# 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

Paper accepted at the 19th Conference of the European Chapter of the
Association for Computational Linguistics (EACL 2026), main conference.

**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 analyse the feasibility of using Large Language Models (LLM) to aid the ontology alignment problem. LLMs are used only in the validation of a subset of correspondences for which there is high uncertainty. 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 prompt templates. Using LLMs as Oracles resulted in strong performance in the OAEI 2025, achieving the top-2 overall rank in the *bio-ml* track.

**Keywords:** knowledge graph alignment, ontology matching, large language models.

**Supplemental Material:** Source code and relevant resources for the experiments conducted in this paper are available in Zenodo: https://doi.org/10.5281/zenodo.15394653. The source code for the experiments with different LLMs and prompts as diagnostic tools is available at https://github.com/city-artificial-intelligence/rai-ukraine-kga-llm. The source code for LogMapLLM's integrated pipeline is available in this GitHub repository: https://github.com/city-artificial-intelligence/logmap-llm.

## 1 Introduction

Ontology alignment [11] plays a crucial role in integrating diverse data sources across domains. While numerous ontology matching systems exist (*e.g.*, [39]), 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 encoders of large amounts of data. LLMs have shown potential to be useful within an ontology alignment pipeline (*e.g.*, [42]). Nevertheless, LLMs are computationally or financially costly, and an unlimited use is not feasible.

In this paper, we have extended the state-of-the-art ontology matching system `LogMap` [24] to perform calls to an LLM-based Oracle. The LLM-based Oracle is used to validate a subset of

[†] These authors contributed equally to this work

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

correspondences where LogMap 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 about the entities in the mappings in question. We selected the GPT-4o Mini model (OpenAI) and a range of Google Gemini Flash models for our experiments, due to their good performance in recent LLM leaderboards.

To analyse the suitability of the LLM-based Oracles, we have conducted an extensive evaluation with the *anatomy* [10], *largebio* [26], and *bio-ml* [21] datasets of the Ontology Alignment Evaluation Initiative (OAEI) [40, 41], 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-prompt combinations based on the choice of five LLM implementations and six prompt templates. We have also evaluated the contribution of the LLM-based Oracles to the overall matching task by comparing the results with LogMap (automatic mode) and simulated Oracles with variable error rates [30]. We also report experiments for the *anatomy* dataset with the open-weight models Mistral, Llama and Qwen.

In contrast with other state-of-the-art systems that rely heavily on LLMs, our approach is designed to only use the LLM-based Oracle in very specific cases. Hence, the use of LLMs is more accessible without the need for substantial computational infrastructure or financial resources. The following points highlight the main contributions and novel aspects of this work. *(i)* We investigate the effect of incorporating the ontology context of the entities into prompt design, an aspect that has not been thoroughly examined in the ontology alignment literature. *(ii)* To our knowledge, while LLMs are increasingly applied in ontology alignment pipelines, their use as Oracles has been unexplored in the state-of-the-art. *(iii)* We provide a comprehensive evaluation that offers novel insights into the use of LLMs as diagnostic engines for ontology alignment, including a transparent and fine-grained analysis of the LLM contribution. *(iv)* The combination of LogMap with an LLM-based Oracle achieved top-2 overall results in the OAEI 2025 *bio-ml* track.

The paper is organised as follows. Section 2 introduces the necessary background. The relevant related work is provided in Section 3. Section 4 presents our method and system pipeline. Evaluation results are analysed in Section 5. Conclusions, future work and limitations are discussed in Section 6 and Section 7.

## 2 Preliminaries

An ontology alignment is the process of finding correspondences or a *mapping* $\mathcal{M}$ among the entities (ontology classes, properties or instances) 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, in this paper, we refer to an equivalence mapping ($\equiv$) 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 evaluating and comparing 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 [24] 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* [30]. During the mapping selection stage, LogMap identifies a subset of mappings $\mathcal{M}_{ask}$ for which it is uncertain and would prefer 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 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 [30].

**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 w.r.t. 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) [47], as follows, to evaluate the effectiveness of an Oracle at diagnosing mappings in $\mathcal{M}_{ask}$, where TP, FN, TN, and FP stand for the usual true positive, false negative, true negative, and false positive counts, respectively, such that:

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

**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 yes/no or True/False decisions. 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 [39]. Widely-used traditional matchers include LogMap [24] and AgreementMakerLight (AML) [12], 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 expert feedback to resolve low-confidence correspondences, but at the cost of extensive user effort and time [30].

