Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 246 (2024) 1289–1298

Procedia Computer Science

www.elsevier.com/locate/procedia

28th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2024)

Prompting is not all you need

Evaluating GPT-4 performance on a real-world ontology alignment use case

Giulio Macilenti[a,*], Armando Stellato[a], Manuel Fiorelli[a]

[a]*Tor Vergata University of Rome, Via del Politecnico 1, Rome 00133, Italy*

Abstract

Ontology Alignment (OA) is a complex, demanding and error-prone task, requiring the intervention of domain and Semantic Web experts. Automating the alignment process thus becomes a must-do, especially when involving large datasets, to at least produce a first input for human experts. Automated ontology alignment could benefit from the outstanding language ability of Large Language Models (LLMs), which could implicitly provide the background knowledge that has been the Achilles' heel of traditional alignment systems. However, this requires a correct evaluation of the performance of LLMs and understanding the best way to incorporate them into more specific tools. In this paper, we show that a naive prompting approach on the popular GPT-4 model could face several problems when transferred to real-world use cases. To this end, we replicated the methods of Norouzi et al. (2023), applied to the OAEI 2022 conference track, on a reference alignment between a pair of datasets (reduced versions of two popular thesauri: European Commission's EuroVoc and TESEO, from the Italian Senate of the Republic), which has never been tested in OAEI evaluation campaigns. This reference alignment has several features common to real-world use cases: it is has a larger size than those considered in the study we replicated, it is not published online and is therefore not subject to data contamination and it involves relations between concepts that are more complex than simple equivalence. The replicated methods achieved a significantly lower performance on our reference alignment than on the OAEI 2022 conference track, suggesting that size, data contamination, and semantic complexity need to be considered when using LLMs for the alignment task.

Keywords: Semantic technologies, Ontology alignment, Large language models.

1. Introduction

Ontology Alignment (OA) is a task that aims to identify semantic correspondences between ontologies[1, 2]. This is central to the present stage of development of semantic technologies, where various ontologies (and, more in general, datasets of various nature: thesauri, vocabularies, code lists, or any form of instance data) can describe overlapping domains: in such a global scenario where data proliferation should be a blessing rather than a problem, alignments

* Corresponding author.
E-mail address: giulio.macilenti@students.uniroma2.eu

1877-0509
10.1016/j.procs.2024.09.557

play a pivotal role in keeping interoperability between different resources. Even though OA is not a new task, it still takes a lot of effort: most of the time there is a need for domain experts to work on it, and this is time-consuming and even subject to human bias.
Meanwhile, Large Language Models (LLMs) have emerged as extremely powerful tools that can solve many NLP tasks, attracting the interest of the scientific community and the masses. The surprising fact is that LLMs can solve tasks they have not been explicitly trained on simply because of their understanding of language, guided by user instructions given in the form of a prompt[3, 4]. In fact, there are several ways to use LLMs. It is possible to fine-tune a pre-trained model for a specific task, using data related to the task to update the model (i.e., its parameters). But at the same time, it is possible to simply write the instructions in a prompt formulated in natural language with few (few-shot) or no (zero-shot) previous examples [5], and send them to the LLM. The fact that these models perform well even in this astonishingly simple setting has attracted much attention to the necessity of building the best prompt for given tasks, known as prompt engineering.
Given all these facts, and keeping in mind the reciprocal importance that semantic and linguistic knowledge have on each other[6, 7], it is natural to ask if LLMs can provide meaningful alignments and help automating the OA process in a few-shot or zero-shot scenario. Some experiments have already been made following this idea [8, 9], while other approaches try to integrate LLMs in more complex architectures, as one of several steps towards the final goal [10].
In this paper, we aim at providing further ground truth towards understanding whether prompting is sufficient to obtain a good performance on real-world use cases of OA. Some aspects must be considered: first of all data contamination is an important issue, that can lead to overestimating the performance of LLMs [11, 12]. Second, when dealing with OA one should consider the dependence of the results on ontology size and on the complexity of the semantic relations involved, which is not so easy to quantify a priori.
With these open questions in mind, we replicated the prompting approach provided in [9] on different pair of datasets. The original study achieved good performances with this approach on the ontologies in the conference track of the evaluation campaigns organized by OAEI (Ontology Alignment Evaluation Initiative)[1]. However, we show that this is not the case when we apply this approach to a benchmark alignment between two thesauri (two reduced versions of the EuroVoc[2] and TESEO[3] thesauri with 300 concepts each) that is not available on the web and is remarkably larger [13]. We think that our work provides more insight into the problems that can affect the performance of an LLM when it is used to align real-world ontologies.

