Large Language Models as Oracles for Ontology Alignment

Anonymous ACL submission

Abstract

There are many methods and systems to tackle the ontology alignment problem, yet a major challenge persists in producing high-quality mappings among a set of input ontologies. Adopting a human-in-the-loop approach during the alignment process has become essential in applications requiring very accurate mappings. However, user involvement is expensive when dealing with large ontologies. In this paper, we evaluate the feasibility of using Large Language Models (LLM) to aid the ontology alignment problem. The use of LLMs is focused only on the validation of a subset of correspondences where an ontology alignment system is very uncertain. We have conducted an extensive analysis over several tasks of the Ontology Alignment Evaluation Initiative (OAEI), reporting in this paper the performance of several state-of-the-art LLMs using different ontology-driven prompt templates. LLM efficacy is also assessed against simulated Oracles with varying error rates.

1 Introduction

Ontology alignment (Euzenat and Shvaiko, 2013) plays a crucial role in integrating diverse data sources across domains. While numerous ontology matching systems exist (*e.g.*, (Pour et al., 2024)), systems capable of producing high-quality correspondences among the input ontologies are still needed, especially in applications where high confidence 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 potential to be used within an ontology alignment pipeline (*e.g.*, (Qiang et al., 2024)). Nevertheless, LLMs are also computationally or financially costly, and an unlimited use may not be possible.

In this paper, we have extended the state-of-the-art system `LogMap` (Jiménez-Ruiz et al., 2012a) 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 LLM is invoked only for complex cases where traditional alignment techniques may be insufficient. 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. We selected the GPT-4o Mini model (OpenAI) and a range of Google Gemini Flash models for our experiments, due to their good balance between performance and cost.

To analyse the suitability of the LLM-based Oracles, we have conducted an extensive evaluation with the *anatomy* (Dragisic et al., 2017), *largebio* (Jiménez-Ruiz et al., 2012b), and *bio-ml* (He et al., 2022b) datasets of the OAEI campaign (Pour et al., 2021, 2024), involving a total of nine matching tasks. These datasets are complex and have become a reference in the research 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 (Li et al., 2019). We also report experiments for the *anatomy* dataset with the open-weight models Mistral, Llama and Qwen, exploring their suitability as Oracles.

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 compu-

tational 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.

2 Preliminaries

Next, we introduce the background underpinning our methods and experiments.

Ontology alignment. Ontology alignment is the process of finding correspondences or *mappings* among the entities[1] 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 number between 0 and 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 $\mathcal{O}_1$ (source) and $\mathcal{O}_2$ (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` (Jiménez-Ruiz et al., 2012a) as the baseline for our experiments due to its flexibility to be adapted to different evaluation scenarios. `LogMap` can operate in a fully automatic mode or allow interaction with an *Oracle* (Li et al., 2019). `LogMap`, during the mapping selection stage, identifies a subset of mappings $\mathcal{M}_{ask}$ 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 $\mathcal{M}_{ask}$. Appendix A provides additional information about `LogMap`, including the workflow it follows when allowing interaction.

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 (Li et al., 2019).

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 $\mathcal{M}_{RA}$:

$$Pr = \frac{|\mathcal{M}_S \cap \mathcal{M}_{RA}|}{|\mathcal{M}_S|}, \; Re = \frac{|\mathcal{M}_S \cap \mathcal{M}_{RA}|}{|\mathcal{M}_{RA}|},$$

$$F = 2 \cdot \frac{Pr \cdot Re}{Pr + Re}$$

We use *Sensitivity* (Se), *Specificity* (Sp), and *Youden's index* (YI) (Youden, 1950) to evaluate the effectiveness of an Oracle to diagnose positive (sensitivity) and negative (specificity) mappings in $\mathcal{M}_{ask}$ with respect to $\mathcal{M}_{RA}$.

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

TP, FN, TN, and FP stand for true positives, false negatives, true negatives, and false positives.

LLM prompting. LLMs like GPT-4 are pre-trained 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 – "Does 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

The Ontology Alignment Evaluation Initiative (OAEI) has driven progress since 2004 by providing standardised benchmarks and evaluation protocols for matching systems (Pour et al., 2024). Widely used (traditional) matchers include `LogMap` (Jiménez-Ruiz et al., 2012a) and AgreementMakerLight (AML) (Faria et al., 2025), each leveraging different combinations of lexical, structural, and background-knowledge techniques. Human validation has long been recognised as critical for high-precision mappings. Early frameworks combined automated matching with domain

[1] Ontology classes, properties or instances.

expert feedback to resolve low-confidence correspondences, but at the cost of extensive user effort and time (Li et al., 2019).

In recent years, a new generation of systems leveraging machine learning and (large) language models has emerged. The OAEI Bio-ML track (He et al., 2022b) 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.*, (Kolyvakis et al., 2018; Nkisi-Orji et al., 2018; Iyer et al., 2021)). Knowledge graph embedding systems like OWL2Vec (Chen et al., 2021a) were also leveraged in combination with a machine learning architecture to learn and validate ontology alignments (*e.g.*, (Chen et al., 2021b; Hao et al., 2023)). 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 (He et al., 2022a), BioGITOM (Oulefki et al., 2024), and the Matcha family (Faria et al., 2023). Nevertheless, recent developments in the field are increasingly driven by approaches based on Large Language Models. Saki Norouzi et al. (Norouzi et al., 2023) and He et al. (He et al., 2023) performed exploratory studies about the potential of LLMs to the ontology alignment task, while Amini et al. (Amini et al., 2024) extended the exploration to discover complex alignments beyond equivalence or subsumption. Systems like OLaLa (Hertling and Paulheim, 2023), LLMs4OM (Giglou et al., 2024), MILA (Taboada et al., 2025), Agent-OM (Qiang et al., 2024), and HybridOM (Totoian et al., 2024) 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.*, (Hertling and Paulheim, 2023; Giglou et al., 2024; Taboada et al., 2025)). 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 (Qiang et al., 2024). Recent work has explored the use of LLMs to focus on the alignment of instance data within the knowledge graphs (*e.g.*, (Zhang et al., 2024; Dernbach et al., 2024)).

Our approach builds upon `LogMap` (Jiménez-Ruiz et al., 2012a) 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 (Taboada et al., 2025) 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 a 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. 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 1. 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.

