Actively Learning $\mathcal{EL}$ Terminologies from Large Language Models

Matteo Magnini, Riccardo Squarcialupi, Martin T. Sterri, Ana Ozaki

matteo.magnini@unibo.it, riccard.squarcialupi@studio.unibo.it, martin.sterri@student.uib.no, anaoz@uio.no

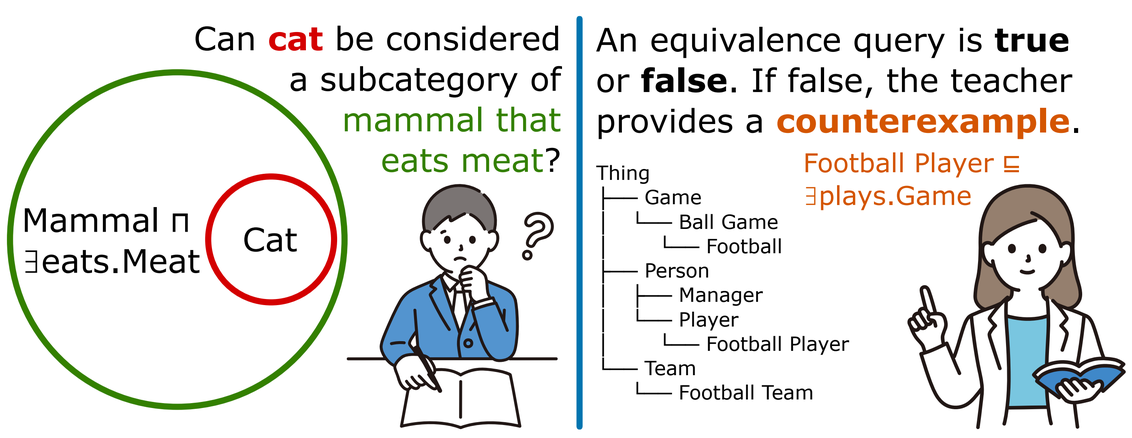
ALMA MATER STUDIORUM UNIVERSITÀ DI BOLOGNA · UNIVERSITY OF BERGEN · UNIVERSITY OF OSLO

Actively Learning & Vision

In the active learning framework a **learner** attempts to learn some kind of knowledge by posing questions to a **teacher**.

Questions made by the learner are:

- **membership** queries → ask whether **concept inclusions** are **true** or **false**;
- **equivalence** queries → ask whether the **idea** of the learner about the knowledge of the teacher is **correct** or **not**.



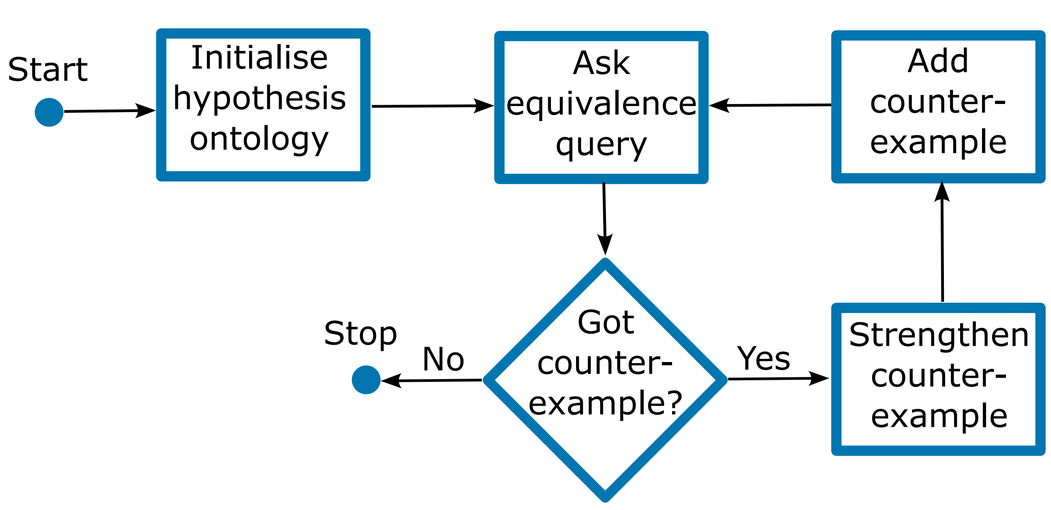
We consider the case in which:

- the knowledge is expressed as an $\mathcal{EL}$ **terminology**;
- membership queries consist in asking if an axiom is a consequence of the onlology;
- equivalence queries are instead **simulated** by a sample with concept inclusions labelled as positive or negative.

Our intention is to use a **large language model** (LLM) as a teacher for actively learning ontologies and evaluate the results.

The Angluin's **exact learning** framework makes use of active learning when membership queries are allowed. We developed a tool, **ExactLearner+LLM**, that implements Angluin's framework using LLMs as teacher.

ExactLearner+LLM algorithm



We approximate equivalence queries through **sampling**. Rather than asking the LLM to validate the **hypothesis** directly, we randomly generate $\mathcal{EL}$ concept inclusions and query the model on whether each of them is true or false.

The algorithm checks if the classification of the examples match with the information in the hypothesis:

- true inclusions must be **logical consequences**;
- false ones must not.

