

Actively Learning Ontologies from LLMs: First Results (Extended Abstract)

Matteo Magnini¹, Ana Ozaki^{2,3} and Riccardo Squarcialupi¹

¹Department of Computer Science and Engineering, University of Bologna, Via dell'Università 50, Cesena, Italy

²Department of Informatics, University of Oslo, Gaustadalléen 23B, Oslo, Norway

³Department of Informatics, University of Bergen, Thormøhlensgate 55, Bergen, Norway

Abstract

In *active learning* a learner attempts to acquire some kind of knowledge by posing questions to a teacher. Here we consider that the teacher is a language model and study the case in which the knowledge is expressed as an ontology. To evaluate the approach, we present first results testing logical consistency and the performance of GPT and other language models when answering whether concept inclusions from existing $\mathcal{EL}$ ontologies are ‘true’ or ‘false’.

1. Introduction

Large language models (LLMs) have reached a point where they have accumulated so much information and improved on their question/answering capability that we are now willing to interact and learn from them. Prompts to these models vary from questions about general knowledge such as basic definitions and historical events, to more domain specific questions, e.g., scientific facts related to health and medicine. What can we learn from LLMs? And, since it is known that they can give false information, is there an automated way of discovering whether responses are incorrect or at least inconsistent?

In this work we explore an active learning approach to learn from LLMs. In active learning [1], a learner attempts to learn some kind of knowledge by posing questions to a teacher. The questions made by the learner are called *membership queries* and are answered with ‘yes’ or ‘no’ (or equivalently, with ‘true’ or ‘false’) [2]. Here we consider that the teacher is an LLM and study the case in which the knowledge to be learned is expressed as an ontology. We use the Manchester OWL Syntax [3] in our prompts, as this syntax is closer to natural language. We present preliminary results showing the performance of GPT and other language models when answering whether concept inclusions created by an ontology engineer on prototypical $\mathcal{EL}$ ontologies are ‘true’ or ‘false’.

 DL 2024: 37th International Workshop on Description Logics, June 18–21, 2024, Bergen, Norway

 matteo.magnini@unibo.it (M. Magnini); anaoz@uio.no (A. Ozaki); riccard.squarcialupi@studio.unibo.it (R. Squarcialupi)



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

2. Probing Language Models

Here we briefly describe challenges encountered when probing LLMs with ontology axioms and how we handled them.

Input Format and Unexpected Responses One important factor is the format of the query. To systematically query an LLM with the goal of learning an ontology, it is useful to standardise the questions. For the membership queries task, we investigate the use of the Manchester OWL syntax [3], as this is an ontology syntax designed to be closer to natural language. Another aspect to consider is that, in principle, there are no constraints in the answers returned by the language model. An LLM may answer with an arbitrary and unexpected response, even if the expected answer is just a single word like in the case of membership queries in the exact learning model. To mitigate this issue, one can explicitly tell the LLM to answer with ‘true’ or ‘false’. This particular request can be done in the question itself (e.g., appending “Answer with ‘true’ or ‘false’.” after the query) or by exploiting some hyper-parameters of the API of the LLM. In the second case, one can use a *system prompt* – a.k.a., integrated text into each query within the chat session – to enrich the model with additional information and useful to harness the response. We highlight that there are also other hyper-parameters that could help driving the LLM’s response into the desired format (e.g., maximum number of tokens, temperature, etc.). Even with all these precautions the model may return an unexpected response. For example: (i) the answer can have more text than just ‘true’ or ‘false’, (ii) both ‘true’ and ‘false’ can appear in the answer, (iii) the answer does not have ‘true’ nor ‘false’. While in the first scenario a trivial parsing would determine the correct classification, in the remaining cases, since there is some ambiguity, we considered a third value, which we called ‘unknown’.