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

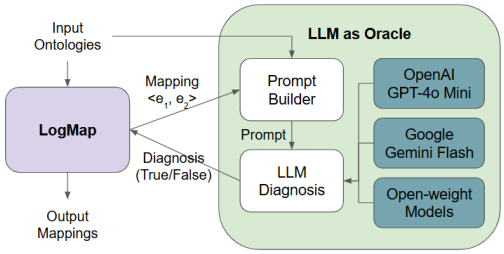

Figure 1: LLM-in-the-loop as Oracle.

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 (Jiménez-Ruiz et al., 2011). 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 hypothesise 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. Below, we show the populated prompts for the mapping ⟨`mouse:MA_0001771` (alveolus epithelium), `human:NCI_C12867` (Alveolar_Epithelium) ⟩.

Structured prompts. This type of prompt uses structured information with less common natural language expressions. Listing 1 shows the prompt $\mathbf{P}$ where the entities to be aligned and their context are provided as a list of elements. Listing 4 in Appendix B shows the structured prompt with extended context $\mathbf{P}_{\text{EC}}$.

```
Analyse 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 $\mathbf{P}$.

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. Listing 2 shows this type of prompt ($\mathbf{P}^{\text{NLF}}$), while Listing 5 in Appendix B includes the version with extended context $\mathbf{P}^{\text{NLF}}_{\text{EC}}$.

```
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 2: $\mathbf{P}^{\text{NLF}}$ Prompt (natural-language friendly).