In recent years, a new generation of systems leveraging Machine Learning (ML) and (large) language models has emerged. The OAEI Bio-ML track [21] 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.*, [29, 33, 23]). Knowledge graph embedding systems like OWL2Vec* [6] were also deployed in combination with ML to learn and validate ontology alignment (*e.g.*, [7, 18]). Systems relying on BERT-based models have become popular, given their flexibility to fine-tuning for specific tasks like ontology alignment. Prominent examples include BERTMap [19], BioGITOM [38], and the Matcha family [13]. Recent developments in the field are increasingly driven by approaches based on LLMs. Saki Norouzi et al. [34] and He et al. [20] performed exploratory studies about the potential of LLMs at ontology alignment. Amini et al. [2] extended the exploration to discover complex alignments beyond equivalence or subsumption. Systems like OLaLa [22], LLMs4OM [17], MILA [44], Agent-OM [42], KROMA [32] and HybridOM [45] have integrated LLMs in their architectures. A common technique has been to use retrieval methods to select top-k candidates for each entity, then asking the LLM to select the best among these candidates

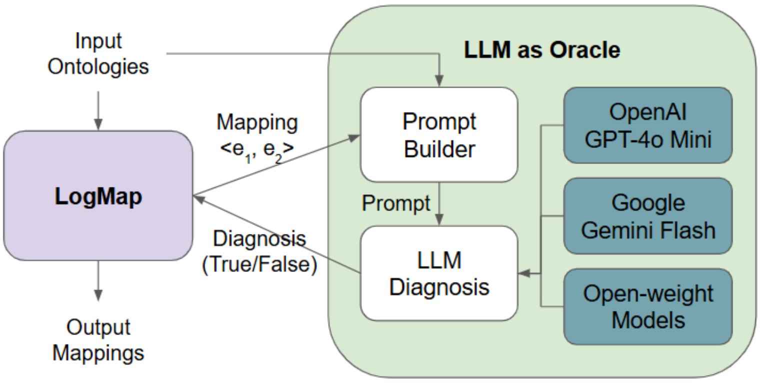

Figure 1: LLM-in-the-loop as an Oracle to diagnose challenging matches in ontology alignment.

(*e.g.*, [22, 17, 44]). By contrast, HybridOM uses an LLM to generate additional lexical descriptions of the entities involved in candidate correspondences. Recent approaches have explored the use of LLMs to focus on the alignment of instance data within knowledge graphs (*e.g.*, [48, 9]). Agent-OM proposed the use of autonomous LLMs to orchestrate multiple matching subtasks, indicating potential for agentic AI workflows to be adopted for ontology matching [42].

Our approach builds upon LogMap [24] 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 [44] and KROMA [32] have also focused on the limitation of the number of queries, leading to a reduction in the required computation times.

# 4 Methods: LLMs as Oracle

As introduced in Section 2, we build upon the system LogMap. In addition to predicting a set of output mappings, LogMap also identifies a subset of uncertain mappings ($\mathcal{M}_{ask}$), which 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 to the mappings in $\mathcal{M}_{ask}$. These mappings are not trivial as they typically involve entities with different labels and/or contexts, and they are better suited as a challenge to the performance of LLMs. LogMap interacts with the Oracle on-demand for each mapping $\langle e_1, e_2 \rangle \in \mathcal{M}_{ask}$. 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 candidate mapping $\langle e_1, e_2 \rangle$. The ontologies provide lexical representations (*e.g.*, labels and synonyms), as well as context (*e.g.*, parent classes) for $e_1$ and $e_2$. According to the locality principle [27], mappings should link entities that have similar contexts. Hence, a basic prompt should include at least the lexical representation of entities $e_1$ and $e_2$, and that of one of their 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 parents 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 also evaluate combinations of the above characteristics. We refer to as $\mathbf{P}^{\text{NLF}}_{\text{EC+S}}$ the prompt using all **NLF**, **EC** and **S** characteristics.