2. Related work

OA is a well explored research area, and through the years various techniques have been developed in the field, resulting in a vast literature[1, 2, 14, 15]. Classical methods are rule-based, and mainly rely on lexical matching, in combination with structural information and filters that control logical conflicts between potential mappings. Among them, it is worth mentioning LogMap and AgreementMakerLight, which obtain state-of-the-art performance in many equivalence matching tasks[16, 17].
More recently, machine learning has pushed forward a new generation of models, like DeepAlignment and OntoEmma [18, 19] which leverage the representation of words in a high-dimensional vector space, called embeddings: the distance in such a vector space resembles the similarity in the meaning of words. Then, with the rise of the transformer architecture[21], new systems were developed that used fine-tuning of pre-trained models for the alignment tasks. For example, BERTMap [20] predicts mappings using a classifier based on fine-tuning the contextual embedding model BERT on text semantics corpora extracted from ontologies.
The final step of the story, to this date, is the arrival of instruction-tuned LLMs: these extremely powerful models consist of billions of parameters and have the ability to solve an enormous variety of tasks starting from generic and massive training. They have achieved impressive performance, for example, in text generation and summarization, question answering and code generation [5, 8, 22, 23].
Since the breakthrough of LLMs is recent, there is a relatively small literature that considers their few or zero-shot

[1] https://oaei.ontologymatching.org/
[2] http://eurovoc.europa.eu/
[3] http://www.senato.it/3235?testo_ generico=745

application for OA.
The article we used as our main reference [9] focuses on a simple conversational approach, where ontologies from the OAEI conference track are lexicalized, arranged as formatted text and then sent to ChatGPT4 in prompts with different structure. The resulting alignments have an F1 score that is lower than state-of-the-art, because they have high recall and low precision. The same authors recently published another work, where they propose to use prompting with GPT-4 to find matches that are more complex than simple equivalence [24]. At the same time, He et. al. [25] tested the zero-shot performance of Flan-T5-XXL and GPT-3.5-turbo models on two challenging subsets from the NCIT-DOID and the SNOMED-FMA (Body) equivalence matching datasets, both part of the Bio-ML track of OAEI. Their results confirm that there is a weakness related to precision with the GPT model, but according to the authors the application of LLMs to OA is still promising. Both papers state that there are various problems to solve: the design of appropriate prompts, the overall framework and the incorporation of ontology contexts being the most notable.
Some of these issues have been addressed by Hertling and Paulheim [10]. They proposed *OLaLa*, an ontology matching system that is built on top of open-source large language models. *OLaLa* uses sentence BERT models (SBERT) to generate matching candidates that are then sent to the LLM as a binary decision: the model then chooses whether the proposed match is correct or not. This results in an improved precision and a lower computational cost, and the model achieves good performance on various OAEI tracks [26].
Another approach that seems to be promising is the development of LLM-based autonomous agents, where the LLM is used as a controller that can perform various functions. For example, in [27] Qiang et al. propose LLMA Dialogue Model, in which multiple agents negotiate the correspondence between two knowledge sets. The system is evaluated on the OAEI Anatomy dataset, showing a good performance as well as advantages in generalization and interpretability. In [28] Agent-OM, a more sophisticated agent-powered LLM-based framework is described and evaluated on three OAEI tracks (Conference, Anatomy and MSE).
Together with the literature focused on the development of OA systems, it is important for our work to also consider the numerous studies that highlight problems in the correct evaluation of the performance of LLMs. From our point of view, the most relevant issue seems to be data contamination[11, 12, 29, 30]. In fact, there is some raising concern whether LLMs performances could be overestimated because of their enormous training sets: testing them on information available on the Web could results in performance that is higher than what can be expected in real-world use cases just because the LLMs have already seen the exact example on which is being tested during pre-training.
Finally, it is important to keep in mind that many ontology alignment systems require a large amount of information to be provided in a prompt and processed by the LLM. In fact, it is unclear whether LLMs retain all the text of the prompt at the same level of specificity: in [31] it is reported that when dealing with long prompts LLMs focus more on data contained at the beginning or at the end of the prompts, while the center of the messages is overlooked.