NLF prompts with synonyms. Although LLMs may inherently encode synonyms and lexical variations related to the ontology entities, the $\mathbf{P}^{\text{NLF}}_{\text{S}}$ and $\mathbf{P}^{\text{NLF}}_{\text{EC+S}}$ prompts are designed to analyse the impact of explicitly including synonyms for both the entities in a given correspondence and their associated context. The $\mathbf{P}^{\text{NLF}}_{\text{S}}$ prompt is given in Listing 3, while Listing 6 in Appendix B provides its variant with extended context ($\mathbf{P}^{\text{NLF}}_{\text{EC+S}}$).

```
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 3: $\mathbf{P}^{\text{NLF}}_{\text{S}}$ Prompt (natural-language friendly with synonyms).

System prompts. In addition to mapping templates, it is possible to add, for each LLM session, a short 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 (*base*); *(ii)* to explain its decision in a natural-language friendly manner (*explainable*); *(iii)* emphasizing the use of hierarchical and semantic context (*hierarchical*); and *(iv)* to leverage explicitly provided synonyms and parent-class semantics (*lexical*). Listing 7 in Appendix B includes the specific system prompts.

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.

In order to achieve binary (True/False) diagnostic classification, we 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 is, if the output is not parsed correctly (e.g., neither True nor False), we resend the same request.

We used the Chat Completions API (OpenAI, 2025a) for GPT-4o Mini, and the OpenAI's SDK endpoint for the Gemini Models (Gemini, 2025c). Response latency typically ranged between 1 and 3 seconds, depending on prompt complexity and model characteristics. This enabled high-throughput querying, especially when requests were 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 (Gemini, 2025a), whereas OpenAI's API begins with a limit of 500 RPM and a daily quota of 10,000 requests (OpenAI, 2025b), 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 for Gemini Flash 1.5 and 2.0 Lite, up to \$0.15 for GPT-4o Mini and Gemini Flash 2.5 Preview (OpenAI, 2025b; Gemini, 2025b). This results in an average cost per 1,000 requests ranging from approximately \$0.01 to \$0.04.

Open-weight models. Our end-to-end evaluation focuses on commercial LLMs due to their ease of integration via their APIs and cost-effective performance. This choice, however, may limit reproducibility and accessibility for users or institutions that prefer or require open-weight alternatives. Hence, we also performed a preliminary evaluation with the open-weight models Mistral, Llama (Meta), and Qwen (Alibaba Cloud). Mistral Small (2402, approx. 24b) and Llama 3-70b Instruct were accessed via the Amazon Bedrock API (Bedrock, 2025b,a). Mistral Small typically responded to mapping requests in around 2 seconds, while Llama 3-70B took between 5 and 8 seconds. The average cost per 1,000 requests was \$0.15 for Mistral Small-2402 and \$0.40 for Llama3-70b. Qwen3 models (1.7b and 8b) (Yang et al., 2025) were run locally on a standard laptop equipped with an integrated M2 GPU. Due to the limitations of the local setup, the average latency per request exceeded 125 seconds.

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 (Li et al., 2019) 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 ($\mathtt{Or}^{LLM}$) against Oracles with error rates ranging from 0% (*i.e.*, perfect Oracle, $\mathtt{Or}^{0}$) to 30% (*i.e.*, $\mathtt{Or}^{30}$). 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 (Li et al., 2019).

5 Evaluation

Our experiments, conducted on a standard laptop, leveraged the selected LLM models as detailed in Section 4.2. We used the *anatomy* (Dragisic et al., 2017), *largebio* (Jiménez-Ruiz et al., 2012b), and *bio-ml* (He et al., 2022b) datasets provided by

OAEI track	Matching task	$\lvert\mathcal{O}_1\rvert$	$\lvert\mathcal{O}_2\rvert$	$\lvert\mathcal{M}_{RA}\rvert$
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	79,049	122,521	9,008
	SNOMED-NCI	122,521	66,919	18,844

Table 1: Statistics of the used OAEI datasets. Ontology size is given in terms of the number of entities. $\mathcal{M}_{RA}$ is the reference alignment of the matching task.

the OAEI evaluation initiative (Pour et al., 2021, 2024).[2] As shown in Table 1, we covered a total of nine ontology matching tasks, involving ontologies of diverse sizes. Note that these ontologies contain mostly concepts. The reference alignments ($\mathcal{M}_{RA}$) 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 relies on public resources like MONDO (Shefchek et al., 2020) and UMLS (Bodenreider, 2004).

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 2 shows a summary of the correctness (Youden's index, YI) of the LLM-based Oracles for each of the selected LLMs and prompt templates. Detailed results per matching task are provided in Appendix C.

Oracles relying on the Gemini 2.5 Flash model led to the best results on average, as summarised in Figure 2. The best results were achieved by the combination of Gemini 2.5 Flash and $\mathbf{P}_{\mathrm{S}}^{\mathrm{NLF}}$ prompts ($\texttt{Or}^{LLM}_{GF2.5}$). Table 2 compares the performance of `LogMap` (automatic mode) and $\texttt{Or}^{LLM}_{GF2.5}$ 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 $\texttt{Or}^{LLM}_{GF2.5}$ achieves significantly better results, with an average YI value exceeding 0.5.

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. While no standard cut-off value exists for YI, some statistical papers use 0.3 / 0.5 / 0.7 as representative values for low / moderate / high. Due to the complexity of the mappings in $\mathcal{M}_{ask}$, moderate YI values are to be expected.

[2]The *largebio* track was last held in 2021 (Pour et al., 2021), but it will be reinstated in the near future.

Matching task	$\lvert\mathcal{M}_{ask}\rvert$		`LogMap` on $\mathcal{M}_{ask}$			$\texttt{Or}^{LLM}_{GF2.5}$ 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 ($\texttt{Or}^{LLM}_{GF2.5}$, using Gemini 2.5 Flash and $\mathbf{P}_{\mathrm{S}}^{\mathrm{NLF}}$ 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.

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.

