LLMs4OM: Matching Ontologies with Large Language Models

No Author Given

No Institute Given

Abstract. Ontology Matching (OM), is a critical task in knowledge integration, where aligning heterogeneous ontologies facilitates data interoperability and knowledge sharing. Traditional OM systems often rely on expert knowledge or predictive models, with limited exploration of the potential of Large Language Models (LLMs). We present the LLMs4OM framework, a novel approach to evaluate the effectiveness of LLMs in OM tasks. This framework utilizes two modules for retrieval and matching, respectively, enhanced by zero-shot prompting across three ontology representations: concept, concept-parent, and concept-children. Through comprehensive evaluations using 20 OM datasets from various domains, we demonstrate that LLMs, under the LLMs4OM framework, can match and even surpass the performance of traditional OM systems, particularly in complex matching scenarios. Our results highlight the potential of LLMs to significantly contribute to the field of OM.

Keywords: Ontology Matching · Ontology Alignment · Large Language Models · Retrieval Augmented Generation · Zero-Shot Testing

1 Introduction

In the dynamic field of information and data management, ensuring the interoperability and integration of varied knowledge systems is critical. Ontologies play a key role in achieving semantic interoperability by providing a structured, understandable framework for both humans and machines [44,31]. However, the proliferation of new ontologies presents challenges in aligning them for seamless communication across different systems [51,34]. Ontology matching (OM) emerges as a vital solution, automating the discovery of correspondences across ontologies [11]. The emergence of Large Language Models (LLMs) in natural language processing has revolutionized the traditional boundaries between human and machine understanding of language, making LLMs highly relevant for OM tasks. Despite initial efforts to apply LLMs to OM [21,30], the rapid development of these models calls for an in-depth exploration of their potential in OM, which this study aims to provide, emphasizing the importance of OM and the promising capabilities of LLMs in addressing its challenges.

To pursue this objective, we present the LLMs4OM framework, which assesses diverse LLMs across various tracks and domains introduced within the

Ontology Alignment Evaluation Initiative (OAEI) [40]. OM aims to map concepts between source $C_{source} \in O_{source}$ and target $C_{target} \in O_{target}$ ontologies. Formally, the task is to identify for any $C_s \in C_{source}$, possible $C_t \in C_{target}$ that $(C_s, C_t, S_{C_s \equiv C_t})$, where $S \in [0,1]$ represents the likelihood of equivalence $C_s \equiv C_t$ [10].

An initial exploratory study using ChatGPT-4 [30], demonstrated the OM task via a conversational, naive approach, where ontologies O_{source} and O_{target} were fully inputted into the LLM to solicit matchings. This approach, however, highlighted two primary drawbacks: i) the limited context length LLMs can process, which may be exceeded by larger ontologies, and ii) the increased likelihood of erroneous or "hallucinated" responses due to the volume of information provided. To address these, LLMs4OM employs a dual-module strategy: first, using the Retrieval-Augmented Generation (RAG) [26] for candidate selection for a given query C_{source} from a knowledge base of C_{target}, and then LLM-based matching, in a second module, for finer accuracy. This approach mitigates the limitations of direct LLM prompting by optimizing for the specific challenges of OM, demonstrating a strategic advancement in leveraging LLMs for ontology alignment. In our study using the LLMs4OM framework, we conduct extensive evaluations, beginning with the RAG module where we explore four retrieval methods: TFIDF [38], sentence-BERT [37], SPECTER2 [42], and OpenAI text-embedding-ada [33]. Subsequently, within the LLM module, we pair these retrieval techniques with seven state-of-the-art LLMs: LLaMA-2 [46], GPT-3.5 [32], Mistral [22], Vicuna [52], MPT [45], Falcon [2], and Mamba [16], to assess their combined effectiveness. Furthermore, a detailed analysis based on our large-scale experiments framed three main research questions (**RQ**s). **RQ1**: What impact do the three concept representations (concept, concept-parent, concept-children), respectively have on improving matching efficacy? **RQ2**: For the RAG module, which retriever performs best per track? (*RQ2.1*) Additionally, how does recall vary in the retrieval module across our different retrieval techniques employed? (*RQ2.2*) **RQ3**: Which LLM performs best per track? (*RQ3.1*) Furthermore, how does the performance of various LLMs differ across the three concept representations for the OM tracks? (*RQ3.2*).