Correctness and Logical Consistency We also need to deal with challenges regarding the correctness of the responses (assuming the format of the responses returned by the language model are as expected, see Section 2). Actively learning ontologies has been investigated for various fragments of $\mathcal{EL}$ [4, 5, 6], though, without using LLMs as teachers. If an LLM is playing the role of the teacher then there is no guarantee that the responses are correct [7] (in the sense of reflecting the ‘truth’ about the real world) and, moreover, that they are logically consistent with any $\mathcal{EL}$ ontology. Indeed, it is known that LLMs can learn statistical features instead of performing logical reasoning [8]. So, we need to consider the following kinds of errors:

1. $C \sqsubseteq D$ should be ‘false’ (cf. the real world) but the LLM answers ‘true’;
2. $C \sqsubseteq D$ should be ‘true’ (cf. the real world) but the LLM answers ‘false’;
3. all concept inclusions in $\mathcal{T} = \{C_1 \sqsubseteq D_1, \dots, C_n \sqsubseteq D_n\}$ are answered with ‘true’, $\mathcal{T} \models C \sqsubseteq D$ but $C \sqsubseteq D$ is classified as ‘false’.

The last case is a logical inconsistency. One strategy to handle this issue is to consider the *closure* under logical consequence [9]. That is, in Point 3, one could consider $C \sqsubseteq D$ as ‘true’.

3. Experiments

The experiments consist in performing a number of membership queries with multiple LLMs on prototypical ontologies. These are small ontologies taken from ontology repositories and used for experiments in the ExactLearner project [6]^{1 2}, which focuses on $\mathcal{EL}$ ontologies. In all ontologies considered, the logical closure is finite. We consider the following ontologies:

1. *Animals* contains knowledge related to the animal realm, including actual animals, sub-phyla, classes, orders, etc. The ontology has 12 (explicit) logical axioms in $\mathcal{EL}$ and 20 logical axioms in the logical closure (that is, taking into account inferred axioms).
2. *Cell* provides information about different cells based on their type, development stage and organism. The ontology has 24 logical axioms in $\mathcal{EL}$ and 24 in the logical closure.
3. *Football* is a minimal ontology that describes the relations between football game, teams, players and managers. It has 9 logical axioms in $\mathcal{EL}$ and 12 in the logical closure.
4. *Generations* describes the members and relations within a family. This ontology has 18 (explicit) logical axioms in $\mathcal{EL}$ and 42 in the logical closure.
5. *University* is a small ontology, focusing on the professor role, with 4 logical axioms in the logical axioms in $\mathcal{EL}$ and 8 in the logical closure.

We use a total of 5 LLMs: Open AI’s GPT 3.5 Turbo [10], Mistral [11], Mixtral [12] and two Llama 2 [13] models (we use Ollama’s API³). Both Mistral and Mixtral are open models. Llama 2 is free of charge for research while GPT can be expensive as it charges for each query⁴.

For each logical axiom in an ontology we generate a membership query to an LLM using the Manchester OWL syntax. The goal is to test how well an LLM can correctly answer to membership queries on different domains and without any fine-tuning, where ‘correctly’ means that it answers ‘true’ for the axioms in the ontology (even though ontologies may not match with the real world, we expect them to be mostly correct). The results are in Table 1.

We generate all the inferred axioms using the HermiT [14] reasoner (as mentioned above, the logical closure of the ontologies is finite) and we repeat the experiments with the new ontologies. Probing the LLMs on ontologies with inferred axioms is useful to test logical consistency. While it is possible that the LLMs could have seen these ontologies during their training (since they are available online), it is unlikely that this is the case for the inferred axioms, since they are not explicitly present in the ontologies. The results are in Table 2.

We perform a third experiment where we actively learn ontologies by means of a naive learning algorithm where all concept inclusions of the form $A \sqsubseteq B$ with A, B concept names in a given signature are asked (the ontologies have complex $\mathcal{EL}$ concepts, but in this experiment we only considered concept names to reduce the number of membership queries). The results

¹<https://github.com/bkonev/ExactLearner/>

²Generations, University, and Cell were also part of the Protégé Ontology Library. Not maintained anymore at https://protegewiki.stanford.edu/wiki/Protege_Ontology_Library but still accessible via web archive at https://web.archive.org/web/20210226123540/https://protegewiki.stanford.edu/wiki/Protege_Ontology_Library

³<https://github.com/ollama/ollama>

⁴The source code of the experiments is publicly available <https://github.com/MatteoMagnini/ExactLearner>

