GenOM: Ontology Matching with Description Generation and Large Language Model

Yiping Song[1,*], Jiaoyan Chen[1] and Renate A. Schmidt[1]

[1]Department of Computer Science, The University of Manchester

Abstract

Ontology matching (OM) plays an essential role in enabling semantic interoperability and integration across heterogeneous knowledge sources, particularly in the biomedical domain which contains numerous complex concepts related to diseases and pharmaceuticals. This paper introduces **GenOM**, a large language model (LLM)-based ontology alignment framework, which enriches the semantic representations of ontology concepts via generating textual definitions, retrieves alignment candidates with an embedding model, and incorporates exact matching-based tools to improve precision. Extensive experiments conducted on the OAEI Bio-ML track demonstrate that GenOM can often achieve competitive performance, surpassing many baselines including traditional OM systems and recent LLM-based methods. Further ablation studies confirm the effectiveness of semantic enrichment and few-shot prompting, highlighting the framework's robustness and adaptability.

Keywords

Ontology Matching, Large Language Model, Semantic Embedding, Definition Generation

1. Introduction

In recent years, the rapid growth of domain-specific ontologies has led to a growing need for semantic interoperability and knowledge integration across diverse knowledge systems [1]. Ontologies are developed to formally represent concepts and relationships in a domain, yet they are often constructed in isolation, following distinct modelling choices, terminological conventions, and structural assumptions. This independence has resulted in significant heterogeneity across ontologies, which poses considerable challenges to the integration and reuse of knowledge.

Ontology Matching (OM) as known as ontology alignment, the task of identifying semantic correspondences between entities in different ontologies, has therefore become a crucial area of research [2, 3]. By establishing links such as equivalence (indicating that two concepts represent the same or highly similar meaning) or subsumption (where one concept is a more general or specific variant of the other), ontology alignment facilitates accurate knowledge translation and consistent information exchange between systems. However, matching concepts across ontologies is far from straightforward. Three major sources of heterogeneity commonly hinder this process: (1) Terminological differences, where the same concept may be described using different labels or synonyms; (2) Structural differences, reflecting the varying levels of complexity in ontology design—from deeply nested hierarchies to flat enumerative lists; (3) Granularity differences, where the same domain knowledge may be captured with differing levels of detail or abstraction.

These variations significantly increase the cognitive and computational burden of OM. The challenge is further magnified by the rapidly growing scale of modern ontologies. For example, SNOMED-CT [4], a widely adopted clinical terminology, contains several hundred thousand medical concepts. As ontologies continue to expand in size and complexity, manual alignment methods become increasingly

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

*Corresponding author.

yiping.song@postgrad.manchester.ac.uk (Y. Song); jiaoyan.chen@manchester.ac.uk (J. Chen); Renate.Schmidt@manchester.ac.uk (R. A. Schmidt)

https://orcid.org/0009-0007-4638-7019 (Y. Song); https://orcid.org/0000-0003-4643-6750 (J. Chen); https://orcid.org/0000-0002-6673-3333 (R. A. Schmidt)

infeasible, underscoring the necessity for automated or semi-automated alignment techniques capable of operating at scale.

Traditional OM systems, such as LogMap [5] and AML [6], primarily rely on string matching, indexing and structure matching technologies. They often fall short in capturing or fully utilising the underlying semantic information of concepts. With the advent of large language models (LLMs), their remarkable capabilities in text understanding and generalisation have attracted significant attention. Recently, several ontology alignment systems have begun to incorporate LLMs to better capture the semantics of complex concepts across heterogeneous ontologies. These approaches typically leverage LLMs for identifying semantic similarities between concept pairs via embedding, and/or making direct alignment decisions via generation. For example, LLM4OM [7] employs ChatGPT and OpenAI embeddings for pairwise matching (see Section 2 for more related work analysis). However, current LLM-based OM approaches still exhibit notable limitations. Some of them, despite leveraging powerful language models, struggles to deliver satisfactory performance on more complex OM tasks. Some others can achieve promising results on certain benchmarks, but they may rely on LLMs with very large-scale parameters (e.g., 70B LLaMA-2 used in Olala [8]) that imposes substantial computational demands, raising concerns about scalability.

To address these limitations, we propose a novel ontology matching framework named **GenOM** which utilises LLMs and extended textual definitions of concepts. The framework begins by extracting both lexical and structural information from the source and target ontologies. This information is then semantically enhanced using an LLM, resulting in more informative and context-aware concept descriptions. Subsequently, these enriched representations are embedded into vector space, enabling the retrieval of candidate alignments based on semantic similarity. A lightweight 7B-parameter LLM is employed to assess the equivalence of candidate pairs through a classification-based approach, while traditional exact matching techniques are incorporated to supplement and refine the alignment results. The effectiveness of GenOM is demonstrated through comprehensive experiments evaluated using standard metrics such as precision, recall, F1-score, mean reciprocal rank (MRR), and Hit@K. The model achieves competitive performance compared to several state-of-the-art ontology alignment systems, highlighting its robustness and practical applicability. Extensive ablation studies were also conducted to demonstrate the effectiveness of the proposed framework across multiple dimensions.