This study's empirical tests are varied not only in their approach but also in the range of ontological knowledge domains they cover. We evaluate LLMs across six tracks of the OAEI campaign, encompassing 20 datasets in total. The primary contributions of this paper are threefold: 1) Introduction of the LLMs4OM, an end-to-end framework that utilizes LLMs for OM; 2) A thorough empirical evaluation of seven state-of-the-art domain-independent LLMs and four retrieval models for their suitability to the various OM tasks; and 3) The source code implementation of the LLMs4OM framework released here https://anonymous.4open.science/r/LLMs4OM.

2   Related Work

Ontology matching, a well-explored research area, has seen diverse methodologies, from traditional techniques [23,13,39,41,1] to recent transformer-based

methods [35,47,9,14,43,4,48,15,14,18,30,21,19], each contributing to advancements in the field. Despite the proven effectiveness of conventional approaches, this work focuses on classifying ontology matching systems, especially those utilizing transformers [49], into three categories based on their research goals: unsupervised learning, supervised learning, and LLM-based approaches.

Unsupervised learning methods in ontology matching often use embeddings for similarity assessments. Techniques such as TTEXTO [35], PropMatch [43], AMD [47], and Matcha [14] primarily leverage BERT [7] variants (e.g., RoBERTa [28], sentence-BERT) to generate ontology embeddings for these calculations. Additionally, some methods combine transformer models with multiple representations: TEXTO integrates GloVe [36] with BERT, AMD pairs knowledge graph embeddings with BERT, GraphMatcher [9] combines universal sentence encoder [5] with graph learning techniques, and PropMatch uses sentence-BERT with TFIDF for enhanced matching accuracy. Supervised OM methods predominantly fine-tune transformer models. Truveta Mapper [4] utilizes ByT5 [50] on the Bio-ML track, employing a sequence-to-sequence approach. LaKERMap [48] focuses on domain-specific tuning with Bio-ClinicalBERT [3]. SORBETmatcher [15] combines BERT with random walks and regression loss for ontology embeddings. Matcha-DL [14] uses sentence-BERT in a semi-supervised setup with a dense network for candidate ranking. BERTMap [18] integrates unsupervised and semi-supervised strategies by initially fine-tuning BERT on ontology texts, and then refining mappings based on ontology structure.

Research on larger parameter models [30,19,21] reveals significant strategies for ontology matching (OM). [30] leverages prompt templates with LLMs to input source and target ontologies, showcasing OM potential. OLaLa [21] utilizes LLaMA-2 models and BERT retrievers to extract top-k matches from target ontologies for LLM prompts, refining final alignments with a precision matcher and filters. LLMap [19] investigates Flan-T5 [6] and GPT-3.5's zero-shot capabilities, focusing on concept labels and structural contexts.

3 LLMs4OM – Methodological Framework

The LLMs4OM framework offers a RAG approach within various LLMs for OM. LLMs4OM uses O_{source} as query $Q(O_{source})$ to retrieve possible matches for for any $C_s \in C_{source}$ from $C_{target} \in O_{target}$. Where, C_{target} is stored in the knowledge base $KB(O_{target})$. Later, C_s and obtained $C_t \in C_{target}$ are used to query the LLM to check whether the (C_s, C_t) pair is a match. As shown in Figure 1, the framework comprises four main steps: 1) Concept representation, 2) Retriever model, 3) LLM, and 4) Post-processing.

1) Concept representation. Within this module, we process the ontologies, to extract the child, parent, and concept-specific representations of ontology elements. These representations will be utilized to generate three distinct input representations: i) Concept (C), a foundational representation that encapsulates the core characteristics of a standalone concept within the ontology, ii) Concept-Parent (CP), extending beyond individual concepts, this representation
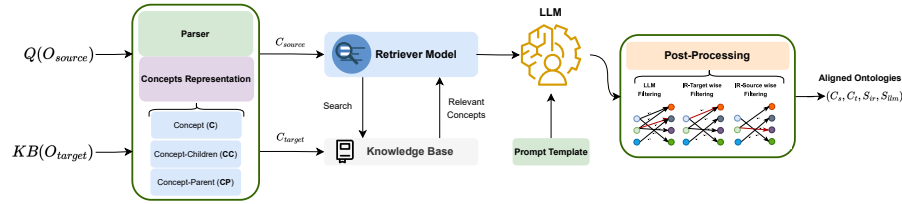
Fig. 1. Overview on LLMs4OM as an end-to-end framework for OM.

establishes the hierarchical relationships by incorporating information about the parent concepts, and iii) Concept-Children (CC) complementing the CP representation which focuses on the descendants of a given concept. These variant representations ensure a comprehensive understanding of ontologies, capturing both individual concepts and their hierarchical relationships, thus supporting the complete nature of ontologies. Subsequently, minor preprocessing is carried out to acquire clean textual data by converting representations to lowercase and removing punctuation. Finally, the parents in CP and children in CC are merged to create a list of parents and children associated with concepts, respectively.

