Large Language Models as Oracles for Ontology Alignment

Sviatoslav Lushnei[1], Dmytro Shumskyi[1], Severyn Shykula[1],
Ernesto Jimenéz-Ruiz*[2], and Artur d'Avila Garcez[2]

[1] *Ukrainian Catholic University, Ukraine*
[2] *City St George's, University of London, UK*

August 2025

Abstract

Ontology alignment plays a crucial role in integrating diverse data sources across domains. There is a large plethora of systems that tackle the ontology alignment problem, yet challenges persist in producing highly quality correspondences among a set of input ontologies. Human-in-the-loop during the alignment process is essential in applications requiring very accurate mappings. User involvement is, however, expensive when dealing with large ontologies. In this paper, we explore the feasibility of using Large Language Models (LLM) as an alternative to the domain expert. The use of the LLM focuses only on the validation of the subset of correspondences where an ontology alignment system is very uncertain. We have conducted an extensive evaluation over several matching tasks of the Ontology Alignment Evaluation Initiative (OAEI), analysing the performance of several state-of-the-art LLMs using different ontology-driven prompt templates. The LLM results are also compared against simulated Oracles with variable error rates.

Keywords: knowledge graph alignment, large language models, oracle

Supplemental Material: Source codes and relevant resources for the evaluation conducted in this paper are available in this Zenodo repository: https://doi.org/10.5281/zenodo.15394653. Latest codes are available in this GitHub repository: https://github.com/city-artificial-intelligence/rai-ukraine-kga-llm

1 Introduction

Ontology alignment [9] plays a crucial role in integrating diverse data sources across domains. While numerous ontology matching systems exist (*e.g.*, [31]), systems capable of producing highly quality correspondences among the input ontologies are still needed, especially in applications where accuracy is paramount. One way to address this issue is through user interaction to manually verify uncertain mappings; however, this approach is often time-consuming and expensive. An alternative is to leverage Large Language Models (LLMs) as they encode large amounts of (parametric) knowledge, and they have shown their potential to be used within an ontology alignment pipeline (*e.g.*, [32]). Nevertheless, LLMs are also financially costly, and an unlimited use may not be possible depending on the available resources.

In this paper, we have extended the state-of-the-art system `LogMap` [21, 20] to perform calls to an LLM-based Oracle. The LLM-based Oracle is used to validate a subset of the correspondences where `LogMap`, a traditional ontology alignment system, is uncertain. Thus, the power of the LLM is centred on complex cases where traditional ontology alignment techniques are not sufficient. We selected the GPT-4o Mini model[1] (OpenAI) and a range of Google Gemini Flash models[2] (v1.5, 2.0, 2.0 Lite, and 2.5 Preview) for our experiments, due to their optimal

*Corresponding author: ernesto.jimenez-ruiz@citystgeorges.ac.uk

[1] https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/

[2] https://deepmind.google/technologies/gemini/

balance between performance and cost. Each model offers competitive pricing while maintaining high-quality outputs, making them suitable for ontology alignment tasks. The calls to the LLM-based oracle are performed via ontology-driven prompts that exploit different levels of lexical and contextual information of the entities in the mappings to be assessed. The use of multiple Gemini variants also allowed us to analyse the evolution of the LLM model.

To analyse the suitability of the LLM-based Oracles, we have conducted an extensive evaluation with the *anatomy* [8], *largebio* [22], and *bio-ml* [16] datasets of the OAEI campaign [30, 31], involving a total of nine matching tasks. These datasets are complex and have become a reference within the community. We have assessed the diagnostic capabilities of thirty different LLM-based Oracles, depending on the choice among the five LLMs and six prompt templates. We have also evaluated the contribution of the LLM-based Oracles on the overall matching task by comparing the results with `LogMap` (automatic mode) and simulated Oracles with variable error rates [25].

In contrast to other state-of-the-art systems that heavily rely on LLMs, our approach is designed to be cost-effective, where the LLM-based Oracle is only used in very specific cases, limiting the number of calls. Hence, the use of LLMs is more accessible without the need for substantial computational infrastructure or financial resources. Additionally, we conduct a more fine-grained analysis of *(i)* the diagnostics capabilities of the LLM-based Oracle, *(ii)* its overall contribution relative to simulated Oracles, and *(iii)* the effect of different ontology-driven prompt templates, including both structured formats and those optimized for natural language interaction.

The rest of the paper is organised as follows. Section 2 introduces the necessary background concepts. The relevant related work is provided in Section 3. Section 4 presents the core methods of our approach. Evaluation results are analysed and discussed in Section 5. Finally, conclusions and some lines of future work are given in Section 6.

2 Preliminaries

This section introduces the essential background that underpins our methods and experiments. Traditionally, in the Ontology Matching community, there has been a distinction between schema and instance matching [9, 34], where schema matching focuses on the alignment of the terminological or conceptual knowledge of the ontology, while instance matching deals with the challenges concerning the ontology data. Currently, a similar distinction can be seen in the literature between *ontology alignment* and *knowledge graph alignment*. In the Semantic Web, however, there is not a strong difference between ontologies and knowledge graphs, as it is understood that knowledge graphs are semantically rich and described according to the vocabulary of an ontology [18].

Ontology alignment. Ontology alignment is the process of finding correspondences or *mappings* among the entities[3] of two or more ontologies. A *mapping* involving two entities is typically represented as a 4-tuple $\langle e_1, e_2, r, c\rangle$ where e_1 and e_2 are entities of the ontologies $\mathcal{O}_1$ and $\mathcal{O}_2$, respectively; r is a semantic relation, typically one of $\{\sqsubseteq, \sqsupseteq, \equiv\}$; and c is a confidence value, usually, a real number within the interval $(0, 1]$. For simplicity, we refer to an equivalence ($\equiv$) mapping as a pair $\langle e_1, e_2\rangle$.

Alignment task. In the OAEI, an alignment or matching task is composed of a pair of ontologies $\mathcal{O}_1$ (typically called source) and $\mathcal{O}_2$ (typically called target), and an associated *reference alignment* $\mathcal{M}^{RA}$. An $\mathcal{M}^{RA}$, although it may not be perfect, serves as a guide to evaluate and compare alignment systems.

Alignment system. An ontology *alignment system* is a program that, given as input an alignment task, generates an ontology alignment $\mathcal{M}^S$. We have selected the state-of-the-art alignment system `LogMap` [21, 20] as the baseline for our experiments due to its flexibility to

[3] We refer to ontology classes, properties and instances as entities.

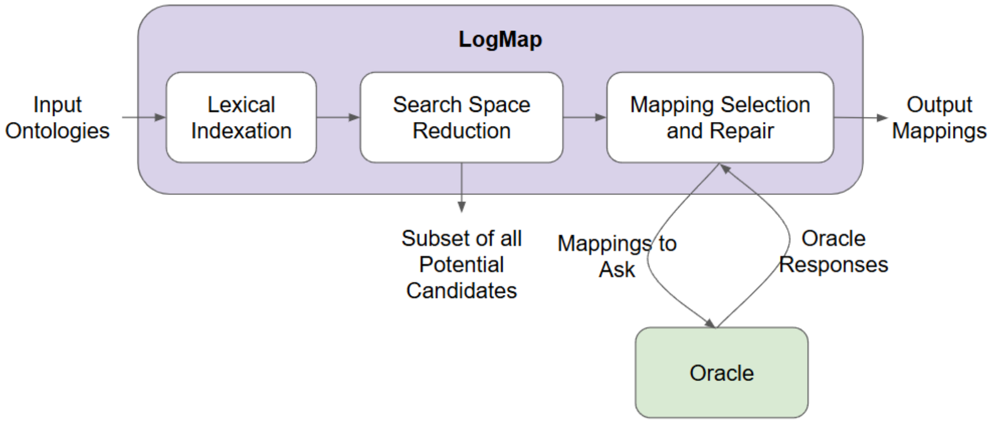

Figure 1: Workflow of the ontology alignment system `LogMap` with calls to an Oracle.

be adapted to different evaluation scenarios. `LogMap` can operate in a fully automatic mode or allow interaction with an *Oracle* [25]. Figure 1 shows the workflow followed by `LogMap` when allowing interaction. `LogMap`, during the mapping selection stage, identifies a subset of mappings $\mathcal{M}_{ask}$ (mappings to ask) for which it is uncertain and prefers to leverage the expertise of the Oracle. If the Oracle is not available, `LogMap` performs automatic decisions over $\mathcal{M}_{ask}$.

Oracle. We define an Oracle in our setting as an external party that can assess the correctness of a given mapping $\langle e_1, e_2 \rangle$. An Oracle can be a domain expert or an automated engine that exploits background knowledge. Additionally, the OAEI's interactive matching task simulates domain experts with different error rates via Oracles relying on the reference alignment of the alignment task and randomly generating erroneous replies according to the selected error rate [25].