2. Related Work

At present, OM approaches can be broadly categorised into four main types: traditional knowledge-based systems, machine learning-based systems, pre-trained language model-based systems and more recently, large language models (LLMs)-based systems. Traditional systems, such as LogMap and AML, rely primarily on lexical similarity, structural heuristics, and external resources like UMLS or WordNet [9]. LogMap [5] extracts class names and searches for matches via external lexicons while addressing logical inconsistencies by selecting alignments with higher confidence scores. AML [6] employs multiple matching strategies, including exact and character-based matchers, and has shown strong performance on medical datasets. However, these systems depend heavily on curated lexicons and often fail to fully capture and utilise complex semantics of different kinds.

With the rise of machine learning, several approaches have emerged that model alignment as a classification or (embedding-based) similarity learning task [10, 11, 12, 13, 14, 15]. For instance, DeepAlignment [11] vectorises class names and computes Euclidean distances to assess similarity, while the CNN-based system [15] use character-level embeddings and hierarchical context to train binary classifiers for equivalence detection. Although these methods offer improvements over traditional techniques, they often require large annotated datasets and extensive parameter tuning. Moreover, their domain-specific nature limits transferability across ontologies in different fields.

Using encoder-based pre-trained language models like BERT [16] have leveraged the representational power of contextual embeddings and a memorization based on large-scale parameters learned from corpora to address some of these limitations. BioSTransformers [17] adopt a Siamese architecture

based on domain-specific BERT models to compute semantic similarity between biomedical concepts. Built on the BERT architecture [18], BERTMap fine-tunes a domain-specific BERT model on ontology alignment corpora, enabling it to capture subtle semantic differences even when lexical overlap is low. This approach has demonstrated strong results, particularly in biomedical applications. BERTSubs [19] takes a similar architecture as BERTMap but focuses on the subsumption relationship.

More recently, LLMs such as GPT-3.5, GPT-4, and T5-XXXL have been applied to ontology alignment, offering stronger generalisation and semantic reasoning capabilities. Several studies have explored prompt-based querying and retrieval-augmented generation to support alignment tasks. For example, Norouz et al. [20] used GPT-4 to align ontology via prompts, observing high recall but reduced precision due to the model incorrectly classifying subclass relationships as equivalence relations. Yuan et al. [21] tested both open- and closed-source models on medical alignment tasks, using structural context to enhance predictions. Other work has introduced hybrid frameworks combining vector similarity retrieval (e.g., using SBERT) with LLM verification stages [7, 22], aiming to reduce hallucination and improve alignment quality. The Olala system [8] further integrates embedding-based candidate filtering and post-processing with LLaMA-2 for final alignment decisions. While LLM-based approaches have shown considerable potential, they still face challenges including scalability to large ontologies such as SNOMED CT, struggles to deliver satisfactory performance on more complex OM tasks and computational cost, particularly when relying on closed-source or very large models.

3. Methodology

3.1. Task Formulation

The OM task can be formally defined as follows. Given two ontologies, referred to as the *source ontology* O_{s} and the *target ontology* O_{t}, let C_s denote the set of named concepts in O_{s} and C_t denote the set of named concepts in O_{t}. The objective is to identify a set of mappings, where each mapping consists of a concept pair (c_s, c_t), with $c_s \in C_s$ and $c_t \in C_t$, that are considered semantically related. Formally, the alignment output is represented as:

$$M = \{(c_s, c_t, \alpha) \mid c_s \in C_s,\ c_t \in C_t,\ \alpha \in [0, 1]\} \tag{1}$$

where α denotes the confidence score, quantifying the degree of semantic equivalence between c_s and c_t.

This score often serves as a basis for selecting high-confidence mappings.

3.2. System Architecture

As shown in Figure 1, GenOM comprises five main components:

1. **Ontology Data Extraction**: Structural and lexical information is extracted for each named concept from both the source and target ontologies. This includes labels, synonyms, parent concepts, and axioms of concept equivalence.
2. **Definition Generation**: An LLM is prompted to generate natural language definitions or paraphrased descriptions of each concept, based on the extracted information. This step enhances the semantic representation of concepts, especially for those lacking explicit textual definitions.
3. **Candidate Mapping Generation**: Using the embedding model, pairwise cosine similarity scores are computed between the source and target concepts to generate top-k candidate mappings.
4. **LLM-Based Equivalence Judgement**: For each candidate mapping, an LLM is queried to determine whether the mapping's two concepts are semantically equivalent, with prompts that incorporate enriched definitions and structural context.
5. **Post-processing and Result Fusion**: Filtering is performed based on both the probability distribution of the LLM outputs and the cosine similarity scores, retaining only high-confidence alignment results. In addition, exact matching modules are applied to recall highly confident matches based on identical labels.