2) Retriever model. First, an embedding extractor model operates by extracting embeddings for $C_{target} \in O_{target}$ and forming an embedding knowledge base for all C_{target}. Next, for a given $C_s \in C_{source}$, using the embedding extractor model, a C_s embedding is generated to calculate cosine similarity across all C_{target}, to identify top_k most similar candidates for alignments. The retrieval model will result in a $\{(C_s, C_{t_1}), ..., (C_s, C_{t_k})\}$ pairs with similarity score of S_{ir} per pair. For other input representations, C can be substituted with CC to include children or with CP to integrate the parent of C within the representations for the retrieval model.

3) LLM. Using obtained $\{(C_s, C_{t_1}), ..., (C_s, C_{t_k})\}$ pairs from the retrieval model, each pair is verbalized as text and replaced in the prompt template to input LLMs. Subsequently, employing the LLM prompting technique [27], inputs are categorized into "yes" and "no" classes using label words such as yes/true/right for the "yes" class and no/false/wrong for the "no" class. Later, the confidence score of S_{llm} is derived from the probabilities assigned to the "yes" and "no" classes label words corresponding to the obtained pairs. The following prompt template is designed to use C, CC, or CP representation of ontology concepts.

> *Classify if two concepts refer to the same real-world entity or not (answer only yes or no).\n### First concept:\n{C_s}\n[Parents|Children]: {$CP|CC$}\n### Second concept:\n{C_t}\n [Parents|Children]: {$CP|CC$}\n### Answer:*

Where in the template {C_s} and {C_t} are placeholders for pair concepts. The notation *"[Parents|Children]: {$CP|CC$}"* offers flexibility in representing ontology concepts, allowing for the inclusion of either parent or children concepts via CP and CC representations.

4) Post-processing. After obtaining the retrieval model similarity score of S_{ir} and LLM's confidence scores of S_{llm} for "yes" and "no" classes, we conducted hy-

brid post-processing to obtain final pairs that match among $(C_s, C_{t_1}), ..., (C_s, C_{t_k})$). The hybrid post-processing involves three steps:

1. *Confidence-driven filtering by LLM*: First, predicted pairs with the "no" class are disregarded. Then, pairs with $S_{llm} > 0.7$ for the "yes" class are retained.
2. *The high precision matcher*: This step applies to the retrieval model similarity score using $S_{ir} > 0.9$. The resulting output consists of exact matches.
3. *Cardinality-based filtering*: Implemented to prevent multiple matches per C_{source} or C_{target} concepts.

This yields $(C_s, C_t, S_{ir}, S_{llm})$ as the set matching between concepts.

4   LLMs4OM – Ontology Matching Evaluations

Evaluation Datasets – OAEI Tracks & Tasks. We selected five tracks from the OAEI campaign, covering diverse domains, and utilized three setups, i.e. concept, concept-children, and concept-parent, for our experiment. These configurations aim to identify the most effective ontology representation for OM, particularly focusing on the equivalence matching problem. The chosen tracks includes: Anatomy) Anatomy [8] (Mouse-Human), biodiv) Biodiversity and Ecology [24] (8 tasks), Phenotype) Disease and Phenotype [17] (DOID-ORDO and HP-MP), CommonKG) Common Knowledge Graphs [12] (Nell-DBpedia and YAGO-Wikidata), Bio-ML) Biomedical Machine Learning [20] (5 tasks), and MSE) Material Sciences and Engineering [29] (MI-EMMO and MI-MatOnto) OAEI tracks which resulted in 20 tasks/datasets.

Evaluation Models – Retrievers & LLMs. As already introduced earlier, in this work, this study evaluates 7 state-of-the-art LLMs across 4 retriever models using the LLMs4OM framework. We assess retrieval models including TFIDF [38], sentence-BERT [37], SPECTER2 [42], and OpenAI text-embedding-ada [33]. Afterward, we combine these with LLMs (number of parameters written in parenthesis) such as LLaMA-2 (7B) [46], GPT-3.5 (174B) [32], Mistral (7B) [22], Vicuna (7B) [52], MPT (7B) [45], Falcon (7B) [2], and Mamba (2.8B) [16] to measure their effectiveness for OM.