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 uncommon natural language expressions. Listing 1 shows the $\mathbf{P}$ prompt, where the entities and their context are listed. Listing 4 in Appendix B shows the structured prompt with extended context $\mathbf{P}_{\text{EC}}$.

```
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 ($\mathbf{P}$)

**Natural-language friendly prompts.** These prompts are based on the assumption that, given that LLMs are trained on large corpora of human-generated text, formulating questions in a more human-like way is expected to 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).

**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. A $\mathbf{P}^{\text{NLF}}_{\text{S}}$ prompt is shown 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 the above mapping templates, it is possible to add, for each LLM session, a short message that frames the model's overall role and answering 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)* as 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.

| Model | Cost / 1M Tokens | | Request Limits | | Cost | Latency (s) |
|---|---|---|---|---|---|---|
| | Input | Output | Minute | Day | 1k requests | |
| **Qwen3-8b (local)** | - | - | 0.5 | <1,000 | - | >125 |
| **Mistral Small-2402** | $1.00 | $3.00 | 400 | 576,000 | $0.15–$0.23 | 6–10 |
| **Llama 3-70b** | $2.65 | $3.50 | 800 | 1,152,000 | $0.35–$0.62 | 7–20 |
| **Gemini 1.5 Flash** | $0.08 | $0.30 | 2,000 | 2,880,000 | $0.010–$0.018 | 6–10 |
| **Gemini 2.0 Flash** | $0.10 | $0.40 | 2,000 | 2,880,000 | $0.014–$0.024 | 4–7 |
| **Gemini 2.0 Flash-Lite** | $0.08 | $0.30 | 4,000 | 5,760,000 | $0.010–$0.018 | 5.5–7.5 |
| **Gemini 2.5 Flash** | $0.15 | $0.60 | 1,000 | 10,000 | $0.018–$0.033 | 6.5–8 |
| **GPT-4o Mini** | $0.15 | $0.60 | 500 | 10,000 | $0.025–$0.04 | 4–14 |

Table 1: Latency and cost of evaluated LLMs. Each request typically consumed between 100 and 250 input tokens and 5 to 10 output tokens.

## 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 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 [36] for GPT-4o Mini, and the OpenAI's SDK endpoint for the Gemini Models [16]. Response latency typically remains within a few 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 [14], whereas OpenAI's API begins with a limit of 500 RPM and a daily quota of 10,000 requests [37], 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 per model [37, 15]. The average cost per 1,000 requests ranged from approximately $0.01 to $0.04. Table 1 summarises the cost and latency of the evaluated LLM models.

**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 [4, 3]. Mistral Small typically responded to mapping requests in less than 10 seconds, while Llama 3-70b took between 7 and 20 seconds. The cost per 1,000 requests was under $0.23 for Mistral Small-2402 and $0.62 for Llama3-70b. Qwen3 models (1.7b and 8b) [46] 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. Table 1 also summarises the cost and latencies for the evaluated open-weight models.

| 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 2: 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.