Evaluation metrics. We use the standard evaluation metrics *Precision* (Pr), *Recall* (Re), and *F-score* (F) to evaluate an alignment $\mathcal{M}^S$, computed by a system, against a reference alignment $\mathcal{M}^{RA}$:

$$Pr = \frac{|\mathcal{M}^S \cap \mathcal{M}^{RA}|}{|\mathcal{M}^S|}, \; Re = \frac{|\mathcal{M}^S \cap \mathcal{M}^{RA}|}{|\mathcal{M}^{RA}|}, \; F = 2 \cdot \frac{Pr \cdot Re}{Pr + Re} \tag{1}$$

We use *Sensitivity* (Se), *Specificity* (Sp), and *Youden's index* (YI) [37] to evaluate the effectiveness of an Oracle to diagnose positive (sensitivity) and negative (specificity) mappings in $\mathcal{M}_{ask}$ with respect to $\mathcal{M}^{RA}$.

$$Se = \frac{|TP|}{|TP| + |FN|}, \; Sp = \frac{|TN|}{|FP| + |TN|}, \; YI = Se + Sp - 1 \tag{2}$$

TP (true positives) are mappings predicted as correct and in $\mathcal{M}^{RA}$, FN (false negatives) are mappings predicted as incorrect but in $\mathcal{M}^{RA}$, TN (true negatives) are mappings predicted as incorrect and not present in $\mathcal{M}^{RA}$, and FP (false positives) are mappings predicted as correct but not present in $\mathcal{M}^{RA}$.

LLM prompting. LLMs like GPT-4 are pretrained on vast text corpora. They are commonly used in a few-shot or zero-shot setting via prompts. Prompts can exploit the generative capabilities of the LLM or ask for specific binary decision-making. In the ontology alignment setting, a mapping $\langle e_1, e_2 \rangle$ can be transformed into a binary question to the LLM – "Do e_1 represent the same entity as e_2? (True/False)" – possibly enriched with ontology context (*e.g.*, parent classes or synonyms). This approach allows the LLM to be used as a lightweight semantic Oracle.

3 Related work

Ontology matching (or alignment) is the task of identifying correspondences between semantically related entities in two or more ontologies, with the goal of reducing terminological and conceptual heterogeneity in semantic applications [9]. The Ontology Alignment Evaluation Initiative (OAEI) has driven progress since 2004 by providing standardized benchmarks and evaluation protocols for matching systems [31]. Widely used (traditional) matchers include LogMap [21] and AgreementMakerLight (AML) [10], each leveraging different combinations of lexical, structural, and background-knowledge techniques. Human validation has long been recognized as critical for high-precision mappings. Early frameworks combined automated matching with domain expert feedback to resolve low-confidence correspondences, but at the cost of extensive user effort and time [25]. Other efforts attempted to use external resources like BioPortal as an Oracle (*e.g.*, [6]).

In recent years, a new generation of systems leveraging machine learning and (large) language models has emerged. The OAEI Bio-ML track [16] was established to foster participation in the OAEI and to facilitate the systematic evaluation of these systems. Early approaches showed promising results applying word embeddings to the ontology alignment task (*e.g.*, [24, 26, 19]). Knowledge graph embeddings systems like OWL2Vec [4] were also leveraged in combination with a machine learning architecture to learn and validate ontology alignments (*e.g.*, [5, 13]). Systems relying on BERT-based models have also become popular, given their flexibility to be fine-tuned to specific tasks like ontology alignment. Prominent examples include BERTMap [14], BioGITOML [29], and the Matcha family [11]. Nevertheless, recent developments in the field are increasingly driven by approaches based on Large Language Models. Saki Norouzi et al. [27] and He et al. [15] performed exploratory studies about the potential of LLMs to the ontology alignment task, while Amini et al. [2] extended the exploration to discover complex alignments beyond equivalence or subsumption. Systems like OLaLa [17], LLMs4OM [12], MILA [35], Agent-OM [32], and HybridOM [36] have integrated LLMs within their architectures. A common technique followed in the literature is to use retrieval methods to select top-k candidates for each entity, and then ask the LLM to select the best one among these candidates (*e.g.*, [17, 12, 35]). In contrast, HybridOM uses the LLM to generate additional lexical descriptions of the entities involved in candidate correspondences. Agent-OM leverages autonomous LLM agents to orchestrate multiple matching subtasks, illustrating the potential of agentive workflows in ontology matching [32]. Recent work has explored the use of LLMs to focus on the alignment of the assertional content (*i.e.*, instance data) within the knowledge graphs (*e.g.*, [38, 7]).

Our approach builds upon `LogMap` [21, 20] and employs the LLM as an Oracle to assess a targeted subset of mappings. Rather than attempting to evaluate a large set of candidate correspondences, we focus on the validation of mappings where `LogMap` is uncertain. Systems like MILA [35] have also focused on the limitation of the number of queries, leading to a reduction in the required computation times. We also investigate the effect of incorporating the ontology context of the entities into the prompt design, an aspect that has not been thoroughly examined in the existing literature. In addition, we conduct an extensive evaluation of the LLM-based Oracles, examining their effectiveness as diagnostic tools. We further assess their performance by comparing results with simulated Oracles with varying levels of error rate.

4 Methods

As introduced in Section 2, we build upon the system `LogMap`. `LogMap`, in addition to predicting the set of output mappings, also identifies a subset of mappings that are challenging (*i.e.*, $\mathcal{M}_{ask}$) and can optionally be given to an Oracle (recall Figure 1). In this paper, we have extended the architecture of `LogMap` to use a state-of-the-art *LLM as an Oracle* as depicted in Figure 2. We restrict the use of the LLM on the $\mathcal{M}_{ask}$ mappings. These mappings are not trivial, as they typically involve entities with different labels and/or contexts, and they are better suited to challenge the performance of LLMs. `LogMap` interacts with the Oracle on-demand for each of mapping $\langle e_1, e_2 \rangle$ where it is uncertain (*i.e.*, $\langle e_1, e_2 \rangle \in \mathcal{M}_{ask}$). The following subsections detail the internal steps involved in the interaction with the LLM-based Oracle.

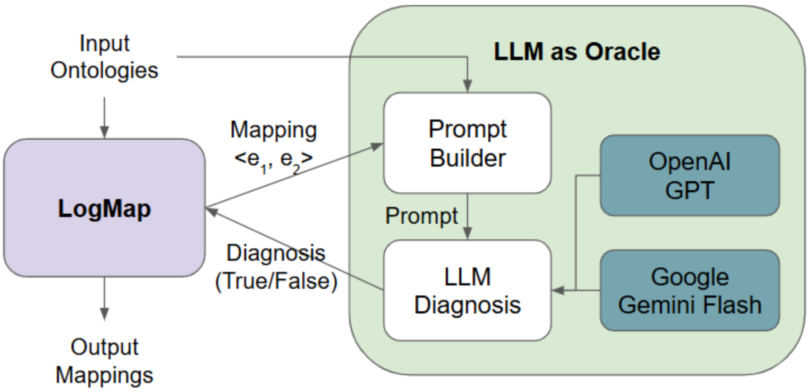

Figure 2: LLM-in-the-loop as Oracle in Ontology Alignment.

4.1 Ontology-driven prompt builder

The first step in the interaction with the LLM-based Oracle is the creation of an ontology-driven prompt to ask about the correctness of a given candidate mapping $\langle e_1, e_2 \rangle$. The respective ontologies provide the lexical representations (*i.e.*, `label(e)`), synonyms (*i.e.*, `synonyms(e)`), as well as the context for e_1 and e_2 (*e.g.*, `direct_parent(e)`, `parents(e)`). According to the locality principle, mappings should link entities that have similar neighbourhoods [23]. Hence, a basic prompt should include at least the lexical representation of the entities e_1 and e_2, and the lexical representation of one of the directly connected entities.

We have designed six different prompt templates combining three characteristics: *(i)* using similar sentences to how humans write (natural language-friendly, **NLF**), *(ii)* inclusion of extended context (**EC**), and *(iii)* inclusion of synonyms (**S**). Prompts without an extended context only include one of the direct ancestors for classes and properties, and one of the direct types for individuals. While the prompts with an extended context include two levels of parent classes. We hypothesize that enabling the **NLF**, **EC**, and **S** characteristics in the prompts will yield better results. For example, we refer to a prompt template with all the above characteristics as $\mathbf{P}^{\text{NLF}}_{\text{EC+S}}$. For each mapping $\langle e_1, e_2 \rangle$ to be assessed, we dynamically populate each of the prompt templates according to the entities in the mapping and their associated ontology information. Next, we show the populated prompts for the mapping $\langle$`mouse:MA_0001771` (alveolus epithelium), `human:NCI_C12867` (Alveolar_Epithelium) $\rangle$.

