Assessing the Capability of Large Language Models for Domain-Specific Ontology Generation

Anna Sofia Lippolis, Mohammad Javad Saeedizade,
Robin Keskisärkkä, Aldo Gangemi,
Eva Blomqvist, and Andrea Giovanni Nuzzolese

Assessing the Capability of Large Language Models for Domain-Specific Ontology Generation

Anna Sofia Lippolis*

Mohammad Javad Saeedizade*

Robin Keskisärkkä, Aldo Gangemi,
Eva Blomqvist, and Andrea Giovanni Nuzzolese

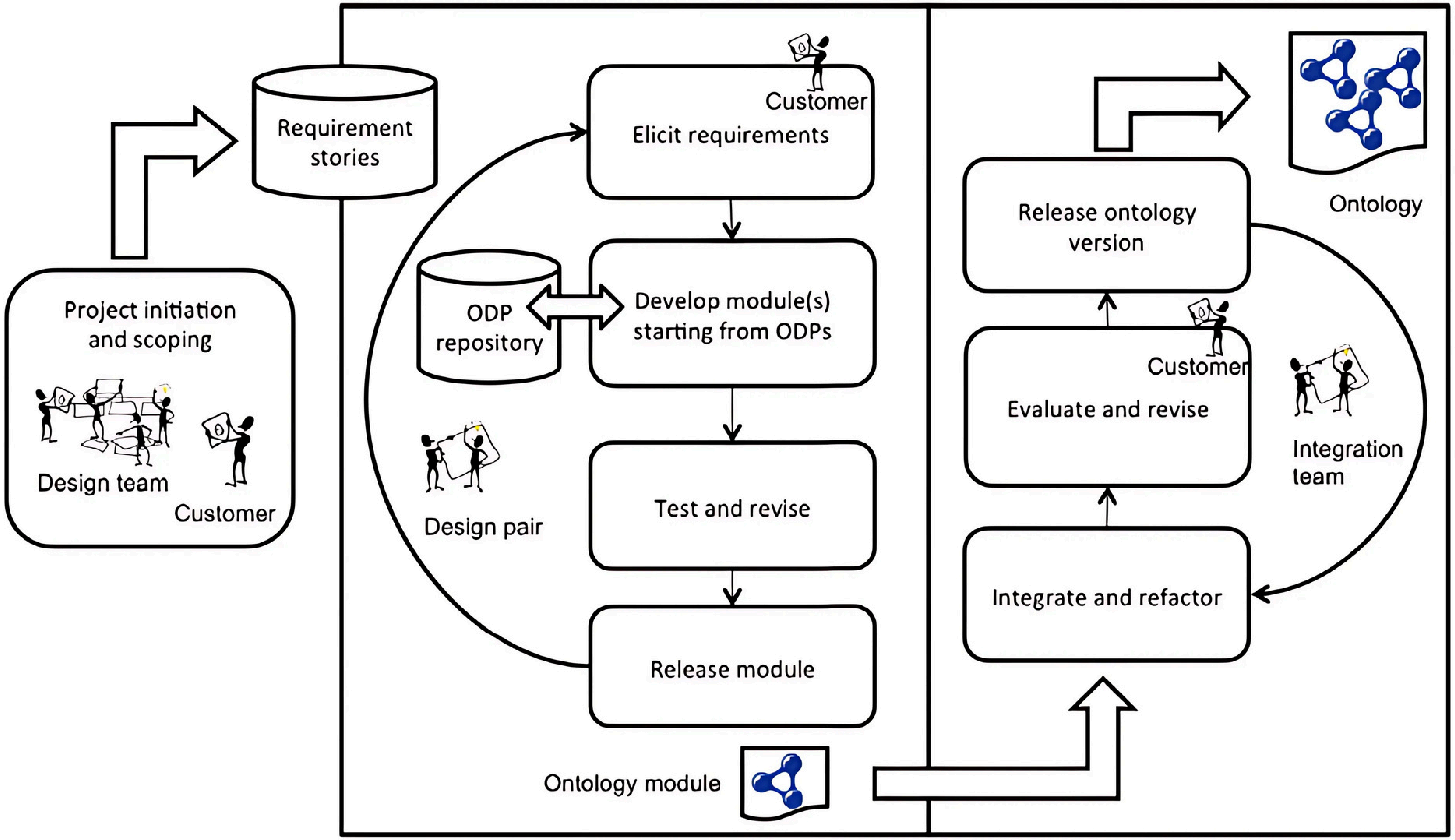
*equal contribution

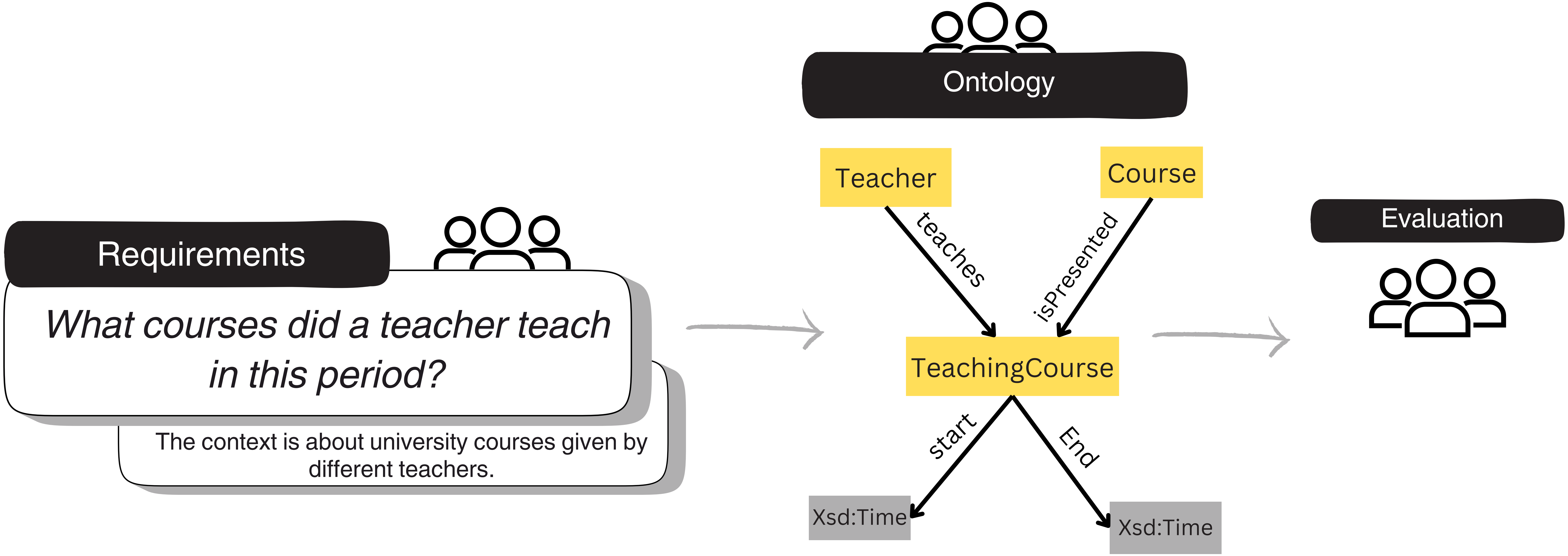
CNR - Istituto di Scienze e Tecnologie della Cognizione

ALMA MATER STUDIORUM UNIVERSITÀ DI BOLOGNA

LINKÖPING UNIVERSITY

Ontology engineering is a labour-intensive, expertise-driven task...