### 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 [30] 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 [30].

## 5 Experimental evaluation

Our experiments were conducted on a standard laptop with the selected LLM models as detailed in Section 4.2. All the experiments reported here were obtained with a budget of less than $50. We used the *anatomy* [10], *largebio* [26], and *bio-ml* [21] datasets provided by the OAEI evaluation initiative [40, 39]. As shown in Table 2, we covered a total of nine ontology matching tasks, involving ontologies of diverse sizes containing mostly concepts. The reference alignments ($\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 [43] and UMLS [5].

**Diagnostic capability.** 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 with the LLM models referred to in Section 4.2. Figure 2 shows the Youden's index (YI) as a measure of the correctness of the LLM-based Oracles for each LLM and prompt template combination. 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, which we refer to as $\texttt{Or}^{LLM}_{GF2.5}$. Table 3 compares the performance of LogMap (automatic mode) and the best model combination $\texttt{Or}^{LLM}_{GF2.5}$ 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 captures the effectiveness of an Oracle at identifying positive (sensitivity) and negative (specificity) mappings. A YI value of 1.0 indicates optimal performance. While no standard cut-off values exist for YI, some papers use 0.3, 0.5 and 0.7 as representative values

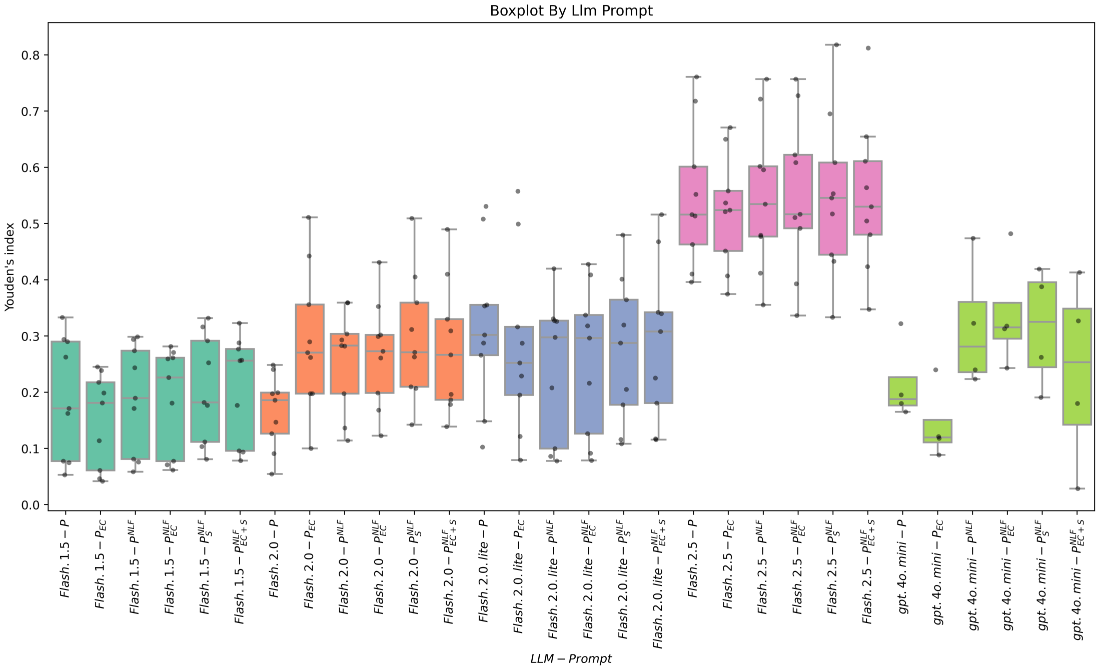

Figure 2: Summary of the diagnostic results (Youden's index) for the LLM-based Oracles.

| **Matching task** | $\lvert\mathcal{M}_{ask}\rvert$ | | **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 3: Comparison of `LogMap` (automatic mode) against the best LLM-based Oracle ($\mathtt{Or}^{LLM}_{GF2.5}$, using Gemini 2.5 Flash and $\mathbf{P}^{\mathrm{NLF}}_{\mathrm{S}}$ prompts) to diagnose the correctness of $\mathcal{M}_{ask}$. P is the number of real positives in $\mathcal{M}_{ask}$, N the number of real negatives, Se denotes Sensitivity, Sp Specificity and YI the Youden's index.

for low, moderate and high effectiveness, respectively (*e.g.*, [31]). Due to the complexity of the mappings in $\mathcal{M}_{ask}$, moderate YI values can be expected of an Oracle.

**Impact of the prompt template.** Figure 2 also illustrates the impact of using different prompt templates. Results vary across LLM models. Natural-language friendly prompts produce more consistent behaviour, while incorporating extended context and synonyms has a positive impact. Overall, for the Gemini 2.0 Flash and 2.5 Flash models, the most effective prompts were $\mathbf{P}^{\mathrm{NLF}}_{\mathrm{S}}$ (natural-language friendly with synonyms). We emphasise 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

| **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 4: 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 all matching tasks. Pr denotes Precision, Re Recall and F is the F-score.

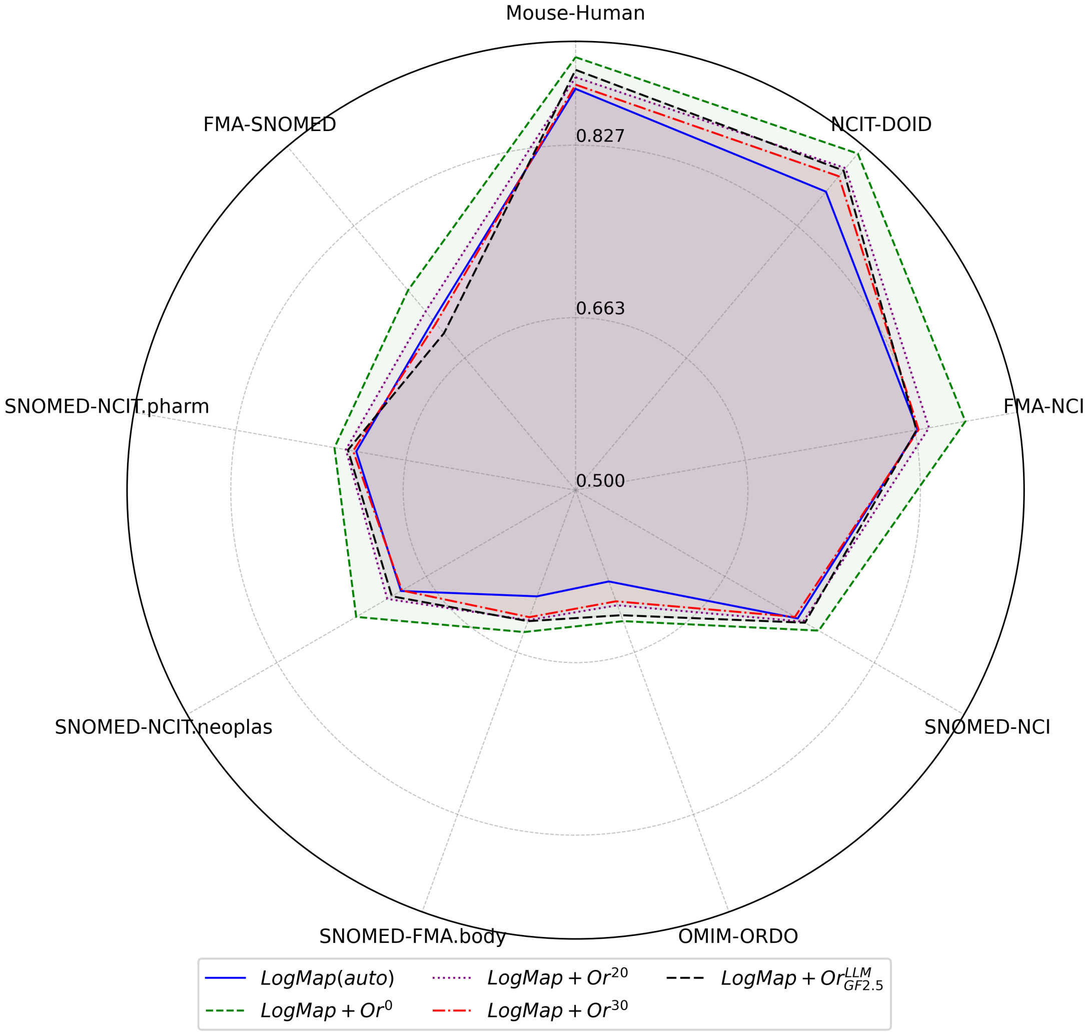