Structured prompts. This type of prompt uses structured information with less common natura l language expressions. Listings 1-2 show the prompts $\mathbf{P}$ and $\mathbf{P}_{\text{EC}}$ where the entities to be aligned and their context (extended with $\mathbf{P}_{\text{EC}}$) are provided as a list of elements.

```
Analyze the following entities, each originating from a distinct ontology. Your task is to assess
whether they represent the **same ontological concept**, considering both their semantic meaning
and hierarchical position.

1. Source entity: "alveolus epithelium"
    - Direct ontological parent: lung epithelium

2. Target entity: "Alveolar_Epithelium"
    - Direct ontological parent: Epithelium

Are these entities **ontologically equivalent** within their respective ontologies? Respond with
"True" or "False".
```

Listing 1: Basic prompt without any of the characteristics enabled ($\mathbf{P}$)

```
Analyze the following entities, each originating from a distinct ontology. Each is represented by
its **ontological lineage**, capturing its hierarchical placement from the most general to the
most specific level.

1. Source entity ontological lineage:
  Level 0: alveolus epithelium
  Level 1: lung epithelium
  Level 2: respiratory system epithelium

2. Target entity ontological lineage:
  Level 0: Alveolar_Epithelium
  Level 1: Epithelium
  Level 2: Epithelial_Tissue, Normal_Tissue

Based on their **ontological positioning, hierarchical relationships, and semantic alignment**, do
these entities represent the **same ontological concept**? Respond with "True" or "False".
```

Listing 2: $\mathbf{P}_{\mathrm{EC}}$ prompt (non natural language-friendly with extended context).

Natural-language friendly prompts. These prompts are based on the assumption that, given LLMs are trained on large corpora of human-generated text, formulating questions in a more natural language will likely yield more accurate results. Listings 3-4 show the examples of the $\mathbf{P}^{\mathrm{NLF}}$ and $\mathbf{P}_{\mathrm{EC}}^{\mathrm{NLF}}$ prompts.

```
We have two entities from different ontologies.

The first one is "alveolus epithelium", which belongs to the broader category "lung epithelium"

The second one is "Alveolar_Epithelium", which belongs to the broader category "Epithelium"

Do they mean the same thing? Respond with "True" or "False".
```

Listing 3: $\mathbf{P}^{\mathrm{NLF}}$ Prompt (natural-language friendly).

```
We have two entities from different ontologies.

The first one is "alveolus epithelium", which belongs to the broader category "lung epithelium",
under the even broader category "respiratory system epithelium"

The second one is "Alveolar_Epithelium", which belongs to the broader category "Epithelium", under
the even broader category "Epithelial_Tissue, Normal_Tissue"

Do they mean the same thing? Respond with "True" or "False".
```

Listing 4: $\mathbf{P}_{\mathrm{EC}}^{\mathrm{NLF}}$ Prompt (natural-language friendly with extended context).

NLF prompts with synonyms. Although LLMs may inherently encode synonyms and lexical variations related to the ontology entities, the $\mathbf{P}_{\mathrm{S}}^{\mathrm{NLF}}$ and $\mathbf{P}_{\mathrm{EC+S}}^{\mathrm{NLF}}$ prompts are designed to analyse the impact of explicitly including synonyms for both the entities in a given correspondence and their associated context. Examples of the $\mathbf{P}_{\mathrm{S}}^{\mathrm{NLF}}$ and $\mathbf{P}_{\mathrm{EC+S}}^{\mathrm{NLF}}$ prompts are provided in Listings 5–6, respectively.

```
We have two entities from different ontologies.

The first one is "alveolus epithelium", which falls under the category "lung epithelium".

The second one is "Alveolar_Epithelium", also known as "Lung Alveolar Epithelia", "Alveolar
Epithelium", "Epithelia of lung alveoli", which falls under the category "Epithelium".

Do they mean the same thing? Respond with "True" or "False".
```

Listing 5: $\mathbf{P}_{\mathrm{S}}^{\mathrm{NLF}}$ Prompt (natural-language friendly with synonyms).

```
We have two entities from different ontologies.

The first one is "alveolus epithelium", belongs to broader category "lung epithelium", under the
even broader category "respiratory system epithelium" (also known as "respiratory system mucosa").

The second one is "Alveolar_Epithelium", also known as "Alveolar Epithelium", "Lung Alveolar
Epithelia", "Epithelia of lung alveoli", belongs to broader category "Epithelium" (also known as
"Epithelium", "epithelium"), under the even broader category "Epithelial_Tissue, Normal_Tissue".

Do they mean the same thing? Respond with "True" or "False".
```

Listing 6: $\mathbf{P}^{\text{NLF}}_{\text{EC+S}}$ Prompt (natural-language friendly with synonyms and extended context).

System prompts. In addition to mapping specific templates, it is possible to add, for each LLM session, a short system message that frames the model's overall role and reasoning style before it sees any individual mapping question. We experimented with sessions using no system prompt as well as various system prompt variants, positioning the LLM as follows: *(i)* an ontology matching expert to ensure precision; *(ii)* a biomedical ontology specialist, emphasizing hierarchical and semantic context; *(iii)* to leverage explicitly provided synonyms and parent-class semantics; and *(iv)* to explain its decision in a natural-language friendly manner.

4.2 LLM-based diagnosis

Our selected LLMs include GPT-4o Mini (OpenAI) and a range of Google Gemini Flash models (v1.5, 2.0, 2.0 Lite, and 2.5 Preview). These models were chosen based on their optimal balance of cost-effectiveness, response latency, scalability, reliability, and output quality, as compared to other commercial APIs and open-weight alternatives. Furthermore, these LLMs expose a consistent client interface, enabling straightforward integration into our system. Support for lightweight models such as GPT-4o Mini and Gemini 2.0 Flash-Lite ensures accessibility for researchers operating under constrained budgets. At the same time, including a progression of Gemini Flash versions (from v1.5 to v2.5) allows us to observe how model improvements over time impact diagnostic performance in the ontology alignment task.

For correct diagnostics to achieve binary (True/False) classification, we unified the interfaces and used a structured output feature. We define a Boolean answer that will be a decider of the zero-shot question that we ask the LLM. To enhance robustness, we incorporated a validation and retry mechanism that minimizes hallucinations or improperly formatted outputs. This guarantees consistent results and eliminates the need for manual output parsing.

We used the Chat Completions API[4] for OpenAI models, and the OpenAI's SDK endpoint for the Gemini Models.[5] Response latency typically remains within a few seconds, depending on prompt complexity and model characteristics. On average, responses are received in under three seconds, enabling high-throughput querying when executed in parallel. However, API rate limits imposed practical constraints on experimentation. The Gemini API permits up to 2,000 requests per minute (RPM) by default,[6] whereas OpenAI's API begins with a limit of 500 RPM and a daily quota of 10,000 requests,[7] thereby restricting the overall throughput of our experiments. Regarding the cost of experiments, token usage is a key factor. Each request typically consumes between 100 and 250 input tokens, depending on the complexity and detail of the prompt. Pricing per million input tokens varied significantly, from \$0.075[8] for Gemini Flash 1.5 and 2.0 Lite, up to \$0.15[9] for GPT-4o Mini and Gemini Flash 2.5 Preview. This results in an average cost per 1,000 requests ranging from approximately \$0.01 to \$0.04.

[4]https://platform.openai.com/docs/guides/text?api-mode=chat
[5]https://ai.google.dev/gemini-api/docs/openai
[6]https://ai.google.dev/gemini-api/docs/rate-limits
[7]https://platform.openai.com/docs/models/gpt-4o-mini
[8]https://ai.google.dev/gemini-api/docs/pricing

OAEI track	Matching task	$\|\mathcal{O}_1\|$	$\|\mathcal{O}_2\|$	$\|\mathcal{M}^{RA}\|$
Anatomy	**Mouse-Human**	2,755	3,313	1,516
Bio-ML	**NCIT-DOID**	15,991	8,516	4,686
	OMIM-ORDO	9,662	9,320	3,721
	SNOMED-FMA.body	34,562	89,180	7,256
	SNOMED-NCIT.neoplas	23,116	20,497	3,804
	SNOMED-NCIT.pharm	29,646	22,387	5,803
Largebio	**FMA-NCI**	79,049	66,919	3,024
	FMA-SNOMED	79049	122,521	9,008
	SNOMED-NCI	122,521	66,919	18,844

Table 1: Statistics of the used OAEI datasets. Ontology size represents the number of entities.