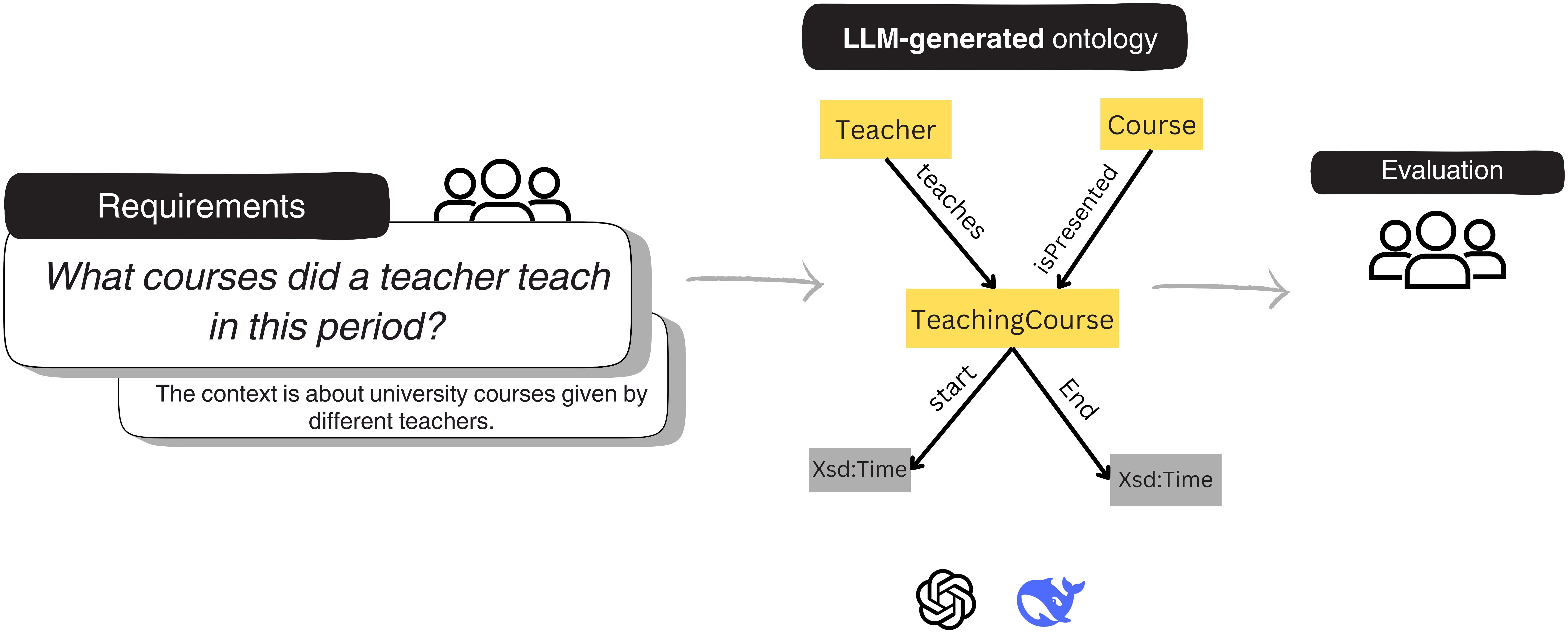



Requirements
What courses did a teacher teach in this period?
The context is about university courses given by different teachers.
Ontology
Teacher
Course
teaches