LLMs4OM Results. For each track, retriever models with $top_k = 5$ are evaluated across proposed concept representations and results are reported in Figure 2. The assessment includes 7 LLMs with C, CC, and CP input representations, along with retrievers like *text-embedding-ada* and *sentence-BERT*, detailed in Table 1. Approximately 50 runs per dataset were conducted, providing foundational results for further analysis (the complete results are indicated in supplementary material). We focus on zero-shot evaluations of LLMs and retrieval models in addressing our research questions.
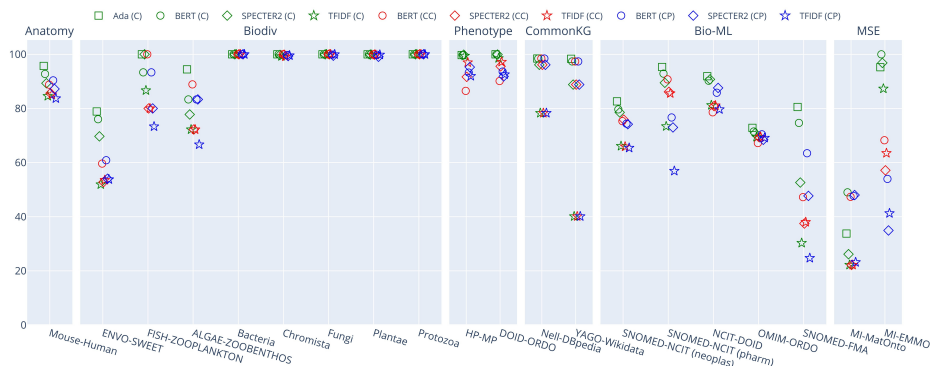
Fig. 2. Comparing retrieval models using recall and $top_k = 5$.

RQ1: What impact do the three concept representations, respectively have on improving matching efficacy? We address this question by analyzing the findings presented in Figure 2, demonstrating the superiority of the C representation across all 20 tasks for retrieval models using the proposed method. Additionally, in Table 1, we find C excelling in 6 tasks, while CP outperforms in 9 tasks. Notably, based on observation of results in Figure 2, SNOMED-FMA tasks from BIO-ML show high sensitivity to input representation. Furthermore, the inclusion of information about concepts, i.e. parents or children, shifts representations towards other concepts within the ontology. However, such information proves to be valuable for LLMs in enhancing their understanding of concepts, as evidenced in the results.

RQ2.1. [Retrieval module] Which retriever performs best per track? Given the results in Figure 2, we analyze this question. Across tracks like ANATOMY, BIODIV, PHENOTYPE, COMMONKG, and BIO-ML, OpenAI *text-embedding-ada* consistently outperforms. However, in MSE track tasks, *sentence-BERT* emerges as the standout performer. Specifically, for the challenging MI-MatOnto task, *sentence-BERT* achieves a 49.00% recall. Combining top retrievers, *text-embedding-ada* and *sentence-BERT*, with LLMs, as shown in Table 1, highlights *sentence-BERT*'s suitability for PHENOTYPE and MSE tracks, while *text-embedding-ada* excels in the remaining 4 tracks. These findings underscore the importance of selecting appropriate retrievers tailored to specific task requirements in LLMs4OM.

RQ2.2: [Retrieval module] How does recall vary in the retrieval module across our different retrieval techniques employed? We investigate this question by comparing retriever models across different values of $top_k \in [5, 10, 20]$. On average, for $top_k = 5$, the retrieval models achieve a recall of 82.09%, increasing to 84.66% for $top_k = 10$, and further to 86.82% for $top_k = 20$. Specifically, given the results in Figure 2, when considering $top_k = 5$, the *text-embedding-ada* retriever achieves a recall of 90.88%, followed by *sentence-BERT* with 86.09%, *SPECTER2* with 82.10%, and *TFIDF* with 75.15%, highlighting the superior performance of *text-embedding-ada* and *sentence-BERT*. However, it's important to note that higher values of top_k lead to increased time complex-

Table 1. Best zero-shot performer model results across 20 tasks, using 7 LLMs, 3 concept representations (C, CP, CC), and 2 retriever models. Bold denotes where LLM4OM using respective models outperforms OAEI 2023 OM systems. The "OAEI" column displays the top F1-score from OAEI 2023.