4.3 Impact of the Oracle

The diagnosis performed by the Oracle over the mapping set $\mathcal{M}_{ask}$ may have an impact on the overall `LogMap` performance as it may lead to the acceptance or rejection of additional mappings. The authors in [25] simulated Oracles with different error rates and performed an extensive analysis of the impact and error propagation of the Oracle decisions. In this work, we have followed a similar approach to evaluate the LLM-based Oracle ($\texttt{Or}^{LLM}$) against Oracles with error rates ranging from 0% (*i.e.*, perfect Oracle, $\texttt{Or}^{0}$) to 30% (*i.e.*, $\texttt{Or}^{30}$). The simulated Oracles rely on the reference alignment of the relevant matching task and generate erroneous replies with the probability of their associated error rate. These Oracles with uniformly distributed errors do not realistically represent how a domain expert would behave, but they serve our purpose to assess the performance of the LLM-based Oracle in comparison with potential domain experts that are likely to make mistakes [25].

5 Evaluation

Our experiments, conducted on a standard laptop, leveraged the APIs of the selected LLM models as detailed in Section 4.2. We used the *anatomy* [8], *largebio* [22], and *bio-ml* [16] datasets provided by the OAEI evaluation initiative [30, 31].[9] As shown in Table 1, we covered a total of nine ontology matching tasks, involving ontologies from a few thousand to several hundred thousand entities. The reference alignments of these matching tasks have different sources. In *anatomy*, the reference alignment has been manually curated, while in *bio-ml* and *largebio* the reference alignment has been automatically computed by reusing public resources like MONDO [33] and UMLS [3, 23].

Diagnostic capabilities. We tested, over the 9 matching tasks, a total of 30 LLM-based Oracles ($\texttt{Or}^{LLM}$), combining the six prompt templates introduced in Section 4.1 and the LLM models presented in Section 4.2. Figure 3 shows the correctness (Youden's index, YI) of the LLM-based Oracles in assessing the mappings in $\mathcal{M}_{ask}$ (*i.e.*, the subset of mappings identified by `LogMap` as uncertain). The YI index effectively captures the effectiveness of an Oracle to identify positive (sensitivity) and negative (specificity) mappings. A YI value equivalent to 1.0 indicates optimal performance. However, due to the complexity of the mappings in $\mathcal{M}_{ask}$, lower YI values are to be expected. For example, Figure 4 shows the average YI values across ontology matching tasks. It is worth emphasizing that different ontologies and matching tasks pose varying challenges due to lexical and structural differences. Our tested prompts aim to capture this by leveraging both structural and lexical information in the input ontologies. There is also a dependency on the selected $\mathcal{M}_{ask}$ mappings by `LogMap`, which may also be more complex in some tasks than others (*e.g.*, mappings involving isolated entities and/or with scarce synonyms). Oracles relying on the Gemini Flash 2.5 models led to the best results on average, as

[9]The *largebio* track was last held in 2021 [30], but it will be reinstated in the near future.

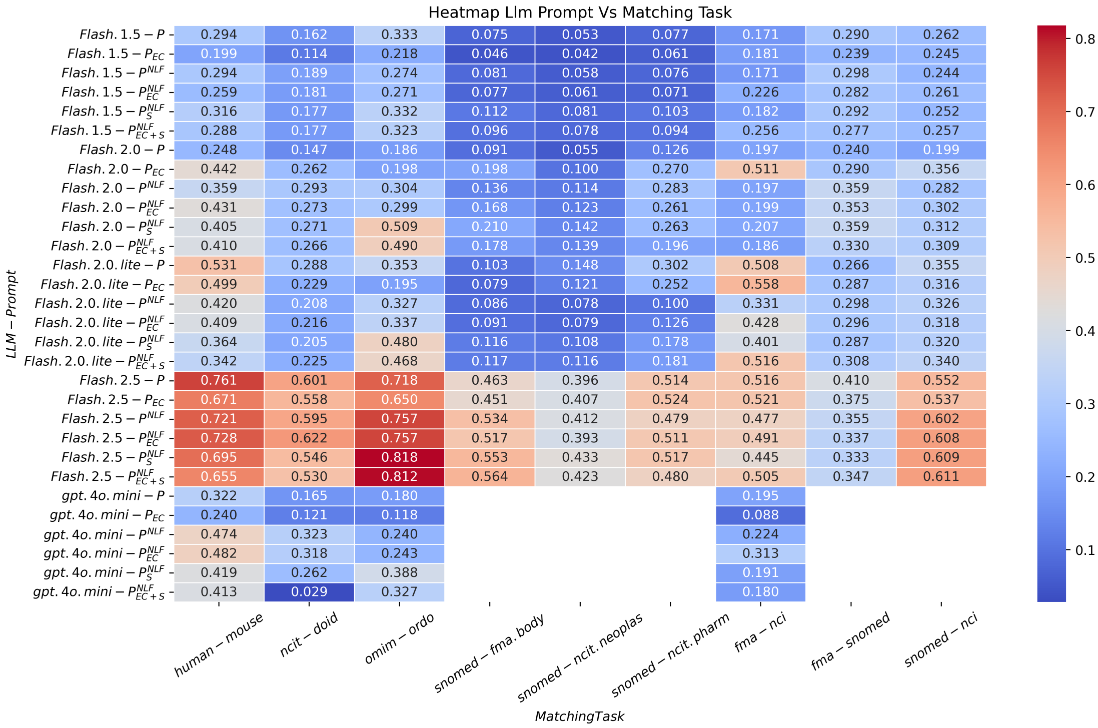

Figure 3: Diagnostic results (Youden's index) by the LLM-based Oracles over the selected ontology matching tasks. For example, *Flash 2.5*-$\mathbf{P}^{\mathrm{NLF}}_{\mathrm{EC+S}}$ represents the LLM-based Oracle relying on the Gemini Flash 2.5 model and evaluated with the natural-language friendly (NLF) prompts with extended context (EC) and synonyms (S). We only completed a subset of experiments with GPT-4o Mini as a reference.

Matching task	$\|\mathcal{M}_{ask}\|$		`LogMap` on $\mathcal{M}_{ask}$			$\mathtt{Or}^{LLM}_{GF2.5}$ on $\mathcal{M}_{ask}$		
	P	**N**	**Se**	**Sp**	**YI**	**Se**	**Sp**	**YI**
Mouse-Human	165	94	1.000	0.000	0.000	0.951	0.744	**0.695**
NCIT-DOID	364	492	1.000	0.000	0.000	0.849	0.697	**0.546**
OMIM-ORDO	172	227	0.343	0.551	-0.106	0.942	0.876	**0.818**
SNOMED-FMA.body	369	619	0.881	0.149	0.029	0.884	0.669	**0.553**
SNOMED-NCIT.neoplas	704	601	0.984	0.067	0.051	0.840	0.593	**0.433**
SNOMED-NCIT.pharm	297	260	0.929	0.065	-0.005	0.848	0.669	**0.517**
FMA-NCI	410	475	0.705	0.726	0.431	0.761	0.684	**0.445**
FMA-SNOMED	831	621	0.941	0.225	0.166	0.480	0.853	**0.333**
SNOMED-NCI	1450	1128	0.887	0.395	0.281	0.846	0.763	**0.609**
Average	529	502	0.852	0.242	0.094	0.822	0.728	**0.550**

Table 2: Comparison of `LogMap` (automatic mode) against the best LLM-based Oracle ($\mathtt{Or}^{LLM}_{GF2.5}$, using Gemini Flash 2.5 and $\mathbf{P}^{\mathrm{NLF}}_{\mathrm{S}}$ prompts) to diagnose the correctness of $\mathcal{M}_{ask}$. P=real positives in $\mathcal{M}_{ask}$, N=real negatives in $\mathcal{M}_{ask}$, Se=Sensitivity, Sp=Specificity, YI=Youden's index.

summarised in Figure 5. The best results were achieved by the combination of Gemini Flash 2.5 and $\mathbf{P}^{\mathrm{NLF}}_{\mathrm{S}}$ prompts ($\mathtt{Or}^{LLM}_{GF2.5}$). Table 2 compares the performance of `LogMap` (automatic mode) and $\mathtt{Or}^{LLM}_{GF2.5}$ in diagnosing the mappings in $\mathcal{M}_{ask}$. As anticipated, `LogMap` performs poorly as a diagnostic engine for the mappings in $\mathcal{M}_{ask}$, yielding YI values close to 0 (*i.e.*, no discriminative power), whereas $\mathtt{Or}^{LLM}_{GF2.5}$ achieves significantly better results, with an average YI value exceeding 0.5. Good values YI are generally considered to be above 0.5.

