From Matching to Retrieval: A New Role for LLMs in Ontology Alignment

Wenxin Hu[1], Ryutaro Ichise[1]

[1]Institute of Science Tokyo, 2-12-1 Ookayama, Meguro, Tokyo, Japan

Abstract

Ontology alignment (or ontology matching) is essential for knowledge integration, as many real-world applications rely on cross-domain knowledge. Effective alignment enhances data interoperability and facilitates seamless knowledge sharing. With the rise of Large Language Models (LLMs) and their strong natural language understanding, they offer promising potential to improve ontology alignment. While most existing approaches utilize LLMs primarily in matching stage, their potential in the retrieval stage remains underexplored. In this paper, we propose a novel approach that integrates LLMs into the retrieval process, investigating the feasibility of using zero-shot prompting and the models' built-in commonsense to augment traditional retrieval methods. We evaluate our approach on several benchmark tasks from the Ontology Alignment Evaluation Initiative (OAEI). Results show that it can match or even outperform current state-of-the-art systems on certain datasets, underscoring the promise of incorporating LLMs into the retrieval phase of ontology alignment.

Keywords

Ontology Alignment, Large Language Models, Retrieval-Augmented Generation

1. Introduction

In the Semantic Web, individual ontologies are often incomplete and context-dependent, yet real-world applications increasingly demand integration across diverse knowledge domains. While entity alignment in Knowledge Graphs has been well-studied, the related task of Ontology Alignment (OA) — identifying semantic correspondences between entities (e.g., classes or properties) across ontologies — remains comparatively underexplored [1]. Also known as ontology matching, OA plays a critical role in resolving heterogeneity and enabling cross-domain interoperability in knowledge representation and reasoning systems [2]. Traditionally reliant on expert manual work, OA remains a challenge, as automated methods still lack the accuracy needed for broad industrial use [3].

Many automated OA approaches depend on domain-specific fine-tuning using large labeled datasets [4, 5]. This reliance arises because ontologies are typically designed within narrow, implicitly defined contexts, lacking the external background knowledge needed to infer semantic equivalence [6]. In recent years, Large Language Models (LLMs) have shown strong potential to generalize across domains by leveraging their pre-trained commonsense and linguistic knowledge. As a result, several studies have explored the use of LLMs for ontology alignment, primarily focusing on the matching stage through prompt engineering, task formulation, and model selection [7]. A common approach among state-of-the-art LLM-based systems is the *retrieve-then-prompt* pipeline: first, the most relevant target classes are retrieved; then, these are used to prompt the LLM to predict the most likely mapping correspondences [8].

However, one critical stage in the OA pipeline remains underexplored: the retrieval phase, where candidate alignments are selected before final matching. This stage plays a crucial role in determining alignment quality, as the set of retrieved candidates directly constrains what can be matched [9]. Yet, most current systems approach retrieval as a surface-level similarity task, using raw labels or structural cues without optimizing how ontology classes are represented prior to retrieval. In particular, the

OM 2025: The 20th International Workshop on Ontology Matching collocated with the 24th International Semantic Web Conference (ISWC 2025), November 2nd, 2025, Nara, Japan

hu.w.c84c@m.isct.ac.jp (W. Hu); ichise@iee.e.titech.ac.jp (R. Ichise)

0000-0003-3449-5980 (W. Hu); 0000-0001-8474-0150 (R. Ichise)

challenge of concept translation, i.e., expressing ontology classes in a way that LLMs can accurately interpret, remains an open problem [7]. We seek to extend the role of LLMs beyond the final matching step by integrating them into the retrieval phase, to enrich the semantic representation of input classes and improve the overall alignment process. We refer to this integration as *infusion* throughout the paper. Furthermore, a precision-compensated refinement strategy is proposed to improve the resulting mappings. Our hypothesis is that enriching ontology classes with contextual and commonsense knowledge during the retrieval phase, combined with a precision-compensated refinement strategy, enhances the quality and relevance of candidate matches and leads to more accurate and robust alignment results.

To validate this hypothesis, we propose a novel *infuse-retrieve-prompt* framework that enhances the conventional Retrieval-Augmented Generation (RAG) pipeline. In our approach, verbalized ontology classes are first infused with additional knowledge via an LLM to produce enriched textual representations. These are then embedded using SBERT [10], enabling more semantically meaningful candidate retrieval. The retrieved pairs are finally evaluated by prompting an LLM to generate alignment decisions along with confidence scores. To further refine alignment outcomes, we employ a precision-compensated strategy. By adjusting decision thresholds, the system can recover plausible correspondences that might otherwise be discarded due to marginal similarity scores. This approach improves recall while maintaining competitive precision on certain datasets. Our approach is built entirely on open-source LLMs, ensuring full reproducibility and flexibility across different deployment scenarios. We also evaluate the framework across multiple datasets from the 2023 Ontology Alignment Evaluation Initiative.

Our main contributions are as follows:

1. *Infuse-Retrieve-Prompt* framework — a novel LLM-integrated approach that enriches concept representations and advances beyond traditional LLM-only matching paradigms
2. Precision-Compensated Strategy — balancing precision with LLM confidence to recover correspondences near similarity thresholds
3. Evaluation — comprehensive benchmarking against a baseline alignment system and OAEI (Ontology Alignment Evaluation Initiative) benchmarks across multiple tracks.