3. Methodology

3.1. Data

We attempted to replicate the approach described in [9] on a different alignment scenario. In the mentioned paper, the authors used the alignments in the OAEI 2022 `conference track`[4], which consists of 21 pairs of matched ontologies. For their evaluation, the reference alignment known as `ra1` was used, incorporating both properties and classes, therefore they referenced `ra1-M3` OAEI 2022 results for comparison.
For our experiments, we focused on an alignment [13] between the thesaurus of the Italian Senate TESEO (TEsauro SEnato per l'Organizzazione dei documenti parlamentari) and EuroVoc (EU's multilingual thesaurus covering the activities of the European Union). Domain and semantic web specialists have manually aligned reduced versions of the two thesauri, consisting of 300 concepts each. The objective was to obtain a gold standard for the evaluation of alignment systems, still within tractable size. The two reduced datasets present interesting real-world challenges from

[4] https://oaei.ontologymatching.org/2022/conference/index.html

the lexical perspective, as false friends and, in some cases, questionable labels (which can still be statistically canceled by the presence of other labels and context, which is the information that helped the human experts) make the mapping task more difficult. As shown in Table 2, the ontologies in the conference track are noticeably smaller than our reduced thesauri: the largest ontology is indeed composed of 140 classes and 41 properties, while these ontologies have on average only 100 between classes and properties.
The TESEO-EuroVoc alignment which we refer is not published online and is therefore not subject to data contamination. It consists of 101 correspondences, belonging to 4 types of relations, namely: = (equal), > (broader), < (narrower) and related.
In Table 1 we report the number of correspondences for each type of relation.

Table 1. The number of correspondences for each type of relation in the TESEO-EuroVoc alignment

Relation	Correspondences
=	57
>	3
<	35
related	6

In our work we didn't consider the correspondences with the "related" relation, because our initial experiments showed that GPT considers an enormous number of concepts as related, resulting in a dramatic decrease in the precision of the alignment.

Table 2. In this table, taken from the data reported on the OAEI website, we show some properties of the various ontologies that make up the OAEI conference track, used in [9]. As you can see they are significantly smaller than our dataset, built on 300 concepts for each thesaurus.

Name	Type	Number of Classes	Number of Datatype Properties	Number of Object Properties	DL expressivity
Ekaw	Insider	74	0	33	SHIN
Sofsem	Insider	60	18	46	ALCHIF(D)
Sigkdd	Web	49	11	17	ALEI(D)
lasted	Web	140	3	38	ALCIN(D)
Micro	Web	32	9	17	ALCOIN(D)
Confious	Tool	57	5	52	SHIN(D)
Pcs	Tool	23	14	24	ALCIF(D)
OpenConf	Tool	62	21	24	ALCOI(D)
ConfTool	Tool	38	23	13	SIN(D)
Crs	Tool	14	2	15	ALCIF(D)
Cmt	Tool	36	10	49	ALCIN(D)
Cocus	Tool	55	0	35	ALCIF
Paperdyne	Tool	47	21	61	ALCHIN(D)
Edas	Tool	104	20	30	ALCOIN(D)
MyReview	Tool	39	17	49	ALCOIN(D)
Linklings	Tool	37	16	31	SROIQ(D)

3.2. Structure of the prompts and formatting

In this paragraph we give a brief overview of the conversational method used in [9] and replicated on our dataset. First of all we have to clarify how ontologies are given as input to the LLM. The triples are presented as formatted text with the structure: "Relation(Subject,Object)". For example, an original triple such as "Decision disjointWith Person " can be represented as "disjointWith(Decision,Person)". Only the axioms that can be directly expressed as triples are included.
Regarding the prompts they are built on three basic elements, that we report in Fig. 1. There is the <**Problem Definition**>, in which we explain the fact that we are analyzing ontologies (in our case thesauri) and the particular format we are employing, then the <**Ontologies Triples**> represented in such a structured way and finally the