Figure 3: Comparison of `LogMap`, `LogMap` with $\texttt{Or}^{LLM}_{GF2.5}$, and `LogMap` in combination with Oracles with different error rates ($\texttt{Or}^0$, $\texttt{Or}^{20}$, and $\texttt{Or}^{30}$).

some tasks than others (*e.g.*, mappings involving isolated entities and/or with scarce synonyms).

| Matching task | **LogMap** | | | | **LogMapLLM** | | | |
|---|---|---|---|---|---|---|---|---|
| | **Pr** | **Re** | **F** | **Rank** | **Pr** | **Re** | **F** | **Rank** |
| **NCIT-DOID** | 0.843 | 0.893 | 0.867 | #4 | 0.932 | 0.883 | 0.907 | #2 |
| **OMIM-ORDO** | 0.834 | 0.456 | 0.589 | #7 | 0.916 | 0.476 | 0.626 | #4 |
| **SNOMED-FMA.body** | 0.760 | 0.569 | 0.651 | #6 | 0.869 | 0.561 | 0.682 | #4 |
| **SNOMED-NCIT.neoplas** | 0.763 | 0.772 | 0.736 | #4 | 0.821 | 0.747 | 0.782 | #1 |
| **SNOMED-NCIT.pharm** | 0.932 | 0.620 | 0.745 | #4 | 0.979 | 0.621 | 0.760 | #1 |
| **OVERALL** | 0.826 | 0.662 | 0.718 | #5 | 0.903 | 0.658 | 0.751 | #2 |

Table 5: `LogMap` and `LogMapLLM` in the OAEI 2025 *bio-ml* track. Pr=Precision, Re=Recall, F=F-score. Rank represents the position out of 10 participants.

**Overall matching task.** Table 4 shows the results obtained using `LogMap` (automatic mode) compared to `LogMap` integrated with an LLM-based Oracle (as depicted in Figure 1). We selected the top-performing LLM-based Oracles using $\mathbf{P}_{\mathrm{S}}^{\mathrm{NLF}}$ prompts: $\texttt{Or}^{LLM}_{GF2.0}$ (based on Gemini 2.0 Flash) and $\texttt{Or}^{LLM}_{GF2.5}$ (based on Gemini 2.5 Flash). As anticipated, the integration with the LLM-based Oracle yields improved F-scores across all tasks. $\texttt{LogMap+Or}^{LLM}_{GF2.5}$ dominates the *anatomy* and *bio-ml* tasks, while $\texttt{LogMap+Or}^{LLM}_{GF2.0}$ achieves the best results on *largebio*. To better contextualise the effectiveness of the LLM-based Oracle, we compared performance also with simulated Oracles with various error rates, following the approach in [30] as detailed in Section 4.3. Figure 3 compares performance across all nine ontology matching tasks for: `LogMap`, $\texttt{LogMap+Or}^{LLM}_{GF2.5}$, 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 [24, 30], $\texttt{LogMap+Or}^{30}$ still outperforms `LogMap` (without Oracle). The statistical analysis in Appendix E supports these observations.

**Comparison with OAEI systems.** The results of $\texttt{LogMap+Or}^{LLM}_{GF2.5}$ are highly competitive when compared with the state-of-the-art systems participating in the OAEI campaign (see results in the OAEI 2021 [40] for *largebio*, and in the OAEI 2024 [39], and OAEI 2025 [41] for *anatomy* and *bio-ml*). For example, $\texttt{LogMap+Or}^{LLM}_{GF2.5}$ would have ranked top-3 in the 2024 *anatomy* track, top-2 in the 2021 *largebio* track, achieving performance comparable to leading systems such as BertMap [19], Matcha [13], and LogMap-Bio [8] in the 2024 *bio-ml* track. In the OAEI 2025, $\texttt{LogMap+Or}^{LLM}_{GF2.5}$ participated under the name `LogMapLLM`, achieving top-4 results in the *anatomy* track and top-2 results in the *bio-ml* track [41, 28]. Table 5 reports the official OAEI results and ranks achieved by `LogMap` and `LogMapLLM` in the *bio-ml* track.