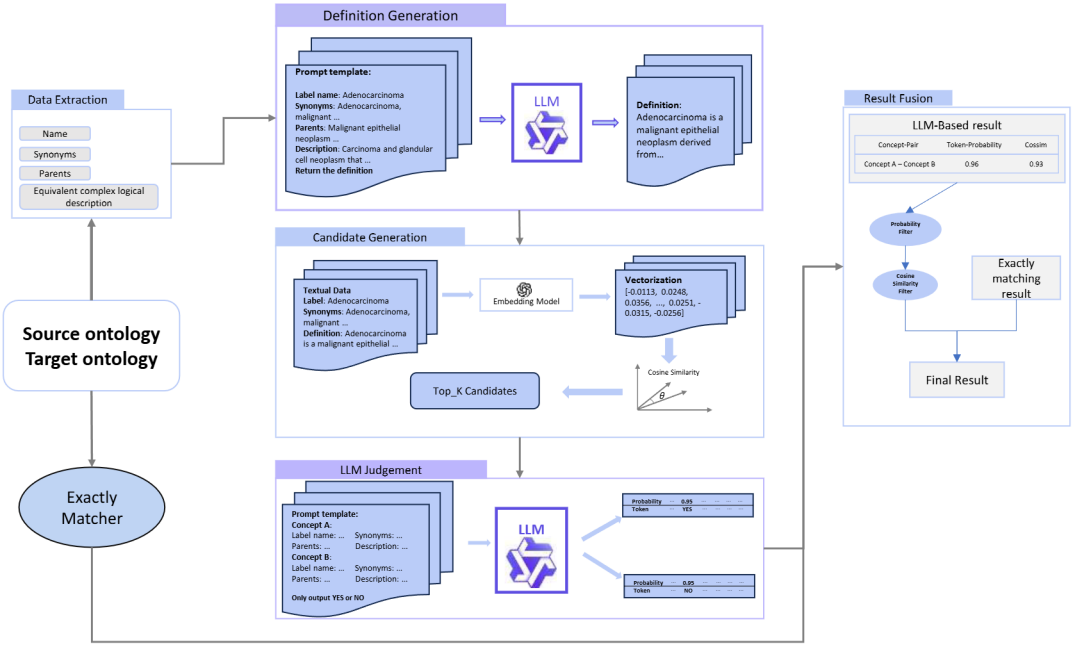

Figure 1: The Architecture of GenOM

By concept definition generation, candidate retrieval, and alignment judgement, GenOM integrates the strengths of embedding-based similarity, exact lexical matching, and LLM reasoning.

3.2.1. Ontology Data Extraction

Both lexical and structural characteristics of each concept from the source and target ontologies are extracted and utilized to support the following alignment process. Specifically, the extracted information includes the concept's *label* (defined by the annotation property *rdfs:label*), a set of *synonyms* (retrieved using the annotation properties listed in Table 1), and its *parent concepts*. For concepts defined using *EquivalentClass* axioms, the built-in `verbalisation` module in DeepOnto [23] is used to convert logical expressions into natural language descriptions (Table 2).

Property	IRI
Label	`http://www.w3.org/2000/01/rdf-schema#label`
Synonym	`http://www.geneontology.org/formats/oboInOwl#hasSynonym`
	`http://www.geneontology.org/formats/oboInOwl#hasExactSynonym`
	`http://www.ebi.ac.uk/efo/alternative_term`
	`http://www.orpha.net/ORDO/Orphanet_#symbol`
	`http://purl.org/sig/ont/fma/synonym`
	`http://www.w3.org/2004/02/skos/core#altLabel`
	`http://www.w3.org/2004/02/skos/core#prefLabel`
	`http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#P108`
	`http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#P90`

Table 1
Annotation property IRIs used for label and synonym extraction

Description Logic Axiom
Product containing only betamethasone and calcipotriol (medicinal product) ≡ *MedicinalProduct* ⊓ ∃ *RoleGroup.*(∃ *hasActiveIngredient.Betamethasone*) ⊓ ∃ *RoleGroup.*(∃ *hasActiveIngredient.Calcipotriol*)
Verbalized Description
Medicinal product (product) that Role group (attribute) something that Has active ingredient (attribute) Betamethasone (substance) and something that Has active ingredient (attribute) Calcipotriol (substance).

Table 2
Example of DL Axiom and Its Verbalized Description

3.2.2. Definition Generation

LLMs encode extensive general purpose and domain knowledge. They are leveraged and integrated with concept information explicitly represented in the ontology to enhance the semantic representation of the concept. In this framework, definitions for medical concepts are generated by prompting an LLM with background information extracted from the ontology, including the concept's *label*, *synonyms*, *parent concepts*, and, where applicable, the verbalised descriptions of logical expressions from *EquivalentClass* axioms.

Table 3 presents the prompt template employed for definition generation. In this template, the LLM is provided with both the available concept-specific information and the name of the source ontology. This additional context helps the LLM recall relevant domain knowledge encoded in its parameters, thereby generating definitions that are more accurate and context-aware. The inclusion of the ontology name in the prompt acts as supplementary guidance, particularly for sparsely annotated concepts.

The amount and richness of information associated with a concept can vary considerably across ontologies. Some concepts are well-described, including multiple synonyms, hierarchical structure, and even formal axioms. In contrast, other concepts may be sparsely defined, often limited to a label with little or no supporting context. In such cases, the LLM's internalised knowledge becomes essential in compensating for missing semantics.

Prompt Templates for Concept Definition Generation
Role: System You are generating a definition for a concept from the $\{O_{\mathrm{s}}\ \ \mathtt{name}\}$ ontology. The definition will be used to align it with candidate concepts in the $\{O_{\mathrm{t}}\ \ \mathtt{name}\}$ ontology. You are a biomedical ontology expert. Your task is to generate a concise, alignment-friendly definition for a given biomedical concept. The definition should be semantically precise, distinguishable from related terms, and suitable for matching across ontologies. Only return the definition.
Role: User Concept: Product containing only betamethasone and calcipotriol (medicinal product) Synonyms: Betamethasone and calcipotriol only product Parents: Product containing betamethasone and calcipotriol (medicinal product) Description: Medicinal product (product) that Role group (attribute) something that Has active ingredient (attribute) Betamethasone (substance) and something that Has active ingredient (attribute) Calcipotriol (substance)
Output: A medicinal product specifically formulated to contain solely betamethasone and calcipotriol as its active ingredients, designed for the treatment or management of specific dermatological conditions.