2. Related Work

Ontology alignment has undergone significant development, evolving from conventional rule-based and string similarity methods to the adoption of machine learning and language model-based techniques [5]. In this section, we outline two primary strands of existing research: 1) conventional and pre-trained model-based ontology alignment systems, and 2) recent systems that leverage LLMs to improve alignment performance.

2.1. Traditional and Pre-trained Model-Based Alignment Systems

Extensive research has been devoted to OA, with the field evolving alongside advancements in computational techniques. Before the emergence of LLMs, OA systems generally fell into three main categories: lexical matching, machine learning-based approaches, and pre-trained language model (PLM)-based systems. Traditional systems such as LogMap [11] and AML [12] rely primarily on lexical similarity and occasionally incorporate logical reasoning. While these systems remain widely used due to their robustness and efficiency across diverse benchmarks, they are inherently limited to surface-level matching and often struggle to capture deeper semantic relationships [5]. To address these limitations, subsequent research introduced machine learning (ML) techniques, as seen in LogMap-ML [4] and DeepAlignment [13]. These methods improved adaptability and alignment quality by learning from labeled data. However, they typically require extensive supervised training and manual feature engineering, making them labor-intensive and less scalable across domains [1]. The emergence of pre-trained language models, such as BERT, introduced a significant advancement by providing rich contextualized representations

learned from large, unlabeled corpora. This enabled systems like BERTMap [5] and KERMIT [14] to outperform traditional ML models in certain alignment scenarios, while reducing reliance on handcrafted features. Despite their advantages, many PLM-based approaches still require fine-tuning on domain-specific alignment data to reach optimal performance. This reintroduces some of the challenges they were meant to alleviate, including manual data preparation and scalability constraints.

2.2. LLM based Ontology Alignment Systems

Recent advances in LLMs have significantly influenced the field of OA, prompting the development of new strategies aimed at improving scalability and semantic accuracy. One widely adopted pipeline is the *retrieve-then-prompt* framework, comprising four main steps: verbalization, candidate retrieval, LLM-based matching, and postprocessing. In the verbalization step, source and target ontologies are transformed into textual representations that capture the semantics of individual concepts, optionally extended with hierarchical context such as parent–child relationships. These representations are then preprocessed for clarity [15]. During candidate retrieval, embedding-based methods are used to generate vector representations of concepts, allowing the model to retrieve top-k target candidates based on similarity to source concepts [16]. In the LLM-based matching phase, retrieved concept pairs are formatted into natural language prompts and processed by an LLM to assess semantic equivalence through binary classification, producing a confidence score [7]. Finally, postprocessing refines the alignment results by applying several filters, such as LLM confidence filtering, cardinality constraints, and exact match detection, to improve overall precision and ensure consistent one-to-one mappings [7, 15]. This strategy effectively reduces the computational cost of exhaustive comparisons while maintaining competitive performance [8].

Prompt engineering has also played a central role in enhancing LLM-based OA systems. Studies have explored various ways of presenting alignment tasks, including pairwise prompts [7, 17], full-ontology prompts [18], and prompts enriched with handcrafted rules or in-context examples [16]. These have been tested in both zero-shot and few-shot settings [19, 15, 16, 20], with findings showing that well-designed zero-shot prompts, especially those providing explicit task instructions, can achieve alignment quality comparable to few-shot methods [16, 20].

Several systems have also experimented with retrieval-augmented generation (RAG) techniques [21], embedding ontologies into a vector space to support more semantically informed retrieval before prompting LLMs for final decision-making [22]. Despite these innovations, most systems continue to rely on conventional candidate representations and pre-retrieval filtering strategies. This often limits the semantic depth of the retrieved candidates, particularly in cases involving sparse aligned ontologies.

To ensure alignment quality, current approaches frequently employ fixed similarity thresholds to control precision during the final mapping stage [7, 15, 8]. While effective in filtering noisy matches, this practice can also lead to the exclusion of semantically valid correspondences that fall just below the threshold, especially when LLM predictions involve nuanced or borderline decisions. Moreover, although LLMs have demonstrated strong capabilities in understanding individual entity descriptions, their potential for improving earlier stages, such as enriching candidate representation with commonsense or contextual signals, remains underexplored.

These gaps point to several opportunities for advancement, particularly in developing more semantically enriched candidate representations and incorporating refinement strategies that reconcile precision with the probabilistic confidence of LLMs. Addressing these limitations holds potential for enhancing the robustness and adaptability of LLM-based ontology alignment pipelines.

3. Methodology

Ontology alignment is the process of identifying semantic correspondences between entities in different ontologies, typically involving relationships such as equivalence, subsumption, or more complex semantic mappings [23]. In this work, we concentrate on identifying equivalence between classes and focus on simple, pairwise alignments, that is, one-to-one correspondences between classes from