isPresented
TeachingCourse
start
End
Xsd:Time
Xsd:Time
Evaluation

The task

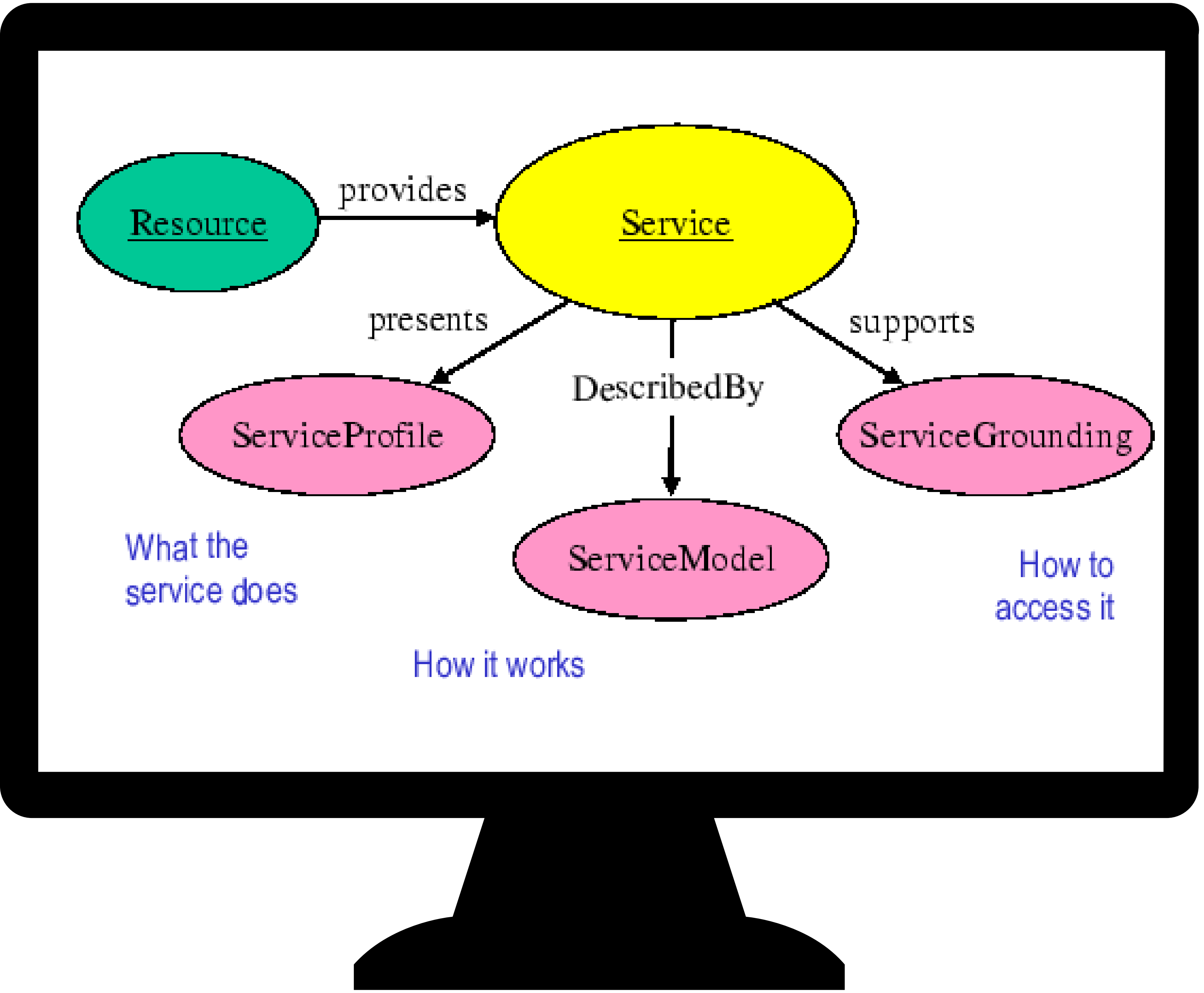
... But recent work has shown promise in automating ontology generation with LLMs

