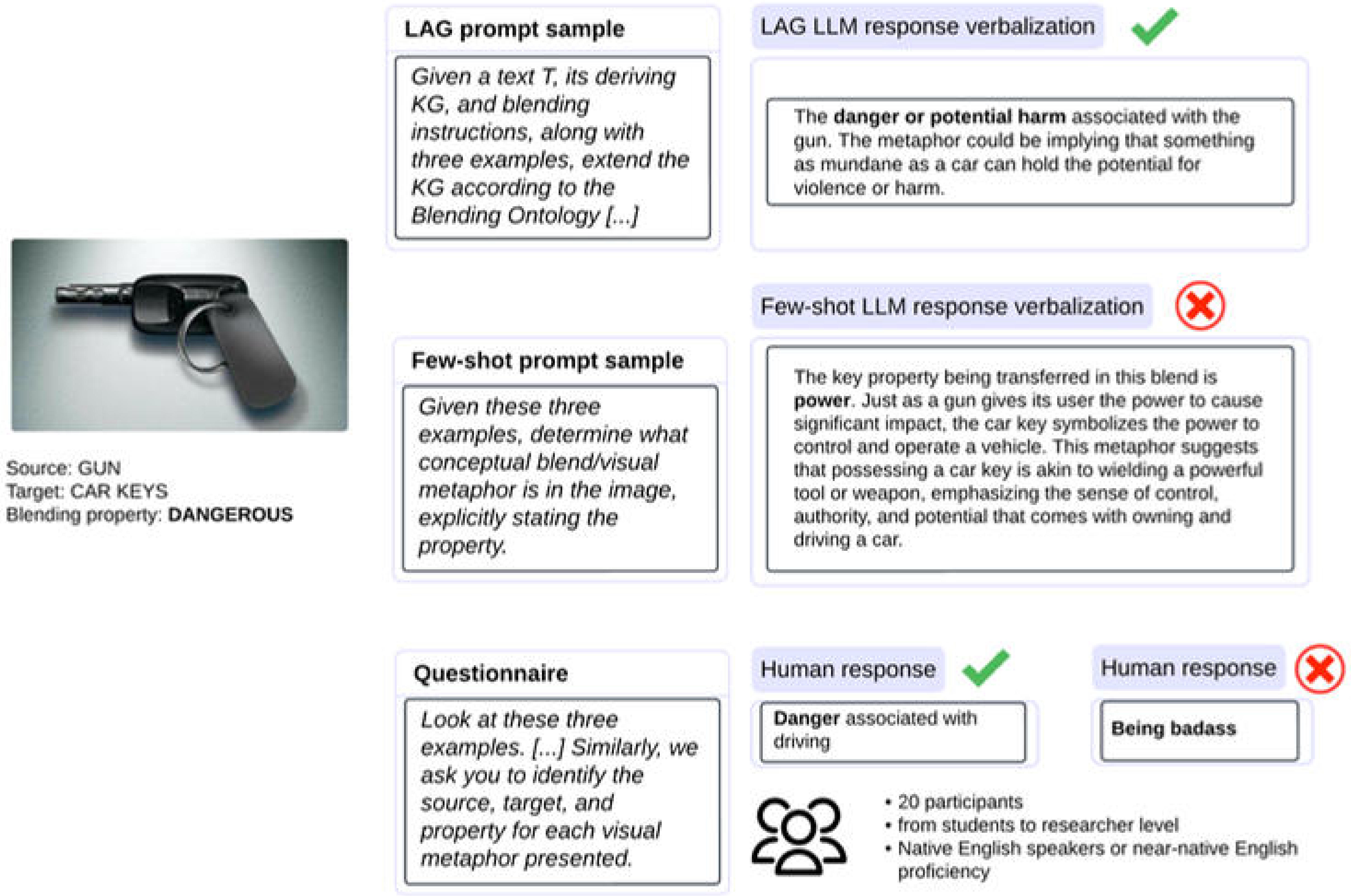
Logic Augmented Generation



Gangemi, A., & Nuzzolese, A. G. (2025). Logic augmented generation. Journal of Web Semantics, 85, 100859.

Logic Augmented Generation

- Multimodal
- LAG treats the LLM as a reactive KG generator
- Enriches with implicit knowledge to an extended KG under logical constraints



Source: GUN
Target: CAR KEYS
Blending property: DANGEROUS
LAG prompt sample
Given a text T, its deriving KG, and blending instructions, along with three examples, extend the KG according to the Blending Ontology [...]
Few-shot prompt sample
Given these three examples, determine what conceptual blend/visual metaphor is in the image, explicitly stating the property.
Questionnaire
Look at these three examples. [...] Similarly, we ask you to identify the source, target, and property for each visual metaphor presented.
LAG LLM response verbalization
The danger or potential harm associated with the gun. The metaphor could be implying that something as mundane as a car can hold the potential for violence or harm.
Few-shot LLM response verbalization
The key property being transferred in this blend is power. Just as a gun gives its user the power to cause significant impact, the car key symbolizes the power to control and operate a vehicle. This metaphor suggests that possessing a car key is akin to wielding a powerful tool or weapon, emphasizing the sense of control, authority, and potential that comes with owning and driving a car.
Human response
Danger associated with driving
Human response
Being badass
• 20 participants
• from students to researcher level
• Native English speakers or near-native English proficiency