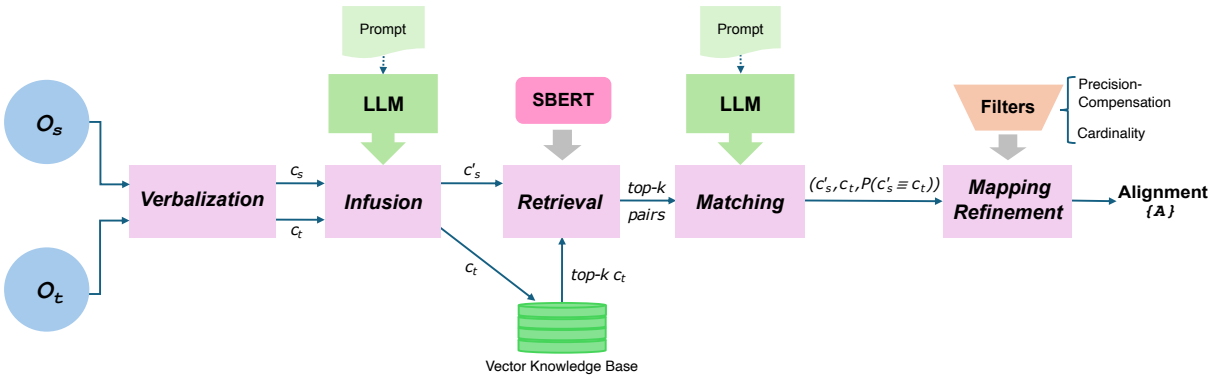

Figure 1: An overview of the *infuse-retrieve-prompt* framework.

two ontologies [24]. The alignment aims to determine whether two classes, $c_s \in O_s$ and $c_t \in O_t$, are semantically equivalent, represented as the triple $(c_s, c_t, P(c_s \equiv c_t))$. Here, $P \in [0, 1]$ denotes the confidence score, with higher values indicating a stronger semantic match between c_s and c_t [5]. To achieve this, we propose a novel *infuse-retrieve-prompt* pipeline for ontology alignment. Figure 1 illustrates the overall framework, which comprises five key components: verbalization, infusion, retrieval, matching, and mapping refinement. The approach enriches the semantic representations of ontology classes using LLMs, retrieves candidate matches based on embedding similarity, and refines the final alignments through LLM-based matching supported by filtering mechanisms, including a precision-compensated strategy. Each component is detailed in the following subsections with illustrative examples.

3.1. Verbalization

At the initial stage, matching candidates must be verbalized from the two input ontologies, O_s and O_t, to prepare them for processing by the LLM in later phases. This involves two steps: (1) extracting ontology classes, and (2) verbalizing these classes into textual form.

First, the input ontologies are parsed to extract standalone classes. Structural information (e.g., hierarchical relations) is not considered at this point, as previous research has shown that incorporating structural context in LLM prompts does not significantly improve ontology alignment performance [2]. Since both the LLM and the SBERT retriever operate exclusively on textual input, each class must be converted into a natural language representation. To ensure that this representation remains focused and avoids introducing noise that might mislead the retriever [15], we limit the verbalization to concise, unambiguous textual elements, specifically, the class labels. Depending on the ontology format (OWL or SKOS), these are extracted using properties such as *skos:prefLabel* or *rdfs:label*, while the class IRIs are retained as unique identifiers.

3.2. Infusion

While evaluating existing ontology alignment pipelines, we identified a common limitation: concept labels are often abbreviated, under-specified, or contextually sparse, which hinders accurate candidate retrieval and poses challenges for LLMs to make correct decision during the matching stage. To mitigate this, we introduce an LLM-based infusion module that enriches concept representations by leveraging natural language understanding and commonsense knowledge.

The infusion process begins with source concept $c_s \in O_s$, which are submitted to an LLM to evaluate their clarity and interpretability. If a label is found to be overly brief or ambiguous, the LLM generates a more informative reformulation, such as expanding abbreviations into their full expressions, to provide additional context for retrieval. To prevent noise, the LLM is guided to produce only minimal and

plausible changes, avoiding speculative elaboration. This process yields enriched candidate sets:

$$c'_s = c_s \cup E$$

where E is a set of clarified or expanded expressions (can be empty). These enriched labels, in combination with the originals, are then passed to the retriever, offering a broader semantic surface, especially in cases where the original labels lack context. Importantly, although the enriched variants are semantically close to the originals, the added clarity helps ensure that potentially relevant candidates are not missed. Beyond retrieval, these enriched inputs also provide downstream benefits in the LLM-based matching stage. Since the final alignment relies on prompting the LLM to assess semantic equivalence between concept pairs, clearer and more descriptive labels facilitate better contextual understanding, improving the reliability of the LLM's mapping decisions.

To reduce computational overhead, only source concepts undergo enrichment. The module operates in a zero-shot setting using a structured prompt template, without any fine-tuning or in-domain examples. To provide the LLM with clearer guidance, we explicitly include two concrete examples (e.g., chemical symbols or technical shorthand), following prior work showing that explicit task instructions can improve LLM-generated outputs [16, 20]. The template presents input labels and task instructions in a controlled format, as detailed below:

> You are a domain-aware assistant specialized in {context}. When given an input of a text or phrase, your task is to:
>
> Evaluate and check the length of the following text: {concept}
>
> 1. If the input length is 3 characters or fewer, treat it likely as abbreviation (e.g. chemical symbol or technical shorthand) and expand it to its exact full form.
> 2. If the input length is more than 3 characters, evaluate if it is a clear and valid word or phrase; if so, return it unchanged; if not, return its exact and clear form.
> 3. If the meaning is unknown, return the original text.
>
> Respond with only the word or phrase — no explanations.

3.3. Retrieval

Due to the limited context window that LLMs can process [18], it is not feasible to analyze entire ontologies at once. A common solution is to focus on each concept individually and retrieve a set of potentially related concepts as alignment candidates. At this stage, the primary objective is to maximize recall, the ability to include as many correct matches as possible, regardless of precision.