<**Objective**>, where the actual description of the task to be performed is given.
In [9] the authors built seven different prompts rearranging and modifying this basic structure. For example the <**Objective**> part can be improved to obtain a more accurate matching process, or the order in which the triples are listed can vary. In most cases the information is split into two separate prompts following each other. The most important changes are in the prompt number 7 where, after the <**Problem Definition**> and <**Ontologies Triples**>, a repetitive instruction for each class/property in the first ontology to match is sent to the LLM. This clearly results in computational cost problems when the size of the considered ontologies grows. The details of the structure of the various prompts are reported in Fig. 2.
Regarding our experiments we closely followed this methodology. The sole difference is that we explicitly stated that the relations involved in the alignment process had to be equal, broader, or narrower. Without this addition the model only identifies matches between concepts deemed equal; this way, instead, it produces results across all the types of relations. Regarding prompt 6, that changes the order of triples to prioritize root class entities, since we are dealing with thesauri we chose to arrange the concepts following an order in which the roots of the trees defined by the broader relations involved in each thesaurus were first.

<Problem Definition>
In this task, we are given two ontologies in the form of Relation(Subject, Object), which consist of classes and properties.
<Ontologies Triples>
Ontology 1:
Ontology 1 Triples
Ontology 2:
Ontology 2 Triples
<Objective>
Our objective is to provide ontology mapping for the provided ontologies based on their semantic similarities.

Fig. 1. In this figure, taken from [9], we show the basic structure of the prompts.

4. Results and analysis

Before using the various prompts on our dataset we replicated the experiments on the OAEI 2022 conference track ontologies, to ensure that we were using their methodology properly. We obtained performances consistent with those reported in [9], and so we evaluated that a comparison with the results on our reference alignment was meaningful.
This comparison is outlined in Table 3. It is clear that there is a significant drop in performance when passing from the OAEI 2022 conference track dataset to our TESEO-EuroVoc alignment. Since the performance changes when repeating the experiment with the same prompt, we report our results averaged over five realizations, and we add the standard deviation.
It is important to note that in our experiment there is not just a performance drop in comparison with [9]; even the top performing prompts change. In our case, prompts 3,5 and 6 have higher performances, while prompts 1 and 7 were the best in the other experiment.

To get a better insight on the causes of the performance drop that we observed, we focused on one specific feature of our dataset that was missing in the OAEI conference ontologies: the presence in our TESEO-EuroVoc alignment of many correspondences that are not of simple equivalences (=) but refer to a narrower-broader (<,>) relation between concepts. Analyzing the results, we discovered that, despite trying, GPT-4 can't grasp this type of relations, so its performance is always zero on them. So, we repeated the evaluation only on the set of equivalence correspondences in the alignment. As you can see from Fig. 3 the performance improves this way, although not reaching the levels of [9].
Furthermore, we considered how many of the correct correspondences produced by GPT-4 were between concepts

P#	Description	Prompt structure
1	Put all the information in a single prompt.	<Problem Definition> <Ontologies Triples> <Objective>
2	Changing the objective of the prompts.	<Problem Definition> <Ontologies Triples>
		Provide a complete and comprehensive matching of the ontologies
3	Changing the objective of the prompt.	<Problem Definition> <Ontologies Triples>
		Match these two ontologies and provide the most accurate matching you can do
4	Separate the class and data/object properties in two consecutive prompts.	<Problem Definition> <Class Triples>
		<Data/Object Triples> <Objective>
5	Following the Exp 2 pattern but changing the objective of the prompt.	<Problem Definition> <Class Triples>
		<Data/Object Triples> Match these two ontologies and provide the most accurate matching you can do
6	Following the Exp 2 pattern but changing the order of triples to prioritizing the root class entities.	<Problem Definition> <Class Triples>
		<Data/Object Triples> <Objective>
7	First, Providing the Ontologies, then asks about the best class/property of the second ontology that can be matched with the class/property of the first one.	<Problem Definition> <Ontologies Triples>
		For a class/property in the first ontology, which class/property in ontology 2 is the best match? <Ontology 2 Triples>

Fig. 2. Here we report the various prompts used in the experiment, as described in [9].