Layer 1: **Detection** - Binary classification - Metrics: Accuracy, F1	*Can the system identify metaphorical language?*	*The most common line of inquiry*
Layer 2: **Analysis** - Source/target identification - Metrics: Semantic similarity (BLEURT), Human expert validation	*Can the system analyze the metaphor's meaning?*	We need to look back at theories
Layer 3: **Reasoning** - Property/blending identification - Metrics: Human expert validation and error analysis	*Can the system justify the analogical mapping?*	We need explainable methods
Layer 4: **Generalization** - Comparison of the performance of the preceding layers across domains and modalities	*How well does the system work across domains/modalities?*	We need data on domains → science

A new dataset

Table 1: Summary of datasets statistics, including instances, percentage of metaphorical sentences, and samples.

Dataset	# Instances	% Met.	# Samples
MOH-X	647	48.7	300
TroFi	3737	43.5	300
WG	447	100	447
BCMTD	49 (conceptual) 49 (scientific) 49 (VUA)	66.6	147
Visual metaphors	51	100	48

LAG applied to metaphor: results

Table 2: Performance comparison of various methods for metaphor detection on MOH-X and TroFi. Best performing results are in bold.

Method	MOH-X		TroFi	
	F1 (%)	Acc. (%)	F1 (%)	Acc.(%)
MetaPRO	84	81	79	70
TSI CMT*	82.5	82.9	66	66.8
LAG	**89.7**	**87.3**	**89.7**	**84.6**

Table 3: Performance metrics for the BCTMD Dataset for metaphor detection. Best performing results are in bold.

Method	Accuracy (%)	F1 Score (%)
LAG	**80.1**	**84.1**
MetaPRO	69.1	69.8
Few-Shot 12	59.0	48.9
Few-Shot 6	52.4	45.2
Few-Shot 3	47.5	42.8
Zero-shot	22.9	33.8

LAG applied to metaphor: ablation

Table 7: Ablation study results for MOH-X, TroFi, and BCMTD datasets. Best results in bold.

Method	MOH-X		TroFi		BCMTD	
	Acc.%	F1%	Acc.%	F1%	Acc.%	F1%
LAG	**87.3**	**89.7**	**84.6**	**89.7**	**80.1**	84.1
No Blending	81.6	87	81.9	86	78.6	**85.2**
No Graph	78.6	82	83.9	87	70	73

LAG applied to metaphor

- Achieves large gains on metaphor detection/understanding, even surpassing human gold on visual metaphors
- Scientific metaphors remain hard, revealing the need for domain-specific treatment
- Error analysis: LLMs show surface association strength but weak relational reasoning; our method improves interpretability and justification of metaphor predictions

Italian National Research Council

Going back to our initial questions...

- **LLMs and knowledge engineering form a loop**: LLMs can accelerate ontology generation and evaluation, while explicit ontologies and knowledge graphs make LLM behaviour more reliable and interpretable.

- **Assisted, not (yet) automated**: LLMs already match or approach human performance on many ontology tasks and reduce perceived difficulty, but still need human oversight to avoid subtle logical and modelling errors.

- **Towards neurosymbolic AI**: Logic-Augmented Generation shows that injecting formal knowledge improves multimodal metaphor understanding and explanation, pointing toward practical neurosymbolic pipelines

- **Reproducibility**: work in progress

Reflections

- We need to study heterogeneous domains
- What other KE tasks should be automated?
- Where should we draw the line in automation (so as to not encourage over-trust)?
- The idea of a plugin for KE?
- What kinds of knowledge (commonsense, spatial, causal, scientific, etc.) are most urgent to encode explicitly for LLM-based systems?
- What new benchmarks and datasets are needed to evaluate relational reasoning and domain-specific metaphor understanding more rigorously?

References

- Lippolis, A. S., Saeedizade, M. J., Keskisärkkä, R., Zuppiroli, S., Ceriani, M., Gangemi, A., ... & Nuzzolese, A. G. (2025, June). *Ontology generation using large language models*. In: European Semantic Web Conference.

- Lippolis, A. S., Saeedizade, M. J., Keskisarkka, R., Gangemi, A., Blomqvist, E., & Nuzzolese, A. G. (2025). *Assessing the Capability of Large Language Models for Domain-Specific Ontology Generation*. In European Semantic Web Conference (ELMKE workshop).

- Lippolis, A. S., Saeedizade, M. J., Keskisärkkä, R., Gangemi, A., Blomqvist, E., & Nuzzolese, A. G. (2025). *Large Language Models Assisting Ontology Evaluation*. In: International Semantic Web Conference.

- Cappa, S., Lippolis, A. S., & Zoia, S. (2025). *Meanings are like Onions: A layered approach to metaphor processing*. In Joint Ontology WOrkshops (JOWO) (EIKE workshop).