To accomplish this, we use a Sentence-BERT (SBERT) embedding model to encode all target ontology concepts c_t, constructing a searchable vector-based knowledge base. Each source concept c'_s is also embedded, and cosine similarity is computed between it and all c_t embeddings. The *top-k* most similar target concepts are retrieved for each source concept, yielding a candidate set of alignment pairs, each associated with a similarity score.

This embedding-based retrieval is crucial for handling large ontologies that exceed the input size limits of current LLMs. Unlike traditional lexical approaches that depend on surface-level textual overlap, SBERT enables semantic similarity matching, capturing related terms even when terms have no textual similarity [25]. As a result, this step generates a high-recall pool of candidates, which lays the foundation for more accurate alignment in downstream matching stages.

3.4. Matching

In this stage, LLMs are employed to assess the plausibility of each candidate alignment. There are two main strategies for presenting candidates to LLMs: (1) evaluating each candidate correspondence independently (binary decision), and (2) presenting all candidate target classes for a given source class simultaneously, allowing the model to choose the best match (multiple-choice decision) [7]. In our

work, we adopt the binary decision strategy. Although the multiple-choice approach can reduce the number of LLM queries, it has been shown to degrade alignment quality [8].

Each retrieved candidate pair (c_s', c_t) retrieved in the previous step is verbalized into natural language and embedded into a prompt template designed for binary classification. Using established prompting techniques [26], the LLM is asked whether the pair refers to the same concept. The model's output is interpreted using predefined label words, with *yes, correct, true...* indicating positive alignments, and *no, incorrect, wrong...* representing negatives.

To optimize performance, we adopt the prompt template that previously achieved the best F1 score [7], tailoring it to our use case by adding relevant context information. The final prompt format is as follows:

> Classify if the following two concepts are the same in the context of {context} (answer only yes or no).
>
> ### First concept:
>
> {source}
>
> ### Second concept:
>
> {target}
>
> ### Answer:

Incorporating confidence scores is essential in LLM-based alignment to mitigate hallucination, filter unreliable mappings, and support thresholding for high-related outputs. To compute a confidence score $P(c_s', c_t)$, we analyze the output distribution of the LLM by extracting generation probabilities for a set of predefined label words (e.g., "Yes" or "No"). After applying a softmax over the full vocabulary, we normalize the highest probability among positive labels by the sum of the top positive and negative label probabilities, following a previous method [7]:

$$\text{Confidence} = \frac{\max(P_{\text{positive}})}{\max(P_{\text{positive}}) + \max(P_{\text{negative}})}$$

If the resulting confidence score exceeds a predefined threshold (e.g., 0.5), the candidate pair is accepted as a valid alignment. This probabilistic scoring approach improves alignment quality by filtering out semantically weak matches, especially those that might have been incorrectly prioritized by the embedding model due to superficial lexical similarity.

3.5. Mapping Refinement

After retrieving a set of scored candidate alignments (c_s', c_t, $P(c_s' \equiv c_t)$), we refine the results through a multi-stage filtering process. First, we apply three standard filters commonly used in state-of-the-art alignment systems. The thresholds are informed by related studies that define the similarity scores corresponding to exact matches and low-confidence predictions for LLMs [7, 15].

- a cardinality filter to enforce one-to-one mappings by removing the mapping with the lower similarity score when multiple candidates share the same source concept,
- a high-precision filter to retain only exact matches with similarity scores above a threshold, and
- a confidence filter to eliminate low-confidence predictions based on LLM output.

In addition to these traditional filters, we introduce a precision-compensated strategy designed to balance precision and recall. Specifically, if an alignment has a high LLM confidence score but a relatively low retrieval similarity score, it may still be accepted. This is particularly useful when the retrieval process fails to capture relevant matches due to limited textual overlap, even though semantically meaningful matches exist. For example, synonyms, paraphrases, or domain-specific terminology may result in low similarity scores, but the LLM's confidence can still indicate that the alignment is correct.

By allowing such alignments to pass through, this strategy helps recover plausible matches that would have otherwise been discarded, ensuring that potentially valuable alignments are not overlooked.

By incorporating this confidence-aware adjustment, the refinement step becomes more robust against shallow or surface-level retrieval errors. This enables the pipeline to recover semantically valid correspondences that would otherwise be filtered out, thereby increasing recall without severely compromising precision. The final alignment output benefits from this multi-faceted filtering logic, which combines the complementary strengths of both retrieval models and LLMs, producing a final alignment set with enhanced accuracy, coverage, and resilience to noise.

4. Experiments

To assess the proposed *infuse-retrieve-prompt* pipeline combined with precision-compensated strategy, we conducted a series of experiments on multiple ontology alignment benchmarks. This section outlines the baseline systems for comparison, the datasets used, the experimental settings, and the evaluation results.