Track	Tasks	Prec	Rec	F1	Best Model	OAEI
ANATOMY	Mouse-Human	90.82	87.46	89.11	GPT-3.5(C)+Ada	94.10
BIODIV	ENVO-SWEET	59.00	51.67	55.09	Mistral(C)+Ada	71.40
	FISH-ZOOPLANKTON	100	80.00	88.88	LLaMA-2(C)+Ada	92.80
	ALGAE-ZOOBENTHOS	100	38.88	**56.00**	Mistral(C)+Ada	50.00
	TAXR-NCBI(Bacteria)	67.96	99.42	**80.74**	GPT-3.5(CP)+Ada	74.80
	TAXR-NCBI(Chromista)	69.87	98.07	**81.61**	GPT-3.5(CP)+Ada	77.30
	TAXR-NCBI(Fungi)	86.97	99.08	**99.63**	GPT-3.5(CP)+Ada	89.10
	TAXR-NCBI(Plantae)	82.59	96.34	**88.94**	GPT-3.5(CP)+Ada	86.60
	TAXR-NCBI(Protozoa)	86.06	98.59	**91.90**	GPT-3.5(CP)+Ada	85.70
PHENOTYPE	DOID-ORDO	85.79	94.26	**89.83**	Mistral(CP)+BERT	75.50
	HP-MP	76.67	95.40	**85.01**	Mistral(CP)+BERT	81.80
CommonKG	Nell-DBpedia	100	89.14	94.26	GPT-3.5(C)+Ada	96.00
	YAGO-Wikidata	100	85.52	92.19	LLaMA-2(C)+Ada	94.00
Bio-ML	NCIT-DOID (disease)	86.19	80.06	83.01	GPT-3.5(C)+Ada	90.80
	OMIM-ORDO (disease)	71.80	57.96	64.14	GPT-3.5(CC)+Ada	71.50
	SNOMED-FMA(body)	21.12	32.60	25.64	GPT-3.5(CP)+Ada	78.50
	SNOMED-NCIT(neoplas)	46.96	52.96	49.47	GPT-3.5(CP)+Ada	77.10
	SNOMED-NCIT(pharm)	81.84	58.19	68.02	GPT-3.5(CC)+Ada	75.20
MSE	MI-EMMO	96.66	92.06	**94.30**	LLaMA-2(CC)+BERT	91.80
	MI-MatOnto	89.70	20.19	32.97	MPT(C)+BERT	33.90

ity and longer waiting times. Consequently, we may opt to sacrifice a 4% average recall with $top_k = 20$ in exchange for reduced waiting times using $top_k = 5$.

RQ3.1: [LLM module] Which LLM performs best per track? We examine this question given the results in Table 1. The **Best Model** column in the table showcases the top-performing models, starting with GPT-3.5, followed by Mistral-7B, LLaMA-2-7B, and finally MPT-7B among the 7 LLMs. The summary of best model results concerning OM systems proposed in OAEI 2023 [40] using F1-score are as follows: for MI-EMMO LLaMA-2-7B 94.30% > Matcha 91.8% [14], for HP-MP Mistral-7B 85.01% > LogMap 81.8% [23], for DOID-ORDO Mistral-7B 89.93% > AML 75.5% [13], for ALGAE-ZOOBENTHOS Mistral-7B 56.00% > OLaLa 50.0% [21], for TAXR-NCBI(Bacteria) GPT-3.5 80.74% > LogMapLt 77.3% [23], for TAXR-NCBI(Fungi) GPT-3.5 99.63% > OLaLa 89.9% [21], for TAXR-NCBI(Plantae) GPT-3.5 88.94% > OLaLa 86.6% [21], and for TAXR-NCBI(Protozoa) GPT-3.5 91.90% > OLaLa 85.7% [21].

RQ3.2: [LLM module] How does the performance of various LLMs differ across the three concept representations for the OM tracks? Using results from Table 1, we find LLMs perform better with additional contexts like parents or children, as seen in tasks across BIODIV, PHENOTYPE, and BIO-ML tracks. In BIODIV, CP consistently boosts LLM performance, especially in TAXR-NCBI tasks. Similarly, PHENOTYPE tasks show improved results with

CP representations, notably in DOID-ORDO (89.83%) and HP-MP (85.01%). However, Bio-ML tasks exhibit mixed outcomes; some like NCIT-DOID perform well (83.01%) with C representation, while others like SNOMED-FMA (25.64%) struggle even with CP representation. In MSE, tasks vary greatly; for example, MI-EMMO achieves 94.30% success with LLaMA-2-7B and CC representation. This highlights the importance of selecting the right model architecture and contextual representation for each task. Overall, this analysis stresses the significance of context in LLMs across diverse domains, emphasizing the need for tailored approaches based on task specifics.

5   Discussion

Benefits of using our RAG technique for OM. Integrating retrieval with LLMs yields benefits. Initially, querying LLM with all pairs led to impractical $O(n^2)$ time complexity, particularly with larger datasets. However, integrating retrieval reduces complexity to linear $O(kn)$, enabling faster processing while preserving LLM-generated confidence scores. Additionally, providing all ontologies at once to the model, as seen in [30], results in mixed outputs, posing challenges in computing matching scores and increasing the risk of high hallucination, especially with larger ontologies.