Impact of the prompt template. Figure 2 also illustrates the impact of using different prompt templates. The behaviour varies across LLM models. Natural-language friendly prompts result in more consistent behaviour, while incorporating extended context and synonyms has a positive impact on some of the matching tasks. Overall, for the Gemini 2.0 Flash and 2.5 Flash models, the most effective prompts were $\mathbf{P}_{\mathrm{S}}^{\mathrm{NLF}}$ (natural-language friendly with synonyms). It is worth emphasising that different ontologies and matching tasks pose varying challenges due to lexical and structural differences (see Appendix C for an overview of the results per matching task). 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).

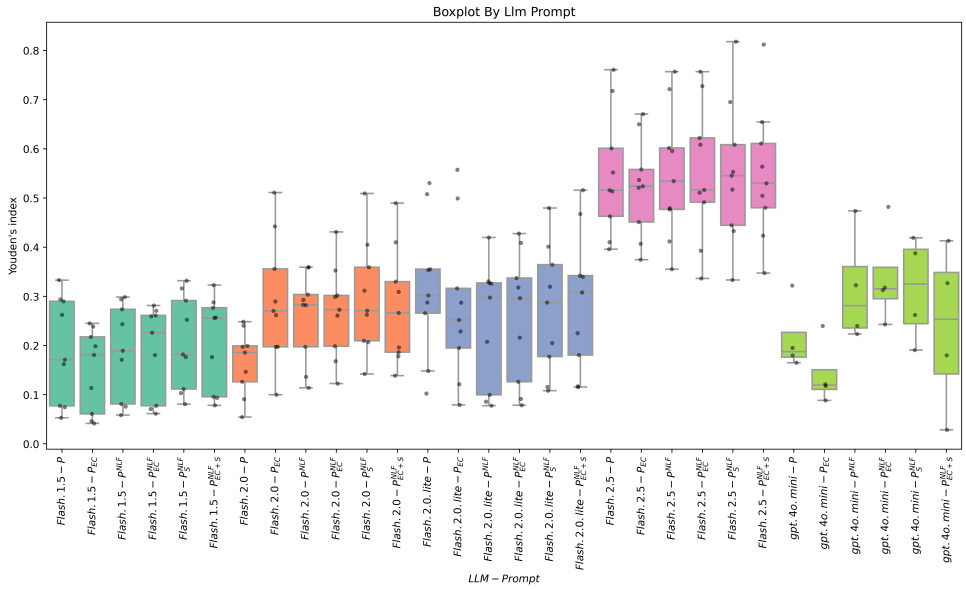

Figure 2: Summary of the diagnostic results (Youden's index) by the LLM-based Oracles.

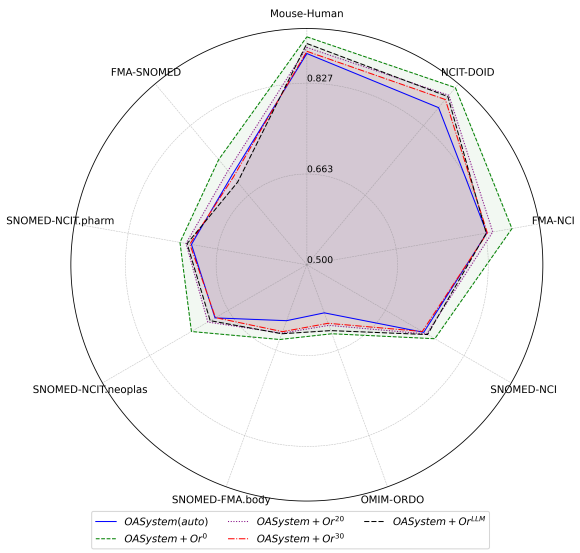

Figure 3: Comparison of LogMap, LogMap with $\mathtt{Or}^{LLM}_{GF2.5}$, and LogMap in combination with Oracles with different error rates ($\mathtt{Or}^{0}$, $\mathtt{Or}^{20}$, and $\mathtt{Or}^{30}$).

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 1). We selected the top-performing LLM-based Oracles using $\mathbf{P}^{\mathrm{NLF}}_{\mathrm{S}}$ prompts: $\mathtt{Or}^{LLM}_{GF2.0}$ (based on Gemini 2.0 Flash) and $\mathtt{Or}^{LLM}_{GF2.5}$ (based on Gemini 2.5 Flash). As anticipated, the integration with the LLM-based Oracle yields improved results across all tasks in terms of F-score. LogMap+$\mathtt{Or}^{LLM}_{GF2.5}$ dominates the *anatomy* and *bio-ml* tasks, while LogMap+$\mathtt{Or}^{LLM}_{GF2.0}$ 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 (Li et al., 2019) as detailed in Section 4.3. Figure 3 compares performance across all nine ontology matching tasks for: LogMap, LogMap+$\mathtt{Or}^{LLM}_{GF2.5}$, and LogMap with the simulated Oracles $\mathtt{Or}^{0}$, $\mathtt{Or}^{20}$, and $\mathtt{Or}^{30}$ (corresponding to error rates of 0%, 20%, and 30%, respectively). We can observe that $\mathtt{Or}^{LLM}_{GF2.5}$ performs similarly to $\mathtt{Or}^{20}$, except in the FMA-SNOMED task (lower F-score) and OMIM-ORDO task (higher F-score). In line with previous studies (Jiménez-Ruiz et al., 2012a; Li et al., 2019), LogMap+$\mathtt{Or}^{30}$ still outperforms LogMap (without Oracle). The statistical analysis in Appendix E supports these observations.

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 LogMap+$\mathtt{Or}^{LLM}_{GF2.5}$ are highly competitive when compared with the state-of-the-art systems participating in the OAEI campaign (see results in the OAEI 2021 (Pour et al., 2021) for *largebio*, and in the OAEI 2024 (Pour et al., 2024) for *anatomy* and *bio-ml*). For example, LogMap+$\mathtt{Or}^{LLM}_{GF2.5}$ would have ranked top-3 in the 2024 *anatomy* track, top-2 in the 2021 *largebio* track, and performance comparable to lead-

ing systems such as BertMap (He et al., 2022a), Matcha (Faria et al., 2023), and LogMap-Bio (Chen et al., 2014) in the 2024 *bio-ml* track. The OAEI 2025 is in progress (OAEI, 2025).

Determinism 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 2.0 Flash and Flash-Lite models, applying all six prompt templates across three matching tasks. Performance variation over four separate runs was negligible (*i.e.*, the observed standard deviation for YI ranged from 0.001 to 0.005, see Appendix D for details). While system prompts did not lead to significant changes, they had a modest impact on average performance, suggesting that framing the LLM context is important (see Appendix D).

Opportunities with Open-weight models. We tested the diagnostic capabilities of Mistral Small-2402, Llama 3-70b, Qwen3-1.7b, and Qwen3-8b. Table 4 shows the results for the *anatomy* matching task using the $\mathbf{P}_{\mathrm{S}}^{\mathrm{NLF}}$ prompts. Qwen3-1.7b, the smallest model evaluated, performed as expected with limited success in diagnosing mappings within $\mathcal{M}_{ask}$. In contrast, Qwen3-8b delivered highly competitive results, surpassing several larger commercial and open-weight models in terms of YI index. Appendix F includes results for Qwen3 models for all prompt templates. Mistral Small-2402 also performed strongly, matching the level of Gemini 2.0 Flash and GPT-4o Mini. Despite its size, Llama 3-70b underperformed expectations. These findings highlight the potential of open-weight models—whether accessed via APIs like Amazon Bedrock or deployed on local infrastructure—to serve effectively as Oracles in ontology alignment tasks. Nonetheless, the choice of model involves balancing trade-offs among performance, latency, cost, and access to infrastructure. Running open-weight models via Amazon Bedrock was generally more costly than commercial APIs, while local deployment significantly increased latency due to the limited resources.

LLM Model	Sensitivity	Specificity	Youden's Index
Mistral Small-2401	0.945	0.547	0.492
Llama 3-70b	0.989	0.359	0.348
Qwen3-1.7b	0.811	0.411	0.222
Qwen3-8b	0.764	0.825	**0.590**
Gemini 1.5 Flash	0.994	0.322	0.316
Gemini 2.0 Flash	0.994	0.411	0.405
Gemini 2.0 Flash-Lite	0.976	0.389	0.364
Gemini 2.5 Flash	0.951	0.744	**0.695**
GPT-4o Mini	0.908	0.511	0.419

Table 4: Diagnostic capabilities of open-weight models over $\mathcal{M}_{ask}$ in the *anatomy* task (top). We also show the performance of commercial models (bottom). All models were evaluated with $\mathbf{P}_{\mathrm{S}}^{\mathrm{NLF}}$ prompts.

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 encouraging, improving the performance of `LogMap`; however, we have also shown that the results are far from a perfect Oracle.

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. Retrieval-augmented generation (RAG) may also enable systems to dynamically access relevant background knowledge, such as BioPortal (Noy et al., 2009), leading to more informed and accurate diagnostic capabilities.

7 Limitations

While our approach demonstrates promising results, several limitations merit discussion.

Missing human evaluation. A user study may also provide interesting insights in comparison with the LLM-based Oracles. However, to make the exercise meaningful, we should involve domain experts in the process. In this paper, we have performed a comparison with simulated Oracles with different error rates, simulating a potential behaviour of a (non-perfect) domain expert.