4.1. Baselines

To evaluate the performance of our proposed approach, we compare it against several representative baselines. These include:

- LLMs4OM[15]: This baseline follows a standard *retrieve-then-prompt* pipeline and serves as the technical foundation for our implementation of the proposed approach. It employs Sentence-BERT embeddings to compute cosine similarity between source and target classes, selecting candidate alignments based solely on the *top-k* most similar pairs. No LLM-based enrichment or transformation is applied prior to retrieval. This baseline serves to isolate and assess the added value of the proposed *infuse-retrieve-prompt* pipeline.
- Two Strong Baselines from OAEI 2023[27]: As competitive baselines, we include two of the top-performing systems from the OAEI 2023 campaign, Matcha and OLaLa, both of which achieved the highest F1 scores on one or two of the datasets used in our experiments. Matcha was selected for its strong overall performance and broader coverage, as it participated in more tracks and includes all the datasets we evaluate. OLaLa, while not attending all tracks, integrates LLMs into its alignment process, making it a relevant point of comparison. Together, these systems represent strong recent advancements in ontology alignment.

Together, these baselines allow us to quantify the improvements contributed by incorporating LLM-supported retrieval and hybrid filtering strategies in our proposed pipeline.

4.2. Datasets

To evaluate our approach, we selected four tracks comprising a total of eight datasets from the OAEI 2023 edition, focusing specifically on the equivalence matching task. The 2023 edition was chosen over 2024 due to the limited public availability of some datasets (e.g., Biodiv datasets) in the latest release. Table 1 provides an overview of the datasets, including the number of classes in the source (c_s) and target (c_t) ontologies, as well as the number of available reference mappings ($Refs$).

The datasets used in this study were selected based on several criteria to ensure a comprehensive and meaningful evaluation. We prioritized cases that met one or more of the following conditions:

1. low performance from both the baseline model LLMs4OM and OAEI systems, indicating inherently difficult alignment tasks;
2. significant performance gaps between LLMs4OM and existing OAEI systems, suggesting potential for improvement;

Table 1
Numbers of source and target classes and reference mappings.

Tracks	Datasets	c_s	c_t	Refs
Biodiv	ENVO-SWEET	6566	10239	806
	FISH-ZOOPLANKTON	145	56	15
	ALGAE-ZOOBENTHOS	108	128	18
CommonKG	Nell-DBpedia	134	137	129
	YAGO-Wikidata	304	304	304
Bio-ML	OMIM-ORDO(disease)	9642	8838	3721
	SNOMED-FMA(body)	24182	64726	7256
MSE	MI-MatOnto	545	825	302

3. diversity in domain, ontology size, and serialization format (e.g., OWL, SKOS), to test generalizability across different contexts.

This selection strategy ensures that our method is tested across a broad range of realistic and difficult alignment scenarios, enabling a robust assessment of its generalizability and performance. Notably, for the ENVO-SWEET dataset, the original version used during the OAEI campaign was not accessible at the time. To ensure reproducibility and data completeness, we instead employed the original full SWEET ontology. While this may introduce slight differences in statistics, it guarantees that all required data are consistently available for future experiments. Additionally, even within the same OAEI edition, multiple versions of a dataset may exist. As a result, the statistics reported in our study may differ slightly from those in other works, depending on the dataset variant used. Such variation can lead to inconsistencies that potentially affect the fairness of direct comparisons with OAEI-reported results or top-performing systems.

4.3. Experiment Settings

Performance was assessed using standard evaluation metrics: Precision (P), Recall (R), and F-measure (F1) [28]. The technical implementation builds upon the LLMs4OM framework [15], which follows a standard *retrieve-then-prompt* architecture for ontology alignment using LLMs. For the final configuration, several key parameters were fixed to ensure consistent and reproducible results. The *top-k* parameter, which determines the number of candidate correspondences retrieved per source concept, was set to 5. This value was chosen based on prior work [15] to as a compromise between maintaining high recall and controlling computational overhead. For candidate generation, we employed the SBERT retriever model *multi-qa-mpnet-base-dot-v1*, which showed the highest recall at $k = 5$ according to the findings in a prior study [7].

For the LLM-related stages, we employed *LLaMA-3.3-70B-Instruct* [29], as previous studies have shown that models of this size provide strong performance and improved task understanding in ontology alignment tasks [20]. To ensure a fair comparison, we re-executed the baseline model LLMs4OM using the same, updated version of LLaMA employed in our approach. During the refinement stage, candidate pairs with a similarity score above 0.9 were directly retained as high-precision matches. Alignments were further filtered using LLM confidence, with only those scoring above 0.7 considered for final inclusion. These thresholds were chosen based on best practices reported in previous studies. Additionally, to balance precision and recall, candidate pairs with very high LLM confidence (greater than 0.9) were also accepted if their retrieval similarity score exceeded 0.8.

Experiments were conducted using two Ubuntu servers: one equipped with 320 GB RAM, an Xeon w5-3435X CPU, and dual 6000 Ada GPUs; the other with 272 GB RAM, an 7402P/24core CPU, and four A6000 GPUs. Tasks were distributed across the servers to optimize workload and resource utilization. Each dataset was executed five times to account for variability in LLM outputs, except for the three largest datasets, ENVO-SWEET, OMIM-ORDO (disease), and SNOMED-FMA (body), which were run once or twice due to their high computational demands. Across all repeated runs, the variance in results

Table 2
Retrieval results

Tracks	Datasets	Our Approach	LLMs4OM
Biodiv	ENVO-SWEET	75.56	75.19
	FISH-ZOOPLANKTON	93.33	93.33
	ALGAE-ZOOBENTHOS	83.33	83.33
CommonKG	Nell-DBpedia	100	98.45
	YAGO-Wikidata	98.36	97.37
Bio-ML	OMIM-ORDO(disease)	72.10	71.51
	SNOMED-FMA(body)	79.89	79.84
MSE	MI-MatOnto	90.40	49.34

was minimal, indicating that the outcomes are stable and representative.