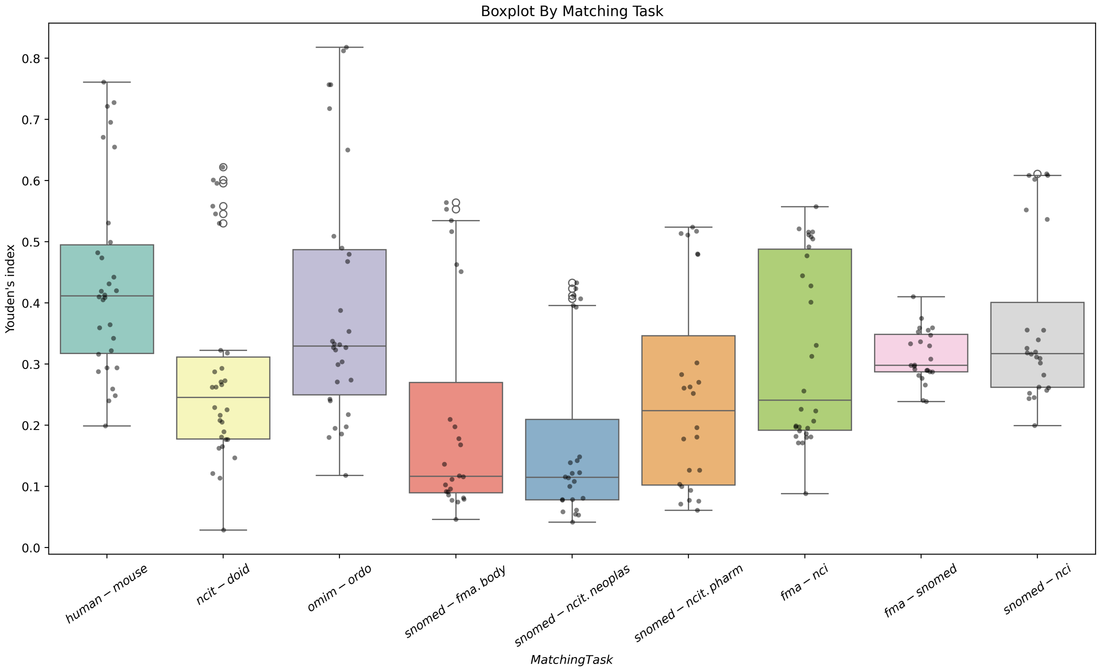

Figure 4: Average YI values for each ontology matching task, reflecting the varying complexity of $\mathcal{M}_{ask}$ across tasks.

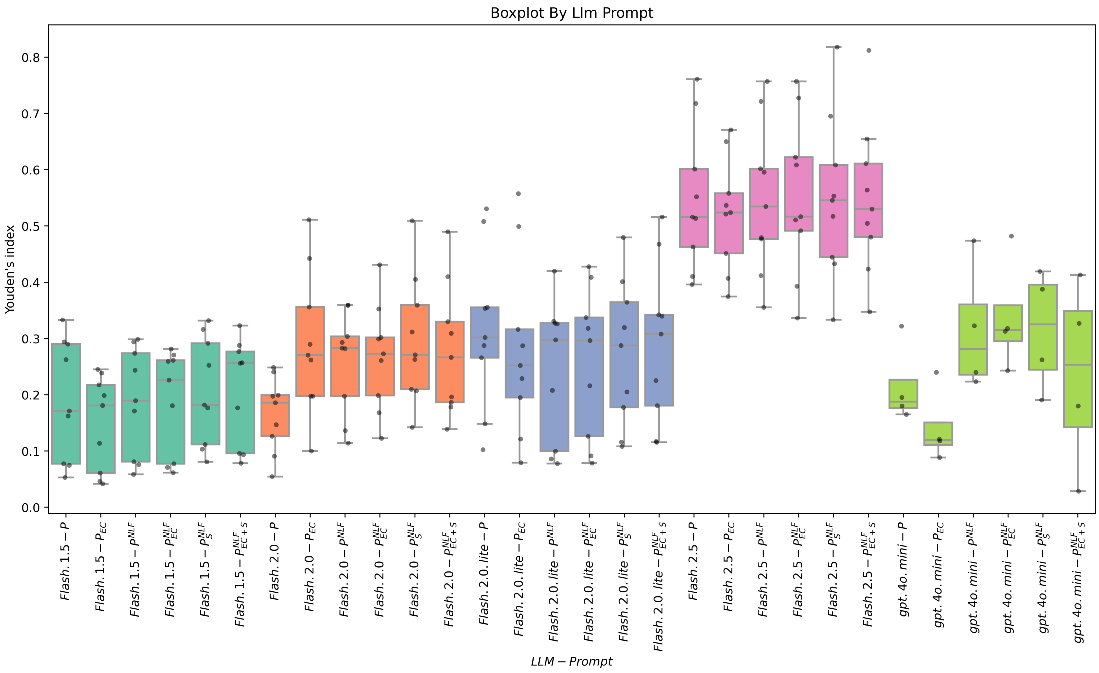

Figure 5: Summary of the diagnostic results (Youden's index) by the LLM-based Oracles. As the results for the Oracles relying on the GPT-4o Mini model are for a subset of the tasks, their boxplots are only indicative for comparison across prompt templates.

Impact of the prompt template and LLM model. Figure 5 illustrates the impact of using different LLMs models and prompt templates. We can appreciate a positive evolution of the Gemini Flash versions, with Gemini Flash 2.5 showing superior performance when assessing the correctness of $\mathcal{M}_{ask}$. The overall performance of the Oracles relying on the GPT-4o Mini models

Matching task	LogMap			LogMap+$\texttt{Or}^{LLM}_{GF2.0}$			LogMap+$\texttt{Or}^{LLM}_{GF2.5}$		
	Pr	**Re**	**F**	**Pr**	**Re**	**F**	**Pr**	**Re**	**F**
Mouse-Human	0.915	0.848	0.880	0.945	0.844	0.892	0.963	0.842	**0.898**
NCIT-DOID	0.845	0.895	0.869	0.875	0.890	0.882	0.907	0.883	**0.895**
OMIM-ORDO	0.874	0.448	0.592	0.882	0.478	0.620	0.914	0.476	**0.626**
SNOMED-FMA.body	0.695	0.538	0.607	0.727	0.543	0.622	0.751	0.545	**0.632**
SNOMED-NCIT.neoplas	0.624	0.774	0.691	0.636	0.763	0.694	0.661	0.747	**0.701**
SNOMED-NCIT.pharm	0.825	0.625	0.711	0.847	0.625	**0.719**	0.855	0.621	**0.719**
FMA-NCI	0.860	0.800	0.829	0.901	0.796	**0.845**	0.853	0.804	0.828
FMA-SNOMED	0.796	0.641	0.710	0.814	0.644	**0.719**	0.854	0.585	0.694
SNOMED-NCI	0.868	0.650	0.743	0.866	0.656	0.747	0.897	0.646	**0.751**
Average	0.811	0.691	0.737	0.833	0.693	**0.749**	0.851	0.683	**0.749**

Table 3: Comparison of LogMap (automatic mode) with LogMap with the best LLM-based Oracles ($\texttt{Or}^{LLM}_{GF2.0}$ and $\texttt{Or}^{LLM}_{GF2.5}$) on the full matching tasks. Pr=Precision, Re=Recall, F=F-score.

is lower. In terms of different prompt templates, the behaviour varies across models. Natural-language friendly prompts result in more consistent behaviour, while incorporating extended context and synonyms has a positive impact in some of the matching tasks. Overall, for the Gemini Flash 2.0 and Flash 2.5 models, the most effective prompts were $\mathbf{P}^{\text{NLF}}_{\text{S}}$ (natural-language friendly with synonyms).

Impact on the overall matching task. Table 3 presents the results obtained using LogMap (automatic mode), compared to a version of LogMap integrated with an LLM-based Oracle (as depicted in Figure 2). We selected the top-performing LLM-based Oracles using $\mathbf{P}^{\text{NLF}}_{\text{S}}$ prompts: $\texttt{Or}^{LLM}_{GF2.0}$ (based on Gemini Flash 2.0) and $\texttt{Or}^{LLM}_{GF2.5}$ (based on Gemini Flash 2.5). As anticipated, the integration with the LLM-based Oracle yields improved results across all tasks in terms of F-score. LogMap+$\texttt{Or}^{LLM}_{GF2.5}$ dominates the *anatomy* and *bio-ml* tasks, while LogMap+$\texttt{Or}^{LLM}_{GF2.0}$ achieves the best results in *largebio*. To better contextualise the effectiveness of the LLM-based Oracle, we compared its performance with simulated Oracles of variable error rates, following the approach in [25] as detailed in Section 4.3. Figure 6 compares performance across all nine ontology matching tasks for: LogMap, LogMap+$\texttt{Or}^{LLM}_{GF2.5}$, and LogMap with the simulated Oracles $\texttt{Or}^{0}$, $\texttt{Or}^{20}$, and $\texttt{Or}^{30}$ (corresponding to error rates of 0%, 20%, and 30%, respectively). We can observe that $\texttt{Or}^{LLM}_{GF2.5}$ performs similarly to $\texttt{Or}^{20}$, except in the FMA-SNOMED task (lower F-score) and OMIM-ORDO task (higher F-score). In line with previous studies [20, 25], LogMap combined with $\texttt{Or}^{30}$ still outperforms LogMap (without Oracle).