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* (Alharbi et al., 2024). Nevertheless, the conducted experiments are still valid, even under the potential assumption of data leakage, as we are comparing the diagnostic capabilities of state-of-the-art models under the same conditions.

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.

Evaluation scope. Our experiments focused on OAEI datasets within three tracks. While these cover a diverse set of biomedical domains and alignment challenges, additional evaluation on other OAEI datasets would be necessary to fully understand the robustness and limitations of our LLM-based Oracle approach—especially OAEI datasets in different domains and involving the matching of properties and instances.

Focus on equivalence mappings. `LogMap` can output both equivalence and subsumption mappings; indeed, it internally represents equivalence mapping as two subsumption mappings. Nevertheless, in this paper, we focus on the most common type of mapping: equivalence. In the future, LLMs can be leveraged to also discover and validate subsumption mappings and complex alignments (*i.e.*, arbitrary ontology axioms mentioning entities of two or more ontologies).

Evaluation with additional LLMs. Our evaluation relies on a subset of available LLM models selected according to limited financial and computational resources. We plan, however, to extend the evaluation with additional models that meet our resource limitations. We also intend to extend the evaluation to test open-weight models on the end-to-end alignment task.

8 Ethical Consideration

AI tools were used only for grammar and minor language edits in this paper. No content creation, idea development, or substantive rewriting was performed by AI. All research design, experiments, analysis, and writing were carried out by the authors.

Supplemental Material Statement. Source code and relevant resources are available in Zenodo: `https://doi.org/10.5281/zenodo.15394653`

References

Reham Alharbi, Jacopo de Berardinis, Paul Groth, Albert Meroño-Peñuela, Elena Simperl, and Valentina Tamma. 2024. Harmonising Generative AI and Semantic Web Technologies (HGAIS 2024). In *Special Session co-located with the 23rd International Semantic Web Conference (ISWC 2024)*, volume 3953 of *CEUR Workshop Proceedings*. CEUR-WS.org. `https://ceur-ws.org/Vol-3953/`.

Reihaneh Amini, Sanaz Saki Norouzi, Pascal Hitzler, and Reza Amini. 2024. Towards Complex Ontology Alignment Using Large Language Models. In *Knowledge Graphs and Semantic Web - 6th International Conference, KGSWC 2024*, volume 15459 of *Lecture Notes in Computer Science*, pages 17–31. Springer.

Amazon Bedrock. 2025a. Meta's Llama in Amazon Bedrock. `https://aws.amazon.com/bedrock/meta/` (accessed September 2025).

Amazon Bedrock. 2025b. Mistral AI in Amazon Bedrock. `https://aws.amazon.com/bedrock/mistral/` (accessed September 2025).

Olivier Bodenreider. 2004. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic acids research*.

Jiaoyan Chen, Pan Hu, Ernesto Jiménez-Ruiz, Ole Magnus Holter, Denvar Antonyrajah, and Ian Horrocks. 2021a. OWL2vec*: Embedding of OWL ontologies. *Machine Learning*, 110(7):1813–1845.

Jiaoyan Chen, Ernesto Jiménez-Ruiz, Ian Horrocks, Denvar Antonyrajah, Ali Hadian, and Jaehun Lee. 2021b. Augmenting Ontology Alignment by Semantic Embedding and Distant Supervision. In *18th Extended Semantic Web Conference (ESWC)*, volume 12731 of *Lecture Notes in Computer Science*, pages 392–408. Springer.

Xi Chen, Weiguo Xia, Ernesto Jiménez-Ruiz, and Valerie V. Cross. 2014. 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*, volume 1272 of *CEUR Workshop Proceedings*, pages 313–316.

Stefan Dernbach, Khushbu Agarwal, Alejandro Zuniga, Michael Henry, and Sutanay Choudhury. 2024. 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*, pages 82–89. AAAI Press.

Zlatan Dragisic, Valentina Ivanova, Huanyu Li, and Patrick Lambrix. 2017. Experiences from the anatomy track in the ontology alignment evaluation initiative. *J. Biomed. Semant.*, 8(1):56:1–56:28.

Jérôme Euzenat and Pavel Shvaiko. 2013. *Ontology Matching, Second Edition*. Springer.

Daniel Faria, Emanuel Santos, Booma Sowkarthiga Balasubramani, Marta C Silva, Francisco M Couto, and Catia Pesquita. 2025. AgreementMakerLight. *Semantic Web*, 16(2):SW–233304.

Daniel Faria, Marta Contreiras Silva, Pedro Cotovio, Lucas Ferraz, Laura Balbi, and Catia Pesquita. 2023. 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*, volume 3591 of *CEUR Workshop Proceedings*, pages 164–169. CEUR-WS.org.

Gemini. 2025a. Gemini API rate limits. `https://ai.google.dev/gemini-api/docs/rate-limits` (accessed August 2025).

Gemini. 2025b. Gemini APIs pricing. `https://ai.google.dev/gemini-api/docs/pricing` (accessed August 2025).

Gemini. 2025c. OpenAI's SDK endpoint. `https://ai.google.dev/gemini-api/docs/openai` (accessed August 2025).

Hamed Babaei Giglou, Jennifer D'Souza, Felix Engel, and Sören Auer. 2024. LLMs4OM: Matching Ontologies with Large Language Models. *CoRR*, abs/2404.10317.

Zhigang Hao, Wolfgang Mayer, Jingbo Xia, Guoliang Li, Li Qin, and Zaiwen Feng. 2023. Ontology alignment with semantic and structural embeddings. *J. Web Semant.*, 78:100798.

Yuan He, Jiaoyan Chen, Denvar Antonyrajah, and Ian Horrocks. 2022a. BERTMap: A BERT-Based Ontology Alignment System. In *Thirty-Sixth AAAI Conference on Artificial Intelligence*, pages 5684–5691. AAAI Press.

Yuan He, Jiaoyan Chen, Hang Dong, and Ian Horrocks. 2023. 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)*, volume 3632 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Yuan He, Jiaoyan Chen, Hang Dong, Ernesto Jiménez-Ruiz, Ali Hadian, and Ian Horrocks. 2022b. Machine Learning-Friendly Biomedical Datasets for Equivalence and Subsumption Ontology Matching. In *21st International Semantic Web Conference (ISWC)*, volume 13489 of *Lecture Notes in Computer Science*, pages 575–591. Springer.

Sven Hertling and Heiko Paulheim. 2023. OLaLa: Ontology Matching with Large Language Models. In *Proceedings of the 12th Knowledge Capture Conference 2023 (K-CAP)*, pages 131–139. ACM.