Table 3. Comparison between the performance reported in [9] on the OAEI 2022 Conference Track and the one we obtained on our TESEO-EuroVoc alignment. The highest results for each column are in bold.

# Prompt	OAEI 2022 Conference Track			TESEO-EuroVoc Alignment		
	Recall	Precision	F1-Score	Recall	Precision	F1-Score
Prompt1	0.52	**0.52**	**0.52**	0.02 ± 0.01	0.04 ± 0.01	0.03 ± 0.01
Prompt2	0.60	0.43	0.49	0.09 ± 0.03	0.06 ± 0.03	0.07 ± 0.02
Prompt3	0.57	0.49	0.51	0.09 ± 0.02	**0.48** ± 0.14	0.10 ± 0.03
Prompt4	0.63	0.37	0.45	0.05 ± 0.02	0.25 ± 0.07	0.08 ± 0.03
Prompt5	0.69	0.37	0.46	0.13 ± 0.04	0.13 ± 0.05	**0.13** ± 0.03
Prompt6	0.61	0.39	0.46	**0.15** ± 0.03	0.09 ± 0.02	0.12 ± 0.02
Prompt7	**0.92**	0.37	**0.52**	0.13 ± 0.03	0.05 ± 0.01	0.07 ± 0.01

that share the same principal label. Always in Fig. 3 you can observe that there are many, because both recall and precision drops remarkably when they are excluded from the evaluation.