Comparison with OAEI systems. The main purpose of the conducted evaluation was not to focus on the absolute values of F-score, but on the relative improvement with respect to LogMap and the comparison with the simulated Oracles. Nevertheless, the results of both LogMap and LogMap+$\texttt{Or}^{LLM}_{GF2.5}$ are highly competitive when compared with the state-of-the-art systems participating in the OAEI campaign (see results in the OAEI 2021 [30] for *largebio*, and in the OAEI 2024 [31] for *anatomy* and *bio-ml*). For example, LogMap+$\texttt{Or}^{LLM}_{GF2.5}$ would have ranked top-3 in the 2024 *anatomy* track, top-2 in the 2021 *largebio* track, and performance comparable to leading systems such as BertMap [14], Matcha [11], and LogMap-Bio [6] in the 2024 *bio-ml* track.

Determinisms of the LLM-based Oracles. The reliability of systems built on LLMs is a critical concern. Thus, we assessed the variability in the performance of the LLM-based Oracle across multiple independent runs, as well as the influence of the system prompt/message (detailed in Section 4.1). In this experiment, we used the Gemini Flash 2.0 and Flash 2.0 Lite models, applying all six prompt templates across three ontology matching tasks. Performance variation

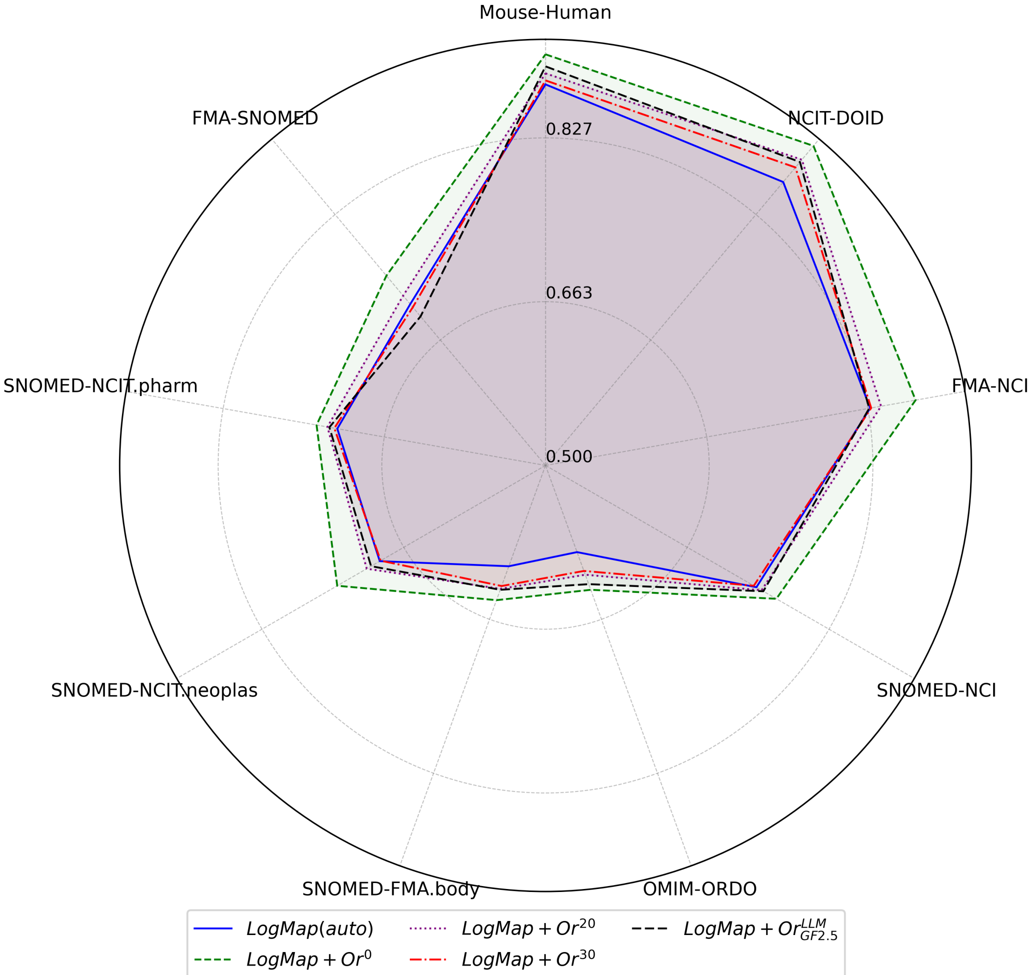

Figure 6: Comparison of LogMap, LogMap with $\texttt{Or}^{LLM}_{GF2.5}$, and LogMap in combination with Oracles with different error rates ($\texttt{Or}^0$, $\texttt{Or}^{20}$, and $\texttt{Or}^{30}$).

over four separate runs was negligible (*i.e.*, the observed standard deviation for the YI ranged from 0.001 to 0.005). While system prompts did not lead to significant changes, they had a modest impact on average performance, suggesting that framing the context for the LLM can offer some benefit.

Statistical analysis on the overall matching task. We conducted t-test and Wilcoxon statistical tests to analyse whether the performance differences reported in Table 3 and Figure 6, among LogMap, LogMap with $\texttt{Or}^{LLM}_{GF2.0}$ and $\texttt{Or}^{LLM}_{GF2.5}$, and LogMap in combination with Oracles with different error rates ($\texttt{Or}^0$, $\texttt{Or}^{20}$, and $\texttt{Or}^{30}$), were significant (p-value < 0.05). Table 4 confirms that LogMap+$\texttt{Or}^0$ and LogMap+$\texttt{Or}^{10}$ lead to significantly better results than both LogMap+$\texttt{Or}^{LLM}_{GF2.0}$ and LogMap+$\texttt{Or}^{LLM}_{GF2.5}$ (*i.e.*, $p < 0.01$ in the "less" and "two-sided" settings with both t-test and Wilcoxon). The comparison of LogMap+$\texttt{Or}^{LLM}_{GF2.0}$ and LogMap+$\texttt{Or}^{LLM}_{GF2.5}$ with LogMap+$\texttt{Or}^{LLM}_{GF2.0}$ and LogMap+$\texttt{Or}^{20}$ yields no significant differences ($p > 0.1$ in the two-sided setting), while the comparison with LogMap+$\texttt{Or}^{30}$ shows significance in the "greater" direction for both tests ($p < 0.05$). LogMap+$\texttt{Or}^{LLM}_{GF2.0}$ and LogMap+$\texttt{Or}^{LLM}_{GF2.5}$ also significantly improve LogMap in automatic mode (*i.e.*, $p < 0.05$ in the 'greater' direction). Overall, the statistical tests support the results and the discussion presented above.

System Comparison	t_test			Wilcoxon		
	greater	less	2-sided	greater	less	2-sided
LogMap+Or$^{LLM}_{GF2.0}$ vs LogMap+Or^{0}	0.9999	0.0001	0.0002	1.0	0.002	0.0039
LogMap+Or$^{LLM}_{GF2.0}$ vs LogMap+Or^{10}	0.9997	0.0003	0.0006	1.0	0.002	0.0039
LogMap+Or$^{LLM}_{GF2.0}$ vs LogMap+Or^{20}	0.9176	0.0824	0.1648	0.8984	0.125	0.25
LogMap+Or$^{LLM}_{GF2.0}$ vs LogMap+Or^{30}	0.0301	0.9699	0.0602	0.0371	0.9727	0.0742
LogMap+Or$^{LLM}_{GF2.0}$ vs LogMap	0.0007	0.9993	0.0014	0.002	1.0	0.0039
LogMap+Or$^{LLM}_{GF2.5}$ vs LogMap+Or^{0}	0.9984	0.0016	0.0033	1.0	0.002	0.0039
LogMap+Or$^{LLM}_{GF2.5}$ vs LogMap+Or^{10}	0.9909	0.0091	0.0183	0.998	0.0039	0.0078
LogMap+Or$^{LLM}_{GF2.5}$ vs LogMap+Or^{20}	0.797	0.203	0.406	0.7871	0.248	0.4961
LogMap+Or$^{LLM}_{GF2.5}$ vs LogMap+Or^{30}	0.0372	0.9628	0.0744	0.0488	0.9629	0.0977
LogMap+Or$^{LLM}_{GF2.5}$ vs LogMap	0.0203	0.9797	0.0405	0.0273	0.9805	0.0547

Table 4: Statistical test results comparing LogMap, LogMap with Or$^{LLM}_{GF2.0}$ and Or$^{LLM}_{GF2.5}$, and LogMap in combination with Oracles with different error rates (Or^{0}, Or^{20}, and Or^{30}). Values represent p-values for t-test and Wilcoxon signed-rank test in *greater*, *less*, and *two-sided* settings.