Vivek Iyer, Arvind Agarwal, and Harshit Kumar. 2021. VeeAlign: Multifaceted Context Representation Using Dual Attention for Ontology Alignment. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 10780–10792. Association for Computational Linguistics.

Ernesto Jiménez-Ruiz, Bernardo Cuenca Grau, Yujiao Zhou, and Ian Horrocks. 2012a. Large-scale interactive ontology matching: Algorithms and implementation. In *20th European Conference on Artificial Intelligence (ECAI)*, volume 242 of *Frontiers in Artificial Intelligence and Applications*, pages 444–449. IOS Press.

Ernesto Jiménez-Ruiz and Bernardo Cuenca Grau. 2011. Logmap: Logic-based and scalable ontology matching. In *10th International Semantic Web Conference (ISWC)*, volume 7031 of *Lecture Notes in Computer Science*, pages 273–288. Springer.

Ernesto Jiménez-Ruiz, Bernardo Cuenca Grau, and Ian Horrocks. 2012b. Exploiting the UMLS metathesaurus in the ontology alignment evaluation initiative. In *Proceedings of the 2nd International Workshop on Exploiting Large Knowledge Repositories*, volume 882 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Ernesto Jiménez-Ruiz, Bernardo Cuenca Grau, Ian Horrocks, and Rafael Berlanga Llavori. 2011. Logic-based assessment of the compatibility of UMLS ontology sources. *J. Biomed. Semant.*, 2(S-1):S2.

Prodromos Kolyvakis, Alexandros Kalousis, and Dimitris Kiritsis. 2018. DeepAlignment: Unsupervised ontology matching with refined word vectors. In *Proceedings of NAACL*, pages 787–798.

Huanyu Li, Zlatan Dragisic, Daniel Faria, Valentina Ivanova, Ernesto Jiménez-Ruiz, Patrick Lambrix, and Catia Pesquita. 2019. User validation in ontology alignment: functional assessment and impact. *Knowl. Eng. Rev.*, 34:e15.

Ikechukwu Nkisi-Orji, Nirmalie Wiratunga, Stewart Massie, Kit-Ying Hui, and Rachel Heaven. 2018. Ontology alignment based on word embedding and random forest classification. In *ECML-PKDD*, pages 557–572. Springer.

Sanaz Saki Norouzi, Mohammad Saeid Mahdavinejad, and Pascal Hitzler. 2023. Conversational ontology alignment with ChatGPT. In *Proceedings of the 18th International Workshop on Ontology Matching co-located with the 22nd International Semantic Web Conference*, volume 3591 of *CEUR Workshop Proceedings*, pages 61–66. CEUR-WS.org.

Natalya Fridman Noy, Nigam H. Shah, Patricia L. Whetzel, Benjamin Dai, Michael Dorf, Nicholas Griffith, Clément Jonquet, Daniel L. Rubin, Margaret-Anne D. Storey, Christopher G. Chute, and Mark A. Musen. 2009. BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Res.*, 37(Web-Server-Issue):170–173.

OAEI. 2025. Ontology Alignment Evaluation Initiative: 2025 results. https://oaei.ontologymatching.org/2025/results/ (accessed October 2025).

OpenAI. 2025a. Chat Completions API. https://platform.openai.com/docs/guides/text?api-mode=chat (accessed August 2025).

OpenAI. 2025b. OpenAI's GPT-4o mini API limits and pricing. https://platform.openai.com/docs/models/gpt-4o-mini (accessed August 2025).

Samira Oulefki, Lamia Berkani, Ladjel Bellatreche, Nassim Boudjenah, and Aïcha Mokhtari. 2024. 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)*, volume 3897 of *CEUR Workshop Proceedings*, pages 104–109. CEUR-WS.org.

Mina Abd Nikooie Pour, Alsayed Algergawy, Florence Amardeilh, Reihaneh Amini, Omaima Fallatah, Daniel Faria, Irini Fundulaki, Ian Harrow, Sven Hertling, Pascal Hitzler, Martin Huschka, Liliana Ibanescu, Ernesto Jiménez-Ruiz, Naouel Karam, Amir Laadhar, Patrick Lambrix, Huanyu Li, Ying Li, Franck Michel, and 13 others. 2021. 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)*, volume 3063, pages 62–108. CEUR-WS.org.

Mina Abd Nikooie Pour, Alsayed Algergawy, Eva Blomqvist, Patrice Buche, Jiaoyan Chen, Adrien Coulet, Julien Cufi, Hang Dong, Daniel Faria, Lucas Ferraz, Pedro Giesteira Cotovio, Yuan He, Sven Hertling, Ian Horrocks, Liliana Ibanescu, Sarika Jain, Ernesto Jiménez-Ruiz, Naouel Karam, Felix Kraus, and 15 others. 2024. 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*, volume 3897 of *CEUR Workshop Proceedings*, pages 64–97. CEUR-WS.org.

Zhangcheng Qiang, Weiqing Wang, and Kerry Taylor. 2024. Agent-OM: Leveraging Large Language Models for Ontology Matching. *VLDB EndowmentVol*, V. 18, No. 3.

Kent A. Shefchek and 1 others. 2020. The Monarch Initiative in 2019: an integrative data and analytic platform connecting phenotypes to genotypes across species. *Nucleic Acids Research*.

Maria Taboada, Diego Martínez, Mohammed Arideh, and Rosa Mosquera. 2025. Ontology matching with Large Language Models and prioritized depth-first search. *Inf. Fusion*, 123:103254.

Marius-Horatiu Totoian, Anca Marginean, Philipp Blohm, and Mir Nawab Hussain. 2024. 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)*, volume 3897 of *CEUR Workshop Proceedings*, pages 138–145. CEUR-WS.org.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

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

Rui Zhang, Yixin Su, Bayu Distiawan Trisedya, Xiaoyan Zhao, Min Yang, Hong Cheng, and Jianzhong Qi. 2024. AutoAlign: Fully Automatic and Effective Knowledge Graph Alignment Enabled by Large Language Models. *IEEE Trans. Knowl. Data Eng.*, 36(6):2357–2371.

A LogMap ontology alignment system

LogMap is an ontology alignment system that combines lexical, structural, and semantic techniques with logical reasoning to efficiently match large ontologies while preserving consistency (Jiménez-Ruiz and Grau, 2011; Jiménez-Ruiz et al., 2012a).[3] It supports interactive alignment and has consistently performed well in the Ontology Alignment Evaluation Initiative (OAEI).

LogMap defines some heuristics based on lexical and structural similarity to decide about the validity of a candidate mapping. However, some mappings are not clear-cut cases, and those are the mappings LogMap selects to ask an Oracle ($\mathcal{M}_{ask}$). Figure 4 shows the workflow followed by LogMap when allowing interaction.

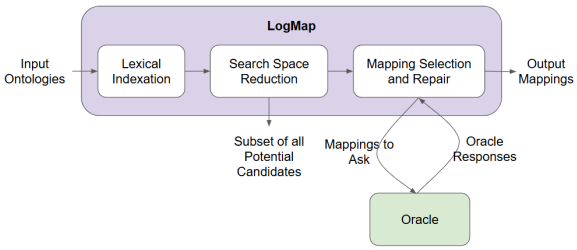

Figure 4: Workflow of the ontology alignment system LogMap with calls to an Oracle.

B Ontology-driven and system prompts

Listings 4-6 show the ontology-driven prompts with extended context. While Listing 7 shows the tested system prompts.