Table 3
The prompt template used for LLM-based definition generation.

3.2.3. Candidate Mapping Generation

To generate candidate concept pairs for alignment, this framework adopts an embedding-based retrieval strategy. Concepts from both the source ontology O_s and the target ontology O_t are first encoded into fixed-size vector representations. Each concept is represented using a combination of its *label*, *synonyms*, and its enriched *definition* produced in the previous step. Figure 2 shows an example of the input text fed into the embedding model. Structural information such as hierarchical relations is deliberately excluded at this stage to reduce complexity in embedding. The framework adopts the `text-embedding-3-small`[1] model for embedding generation. Compared to traditional encoder-based models such as Sentence-BERT [24], this LLM-based embedding model demonstrates superior capability in distinguishing complex biomedical concepts.

Once the embeddings are obtained, a cosine similarity-based retrieval process is applied to identify, for each source concept, the top-k most semantically similar concepts from the target ontology. Candidate selection is based on vector similarity, allowing the system to retrieve a shortlist of potentially equivalent concept pairs. These candidates are subsequently passed to the next stage for semantic equivalence assessment.

Embedding Input Example

Label: Product containing only betamethasone and calcipotriol (medicinal product); Synonyms: Betamethasone and calcipotriol only product; Definition: A medicinal product specifically formulated to contain solely betamethasone and calcipotriol as its active ... ;

Figure 2: Concept information input format for embedding.

3.2.4. LLM-Based Equivalence Judgement

In this stage, an LLM is employed to determine whether each candidate concept pair represents a semantic equivalence. Rather than prompting the model to generate full descriptive justifications, which would be time-consuming and potentially verbose, a lightweight classification strategy is adopted. Specifically, each concept pair is presented via a prompt designed to elicit a binary response — YES if the concepts are equivalent, and NO otherwise.

To support this, a prompt (as shown in Table 4) is constructed with strong instructional guidance, encouraging the model to respond using only a single classification token.

The predicted equivalence score is then computed based on the probability of the YES token, extracted directly from the model's output logits. Specifically, given the model's output logits $\mathbf{z} \in \mathbb{R}^V$ at the final decoding position (where V is the vocabulary size), the softmax function is applied to convert the logits into a probability distribution:

$$P(\text{YES}) = \text{softmax}(\mathbf{z})_{\text{YES}} = \frac{\exp(z_{\text{YES}})}{\sum_{i=1}^{V} \exp(z_i)}$$

Here, z_{YES} denotes the logit corresponding to the token YES. The resulting probability serves as the model's confidence in semantic equivalence for a given concept pair. Concept pairs with $P(\text{YES})$ exceeding a predefined threshold are retained for alignment.

This probability-based scoring approach significantly reduces inference time and simplifies decision-making, as it avoids generating full-length text responses and instead relies on a single-token classification strategy, while maintaining high alignment precision.

[1] https://platform.openai.com/docs/models/text-embedding-3-small

Prompt Template for Equivalence Judgement
System Message: You are an expert in biomedical concept classification. You will be given two biomedical concepts. Based on the information provided, determine whether the two concepts refer to the same real-world entity (ontology matching). Only respond with YES or NO.
User Message: **Concept A** **Name:** {*lateral rectus nerve*} **Synonyms:** {*abducens nerve ...*} **Superclass:** {*peripheral nerve of head and neck (body structure) ...*} **Definition:** {*the lateral rectus nerve, also known as ...*} **Concept B** **Name:** {*abducent nerve [vi]*} **Synonyms:** {*nervus abducens ...*} **Superclass:** {*right posterior crico-arytenoid ligament ...*} **Definition:** {*the abducent nerve [vi] is a branch of the cranial nerve vi that innervates ...'*}}

Table 4
The prompt template used for binary equivalence classification between ontology concepts. Dynamic fields are populated with structured ontology information.

3.2.5. Post-processing and Result Fusion

To ensure the quality and reliability of the final alignment output, a post-processing stage is applied to filter and refine the results generated by the LLM. First, a threshold λ_{prob} is imposed on the token-level probability associated with the YES response. Candidate pairs with confidence scores below this threshold are discarded. In parallel, the cosine similarity scores obtained during candidate generation are also considered, and pairs with lower than λ_{cs} embedding similarity are removed to prevent semantically distant matches from being retained.

After this dual-filtering step, To further enhance precision, outputs from the LLM module were merged with the results of two exact matching systems, `LogMapLt`[2] and `BERTMapLt`[3], which are lightweight versions of their original models, both simplified to include only the string matching component. This fusion combines semantic reasoning with surface-level matching, improving overall coverage while preserving precision. The resulting set constitutes the final alignment output.

4. Experiment and Evaluation

4.1. Datasets and Evaluation Metrics

The experiments were conducted on the `OAEI 2024 Bio-ML`[4] track [25] whose benchmarks are designed for biomedical ontology alignment tasks. The dataset comprises five sub-tasks involving six widely used biomedical ontologies: `Systematized Nomenclature of Medicine - Clinical Terms (SNOMED-CT)`, `National Cancer Institute Thesaurus (NCIT)`, `Foundational Model of Anatomy (FMA)`, `Human Disease Ontology (DOID)`, `Orphanet Rare Disease Ontology (ORDO)`, and `Online Mendelian Inheritance in Man (OMIM)`. The two official evaluation protocols of the Bio-ML track were adopted: *global matching* which focuses on ranking the correct target concept among a list of candidates, and *local ranking* which is to evaluate the system's ability to identify correct mappings among all the possible concept pairs across two ontologies. Precision (P), Recall (R), and F1-score are measured for global matching, and Mean Reciprocal Rank (MRR) and Hit@1 are calculated for local ranking. These metrics provide a comprehensive view of the OM systems.