- Benson et al. (2024): GPT-4 for BFO-compliant outputs
- Ontogenia (Lippolis et al., 2024): Decomposed prompting for African Wildlife domain.
- Saeedizade & Blomqvist (2024): Compared LLMs to student models.
- **Lippolis et al. (2025): Cross-domain CQs, showed o1-preview outperforms.** → **Research track**
- Doumanas et al. (2025): Fine-tuning for specific domains.
- Fathallah (2024): Life sciences focus.

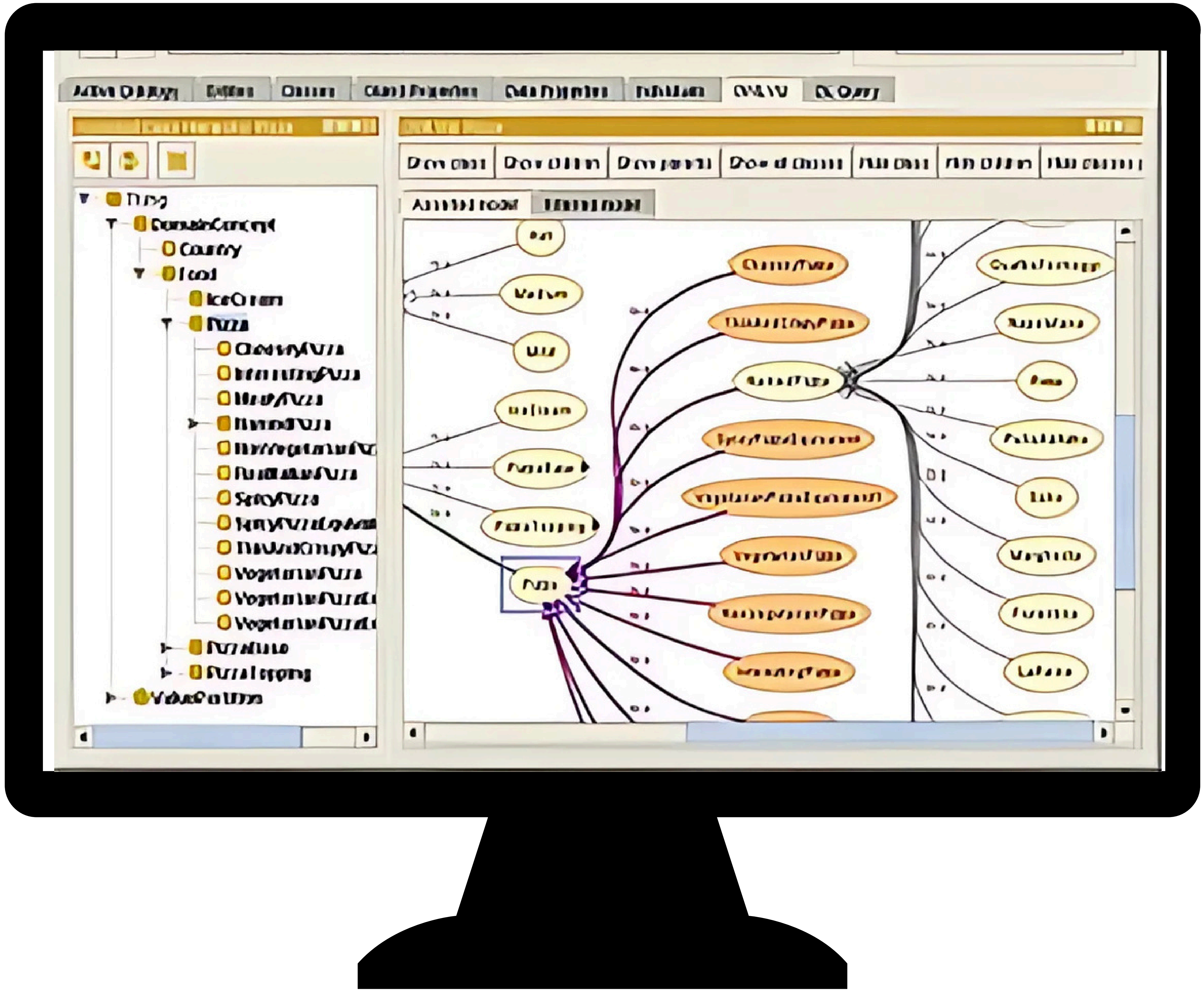

Requirements
What courses did a teacher teach in this period?
The context is about university courses given by different teachers.
LLM-generated ontology
Teacher
Course
teaches
isPresented
TeachingCourse
start
End
Xsd:Time
Xsd:Time
Evaluation