Low performance on the Bio-ML track. LLMs4OM showed low performance compared to traditional methods across the Bio-ML track tasks. Analyzing their performance with two retrievers, we found an average F1-score of 53% with *text-embedding-ada* and around 44% with *sentence-BERT*. Despite strong retriever performance in candidate retrieval (see Figure 2), LLMs' overall performance remains low. There is a general under-performance on this track when LLM solutions have been used, and given the low performance, we tested domain-specific LLM i.e. BioMistral-7B [25] and we obtained the following results (* refers to the best model result from Table 1). NCIT-ORDO 69.04% < 83.01%*, OMIM-ORDO 57.84% < 64.14%*, SNOMED-FMA 33.98% > 25.64%*, SNOMED-NCIT(neoplas) 46.24% > 49.47%*, and SNOMED-NCIT(pharm) 62.00% < 68.02%*. The low performance on all the tasks even with domain-specific LLM, showed a need for a different approach with LLMs for Bio-ML track.

6   Conclusion

The proposed LLMs4OM framework highlights the efficacy of LLMs in OM, specifically in aligning diverse ontologies for knowledge engineering. By rigorously evaluating 20 tasks spanning different domains, our framework shows that LLMs, when combined with retriever models and guided by zero-shot prompting while utilizing C, CP, and CC representations, can surpass traditional OM systems in complex matching scenarios. These findings underscore the significant potential of LLMs in OM, paving the way for further exploration.

Supplemental Material Statement: Our detailed results and codebase are available https://anonymous.4open.science/r/LLMs4OM.

References

1. Algergawy, A., Babalou, S., Klan, F., König-Ries, B.: Ontology modularization with oapt. Journal on Data Semantics **9**(2), 53–83 (September 1 2020). https://doi.org/10.1007/s13740-020-00114-7, https://doi.org/10.1007/s13740-020-00114-7

2. Almazrouei, E., Alobeidli, H., Alshamsi, A., Cappelli, A., Cojocaru, R., Debbah, M., Étienne Goffinet, Hesslow, D., Launay, J., Malartic, Q., Mazzotta, D., Noune, B., Pannier, B., Penedo, G.: The falcon series of open language models (2023)

3. Alsentzer, E., Murphy, J., Boag, W., Weng, W.H., Jin, D., Naumann, T., McDermott, M.: Publicly available clinical BERT embeddings. In: Proceedings of the 2nd Clinical Natural Language Processing Workshop. pp. 72–78. Association for Computational Linguistics, Minneapolis, Minnesota, USA (Jun 2019). https://doi.org/10.18653/v1/W19-1909, https://www.aclweb.org/anthology/W19-1909

4. Amir, M., Baruah, M., Eslamialishah, M., Ehsani, S., Bahramali, A., Naddaf-Sh, S., Zarandioon, S.: Truveta mapper: A zero-shot ontology alignment framework (2023)

5. Cer, D., Yang, Y., yi Kong, S., Hua, N., Limtiaco, N., John, R.S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Sung, Y.H., Strope, B., Kurzweil, R.: Universal sentence encoder (2018)

6. Chung, H.W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., Webson, A., Gu, S.S., Dai, Z., Suzgun, M., Chen, X., Chowdhery, A., Castro-Ros, A., Pellat, M., Robinson, K., Valter, D., Narang, S., Mishra, G., Yu, A., Zhao, V., Huang, Y., Dai, A., Yu, H., Petrov, S., Chi, E.H., Dean, J., Devlin, J., Roberts, A., Zhou, D., Le, Q.V., Wei, J.: Scaling instruction-finetuned language models (2022)

7. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding (2019)

8. Dragisic, Z., Ivanova, V., Li, H., Lambrix, P.: Experiences from the anatomy track in the ontology alignment evaluation initiative. J Biomed Semant **8**(1), 56 (2017). https://doi.org/10.1186/s13326-017-0166-5, https://doi.org/10.1186/s13326-017-0166-5

9. Efeoglu, S.: Graphmatcher: a graph representation learning approach for ontology matching. In: OM@ISWC. CEUR Workshop Proceedings, vol. 3324, pp. 174–180. CEUR-WS.org (2022)

10. Euzenat, J., Meilicke, C., Stuckenschmidt, H., Shvaiko, P., Trojahn, C.: Ontology Alignment Evaluation Initiative: Six Years of Experience, pp. 158–192. Springer Berlin Heidelberg, Berlin, Heidelberg (2011). https://doi.org/10.1007/978-3-642-22630-4_6

11. Euzenat, J., Shvaiko, P.: Ontology Matching. Springer Publishing Company, Incorporated, 2nd edn. (2013)

12. Fallatah, O., Zhang, Z., Hopfgartner, F.: A gold standard dataset for large knowledge graphs matching (December 2020), https://eprints.whiterose.ac.uk/173366/, © 2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (http://creativecommons.org/licenses/by/4.0).