[2] https://github.com/ernestojimenezruiz/logmap-matcher
[3] https://github.com/KRR-Oxford/DeepOnto/tree/main/src/deeponto/align/bertmap
[4] https://krr-oxford.github.io/OAEI-Bio-ML/2024/index.html

4.2. Experiment Setup

To systematically evaluate the proposed framework, we designed distinct experimental settings tailored to each module of the pipeline. **Definition Generation:** For semantic enrichment, the `Qwen2.5-7B-Instruct-1M`[5] LLM was utilised to generate concise definitions based on concept-level contextual information. The generation process was controlled with a temperature of 0.7 and a top_p value of 0.9, ensuring that the definitions remained focused and semantically aligned with the underlying concepts. **Candidate Mapping Generation:** Cosine similarity computations and HNSW-based indexing were implemented via the `faiss` library to efficiently retrieve the top-10 most similar concepts for each source entity. **LLM-based Judgement:** The same `Qwen2.5-7B-Instruct-1M` model was applied to perform binary equivalence classification over the candidate pairs. To accelerate inference, we adopted the `float16` data type. **Post-processing and Result Fusion.** In the final stage, results were filtered using a token probability threshold λ_{prob} of 0.99 and a cosine similarity threshold λ_{cs} of 0.97. These thresholds were initially optimised on the *SNOMED-NCIT (neoplas)* task, then fixed and consistently applied across all the other tasks. Moreover, since BERTMapLt demonstrated stronger performance on the `neoplas` task among the exact matching models, it was selected for application to the remaining tasks as well.

4.3. Overall Results

Table 5 presents the comparison between our proposed method and several state-of-the-art ontology alignment systems on the Bio-ML track. The results demonstrate that our model achieves consistently strong performance across all tasks, ranking among the top three in every case. Moreover, in tasks where our method ranks second, the performance gap with the best-performing system is marginal—for instance, only 0.006 in the *SNOMED-NCIT (pharm)* task and as small as 0.001 in the *SNOMED-FMA (body)* task.

To ensure a fair assessment and avoid overestimating the performance of our model, we excluded the `neoplas` task from the overall evaluation, as it was used during threshold tuning. When evaluated solely on the remaining unseen tasks—where no threshold optimisation was performed—our framework still achieved the highest average F1 score of 0.769. This surpasses the second-best method, BERTMap (0.762), and the third-best, LogMapBio (0.760). These results indicate that the proposed approach not only performs competitively on tuned datasets but also maintains strong and consistent performance across previously unseen tasks, highlighting its robustness and generalisability in biomedical ontology alignment.

In addition, when compared to the other LLM-based OM system *LLM4OM* which leverages ChatGPT-3.5 and OpenAI's embedding model as reported in its original paper, our approach *GenOM* delivers consistently better performance across all evaluated tasks, including the last four tasks where GenOM generalises the hyper parameter settings optimised from the first task.

4.4. Ablation Study

4.4.1. Impact of Definition Enrichment on Local Ranking and Candidate Retrieval

We study the impact of the generated concept definition on each of the two stages: LLM-based equivalence judgement and candidate generation. For LLM-based equivalence judgement, we compare using only the concept label, and using both the label and the generated definition. The local ranking result is shown in Table 6. In particular, both MRR and Hit@1 show noticeable gains in all tasks except for *SNOMED-FMA (body)*, where performance remains comparable.

For the candidate generation stage, we additionally report Hit@5 and Hit@10, as these metrics are essential for determining the appropriate top-k value—that is, how many candidate concepts should be passed to the LLM for equivalence judgement. As shown in Table 7, incorporating definition information led to improvements across all three metrics (Hit@1, Hit@5, Hit@10) in all tasks, except for a slight