The task

Final goal: an assistant for ontologists



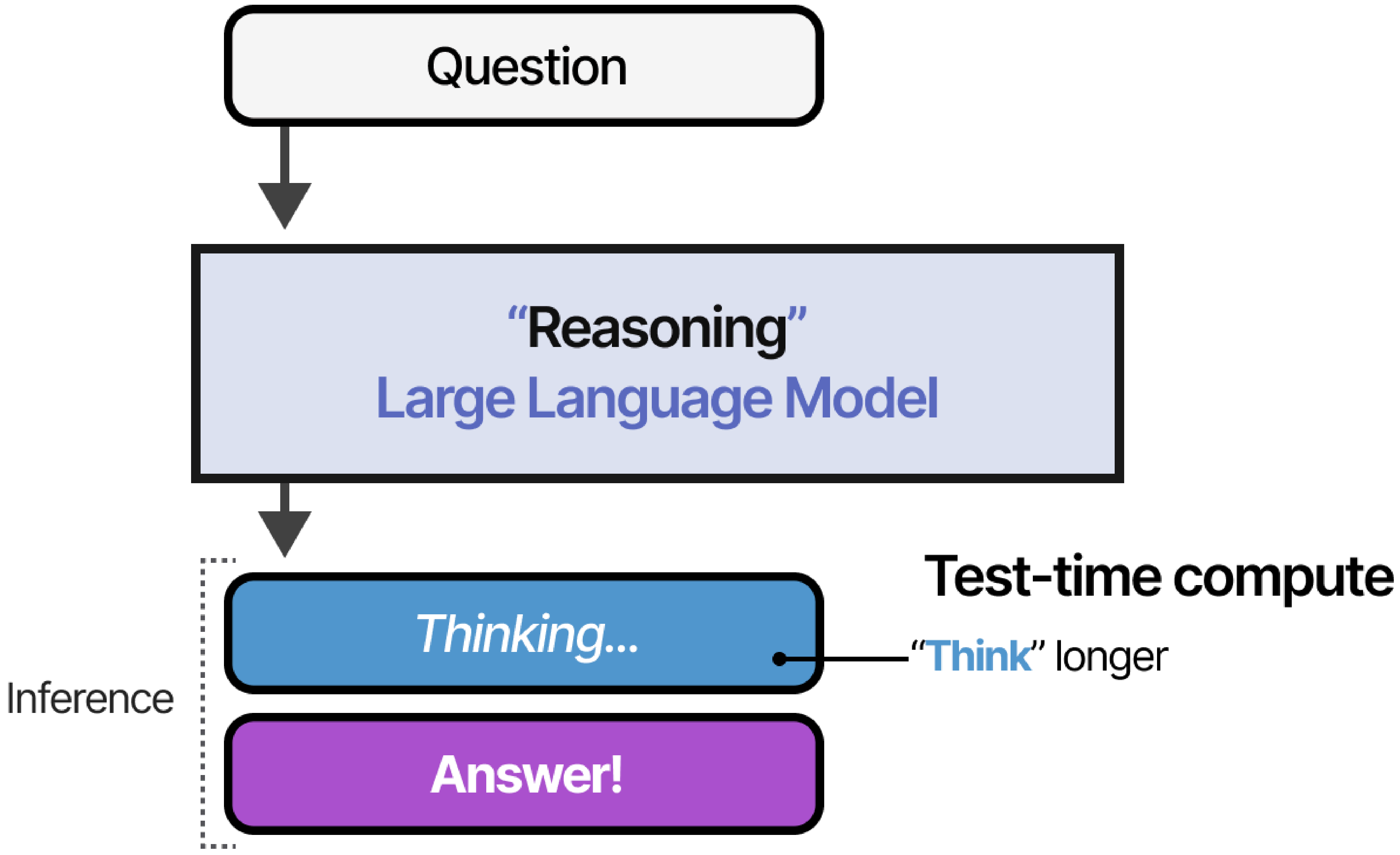
LLM suggestions



Protégé

Some issues remain:

- No testing on several distinct **domain-specific** ontologies
- No distinction between generating from **easy** and **hard** requirements
- No ontology generation assessment specific for **reasoning** models



Our contributions

1

Automated pipeline for **domain-agnostic** ontology generation

2

Benchmark dataset to test on **easy** and **hard** requirements and six different domains

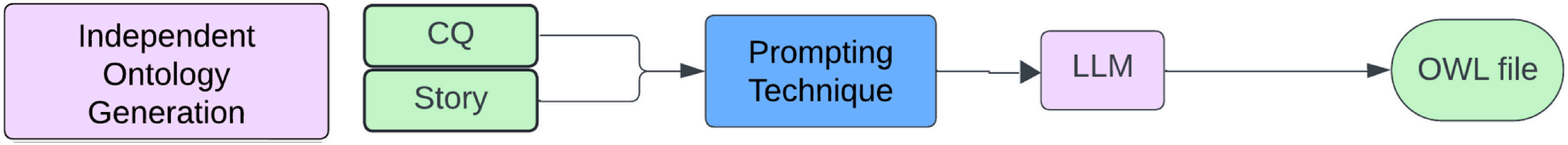
3

Evaluation with two **reasoning** Large Language Models: OpenAI's o1-preview and DeepSeek's R1

Ontology generation

Independent ontology generation

each CQ and its associated ontology story are provided to an LLM through a prompt to generate the corresponding ontology



DeepSeek R1 and OpenAI o1-preview with default hyperparameters

Dataset creation

Easy CQ: if a CQ required at most 2 classes and 1 property

Hard CQ: more than 2 classes and 1 property

*Human developed ontologies, the others are semi-automatically generated

Domain	Total CQs	Easy	Hard
Circular economy*	16	5	11
Music*	16	10	6
Events*	18	8	10
Microbe habitat	15	8	7
Carbon and nytrogen cycling	15	4	11
Water and health*	15	7	8
Total	**95**	**42**	**53**