4.4. Results

This section presents the experimental results, highlighting the performance of the proposed approach in comparison with baseline methods. The evaluation is organized into three parts: (1) retrieval performance following LLM-based infusion, (2) alignment results with the precision-compensated filtering, and (3) an ablation study.

4.4.1. Retrieval Performance.

As recall is the primary metric of interest in the retrieval phase, we report recall values for both our approach and the LLMs4OM baseline. Table 2 summarizes the retrieval recall scores across all datasets. The results show that our method achieves equal or superior recall in all cases, with particularly strong improvements observed in challenging datasets.

In datasets where baseline performance is already relatively high such as FISH-ZOOPLANKTON, ALGAE-ZOOBENTHOS, and SNOMED-FMA, the gains are marginal or non-existent. This suggests that for well-aligned or less complex cases, standard embedding-based retrieval is already sufficient, and the added value of LLM-based infusion is limited. Another possible reason is that, in these cases, we deliberately avoided introducing additional information during concept infusion in order to preserve the original semantics as much as possible, which may have constrained the potential for noticeable retrieval improvements.

However, in more challenging or specialized datasets with lower baseline recall, the improvements are substantial. Notably, in the MI-MatOnto dataset, our method improves recall from 49.34 to 90.40, a gain of over 40 percentage points. This highlights the strength of our approach in handling domain-specific or low-resource scenarios, where standard embedding-based retrieval struggles due to limited training data, specialized vocabulary or sparse reference mappings.

4.4.2. Alignment Results.

Table 3 presents alignment performance (P, R, $F1$) for our approach compared to the LLMs4OM baseline and two strong systems from the OAEI 2023 campaign. For Matcha and OLaLa, we report only their F1 scores due to space constraints. To ensure a fair comparison, we bold the higher F1 score between our approach and the baseline model (LLMs4OM), and underline the best F1 score across all systems. The results highlight the potential of our pipeline, particularly in challenging or low-resource scenarios where traditional systems struggle.

Most notably, our method yields improved recall in several cases, leading to higher F1 scores in multiple datasets. For instance, on the YAGO-Wikidata dataset (CommonKG track), our recall increases from 47.37 (LLMs4OM) to 82.57, boosting F1 from 64.29 to 90.13, surpassing OLaLa (81.00) and approaching Matcha's top F1 of 94.00. Similarly, in MI-MatOnto (MSE track), our recall rises from 20.86 to 56.62, increasing F1 from 33.96 to 68.67, more than doubling the Matcha benchmark of 33.90. These results

Table 3
Alignment results

Tracks	Datasets	Our Approach			LLMs4OM			Matcha	OLaLa
		P	*R*	*F*1	*P*	*R*	*F*1	*F*1	*F*1
Biodiv	ENVO-SWEET	57.14	51.61	54.24	81.40	43.92	**57.05**	—[1]	—
	FISH-ZOOPLANKTON	100	73.33	**84.62**	100	53.33	69.57	41.90	92.80
	ALGAE-ZOOBENTHOS	83.33	27.78	41.67	83.33	27.78	41.67	28.50	50.00
CommonKG	Nell-DBpedia	100	87.60	**93.39**	100	79.07	88.31	93.00	96.00
	YAGO-Wikidata	99.21	82.57	**90.13**	100	47.37	64.29	94.00	81.00
Bio-ML	OMIM-ORDO(disease)	72.47	53.99	**61.88**	88.68	43.38	58.26	61.70	64.90
	SNOMED-FMA(body)	23.09	27.89	**25.27**	42.97	16.68	24.03	64.10	30.40
MSE	MI-MatOnto	87.24	56.62	**68.67**	91.30	20.86	33.96	33.90	—[2]

suggest that LLM-based infusion and enrichment are likely beneficial in domains where concepts are loosely defined or training data is sparse.

On Nell-DBpedia (CommonKG), our F1 improves from 88.31 to 93.39, narrowing the gap with OLaLa (96.00). In FISH-ZOOPLANKTON, our approach also delivers an F1 of 84.62, considerably outperforming Matcha (41.90) but still behind OLaLa (92.80). In contrast, for smaller or simpler datasets such as ALGAE-ZOOBENTHOS, both our method and LLMs4OM perform identically (F1 = 41.67), suggesting limited room for improvement where alignments are already straightforward.

Our method does lag behind existing systems in some cases. For example, on SNOMED-FMA, we achieve an F1 of 25.27, slightly higher than LLMs4OM (24.03) but lower than both Matcha (64.10) and OLaLa (30.40). This points to areas where domain-specific adaptation may still offer benefits.

Overall, while our system does not outperform all baselines across every dataset, it remains competitive without relying on domain-specific tuning or training. In particular, the integration of LLM-based concept enrichment appears to contribute positively in cases involving loosely defined concepts or limited training data. Although performance varies by domain, these results suggest that LLMs can help bridge semantic gaps that traditional alignment methods may miss, offering a complementary approach in complex or under-resourced scenarios.