[5] https://huggingface.co/Qwen/Qwen2.5-7B-Instruct-1M

Task	System	P	R	F1	MRR	H@1
SNOMED-NCIT (Neoplas)	LogMap	0.870	0.586	0.701	NA	NA
	LogMapBio	0.784	0.795	0.771	NA	NA
	LogMapLt	0.951	0.517	0.670	NA	NA
	Matcha	0.838	0.551	0.665	0.889	**0.936**
	BERTMap	0.557	0.762	0.643	**0.954**	0.928
	BERTMapLt	0.831	0.687	0.752	0.891	0.859
	BioSTransMatch	0.289	0.663	0.402	0.846	0.789
	LLM4OM	0.470	0.530	0.495	NA	NA
	GenOM	0.795	0.764	**0.779**	0.934	0.899
SNOMED-NCIT (Pharm)	LogMap	0.966	0.607	0.746	NA	NA
	LogMapBio	0.928	0.611	0.737	NA	NA
	LogMapLt	0.996	0.599	0.748	NA	NA
	Matcha	0.987	0.607	**0.752**	0.936	0.921
	BERTMap	0.971	0.585	0.730	**0.969**	**0.951**
	BERTMapLt	0.981	0.574	0.724	0.849	0.773
	BioSTransMatch	0.584	0.844	0.690	0.943	0.918
	LLM4OM	0.818	0.582	0.680	NA	NA
	GenOM	0.988	0.599	0.746	0.941	0.901
SNOMED-FMA (Body)	LogMap	0.744	0.407	0.526	NA	NA
	LogMapBio	0.827	0.577	0.680	NA	NA
	LogMapLt	0.970	0.542	0.696	NA	NA
	Matcha	0.887	0.502	0.641	**0.950**	**0.935**
	BERTMap	0.979	0.662	**0.790**	0.944	0.920
	BERTMapLt	0.979	0.655	0.785	0.892	0.865
	BioSTransMatch	0.128	0.384	0.192	0.633	0.513
	LLM4OM	0.211	0.326	0.256	NA	NA
	GenOM	0.944	0.678	0.789	0.895	0.844
OMIM-ORDO	LogMap	0.876	0.448	0.593	NA	NA
	LogMapBio	0.866	0.609	**0.715**	NA	NA
	LogMapLt	0.940	0.252	0.397	NA	NA
	Matcha	0.781	0.509	0.617	0.815	0.782
	BERTMap	0.734	0.576	0.646	0.880	0.830
	BERTMapLt	0.834	0.497	0.623	0.766	0.716
	BioSTransMatch	0.312	0.586	0.407	0.741	0.683
	LLM4OM	0.718	0.580	0.641	NA	NA
	GenOM	0.803	0.565	0.664	**0.910**	**0.875**
NCIT-DOID	LogMap	0.934	0.668	0.779	NA	NA
	LogMapBio	0.860	0.962	**0.908**	NA	NA
	LogMapLt	0.983	0.575	0.725	NA	NA
	Matcha	0.882	0.756	0.814	0.902	0.873
	BERTMap	0.888	0.878	0.883	**0.959**	**0.937**
	BERTMapLt	0.919	0.772	0.839	0.890	0.861
	BioSTransMatch	0.657	0.833	0.735	0.900	0.865
	LLM4OM	0.862	0.801	0.830	NA	NA
	GenOM	0.912	0.846	0.878	0.950	0.921

Table 5
Overall performance on the five OAEI 2024 Bio-ML tasks. NA indicates no results as the systems do not supporting the calculation of the metrics. Bold indicates the best performance, while underline indicates the second-best. All the baseline results are from the track's website, where the LogMapLt and the BERTMap series are reimplemented with system submissions, while the other baselines submitted the final result files without system reimplementation, probably using settings optimised for each task.

drop in Hit@10 on the *SNOMED-NCIT (pharm)* task. These results suggest that enriched definitions help retrieve a greater number of correct candidates, thereby increasing the likelihood of including the true match within the top-k shortlist.

4.4.2. Effectiveness Compared to Original Exact Matching Methods

We compare the results of GenOM with the results of the stand-alone exact matching system that GenOM adopts in the final stage (either BERTMapLt or LogMapLt). The BERTMapLt and LogMapLt

Task	MRR	Hit@1
OMIM_ORDO (with)	0.880	0.831
OMIM_ORDO (without)	0.659	0.543
NCIT_DOID (with)	0.889	0.820
NCIT_DOID (without)	0.858	0.782
SNOMED_NCIT_pharm (with)	0.853	0.770
SNOMED_NCIT_pharm (without)	0.777	0.688
SNOMED_NCIT_neoplas (with)	0.863	0.777
SNOMED_NCIT_neoplas (without)	0.824	0.732
SNOMED_FMA_body (with)	0.754	0.631
SNOMED_FMA_body (without)	0.764	0.633

Table 6
Impact of concept definitions on LLM-based Local Ranking (Qwen-2.5-7B)

Task	Hit@1	Hit@5	Hit@10
OMIM_ORDO (with)	0.737	0.879	0.909
OMIM_ORDO (without)	0.724	0.875	0.907
NCIT_DOID (with)	0.886	0.974	0.985
NCIT_DOID (without)	0.786	0.940	0.962
SNOMED_NCIT_pharm (with)	0.747	0.942	0.966
SNOMED_NCIT_pharm (without)	0.712	0.911	0.946
SNOMED_NCIT_neoplas (with)	0.722	0.896	0.932
SNOMED_NCIT_neoplas (without)	0.718	0.896	0.938
SNOMED_FMA_body (with)	0.695	0.912	0.947
SNOMED_FMA_body (without)	0.642	0.883	0.930

Table 7
Impact of concept definitions on embedding-based candidate retrieval (text-embedding-3-small)

results reported in this section are reproduced within the scope of this work, and may therefore differ slightly from the results presented earlier. As shown in Table 8, GenOM consistently outperforms the original exact matchers across all evaluated tasks.

GenOM (BERTMapLt) achieved an average F1 score of 0.771 across the five benchmark tasks, outperforming the original BERTMapLt model, which obtained an average of 0.744. A similar improvement was observed with LogMapLt: while the standalone LogMapLt achieved an average F1 score of only 0.631, the GenOM-enhanced version reached 0.716. These results suggest that while exact matching provides a solid foundation for identifying high-confidence correspondences, it remains limited in capturing more nuanced semantic equivalence. By integrating LLM-based reasoning and enriched conceptual representations, GenOM is able to significantly enhance both coverage and accuracy over the base exact matching techniques. This is reflected in a notable increase in recall: GenOM achieves, on average, an 8% improvement in recall over BERTMapLt, and an even more substantial 24% increase when compared to LogMapLt.