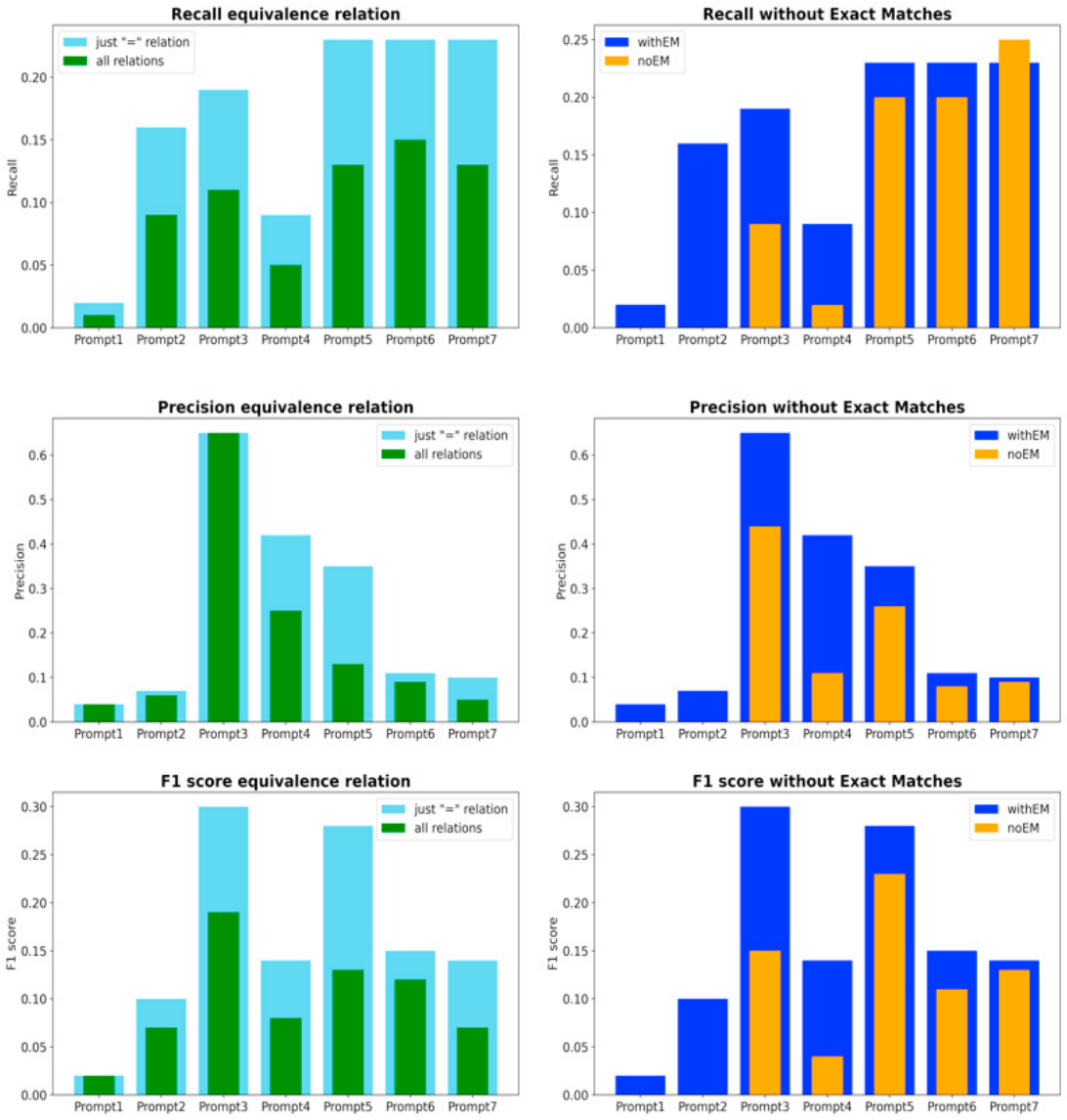

Fig. 3. In this figure we show how recall, precision and F1 score change considering a subset of the reference alignment. In the first column we present the results of the set of equivalence relations of our alignment, while in the second column we report what happens when we remove the concepts with the same principal label from this set.

5. Discussion

We think that our work simply shows, as stated in the title, that *prompting is not all you need* when trying to do ontology alignment using zero or few-shot prompting on pre-trained LLMs. As is highlighted in Table 3 the promising results found in [9] significantly drop when we try to replicate the same approach on a different alignment scenario. Our reference alignment, which comes from real-world thesauri used by the European Commission and the Italian Senate, poses problems that were not considered in the previous study. By the way this does not nullify or invalidate the idea of [9] to use LLMs as an important instrument in ontology alignment nor the various prompts

proposed, but it just suggests that this approach should be combined with other, more specific techniques, as done in [10, 26].
In [9] it was recognized that there is a problem with precision when using LLMs for ontology alignment, but we point out that in our case even the recall is low and the best prompt order is not the same as in the previous study. We are led to think that this means that the result can vary greatly depending on the dataset considered, posing the problem of its reliability. This issue is now getting more and more attention as one of the principal drawbacks of the application of LLMs, especially closed-source ones like GPT[32].

We investigated the factors that can cause the low performance we observed. In our opinion, the following issues are not the unique but the most serious.
1)Size of the two ontologies to be matched: despite the bigger context that can be prompted to GPT-4, it is not clear if the model is able to use optimally all of such information. For example, in [31] it is shown that the LLM is better at answering with data contained at the beginning or at the end of the prompts, while the center of the message is not considered with the same attention.
Our datasets are significantly bigger than those used in [9], resulting in a lower performance, while even other experiments seem to show a similar relation with the size of the ontology involved[25]. It is important to remember that the search space scales like the square of the size the input ontologies and so it is reasonable to expect problems in this regard. But at the same time the largest ontologies are precisely those for which automatic alignment is most useful in real-world contexts.
2)Semantic complexity: in an alignment it is not mandatory that the relation between two concept be just equivalence. Many types of correspondences are possible, with different levels of complexity. In our reference alignment broader, narrower and related correspondences were present. We completely disregarded the related correspondences, because considering them resulted in a complete performance catastrophe, but we retained broader-narrower(>,<) relations. To our knowledge, this analysis has never been performed in the literature, and it results in a significant fact: GPT-4 is never able to produce the right match in this category, despite trying many times. This is important because it points out that the ability to produce a correspondence can strongly depend on the intrinsic complexity of the relations involved.
By the way, as can be seen in Figure 3, focusing just on equal correspondences results in an improvement in the performance but doesn't bring it back to the values reported in [9]. This confirms the fact that semantic complexity is not the only cause of a low performance in our experiment.
3)Data contamination: as already stated, our reference alignment is not present on the web and so is not subject to the possibility of data contamination, namely the fact that the dataset has been used during pre-training of the LLM. This could contribute to lower our results and make them more compliant with the ones one could expect in real-world use cases.