Prompting technique

Few-shot prompting technique

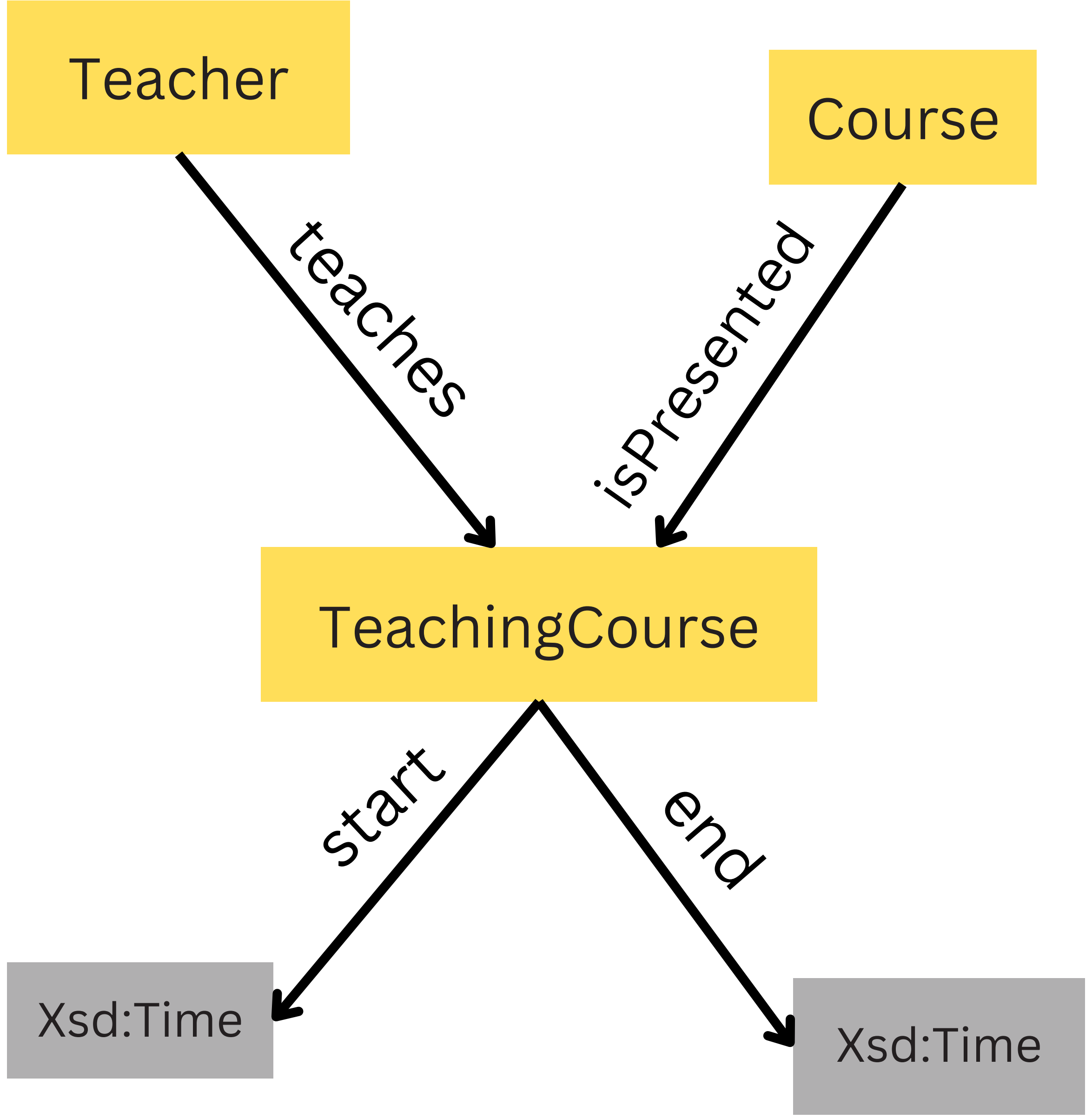
Components:

- Instructions
- Examples
- Task (actual input)

Instructions	You are a helpful assistant...
Examples	Example 1: CQ 1, Story 1, Ontology 1
	Example 2: CQ 2, Story 2, Ontology 2
Task	CQ, Story

Evaluation

Manual annotation by two ontology engineers on the dataset using **CQ verification** (Blomqvist et al., 2012):



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

Competency Question:
What courses did a teacher teach in this period?

```
PREFIX ex: <http://example.org/ontology#>
SELECT ?course
WHERE {
  ?course ex:isPresented ?tc .
  ?teacher ex:teaches ?tc .
  ?tc ex:start ?start ;
      ex:end ?end .
  FILTER(?start >= "09:00:00"^^xsd:time
&& ?end <= "17:00:00"^^xsd:time)
}
```

Answer: the requirements are modelled

Results

High and comparable accuracy with o1-preview and DeepSeek R1

9 and 10 CQs unmodeled respectively (out of 95)