13. Faria, D., Pesquita, C., Santos, E., Palmonari, M., Cruz, I.F., Couto, F.M.: The agreementmakerlight ontology matching system. In: Meersman, R., Panetto, H., Dillon, T., Eder, J., Bellahsene, Z., Ritter, N., De Leenheer, P., Dou, D. (eds.) On the Move to Meaningful Internet Systems: OTM 2013 Conferences. pp. 527–541. Springer Berlin Heidelberg, Berlin, Heidelberg (2013)

14. Faria, D., Silva, M.C., Cotovio, P., Ferraz, L., Balbi, L., Pesquita, C.: Results for matcha and matcha-dl in OAEI 2023. In: OM@ISWC. CEUR Workshop Proceedings, vol. 3591, pp. 164–169. CEUR-WS.org (2023)

15. Gosselin, F., Zouaq, A.: Sorbet: A siamese network for ontology embeddings using a distance-based regression loss and bert. In: Payne, T.R., Presutti, V., Qi, G., Poveda-Villalón, M., Stoilos, G., Hollink, L., Kaoudi, Z., Cheng, G., Li, J. (eds.) The Semantic Web – ISWC 2023. pp. 561–578. Springer Nature Switzerland, Cham (2023)

16. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752 (2023)

17. Harrow, I., Jiménez-Ruiz, E., Splendiani, A., Romacker, M., Woollard, P., Markel, S., Alam-Faruque, Y., Koch, M., Malone, J., Waaler, A.: Matching disease and phenotype ontologies in the ontology alignment evaluation initiative. Journal of Biomedical Semantics **8**(1), 55 (2017). https://doi.org/10.1186/s13326-017-0162-9, https://doi.org/10.1186/s13326-017-0162-9

18. He, Y., Chen, J., Antonyrajah, D., Horrocks, I.: Bertmap: A bert-based ontology alignment system (2022)

19. He, Y., Chen, J., Dong, H., Horrocks, I.: Exploring large language models for ontology alignment (2023)

20. He, Y., Chen, J., Dong, H., Jiménez-Ruiz, E., Hadian, A., Horrocks, I.: Machine learning-friendly biomedical datasets for equivalence and subsumption ontology matching. In: Sattler, U., Hogan, A., Keet, M., Presutti, V., Almeida, J.P.A., Takeda, H., Monnin, P., Pirrò, G., d'Amato, C. (eds.) The Semantic Web – ISWC 2022. pp. 575–591. Springer International Publishing, Cham (2022)

21. Hertling, S., Paulheim, H.: Olala: Ontology matching with large language models. In: Proceedings of the 12th Knowledge Capture Conference 2023. p. 131–139. K-CAP '23, Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3587259.3627571

22. Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L.R., Lachaux, M.A., Stock, P., Scao, T.L., Lavril, T., Wang, T., Lacroix, T., Sayed, W.E.: Mistral 7b (2023)

23. Jiménez-Ruiz, E., Cuenca Grau, B.: Logmap: Logic-based and scalable ontology matching. In: Aroyo, L., Welty, C., Alani, H., Taylor, J., Bernstein, A., Kagal, L., Noy, N., Blomqvist, E. (eds.) The Semantic Web – ISWC 2011. pp. 273–288. Springer Berlin Heidelberg, Berlin, Heidelberg (2011)

24. Karam, N., Khiat, A., Algergawy, A., Sattler, M., Weiland, C., Schmidt, M.: Matching biodiversity and ecology ontologies: challenges and evaluation results. The Knowledge Engineering Review **35**, e9 (2020). https://doi.org/10.1017/S0269888920000132,   https://doi.org/10.1017/S0269888920000132

25. Labrak, Y., Bazoge, A., Morin, E., Gourraud, P.A., Rouvier, M., Dufour, R.: Biomistral: A collection of open-source pretrained large language models for medical domains (2024)

26. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., tau Yih, W., Rocktäschel, T., Riedel, S., Kiela, D.: Retrieval-augmented generation for knowledge-intensive nlp tasks (2021)

27. Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., Neubig, G.: Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing (2021)

28. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach (2019)
29. Nas, E., Huschka, M.: MSE Benchmark. https://github.com/EngyNasr/MSE-Benchmark (2023)
30. Norouzi, S.S., Mahdavinejad, M.S., Hitzler, P.: Conversational ontology alignment with chatgpt. In: OM@ISWC. CEUR Workshop Proceedings, vol. 3591, pp. 61–66. CEUR-WS.org (2023)
31. Noy, N., Mcguinness, D.: Ontology development 101: A guide to creating your first ontology. Knowledge Systems Laboratory **32** (01 2001)
32. OpenAI: Chatgpt. https://openai.com/chat-gpt/ (2023), accessed May 5, 2023
33. OpenAI: New and improved embedding model (January 15 2023), https://openai.com/blog/new-and-improved-embedding-model, retrieved on December 15, 2022
34. Osman, I., Ben Yahia, S., Diallo, G.: Ontology integration: Approaches and challenging issues. Information Fusion **71**, 38–63 (2021). https://doi.org/10.1016/j.inffus.2021.01.007
35. Peng, Y., Alam, M., Bonald, T.: Ontology matching using textual class descriptions. In: OM@ISWC. CEUR Workshop Proceedings, vol. 3591, pp. 67–72. CEUR-WS.org (2023)
36. Pennington, J., Socher, R., Manning, C.: GloVe: Global vectors for word representation. In: Moschitti, A., Pang, B., Daelemans, W. (eds.) Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1532–1543. Association for Computational Linguistics, Doha, Qatar (Oct 2014). https://doi.org/10.3115/v1/D14-1162, https://aclanthology.org/D14-1162
37. Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks (2019)
38. Sammut, C., Webb, G.I. (eds.): TF–IDF, pp. 986–987. Springer US, Boston, MA (2010). https://doi.org/10.1007/978-0-387-30164-8_832
39. Sharma, A., Jain, S.: Lsmatch and lsmatch-multilingual results for OAEI 2023. In: OM@ISWC. CEUR Workshop Proceedings, vol. 3591, pp. 159–163. CEUR-WS.org (2023)
40. Shvaiko, P., Euzenat, J., Jiménez-Ruiz, E., Hassanzadeh, O., Trojahn, C. (eds.): Proceedings of the 18th International Workshop on Ontology Matching co-located with the 22nd International Semantic Web Conference (ISWC 2023), Athens, Greece, November 7, 2023, CEUR Workshop Proceedings, vol. 3591. CEUR-WS.org (2023)
41. da Silva, J., Revoredo, K., Baião, F., Lima, C.: ALIN results for OAEI 2023. In: OM@ISWC. CEUR Workshop Proceedings, vol. 3591, pp. 140–145. CEUR-WS.org (2023)
42. Singh, A., D'Arcy, M., Cohan, A., Downey, D., Feldman, S.: Scirepeval: A multi-format benchmark for scientific document representations. ArXiv **abs/2211.13308** (2022)
43. Sousa, G., Lima, R., Trojahn, C.: Combining word and sentence embeddings with alignment extension for property matching. In: OM@ISWC. CEUR Workshop Proceedings, vol. 3591, pp. 91–96. CEUR-WS.org (2023)
44. Stephan, G. Pascal, H. Andreas, A. Knowledge Representation and Ontologies, pp. 51–105. Springer Berlin Heidelberg, Berlin, Heidelberg (2007). https://doi.org/10.1007/3-540-70894-4_3
45. Team, M.N.: Introducing mpt-7b: A new standard for open-source, commercially usable llms (2023), www.mosaicml.com/blog/mpt-7b, accessed: 2023-05-05

46. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C.C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., at el.: Llama 2: Open foundation and fine-tuned chat models (2023)
47. Wang, Z.: AMD results for OAEI 2023. In: OM@ISWC. CEUR Workshop Proceedings, vol. 3591, pp. 146–153. CEUR-WS.org (2023)
48. Wang, Z.: Contextualized structural self-supervised learning for ontology matching (2023)
49. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., et al.: Huggingface's transformers: State-of-the-art natural language processing (2020)
50. Xue, L., Barua, A., Constant, N., Al-Rfou, R., Narang, S., Kale, M., Roberts, A., Raffel, C.: Byt5: Towards a token-free future with pre-trained byte-to-byte models (2022)
51. Zhang, X., Zhao, C., Wang, X.: A survey on knowledge representation in materials science and engineering: An ontological perspective. Computers in Industry **73**, 8–22 (2015). https://doi.org/10.1016/j.compind.2015.07.005
52. Zheng, L., Chiang, W.L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E.P., Zhang, H., Gonzalez, J.E., Stoica, I.: Judging llm-as-a-judge with mt-bench and chatbot arena (2023)