6. Conclusions and future work

We provided a further step in the evaluation and understanding of LLMs's application to semantic tasks. Our experiment shows that previously reported results[9], obtained with a zero or few-shot approach on the OAEI 2022 conference track, could be overestimated and not compliant with what one should expect in real-world use cases. In fact, when we tested the same prompts of the cited article on an alignment that is bigger and not available online, our TESEO-EuroVoc alignment, the global performance dropped significantly. This could be caused by several reasons: namely, size of the considered thesauri, data contamination and semantic complexity of the relations involved.
With regard to this last reason, we conducted an analysis that shows that semantic complexity is one, but not the only, motivation for the performance drop. In fact, GPT-4 is not able to produce correct correspondences with broader-narrower type of relations, and so the performance raises when restricting the evaluation on the set of equivalence relations, but at the same time even on this smaller set of matches recall, precision and F1 score does not reach values similar to the ones reported in the previous experiment. Moreover we assessed that most of the correspondences found by GPT-4 are between concepts that share the same principal label: this is expected but casts more doubt on the ability of an LLM to solve the task of OA.

Overall, our work confirms that in real-world use cases ontology alignment can be very complicated and that efforts to automate it require special care. LLMs can contribute, but simple prompting seems not to be enough. Furthermore, when evaluating the performance of LLMs on this and other tasks, one must always be careful not to overestimate the results in view of real applications. Size, data contamination, and semantic complexity are significant source of problems that should always be considered.
Future work should focus on two different but interwoven aspects, that is, building more systematic experiments to evaluate what causes the LLMs performance to improve or decrease on one side and integrating their language skills into more complex and task-specific tools on the other.

Acknowledgments

This work has been partially supported by Project ECS 0000024 Rome Technopole, – CUP B83C22002820006, NRP Mission 4 Component 2 Investment 1.5, funded by the European Union – NextGenerationEU and by the KATY project, which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101017453.

References

[1] Ehrig, M. (2006). *Ontology alignment: bridging the semantic gap* (Vol. 4). Springer Science & Business Media.
[2] Shvaiko, P., & Euzenat, J. (2011). Ontology matching: state of the art and future challenges. *IEEE Transactions on knowledge and data engineering*, 25(1), 158-176.
[3] Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., ... & Fedus, W. (2022). Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.
[4] Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., ... & Wen, J. R. (2023). A survey of large language models. *arXiv preprint arXiv:2303.18223*.
[5] Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35, 22199-22213.
[6] Basili, R., Vindigni, M., & Zanzotto, F. M. (2003, October). Integrating ontological and linguistic knowledge for conceptual information extraction. In *Proceedings IEEE/WIC International Conference on Web Intelligence (WI 2003)* (pp. 175-181). IEEE.
[7] Wilks, Y., & Brewster, C. (2009). Natural language processing as a foundation of the semantic web. *Foundations and Trends® in Web Science*, 1(3–4), 199-327.
[8] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33, 1877-1901.
[9] Norouzi, S. S., Mahdavinejad, M. S., & Hitzler, P. (2023). Conversational ontology alignment with chatgpt. *arXiv preprint arXiv:2308.09217*.
[10] Hertling, S., & Paulheim, H. (2023, December). OLaLa: Ontology matching with large language models. *In Proceedings of the 12th Knowledge Capture Conference 2023* (pp. 131-139).
[11] Golchin, S., & Surdeanu, M. (2023). Time travel in llms: Tracing data contamination in large language models. *arXiv preprint arXiv:2308.08493*.
[12] Zhou, K., Zhu, Y., Chen, Z., Chen, W., Zhao, W. X., Chen, X., ... & Han, J. (2023). Don't Make Your LLM an Evaluation Benchmark Cheater. *arXiv preprint arXiv:2311.01964*.
[13] Stellato, A., Turbati, A., Fiorelli, M., Lorenzetti, T., Schmitz, P., Francesconi, E., ... & Batouche, B. (2018, December). Towards the Assessment of Gold-Standard Alignments Between Legal Thesauri. *In JURIX* (pp. 131-140).
[14] Otero-Cerdeira, L., Rodríguez-Martínez, F. J., & Gómez-Rodríguez, A. (2015). Ontology matching: A literature review. *Expert Systems with Applications*, 42(2), 949-971.
[15] Fiorelli, M., Stellato, A., Lorenzetti, T., Schmitz, P., Francesconi, E., Hajlaoui, N., & Batouche, B. (2019). Metadata-driven semantic coordination. In *Metadata and Semantic Research: 13th International Conference, MTSR 2019, Rome, Italy, October 28–31, 2019, Revised Selected Papers* (pp. 16-27). Springer International Publishing.
[16] Jiménez-Ruiz, E., & Cuenca Grau, B. (2011). Logmap: Logic-based and scalable ontology matching. In *The Semantic Web–ISWC 2011: 10th International Semantic Web Conference, Bonn, Germany, October 23-27, 2011, Proceedings, Part I 10* (pp. 273-288). Springer Berlin Heidelberg.
[17] Faria, D., Pesquita, C., Santos, E., Palmonari, M., Cruz, I. F., & Couto, F. M. (2013). *The agreementmakerlight ontology matching system. In On the Move to Meaningful Internet Systems: OTM 2013 Conferences: Confederated International Conferences: CoopIS, DOA-Trusted Cloud, and ODBASE 2013, Graz, Austria, September 9-13, 2013. Proceedings* (pp. 527-541). Springer Berlin Heidelberg.
[18] Kolyvakis, P., Kalousis, A., & Kiritsis, D. (2018, June). Deepalignment: Unsupervised ontology matching with refined word vectors. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (pp. 787-798).