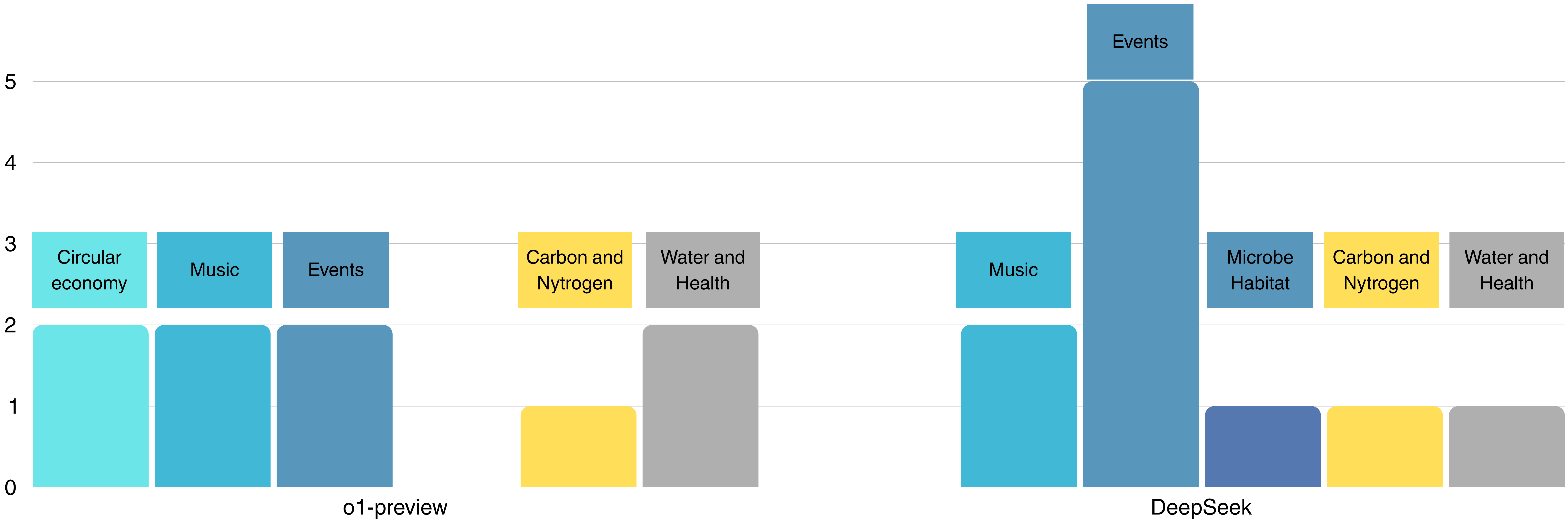
Consistent performance across all domains

This method is applicable rather than being limited to a specific domain

No difference between hard and easy CQs

Question complexity doesn’t necessarily result in lower scores

Error analysis



Error example (Events)

- CQ: Did they travel to reach the place?
    - *they who? what place?*

- CQ: How can we characterise the relation among the participants?
    - *what relation? what kind of characterisation?*

Story: *Ortenz would like to have a system for visualising events (meetings of composers and musicians) in time and space in order to track musicians' careers, their overlap and intersections, gathering trends in time and space, and making emerge patterns of knowledge transmission...*

Limitations and future work

- Only two reasoning models: need to expand more
- Six domains
- Potential dataset leakage
- Need additional evaluation metrics
- The cost of LLMs differs significantly among them and it hinders extensive usage for research
- LLMs don’t generate ontologies with a similar quality to humans: LLM “reasoning” relies on patterns, not actual human-like understanding

Takeaways

- o1-preview and DeepSeek R1 can reliably generate ontology modules across diverse domains.
- Performance is similar across “easy” and “hard” CQs
- Domain-agnostic generalizability
- Few-shot prompting yields better results than sub-task decomposed prompting.
- The main source of errors stems from under/overspecified requirements, not model limitations.

Thank you!
Questions?

Anna Sofia Lippolis, Mohammad Javad Saeedizade,
Robin Keskisärkkä, Aldo Gangemi,
Eva Blomqvist, and Andrea Giovanni Nuzzolese

annasofia.lippolis2@unibo.it
javad.saeedizade@liu.se

Limitations

- Leakage

Data could have been used for training LLMs

- Using reasoning models

Reasource-intensive

- Other metrics

Hallucination, extra components, OOPS!, etc