If the hypothesis fits the classification of the concept inclusions, learning stops; otherwise, the inclusion not fitting the hypothesis serves as a counterexample, as if the LLM had replied no to an equivalence query.

This sampling-based simulation can yield **PAC** guarantees when the sample size

$$|S| \geq \ln(|H|/\delta)/\epsilon$$

is computed from the hypothesis space H ($\mathcal{EL}$ terminologies of bounded structure) and parameters ϵ (error) and δ (confidence).

Probing Language Models

Challenges

- **Input format:** questions standardisation to systematically query an LLM. We investigate the use of the *Manchester OWL syntax* and natural language.
- **Unexpected responses:** LLMs may answer with an arbitrary response. We use custom *system prompts*, a maximum number of *tokens* to mitigate this issue.
- **Correctness & logical consistency:** there is no guarantee that the responses are correct (i.e., true in the real world). Moreover, they may not be logically consistent.

  ↓

  We search for logical inconsistency by creating the **closure under logical consequence** and testing whether something in the closure received false as answer.

Findings

- **Concept inclusions:** A ⊑ B inclusions are much easier to learn than A ⊑ ∃r.B or ∃r.B ⊑ A;
- **Operations:** saturation, desaturation and right decomposition were more successfully applied;
- **Quering:** models achieve better performance when queries are expressed with a natural language prompt and with a customised system prompt.
- **Performance:** among the tested LLMs, Mistral has superior performance. In general, LLMs have better results with terminologies regarding common topics.

Experiments & Results

Ontology	N_C	N_R	Log. Ax.	PAC Sample	Poss. Ax.
Animals	17	4	12	542	6,936
Cell	22	0	24	1,119	10,164
Football	10	3	9	341	1,500
Generations	20	4	18	847	10,800
University	7	3	4	139	588

Table 1
Size of the signature in the tested ontologies, number of logical axioms, PAC sample size (ϵ=0.2, δ=0.1) and the number of possible normalised axioms with the signature. We also tested the Galen medicalontology, by creating modules towork with bigger ontologies (per se, Galen is too big).

Model	Accuracy	Recall	Precision	F1-Score
Llama2 (13b)	0.521	0.71	0.294	0.314
Llama3 (8b)	0.43	0.947	0.218	0.333
Mistral (7b)	0.741	0.747	0.45	0.49
Mixtral (47b)	0.705	0.611	0.547	0.436

Table 4
Results of ExactLearner+LLM grouped by models.

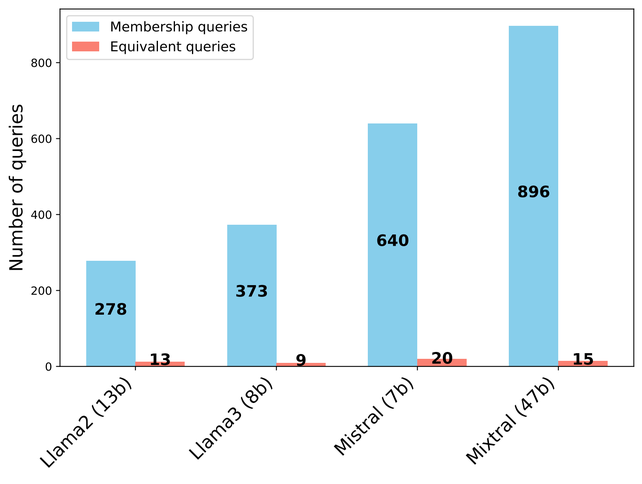


Figure 2
Average number of membership and (simulated) equivalence queries grouped by LLM.

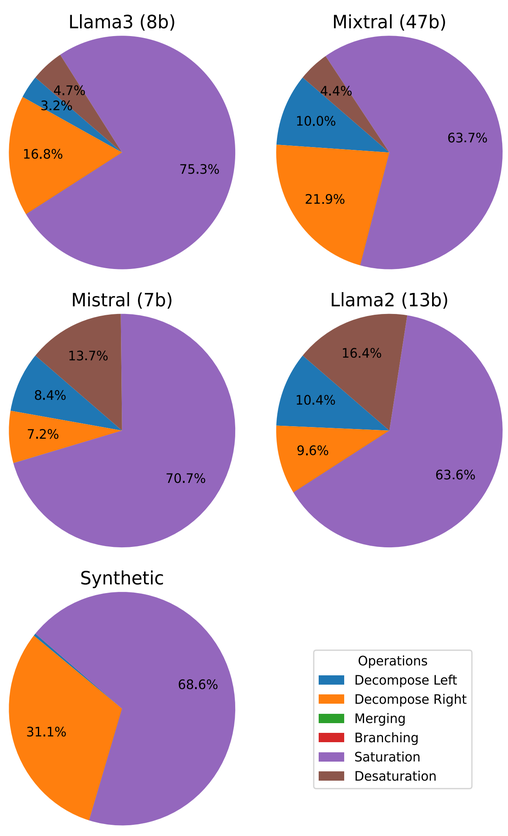


Figure 1
Aggregated results of the operations performed by the learner during the PAC learning of all the ontologies grouped byteacher type.

The application of these operations, that use membership queries, result in concept inclusions that **maximise how informative** they are and also **minimise** their **size**.

The **ExactLearner+LLM** tool and the **experiments** are publicly available on GitHub.