```
Analyse 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 4: $\mathbf{P}_{\text{EC}}$ prompt (non natural language-friendly with extended context).

[3] https://github.com/ernestojimenezruiz/logmap-matcher

```
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 5: $\mathbf{P}_{\text{EC}}^{\text{NLF}}$ Prompt (natural-language friendly with extended context).

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

```
Base = "You are a professional ontology matcher.
You need to answer different questions about
matching ontologies. Be precise."

Explainable = "You are helping researchers
determine if two biomedical terms from different
ontologies refer to the same concept. You'll be
provided with a natural-language description,
possibly including synonyms and parent categories.
Think like a domain expert, but explain your
judgment intuitively. Be precise"

Hierarchical = "You are a biomedical ontology
expert. Your task is to assess whether two given
entities from different biomedical ontologies refer
to the same underlying concept. Consider both their
semantic meaning and hierarchical context,
including parent categories and ontological
lineage. Be precise."

Lexical = "You are a domain expert assisting in
entity alignment across biomedical ontologies. Each
entity may include synonyms and category-level
relationships. Use synonym information and parent
class semantics to decide whether the two entities
mean the same thing. Be precise."
```

Listing 7: System prompts

C Additional supporting results

Figure 7 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). 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 8 shows the average YI values across ontology matching tasks, highlighting the differing levels of complexity in $\mathcal{M}_{ask}$ for each task.

D Experiments on determinism

Figure 5 shows the variations of the YI index across 4 independent runs with all six prompt templates on three ontology matching tasks using Gemini Flash 2.0.

Figure 6 illustrates the effect of different system prompts on the average YI index, revealing prompt-specific variations in performance relative to ontology-driven prompts.

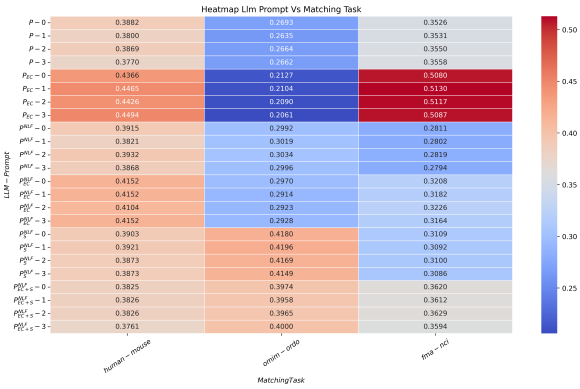

Figure 5: Determinism of Gemini 2.0 Flash across four runs for three matching tasks and all prompt templates.

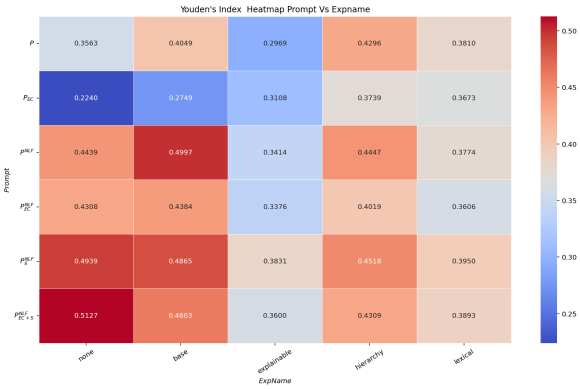

Figure 6: Performance variation of Gemini 2.0 Flash with different system prompts.

E Experiments on statistical analysis

We conducted t-test and Wilcoxon statistical tests to analyse whether the performance differences reported in Table 3 and Figure 3 were significant (p-value < 0.05). Table 5 confirms that LogMap+Or^{0} and LogMap+Or^{10} lead to significantly better results than both

System Comparison	t.test			Wilcoxon		
	greater	less	2-sided	greater	less	2-sided
LogMap+Or$^{LLM}_{GF2.0}$ vs LogMap+Or^{0}	0.999	0.0001	0.0002	1.0	0.002	0.004
LogMap+Or$^{LLM}_{GF2.0}$ vs LogMap+Or^{10}	0.999	0.0003	0.0006	1.0	0.002	0.004