6 Conclusions and future work

The integration and understanding of the power of state-of-the-art LLMs within ontology alignment tasks is still at an early stage. Although the literature has shown promising results, there are still open challenges concerning performance, costs, and the sustainable use of LLMs. In this paper, we have explored the feasibility of integrating an LLM-based Oracle with the state-of-the-art system LogMap, such that the Oracle is only called for a very specific subset of mapping where LogMap is uncertain. To the best of our knowledge, although LLMs are increasingly being used within ontology alignment pipelines, the use of LLMs as Oracles has not been explored in the literature. We have provided an extensive evaluation of LLM-based Oracles as a diagnostic engine, as well as in combination with LogMap on an end-to-end ontology matching task. The obtained results are positive, improving the performance of LogMap; however, we have also shown that the results are far from a perfect Oracle for the tested LLM models and prompt templates.

Future avenues. We foresee several promising directions for future work. One key avenue is to extend the contextual information of the prompts, leveraging additional ontological relationships. Exploring a broader range of prompt formulations and LLM models could also provide deeper insights into the limitations of using LLMs as Oracles. Given the observed variation in performance across different prompts and models, combining multiple LLM-based Oracles through ensemble methods could result in more reliable outcomes and enhanced performance. We also plan to explore few-shot prompts in future work, particularly in tracks like Bio-ML, where a subset of mappings can be used for training. Incorporating retrieval-augmented generation(RAG) may also enable systems to dynamically access relevant background information from ontologies or external resources like BioPortal [28], leading to more informed and accurate decisions, especially in complex cases, as the ones we were dealing with in this paper.

Potential training data leakage. It is important to note that LLMs may have been exposed to existing OAEI benchmarks during pre-training, which could artificially boost their reported accuracy. To support a fair and unbiased evaluation of the new generation of ontology alignment systems relying on LLMs, the ontology matching community should prioritize the creation of new tasks with truly hidden (blind) reference alignments as discussed during the ISWC 2024 special session on *Harmonising Generative AI and Semantic Web Technologies: Opportunities, challenges, and benchmarks* [1].

Resource constraints. Although our selected LLMs strike a balance between cost and quality, financial and infrastructure constraints still pose challenges for widespread adoption of LLM-based Oracles, especially in large-scale or time-sensitive applications. Additionally, commercial model usage often involves rate limits and API changes, which could affect system stability in long-term deployments.

Evaluation scope. Our experiments focused on OAEI datasets within three established tracks. While these cover a diverse set of biomedical domains and alignment challenges, additional evaluation on new or less curated datasets would be necessary to fully understand the robustness and limitations of the LLM-based Oracle approach in unconstrained environments.

Use of proprietary vs. open-source models. Our evaluation focuses on commercial LLMs due to their strong performance, API stability, and ease of integration. However, this choice limits reproducibility and accessibility for users or institutions that prefer or require open-source alternatives. We initially tested open-weight models like Mistral and Llama (Meta), but they showed poorer performance and slower response times. GPT and Gemini models, accessed via APIs, performed better and remained cost-effective. Currently, top-performing models are proprietary, though we see a positive trend with open-source efforts (*e.g.*, OLMo by AllenAI) and open-weight models like Llama. In the near future, we plan to expand our study to include a broader range of both open-source and open-weight LLMs.

Ethical Consideration

We used AI-based tools solely for grammar checking and minor language revisions during the writing of this paper. No content generation, idea development, or substantive rewriting was performed by AI systems. All research design, experimentation, analysis, and writing were conducted by the authors. SL, DS, and SS designed and conducted the evaluation. EJR supervised and designed the experiments. All authors contributed to the proposed solution. All authors contributed to the writing and approved the final manuscript.

Acknowledgement

This research was supported in part by the NYU Center for Responsible AI, and was conducted as part of its RAI for Ukraine program. This work was organized by The Center for Responsible AI at the Tandon School of Engineering at New York University (NYU), with support from the Faculty of Applied Sciences of the Ukrainian Catholic University (UCU) in collaboration with a supervisor from City St George's, University of London. The Ukrainian students conducting the research would like to express their gratitude to their mentor, Ernesto Jimenez-Ruiz, for his guidance throughout the project, as well as to the organizers of the RAI research program.

References

[1] Alharbi, R., de Berardinis, J., Groth, P., Meroño-Peñuela, A., Simperl, E., Tamma, V.: Harmonising Generative AI and Semantic Web Technologies (HGAIS 2024). In: Special Session co-located with the 23rd International Semantic Web Conference (ISWC 2024). CEUR Workshop Proceedings, vol. 3953. CEUR-WS.org (2024), https://ceur-ws.org/Vol-3953/

[2] Amini, R., Norouzi, S.S., Hitzler, P., Amini, R.: Towards Complex Ontology Alignment Using Large Language Models. In: Knowledge Graphs and Semantic Web - 6th International Conference, KGSWC 2024, Paris, France, December 11-13, 2024, Proceedings. Lecture Notes in Computer Science, vol. 15459, pp. 17–31. Springer (2024), https://doi.org/10.1007/978-3-031-81221-7_2

[3] Bodenreider, O.: The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic acids research (2004)

[4] Chen, J., Hu, P., Jiménez-Ruiz, E., Holter, O.M., Antonyrajah, D., Horrocks, I.: OWL2vec*: Embedding of OWL ontologies. Machine Learning **110**(7), 1813–1845 (2021)

[5] Chen, J., Jiménez-Ruiz, E., Horrocks, I., Antonyrajah, D., Hadian, A., Lee, J.: Augmenting Ontology Alignment by Semantic Embedding and Distant Supervision. In: The Semantic Web - 18th International Conference, ESWC 2021, Virtual Event, June 6-10, 2021, Proceedings. Lecture Notes in Computer Science, vol. 12731, pp. 392–408. Springer (2021), https://doi.org/10.1007/978-3-030-77385-4_23

[6] Chen, X., Xia, W., Jiménez-Ruiz, E., Cross, V.V.: Extending an ontology alignment system with BioPortal: a preliminary analysis. In: Proceedings of the ISWC 2014 Posters & Demonstrations Track a track within the 13th International Semantic Web Conference. CEUR Workshop Proceedings, vol. 1272, pp. 313–316 (2014)

[7] Dernbach, S., Agarwal, K., Zuniga, A., Henry, M., Choudhury, S.: GLaM: Fine-Tuning Large Language Models for Domain Knowledge Graph Alignment via Neighborhood Partitioning and Generative Subgraph Encoding. In: Proceedings of the AAAI 2024 Spring Symposium Series, Stanford, CA, USA, March 25-27, 2024. pp. 82–89. AAAI Press (2024), https://doi.org/10.1609/aaaiss.v3i1.31186

[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), 56:1–56:28 (2017), https://doi.org/10.1186/s13326-017-0166-5

[9] Euzenat, J., Shvaiko, P.: Ontology Matching, Second Edition. Springer (2013)

[10] Faria, D., Santos, E., Balasubramani, B.S., Silva, M.C., Couto, F.M., Pesquita, C.: AgreementMakerLight. Semantic Web **16**(2), SW–233304 (2025)

[11] Faria, D., Silva, M.C., Cotovio, P., Ferraz, L., Balbi, L., Pesquita, C.: Results for Matcha and Matcha-DL in OAEI 2023. In: 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, pp. 164–169. CEUR-WS.org (2023), https://ceur-ws.org/Vol-3591/oaei23_paper6.pdf

[12] Giglou, H.B., D'Souza, J., Engel, F., Auer, S.: LLMs4OM: Matching Ontologies with Large Language Models. CoRR **abs/2404.10317** (2024), https://doi.org/10.48550/arXiv.2404.10317

[13] Hao, Z., Mayer, W., Xia, J., Li, G., Qin, L., Feng, Z.: Ontology alignment with semantic and structural embeddings. J. Web Semant. **78**, 100798 (2023), https://doi.org/10.1016/j.websem.2023.100798

[14] He, Y., Chen, J., Antonyrajah, D., Horrocks, I.: BERTMap: A BERT-Based Ontology Alignment System. In: Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022. pp. 5684–5691. AAAI Press (2022), https://doi.org/10.1609/aaai.v36i5.20510

[15] He, Y., Chen, J., Dong, H., Horrocks, I.: Exploring large language models for ontology alignment. In: Proceedings of the ISWC 2023 Posters, Demos and Industry Tracks: From Novel Ideas to Industrial Practice co-located with 22nd International Semantic Web Conference (ISWC 2023), Athens, Greece, November 6-10, 2023. CEUR Workshop Proceedings, vol. 3632. CEUR-WS.org (2023), https://ceur-ws.org/Vol-3632/ISWC2023_paper_427.pdf