Models	Animals			University			Generations			Football			Cell		
	T	F	U	T	F	U	T	F	U	T	F	U	T	F	U
Mistral (7b)	9	1	2	2	0	2	5	10	3	7	2	0	17	1	6
Mixtral (47b)	11	1	0	4	0	0	3	6	9	9	0	0	15	9	0
Llama2 (7b)	11	1	0	4	0	0	16	1	1	9	0	0	24	0	0
Llama2 (13b)	11	1	0	4	0	0	16	1	1	9	0	0	23	1	0
Gpt3.5	10	2	0	4	0	0	13	4	1	9	0	0	21	3	0

Table 1

Results for the experiments testing correctness w.r.t. axioms in the ontologies. Labels T, F and U mean ‘true’, ‘false’ and ‘unknown’ responses count. We indicate the number of parameters in each model in parenthesis (e.g. Mistral has 7 billion). It is not known the number of parameters of GPT 3.5.

Animals				University				Generations				Football				Cell			
T	F	U	L	T	F	U	L	T	F	U	L	T	F	U	L	T	F	U	L
14	2	4	2	5	1	2	0	10	27	5	2	9	3	0	0	18	1	5	0
18	2	0	0	8	0	0	0	19	13	10	0	12	0	0	0	17	7	0	0
20	0	0	0	8	0	0	0	40	1	1	1	12	0	0	0	24	0	0	0
18	2	0	1	7	1	0	0	35	6	1	4	11	1	0	1	21	3	0	0
20	0	0	0	7	1	0	0	36	5	1	0	12	0	0	0	18	6	0	0

Table 2

Results for the experiment testing logical consistency. The number of parameters of each model and the meaning of T, F, U are as in Table 1. L stands for logical inconsistencies (an axiom answered as ‘false’ or ‘unknown’ which can be inferred from the set of the axioms answered as True, see Section 2). Models’ names omitted for better readability (they are the same of Table 1).

Animals			University			Generations			Football			Cell		
A	P	R	A	P	R	A	P	R	A	P	R	A	P	R
0.87	0.52	0.72	0.57	0.67	0.5	0.84	0.71	0.23	0.74	0.44	0.65	0.65	0.48	0.81
0.89	0.57	0.69	0.57	0.48	0.92	0.82	0.64	0.66	0.72	0.43	0.76	0.7	0.32	0.64
0.51	0.2	1	0.24	0.24	1	0.4	0.22	0.88	0.21	0.21	1	0.27	0.18	1
0.73	0.31	0.94	0.45	0.3	0.92	0.63	0.32	0.74	0.44	0.26	0.88	0.44	0.21	0.91
0.71	0.3	1	0.69	0.44	1	0.74	0.41	1	0.68	0.4	1	0.61	0.28	0.91

Table 3

Results for the experiments testing negative examples. Labels A, P and R mean ‘Accuracy’, ‘Precision’ and ‘Recall’ respectively [15]. Models’ names omitted for better readability (they are the same of Table 1).

are in Table 3. We applied the Chi-squared test to check the relationship between the answers of the LLMs and the ontologies, with the null hypothesis being that there is no correlation. We rejected the null hypothesis in every case (p-value lower than 0.05) except the ones in yellow. Mistral/Mixtral were competitive with GPT 3.5 and had better performance in comparison with the Llama 2 models. The LLMs performed well on ontologies with general knowledge (e.g., Animals, Generations). As future work, we would like to build on these first results and extend the experiments to larger ontologies. Moreover, we plan to investigate the task of actively learning ontologies from LLMs using the ExactLearner [6].

Acknowledgements

Ana Ozaki is supported by the Research Council of Norway, project number 316022.