LogMap+Or$^{LLM}_{GF2.0}$ vs LogMap+Or^{20}	0.918	0.082	0.165	0.898	0.125	0.25
LogMap+Or$^{LLM}_{GF2.0}$ vs LogMap+Or^{30}	0.030	0.970	0.060	0.037	0.973	0.074
LogMap+Or$^{LLM}_{GF2.0}$ vs LogMap	0.0007	0.999	0.001	0.002	1.0	0.004
LogMap+Or$^{LLM}_{GF2.5}$ vs LogMap+Or^{0}	0.998	0.002	0.003	1.0	0.002	0.004
LogMap+Or$^{LLM}_{GF2.5}$ vs LogMap+Or^{10}	0.991	0.009	0.018	0.998	0.004	0.008
LogMap+Or$^{LLM}_{GF2.5}$ vs LogMap+Or^{20}	0.797	0.203	0.406	0.787	0.248	0.496
LogMap+Or$^{LLM}_{GF2.5}$ vs LogMap+Or^{30}	0.037	0.963	0.074	0.049	0.963	0.098
LogMap+Or$^{LLM}_{GF2.5}$ vs LogMap	0.020	0.980	0.041	0.027	0.981	0.055

Table 5: 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.

LogMap+Or$^{LLM}_{GF2.0}$ and LogMap+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+Or$^{LLM}_{GF2.0}$ and LogMap+Or$^{LLM}_{GF2.5}$ with LogMap+Or^{20} yields no significant differences ($p > 0.1$ in the two-sided setting). LogMap+Or$^{LLM}_{GF2.0}$ and LogMap+Or$^{LLM}_{GF2.5}$ also significantly improve LogMap+Or^{30} and LogMap in automatic mode (*i.e.*, $p < 0.05$ in the 'greater' direction for both tests). Overall, the statistical tests support the results and the discussion presented in Section 5.

F Additional experiments with Qwen models

Table 6 compares the results of Qwen3-1.7b and Qwen3-8b across the different prompt templates. Qwen3-1.7b produces very low scores, typically diagnosing most mappings as negative. Qwen3-8b performed well with the natural-language-friendly prompt templates. Structured prompts, however, led to poor diagnostic capabilities, similar to Qwen3-1.7b.

Prompt	Qwen3-1.7b on $\mathcal{M}_{ask}$			Qwen3-8b on $\mathcal{M}_{ask}$		
	Se	Sp	YI	Se	Sp	YI
P	0.096	0.972	0.068	0.082	1.000	0.082
$\mathbf{P}_{\text{EC}}$	0.058	0.981	0.039	0.055	0.991	0.045
$\mathbf{P}^{\text{NLF}}$	0.199	0.925	0.123	0.483	0.943	0.426
$\mathbf{P}^{\text{NLF}}_{\text{EC}}$	0.168	0.962	0.130	0.435	0.962	0.397
$\mathbf{P}^{\text{NLF}}_{\text{S}}$	0.411	0.811	0.222	0.825	0.764	**0.590**
$\mathbf{P}^{\text{NLF}}_{\text{EC+S}}$	0,414	0.906	0.320	0.688	0.793	0.481

Table 6: Performance of evaluated Qwen models in the *anatomy* task across the six prompt templates. Se=Sensitivity, Sp=Specificity, YI=Youden's index. NLF=Natural-language friendly, EC=Extended context, S=Synonyms.

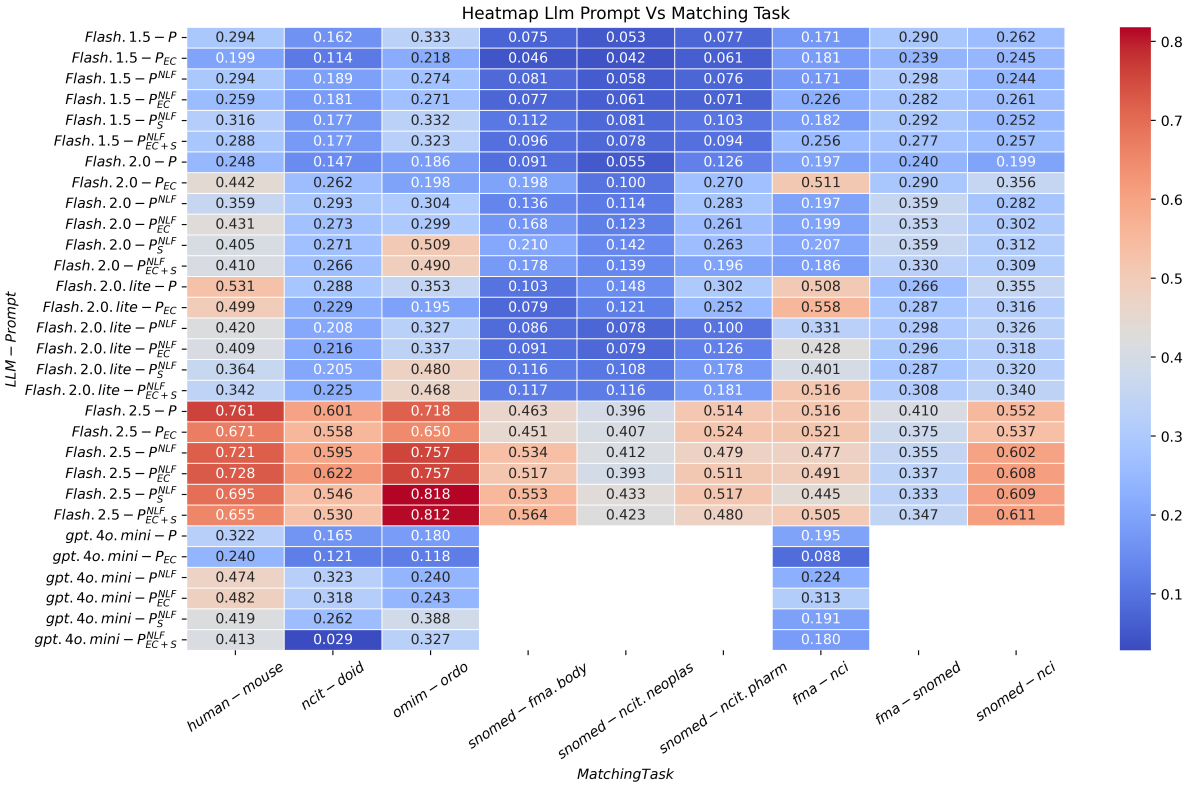

Figure 7: Diagnostic results (Youden's index) by the LLM-based Oracles over the selected ontology matching tasks. For example, *Flash 2.5*-$\mathbf{P}^{\mathrm{NLF}}_{\mathrm{EC+S}}$ represents the LLM-based Oracle relying on the Gemini Flash 2.5 model and evaluated with the natural-language friendly (NLF) prompts with extended context (EC) and synonyms (S). We only completed a subset of experiments with GPT-4o Mini as a reference.

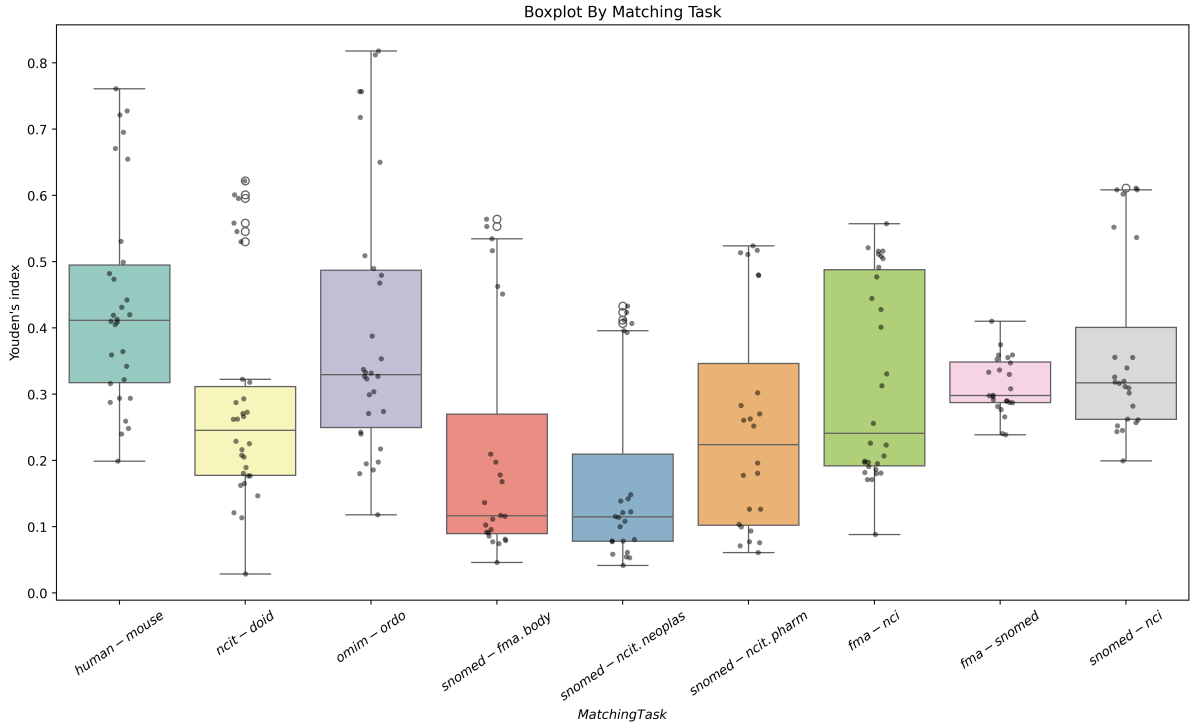

Figure 8: Average YI values per ontology matching task, reflecting the varying complexity of $\mathcal{M}_{ask}$ across tasks.