[19] Wang, L. L., Bhagavatula, C., Neumann, M., Lo, K., Wilhelm, C., & Ammar, W. (2018). Ontology alignment in the biomedical domain using entity definitions and context. *arXiv preprint arXiv:1806.07976*.

[20] He, Y., Chen, J., Antonyrajah, D., & Horrocks, I. (2022, June). BERTMap: a BERT-based ontology alignment system. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 36, No. 5, pp. 5684-5691).

[21] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

[22] Liu, Y., & Lapata, M. (2019). Text summarization with pretrained encoders. *arXiv preprint arXiv:1908.08345*.

[23] Chen, M., Tworek, J., Jun, H., Yuan, Q., de Oliveira Pinto, H. P., Kaplan, J., ... & Zaremba, W. (2021). Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

[24] Amini, R., Norouzi, S. S., Hitzler, P., & Amini, R. (2024). Towards Complex Ontology Alignment using Large Language Models. *arXiv preprint arXiv:2404.10329*.

[25] He, Y., Chen, J., Dong, H., & Horrocks, I. (2023). Exploring large language models for ontology alignment. *arXiv preprint arXiv:2309.07172*.

[26] Hertling, S., & Paulheim, H. (2023). OLaLa Results for OAEI 2023.

[27] Zhang, S., Dong, Y., Zhang, Y., Payne, T. R., & Zhang, J. (2024). Large Language Model Assisted Multi-Agent Dialogue for Ontology Alignment. In *The 23rd International Conference on Autonomous Agents and Multi-Agent Systems*.

[28] Qiang, Z., Wang, W., & Taylor, K. (2023). Agent-OM: Leveraging Large Language Models for Ontology Matching. *arXiv preprint arXiv:2312.00326*.

[29] Ranaldi, L., Ruzzetti, E. S., & Zanzotto, F. M. (2023). PreCog: Exploring the relation between memorization and performance in pre-trained language models. *arXiv preprint arXiv:2305.04673*.

[30] Ranaldi, F., Ruzzetti, E. S., Onorati, D., Ranaldi, L., Giannone, C., Favalli, A., ... & Zanzotto, F. M. (2024). Investigating the Impact of Data Contamination of Large Language Models in Text-to-SQL Translation. *arXiv preprint arXiv:2402.08100*.

[31] Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., & Liang, P. (2024). Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12, 157-173.

[32] Amaro, I., Della Greca, A., Francese, R., Tortora, G., & Tucci, C. (2023, July). AI unreliable answers: A case study on ChatGPT. *In International Conference on Human-Computer Interaction* (pp. 23-40). Cham: Springer Nature Switzerland.