4.4.3. Statistical Significance Testing

To assess whether the observed performance differences are statistically significant, we conduct significance testing on the alignment results, a paired t-test comparing our method to the baseline, which yielded $t = 2.15$, $p = 0.068$. While not statistically significant at the conventional 5% level, the result indicates a positive trend favoring our method. To assess practical relevance, we also computed *Cohen's* $d = 0.76$, indicating a substantial effect size. This suggests that our method provides a meaningful performance gain, even if statistical significance is marginal.

4.4.4. Ablation Study.

Table 4 presents an ablation study assessing the impact of the precision-compensation strategy on alignment performance. Here, P, R, $F1$ represent results with the strategy applied, while P', R', $F1'$ refer to results without it.

The results show that precision-compensation often boosts recall, leading to improved F1 scores, especially in challenging or low-resource settings. For example, in YAGO-Wikidata, recall increases from 73.36 to 82.57, improving F1 from 84.63 to 90.13. In Nell-DBpedia, a similar trend is observed, with F1 rising from 88.79 to 93.39 due to enhanced recall. The strategy proves particularly helpful in difficult domains like SNOMED-FMA, where recall increases from 16.76 to 27.89, despite a drop in

[1]Instead of using the test case subset provided by OAEI for the SWEET ontology (which is not available), we utilized the full version of the ontology. Therefore, a direct comparison with the OAEI-reported values is not entirely meaningful, as they are based on a different dataset scope.
[2]OLaLa didn't attend the MSE track alignment.

Table 4
Performance without applying the precision-compensated filtering

Tracks	Datasets	P	R	$F1$	P'	R'	$F1'$
Biodiv	ENVO-SWEET	57.14	51.61	54.24	71.22	48.51	**57.71**
	FISH-ZOOPLANKTON	100	73.33	**84.62**	100	53.33	69.57
	ALGAE-ZOOBENTHOS	83.33	27.78	41.67	83.33	27.78	41.67
CommonKG	Nell-DBpedia	100	87.60	**93.39**	100	79.84	88.79
	YAGO-Wikidata	99.21	82.57	**90.13**	100	73.36	84.63
Bio-ML	OMIM-ORDO(disease)	72.47	53.99	**61.88**	84.90	43.54	57.56
	SNOMED-FMA(body)	23.09	27.89	**25.27**	42.02	16.76	23.96
MSE	MI-MatOnto	87.24	56.62	68.67	94.89	55.30	**69.87**

precision, resulting in a higher F1 (23.96 ⟶ 25.27). This demonstrates its utility in boosting coverage in semantically complex biomedical ontologies. On the other hand, in already high-precision settings like MI-MatOnto, the strategy offers little benefit, slightly reducing F1 (69.87 ⟶ 68.67), possibly due to marginal over-alignment.

In summary, the precision-compensation strategy effectively improves recall and F1 in complex or sparse alignment scenarios, while having minimal or mixed impact in already well-performing cases.

5. Conclusion and Outlook

In this work, we proposed an LLM-enhanced *infuse-retrieve-prompt* pipeline for ontology alignment. The pipeline first enriches ontology concepts with contextual information using zero-shot LLMs (*infuse*), then retrieves candidate alignments based on the enriched representations (*retrieve*), and finally validates or re-ranks these candidates through prompting (*prompt*). Additionally, we incorporated a precision-compensation strategy to further refine the alignment results. Our experiments show that the proposed approach consistently outperforms the baseline, particularly in recall, leading to improvements in F1 scores. Notably, significant improvements were observed in challenging domains, such as the MI-MatOnto dataset, where both recall and F1 scores showed considerable gains.

Despite these encouraging results, the method presents several limitations, including high computational cost, occasional hallucinations, and reduced precision in certain datasets. These issues are especially pronounced in large-scale or semantically complex ontologies, such as those in the biomedical domain.

To address these challenges, future work will explore techniques for hallucination mitigation, computational efficiency improvements, and integration of structure-aware methods to enhance scalability. While refinement is still needed, our results suggest that LLM-based enrichment offers a promising direction for ontology alignment, particularly in settings where traditional methods fall short.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT in order to: Grammar and spelling check, Paraphrase and reword. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] S. Pan, L. Luo, Y. Wang, C. Chen, J. Wang, X. Wu, Unifying large language models and knowledge graphs: A roadmap, IEEE Transactions on Knowledge and Data Engineering 36 (2024) 3580–3599.

[2] Y. He, J. Chen, H. Dong, I. Horrocks, Exploring large language models for ontology alignment, in: Proceedings of the ISWC Posters, Demos and Industry Tracks, 2023.