[16] 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: 21st International Semantic Web Conference. Lecture Notes in Computer Science, vol. 13489, pp. 575–591. Springer (2022), https://doi.org/10.1007/978-3-031-19433-7_33

[17] Hertling, S., Paulheim, H.: OLaLa: Ontology Matching with Large Language Models. In: Proceedings of the 12th Knowledge Capture Conference 2023, K-CAP 2023, Pensacola, FL, USA, December 5-7, 2023. pp. 131–139. ACM (2023), https://doi.org/10.1145/3587259.3627571

[18] Hogan, A., Blomqvist, E., Cochez, M., d'Amato, C., de Melo, G., Gutierrez, C., Kirrane, S., Gayo, J.E.L., Navigli, R., Neumaier, S., Ngomo, A.N., Polleres, A., Rashid, S.M., Rula, A., Schmelzeisen, L., Sequeda, J.F., Staab, S., Zimmermann, A.: Knowledge graphs. ACM Comput. Surv. **54**(4), 71:1–71:37 (2022), https://doi.org/10.1145/3447772

[19] Iyer, V., Agarwal, A., Kumar, H.: VeeAlign: Multifaceted Context Representation Using Dual Attention for Ontology Alignment. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021. pp. 10780–10792. Association for Computational Linguistics (2021), https://doi.org/10.18653/v1/2021.emnlp-main.842

[20] Jiménez-Ruiz, E., Cuenca Grau, B., Zhou, Y., Horrocks, I.: Large-scale interactive ontology matching: Algorithms and implementation. In: ECAI 2012 - 20th European Conference on Artificial Intelligence. Including Prestigious Applications of Artificial Intelligence (PAIS-2012) System Demonstrations Track, Montpellier, France, August 27-31 , 2012. Frontiers in Artificial Intelligence and Applications, vol. 242, pp. 444–449. IOS Press (2012), https://doi.org/10.3233/978-1-61499-098-7-444

[21] Jiménez-Ruiz, E., Grau, B.C.: 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. Lecture Notes in Computer Science, vol. 7031, pp. 273–288. Springer (2011), https://doi.org/10.1007/978-3-642-25073-6_18

[22] Jiménez-Ruiz, E., Grau, B.C., Horrocks, I.: Exploiting the UMLS metathesaurus in the ontology alignment evaluation initiative. In: Proceedings of the 2nd International Workshop on Exploiting Large Knowledge Repositories, Castellón de la Plana, Spain, September 7, 2012. CEUR Workshop Proceedings, vol. 882. CEUR-WS.org (2012), https://ceur-ws.org/Vol-882/elkr_atsf_2012_paper9.pdf

[23] Jiménez-Ruiz, E., Grau, B.C., Horrocks, I., Llavori, R.B.: Logic-based assessment of the compatibility of UMLS ontology sources. J. Biomed. Semant. **2**(S-1), S2 (2011), http://www.jbiomedsem.com/content/2/S1/S2

[24] Kolyvakis, P., Kalousis, A., Kiritsis, D.: DeepAlignment: Unsupervised ontology matching with refined word vectors. In: Proceedings of NAACL. pp. 787–798 (2018)

[25] Li, H., Dragisic, Z., Faria, D., Ivanova, V., Jiménez-Ruiz, E., Lambrix, P., Pesquita, C.: User validation in ontology alignment: functional assessment and impact. Knowl. Eng. Rev. **34**, e15 (2019), https://doi.org/10.1017/S0269888919000080

[26] Nkisi-Orji, I., Wiratunga, N., Massie, S., Hui, K.Y., Heaven, R.: Ontology alignment based on word embedding and random forest classification. In: ECML-PKDD. pp. 557–572. Springer (2018)

[27] Norouzi, S.S., Mahdavinejad, M.S., Hitzler, P.: Conversational ontology alignment with ChatGPT. In: 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, pp. 61–66. CEUR-WS.org (2023), https://ceur-ws.org/Vol-3591/om2023_STpaper1.pdf

[28] Noy, N.F., Shah, N.H., Whetzel, P.L., Dai, B., Dorf, M., Griffith, N., Jonquet, C., Rubin, D.L., Storey, M.D., Chute, C.G., Musen, M.A.: BioPortal: ontologies and integrated data resources at the click of a mouse. Nucleic Acids Res. **37**(Web-Server-Issue), 170–173 (2009), https://doi.org/10.1093/nar/gkp440

[29] Oulefki, S., Berkani, L., Bellatreche, L., Boudjenah, N., Mokhtari, A.: Results for BioGITOM in OAEI 2024. In: Proceedings of the 19th International Workshop on Ontology Matching co-located with the 23rd International Semantic Web Conference (ISWC 2024), Baltimore, USA, November 11, 2024. CEUR Workshop Proceedings, vol. 3897, pp. 104–109. CEUR-WS.org (2024), https://ceur-ws.org/Vol-3897/oaei2024_paper2.pdf

[30] Pour, M.A.N., Algergawy, A., Amardeilh, F., Amini, R., Fallatah, O., Faria, D., Fundulaki, I., Harrow, I., Hertling, S., Hitzler, P., Huschka, M., Ibanescu, L., Jiménez-Ruiz, E., Karam, N., Laadhar, A., Lambrix, P., Li, H., Li, Y., Michel, F., Nasr, E., Paulheim, H., Pesquita, C., Portisch, J., Roussey, C., Saveta, T., Shvaiko, P., Splendiani, A., Trojahn, C., Vatascinová, J., Yaman, B., Zamazal, O., Zhou, L.: Results of the Ontology Alignment Evaluation Initiative 2021. In: Proceedings of the 16th International Workshop on Ontology Matching co-located with the 20th International Semantic Web Conference (ISWC 2021), Virtual conference, October 25, 2021. CEUR Workshop Proceedings, vol. 3063, pp. 62–108. CEUR-WS.org (2021), https://ceur-ws.org/Vol-3063/oaei21_paper0.pdf

[31] Pour, M.A.N., Algergawy, A., Blomqvist, E., Buche, P., Chen, J., Coulet, A., Cufi, J., Dong, H., Faria, D., Ferraz, L., Cotovio, P.G., He, Y., Hertling, S., Horrocks, I., Ibanescu, L., Jain, S., Jiménez-Ruiz, E., Karam, N., Kraus, F., Lambrix, P., Li, H., Li, Y., Monnin, P., Paulheim, H., Pesquita, C., Sharma, A., Shvaiko, P., Silva, M.C., Sousa, G., Trojahn, C., Vatascinová, J., Yaman, B., Zamazal, O., Zhou, L.: Results of the Ontology Alignment Evaluation Initiative 2024. In: Proceedings of the 19th International Workshop on Ontology Matching co-located with the 23rd International Semantic Web Conference (ISWC 2024), Baltimore, USA, November 11, 2024. CEUR Workshop Proceedings, vol. 3897, pp. 64–97. CEUR-WS.org (2024), https://ceur-ws.org/Vol-3897/oaei2024_paper0.pdf

[32] Qiang, Z., Wang, W., Taylor, K.: Agent-OM: Leveraging Large Language Models for Ontology Matching. VLDB EndowmentVol **V. 18, No. 3** (2024)

[33] Shefchek, K.A., et al.: The Monarch Initiative in 2019: an integrative data and analytic platform connecting phenotypes to genotypes across species. Nucleic Acids Research (2020)

[34] Shvaiko, P., Euzenat, J.: Ontology Matching: State of the Art and Future Challenges. IEEE Trans. Knowl. Data Eng. **25**(1), 158–176 (2013), https://doi.org/10.1109/TKDE.2011.253

[35] Taboada, M., Martínez, D., Arideh, M., Mosquera, R.: Ontology Matching with Large Language Models and Prioritized Depth-First Search. CoRR **abs/2501.11441** (2025), https://doi.org/10.48550/arXiv.2501.11441

[36] Totoian, M., Marginean, A., Blohm, P., Hussain, M.N.: HybridOM: Ontology Matching using Hybrid Search. In: Proceedings of the 19th International Workshop on Ontology Matching co-located with the 23rd International Semantic Web Conference (ISWC 2024), Baltimore, USA, November 11, 2024. CEUR Workshop Proceedings, vol. 3897, pp. 138–145. CEUR-WS.org (2024), https://ceur-ws.org/Vol-3897/oaei2024_paper7.pdf

[37] Youden, W.J.: Index for rating diagnostic tests. Cancer **3**(1), 32–35 (1950)

[38] Zhang, R., Su, Y., Trisedya, B.D., Zhao, X., Yang, M., Cheng, H., Qi, J.: AutoAlign: Fully Automatic and Effective Knowledge Graph Alignment Enabled by Large Language Models. IEEE Trans. Knowl. Data Eng. **36**(6), 2357–2371 (2024), https://doi.org/10.1109/TKDE.2023.3325484