References

- [1] D. Angluin, Computational learning theory: Survey and selected bibliography, in: S. R. Kosaraju, M. Fellows, A. Wigderson, J. A. Ellis (Eds.), *Proceedings of the 24th Annual ACM Symposium on Theory of Computing*, ACM, 1992, pp. 351–369. URL: <https://doi.org/10.1145/129712.129746>. doi:10.1145/129712.129746.
- [2] D. Angluin, Queries and concept learning, *Mach. Learn.* 2 (1987) 319–342. doi:10.1007/BF00116828.
- [3] M. Horridge, N. Drummond, J. Goodwin, A. L. Rector, R. Stevens, H. Wang, The manchester OWL syntax, in: B. C. Grau, P. Hitzler, C. Shankey, E. Wallace (Eds.), *Proceedings of the OWLED*06 Workshop on OWL: Experiences and Directions*, Athens, Georgia, USA, November 10-11, 2006, volume 216 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2006. URL: https://ceur-ws.org/Vol-216/submission_9.pdf.
- [4] B. Konev, C. Lutz, A. Ozaki, F. Wolter, Exact learning of lightweight description logic ontologies, *J. Mach. Learn. Res.* 18 (2017) 201:1–201:63. URL: <http://jmlr.org/papers/v18/16-256.html>.
- [5] A. Ozaki, C. Persia, A. Mazzullo, Learning query inseparable elh ontologies, in: *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020*, New York, NY, USA, February 7-12, 2020, AAAI Press, 2020, pp. 2959–2966. doi:10.1609/AAAI.V34I03.5688.
- [6] M. R. C. Duarte, B. Konev, A. Ozaki, Exactlearner: A tool for exact learning of EL ontologies, in: M. Thielscher, F. Toni, F. Wolter (Eds.), *Principles of Knowledge Representation and Reasoning: Proceedings of the Sixteenth International Conference, KR 2018*, Tempe, Arizona, 30 October - 2 November 2018, AAAI Press, 2018, pp. 409–414. URL: <https://aaai.org/ocs/index.php/KR/KR18/paper/view/18006>.
- [7] M. Funk, S. Hosemann, J. C. Jung, C. Lutz, Towards ontology construction with language models, in: S. Razniewski, J. Kalo, S. Singhanian, J. Z. Pan (Eds.), *Joint proceedings of the 1st workshop on Knowledge Base Construction from Pre-Trained Language Models (KBC-LM) and the 2nd challenge on Language Models for Knowledge Base Construction (LM-KBC) co-located with the 22nd International Semantic Web Conference (ISWC 2023)*, Athens, Greece, November 6, 2023, volume 3577 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023. URL: <https://ceur-ws.org/Vol-3577/paper16.pdf>.
- [8] H. Zhang, L. H. Li, T. Meng, K. Chang, G. V. den Broeck, On the paradox of learning to reason from data, in: *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI 2023*, 19th-25th August 2023, Macao, SAR, China, ijcai.org, 2023, pp. 3365–3373. URL: <https://doi.org/10.24963/ijcai.2023/375>.
- [9] S. Blum, R. Koudijs, A. Ozaki, S. Touileb, Learning horn envelopes via queries from language

models, *International Journal of Approximate Reasoning* (2023) 109026. doi:<https://doi.org/10.1016/j.ijar.2023.109026>.

- [10] 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. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfc4967418bfb8ac142f64a-Paper.pdf.
- [11] A. Q. J. et al., Mistral 7b, CoRR abs/2310.06825 (2023). doi:10.48550/ARXIV.2310.06825. arXiv:2310.06825.
- [12] A. Q. Jiang, A. Sablayrolles, A. Roux, A. Mensch, B. Savary, C. Bamford, D. S. Chaplot, D. de Las Casas, E. B. Hanna, F. Bressand, G. Lengyel, G. Bour, G. Lample, L. R. Lavaud, L. Saulnier, M. Lachaux, P. Stock, S. Subramanian, S. Yang, S. Antoniak, T. L. Scao, T. Gervet, T. Lavril, T. Wang, T. Lacroix, W. E. Sayed, Mixtral of experts, CoRR abs/2401.04088 (2024). doi:10.48550/ARXIV.2401.04088. arXiv:2401.04088.
- [13] H. Touvron, et al., Llama 2: Open foundation and fine-tuned chat models, CoRR abs/2307.09288 (2023). doi:10.48550/ARXIV.2307.09288. arXiv:2307.09288.
- [14] B. Glimm, I. Horrocks, B. Motik, G. Stoilos, Z. Wang, Hermit: An OWL 2 reasoner, *J. Autom. Reason.* 53 (2014) 245–269. doi:10.1007/S10817-014-9305-1.
- [15] M. Grandini, E. Bagli, G. Visani, Metrics for multi-class classification: an overview, CoRR abs/2008.05756 (2020). URL: <https://arxiv.org/abs/2008.05756>. arXiv:2008.05756.