[3] X. L. Dong, Generations of knowledge graphs: The crazy ideas and the business impact, Proc. VLDB Endow. 16 (2023) 4130–4137.
[4] J. Chen, E. Jiménez-Ruiz, I. Horrocks, D. Antonyrajah, A. Hadian, J. Lee, Augmenting ontology alignment by semantic embedding and distant supervision, in: The Semantic Web, Springer International Publishing, 2021, pp. 392–408.
[5] Y. He, J. Chen, D. Antonyrajah, I. Horrocks, Bertmap: A bert-based ontology alignment system, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 36, 2022, pp. 3580–3599.
[6] J. Portisch, M. Hladik, H. Paulheim, Background knowledge in ontology matching: A survey, Semantic Web 15 (2024) 2639–2693.
[7] S. Hertling, H. Paulheim, Olala: Ontology matching with large language models, in: Proceedings of the 12th Knowledge Capture Conference, Association for Computing Machinery, 2023, p. 131–139.
[8] M. Taboada, D. Martinez, M. Arideh, R. Mosquera, Ontology matching with large language models and prioritized depth-first search, Information Fusion 123 (2025) 103254.
[9] P. Shvaiko, J. Euzenat, Ontology matching: State of the art and future challenges, IEEE Transactions on Knowledge and Data Engineering 25 (2013) 158–176.
[10] N. Reimers, I. Gurevych, Sentence-BERT: Sentence embeddings using Siamese BERT-networks, in: Proceedings of the EMNLP-IJCNLP Conference, Association for Computational Linguistics, 2019, pp. 3982–3992.
[11] E. Jiménez-Ruiz, B. Cuenca Grau, Logmap: Logic-based and scalable ontology matching, in: The Semantic Web – ISWC, Springer Berlin Heidelberg, 2011, pp. 273–288.
[12] D. Faria, C. Pesquita, E. Santos, M. Palmonari, I. F. Cruz, F. M. Couto, The AgreementMakerLight Ontology Matching System, in: On the Move to Meaningful Internet Systems: OTM Conferences, Springer Berlin Heidelberg, 2013, pp. 527–541.
[13] P. Kolyvakis, A. Kalousis, D. Kiritsis, DeepAlignment: Unsupervised ontology matching with refined word vectors, in: Proceedings of the NAACL Conference: Human Language Technologies, Association for Computational Linguistics, 2018, pp. 787–798.
[14] S. Hertling, J. Portisch, H. Paulheim, Kermit – a transformer-based approach for knowledge graph matching, 2022. URL: https://arxiv.org/abs/2204.13931.
[15] H. Babaei Giglou, J. D'Souza, F. Engel, S. Auer, LLMs4OM: Matching ontologies with large language models, in: The Semantic Web: ESWC 2024 Satellite Events, Springer Nature Switzerland, 2025, pp. 25–35.
[16] Q. Wang, Z. Gao, R. Xu, Exploring the in-context learning ability of large language model for biomedical concept linking, 2023. URL: https://arxiv.org/abs/2307.01137.
[17] R. Peeters, C. Bizer, Using chatgpt for entity matching, in: New Trends in Database and Information Systems, Springer Nature Switzerland, 2023, pp. 221–230.
[18] S. S. Norouzi, M. S. Mahdavinejad, P. Hitzler, Conversational ontology alignment with chatgpt, 2023. URL: https://arxiv.org/abs/2308.09217.
[19] M. Amir, M. Baruah, M. Eslamialishah, S. Ehsani, A. Bahramali, S. Naddaf-Sh, S. Zarandioon, Truveta mapper: A zero-shot ontology alignment framework, 2023. URL: https://arxiv.org/abs/2301.09767.
[20] Z. van Cauter, N. Yakovets, Ontology-guided knowledge graph construction from maintenance short texts, in: Proceedings of the KaLLM Workshop, Association for Computational Linguistics, 2024, pp. 75–84.
[21] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, D. Kiela, Retrieval-augmented generation for knowledge-intensive nlp tasks, in: Proceedings of the 34th International Conference on Neural Information Processing Systems, Curran Associates Inc., 2020.
[22] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, M. Wang, H. Wang, Retrieval-augmented generation for large language models: A survey, 2024. URL: https://arxiv.org/abs/2312.10997.
[23] J. Euzenat, C. Meilicke, H. Stuckenschmidt, P. Shvaiko, C. Trojahn, Ontology alignment evaluation initiative: Six years of experience, in: Journal on Data Semantics XV, Springer Berlin Heidelberg, 2011, pp. 158–192.

[24] I. Osman, S. Ben Yahia, G. Diallo, Ontology integration: Approaches and challenging issues, Information Fusion 71 (2021) 38–63.
[25] S. Neutel, M. H. T. de Boer, Towards automatic ontology alignment using BERT, in: Proceedings of the AAAI Spring Symposium on Combining Machine Learning and Knowledge Engineering, volume 2846, CEUR-WS.org, 2021.
[26] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, G. Neubig, Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing, ACM Comput. Surv. 55 (2023).
[27] M. A. N. Pour, A. Algergawy, P. Buche, L. J. Castro, J. Chen, A. Coulet, J. Cufi, H. Dong, O. Fallatah, D. Faria, I. Fundulaki, S. Hertling, Y. He, I. Horrocks, M. Huschka, L. Ibanescu, S. Jain, E. Jiménez-Ruiz, N. Karam, P. Lambrix, H. Li, Y. Li, P. Monnin, E. Nasr, H. Paulheim, C. Pesquita, T. Saveta, P. Shvaiko, G. Sousa, C. Trojahn, J. Vatascinová, M. Wu, B. Yaman, O. Zamazal, L. Zhou, Results of the ontology alignment evaluation initiative 2023, in: Ontology Matching Workshop – ISWC, 2023, pp. 97–139.
[28] OAEI, Towards a methodology for evaluating alignment and matching algorithms version 1.0, https://oaei.ontologymatching.org/doc/oaei-methods.1.pdf, 2005. Accessed: 2025-05-6.
[29] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, A. Yang, A. Fan, A. Goyal, A. Hartshorn, A. Yang, A. Mitra, A. Sravankumar, A. Korenev, A. Hinsvark, A. R. et al., The llama 3 herd of models, 2024. URL: https://arxiv.org/abs/2407.21783.