4.4.3. Effectiveness of Few-Shot Prompting

This experiment also investigates the effect of few-shot prompting on the LLM-based equivalence judgement stage. The results are shown inTable 9, where few-shot prompting is set to 2 examples, all results are reported prior to the integration of exact matching, and the threshold for cosine similarity was kept consistent with the earlier setting at 0.97. Aside from the inclusion of few-shot examples, all other settings are identical. The evaluation was conducted without incorporating results from the exact matching module, as the impact of the few-shot strategy tends to be diminished once exact matching is applied. The results indicate that incorporating two-shot examples consistently improves performance across most tasks. Except for the *NCIT-DOID* task, where performance remains unchanged, all other tasks exhibit notable gains in F1 score. This demonstrates that few-shot prompting can effectively guide the LLM towards more accurate classification, especially in borderline cases where single-instance reasoning may be insufficient.

Task	Model Variant	P	R	F1
SNOMED-NCIT-neoplas	GenOM(BERTMapLt)	0.795	0.764	**0.779**
	BERTMapLt	0.831	0.687	0.752
	GenOM(LogMapLt)	0.869	0.655	**0.747**
	LogMapLt	0.952	0.491	0.648
SNOMED-NCIT-pharm	GenOM(BERTMapLt)	0.989	0.600	**0.747**
	BERTMapLt	0.981	0.574	0.724
	GenOM(LogMapLt)	0.988	0.599	**0.746**
	LogMapLt	0.996	0.586	0.738
SNOMED-FMA-body	GenOM(BERTMapLt)	0.944	0.678	**0.789**
	BERTMapLt	0.979	0.655	0.785
	GenOM(LogMapLt)	0.876	0.606	**0.716**
	LogMapLt	0.971	0.527	0.683
NCIT-DOID	GenOM(BERTMapLt)	0.912	0.846	**0.878**
	BERTMapLt	0.919	0.772	0.839
	GenOM(LogMapLt)	0.939	0.753	**0.835**
	LogMapLt	0.955	0.602	0.738
OMIM-ORDO	GenOM(BERTMapLt)	0.803	0.565	**0.664**
	BERTMapLt	0.834	0.497	0.623
	GenOM(LogMapLt)	0.839	0.394	**0.537**
	LogMapLt	0.937	0.215	0.350

Table 8
The results of GenOM and the exact matching systems BERTMapLt and LogMapLt. The results of BERTMapLt and LogMapLt here are reproduced, as a component as of the GenOM framework, and have a small difference as those in Table 5.

Task	Prompt Strategy	P	R	F1
SNOMED-NCIT-neoplas	LLM (Qwen2.5 7B)	0.820	0.308	0.448
	Few-shot	0.815	0.346	**0.486**
SNOMED-NCIT-pharm	LLM (Qwen2.5 7B)	0.94	0.183	0.306
	Few-shot	0.919	0.21	**0.342**
SNOMED-FMA-body	LLM (Qwen2.5 7B)	0.789	0.016	0.113
	Few-shot	0.846	0.166	**0.278**
NCIT-DOID	LLM (Qwen2.5 7B)	0.942	0.474	0.631
	Few-shot	0.935	0.476	**0.632**
OMIM-ORDO	LLM (Qwen2.5 7B)	0.807	0.257	0.390
	Few-shot	0.795	0.271	**0.405**

Table 9
The results of the LLM-based equivalence judgement stage of GenOM, with and without few-shot prompting

5. Conclusion, Discussion and Future Work

This paper presents **GenOM**, a general-purpose framework for ontology alignment that integrates concept semantic enrichment with LLM-based textual definition generation, embedding-based candidate retrieval, LLM prompting-based equivalence judgement, and exact matching in a modular design. The approach demonstrates strong performance across five biomedical ontology alignment tasks of OAEI Bio-ML, outperforming many baselines, and the effectiveness of its important components has been verified via extensive ablation studies. In particular, the framework shows its ability to generalise across datasets while maintaining alignment accuracy, without relying on handcrafted features or extensive task-specific engineering.

Although GemOM has achieved promising performance for equivalence mappings in OM, several challenges remain:

1. It is difficult to consistently assess the degree of equivalence between concept pairs. This challenge affects both the LLM-based judgement stage and the choice of cosine similarity threshold for candidate retrieval.
2. The definition of equivalence which can vary subtly across tasks: concept pairs deemed equivalent in one alignment task may not be considered so in another, leading to inconsistencies in judgement. Accordingly, the optimal similarity threshold becomes task-dependent. For example, a cosine similarity above 0.80 indicates equivalence in some tasks, but some other tasks may require a threshold of 0.95 to ensure equivalence.
3. The alignment performance of LLMs is highly sensitive to the prompt. In evaluation, we observed that vague prompts such as simply asking the model to "determine whether two concepts are equivalent" often fails to elicit correct predictions; in many cases, the LLM almost never produces a "YES" output. This highlights the importance of prompt specificity in steering LLM behaviour and underscores a practical challenge in applying LLMs to alignment in a generalisable way.

For the future work, one key direction is to expand the scope of GenOM to include additional alignment types beyond equivalence, such as subsumption. Another key direction involves addressing the variability in how equivalence is defined across different ontologies and tasks. In many alignment scenarios, the threshold for considering two concepts equivalent may depend on contextual or domain-specific nuances, which are difficult to capture using a fixed similarity score or binary decision. To tackle this, future research will explore task-adaptive alignment criteria, including dynamic threshold selection and prompt-based calibration techniques that allow the LLM to assess the strength or type of correspondence more flexibly. Additionally, incorporating finer-grained semantic similarity measures and confidence estimation strategies could help better reflect the spectrum of equivalence relations observed in practice.

Declaration on Generative AI

During the preparation of this work, the authors used Grammarly in order to grammar and spell check, and improve the text readability. After using the tool, the authors reviewed and edited the content as needed to take full responsibility for the publication's content.

References

[1] S. Staab, R. Studer, Handbook on ontologies, Springer Science & Business Media, 2013.

[2] P. Shvaiko, J. Euzenat, Ontology matching: State of the art and future challenges, IEEE Transactions on Knowledge and Data Engineering 25 (2013) 158–176. doi:10.1109/TKDE.2011.253.

[3] J. Euzenat, P. Shvaiko, Ontology matching, 2nd ed., Springer-Verlag, Heidelberg (DE), 2013.

[4] SNOMED International, SNOMED CT: The Global Clinical Terminology, https://www.snomed.org, 2024. Accessed: 2025-06-23.

[5] E. Jiménez-Ruiz, B. Cuenca Grau, 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, Springer, 2011, pp. 273–288.

[6] 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 2013 Conferences: Confederated International Conferences: CoopIS, DOA-Trusted Cloud, and ODBASE 2013, Graz, Austria, September 9-13, 2013. Proceedings, Springer, 2013, pp. 527–541.

[7] H. B. Giglou, J. D'Souza, F. Engel, S. Auer, Llms4om: Matching ontologies with large language models, arXiv preprint arXiv:2404.10317 (2024).

[8] S. Hertling, H. Paulheim, Olala: Ontology matching with large language models, in: Proceedings of the 12th Knowledge Capture Conference 2023, 2023, pp. 131–139.

[9] S. Anam, Y. S. Kim, B. H. Kang, Q. Liu, Review of ontology matching approaches and challenges, International Journal of Computer Science and Network Solutions 3 (2015) 1–27.

[10] A. Doan, J. Madhavan, P. Domingos, A. Halevy, Ontology matching: A machine learning approach, in: Handbook on ontologies, Springer, 2004, pp. 385–403.
[11] P. Kolyvakis, A. Kalousis, D. Kiritsis, Deepalignment: Unsupervised ontology matching with refined word vectors, in: Proceedings of the 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 1-6 June 2018, 2018.
[12] I. Nkisi-Orji, N. Wiratunga, S. Massie, K.-Y. Hui, R. Heaven, Ontology alignment based on word embedding and random forest classification, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2018, pp. 557–572.
[13] L. L. Wang, C. Bhagavatula, M. Neumann, K. Lo, C. Wilhelm, W. Ammar, Ontology alignment in the biomedical domain using entity definitions and context, in: Proceedings of the BioNLP 2018 workshop, 2018, pp. 47–55.
[14] J. Chen, E. Jiménez-Ruiz, I. Horrocks, D. Antonyrajah, A. Hadian, J. Lee, Augmenting ontology alignment by semantic embedding and distant supervision, in: European Semantic Web Conference, Springer, 2021, pp. 392–408.
[15] A. Bento, A. Zouaq, M. Gagnon, Ontology matching using convolutional neural networks, in: Proceedings of the Twelfth Language Resources and Evaluation Conference, 2020, pp. 5648–5653.
[16] 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. 5684–5691.
[17] S. Menad, W. Laddada, S. Abdeddaïm, L. F. Soualmia, Biostransformers for biomedical ontologies alignment., in: KEOD, 2023, pp. 73–84.
[18] J. Devlin, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).
[19] J. Chen, Y. He, Y. Geng, E. Jiménez-Ruiz, H. Dong, I. Horrocks, Contextual semantic embeddings for ontology subsumption prediction, World Wide Web (WWW) 26 (2023) 2569–2591. URL: https://doi.org/10.1007/s11280-023-01169-9. doi:10.1007/S11280-023-01169-9.
[20] S. S. Norouzi, M. S. Mahdavinejad, P. Hitzler, Conversational ontology alignment with chatgpt, ArXiv abs/2308.09217 (2023). URL: https://api.semanticscholar.org/CorpusID:261031024.
[21] Y. He, J. Chen, H. Dong, I. Horrocks, Exploring large language models for ontology alignment, arXiv preprint arXiv:2309.07172 (2023).
[22] Z. Qiang, W. Wang, K. Taylor, Agent-om: Leveraging llm agents for ontology matching, arXiv preprint arXiv:2312.00326 (2024).
[23] Y. He, J. Chen, H. Dong, I. Horrocks, C. Allocca, T. Kim, B. Sapkota, DeepOnto: A python package for ontology engineering with deep learning, Semantic Web 15 (2024) 1991–2004. URL: download/2024/HeCDHAKS24.pdf. doi:10.3233/SW-243568.
[24] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, arXiv preprint arXiv:1908.10084 (2019).
[25] Y. He, J. Chen, H. Dong, E. Jiménez-Ruiz, A. Hadian, I. Horrocks, Machine learning-friendly biomedical datasets for equivalence and subsumption ontology matching, in: International Semantic Web Conference, Springer, 2022, pp. 575–591.