[1] LogMap-Bio[2] was, on average, the top performer in the *bio-ml* track with an (average) F-score of 0.762, closely followed by `LogMapLLM` with an (average) F-score of 0.751.

**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 performance, suggesting that framing the LLM context is important (see Appendix D).

[1] Note that the F-scores in Table 4 differ from the official OAEI *bio-ml* results, as this track does not consider the complete ground truth for (global matching) evaluation. Further details are available at https://liseda-lab.github.io/OAEI-Bio-ML/2025/index.html.

[2] LogMap-Bio [8] uses BioPortal as a source of mediating (biomedical) ontologies.

| LLM Model | Sensitivity | Specificity | Youden's Index |
|---|---|---|---|
| **Mistral Small-2402** | 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 6: 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.

**Opportunities with Open-weight models.** We tested the diagnostic capabilities of Mistral Small-2402, Llama 3-70b, Qwen3-1.7b, and Qwen3-8b. Table 6 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.

**Experiments on data leakage.** Data leakage is a well-known challenge in the AI community, in general, and in the ontology matching community, in particular, when evaluating LLM-based systems on tasks with publicly available ground truths. We conducted an experiment using the OAEI NCIT–DOID dataset to assess the presence of potential critical data leakage. In the OAEI, ground truths are provided as sets of (correct) URI pairs (*e.g.*, equivalence relations between entities identified by their URIs, *i.e.*, their ontology ids). If the evaluated LLMs had been pretrained on these ground truths, a simple prompt such as "Is URI1 equivalent to URI2?"—without any additional contextual information—would be expected to yield performance comparable to, or better than, the results reported in this study. However, this experiment resulted in a Youden's Index of approximately 0.01, indicating very limited evidence of leakage of the OAEI *bio-ml* ground truths.

# 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 mappings for which `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` and achieving the top-2 results in the OAEI *bio-ml* track. 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 opportunities 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. Automatic prompt tuning and selection tailored to the matching task represents a promising direction for future work. We also plan to investigate few-shot prompts, particularly for tracks such as *bio-ml*, where a subset of mappings can be leveraged for training. Retrieval-augmented generation (RAG) may also enable systems to dynamically access relevant background knowledge, such as BioPortal [35], leading to more informed and accurate diagnostic capabilities.

# 7 Limitations

While our approach demonstrates strong 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.** Although our experiments on data leakage found no strong evidence of leakage of the OAEI ground truths, we cannot guarantee that the evaluated LLMs have not 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 prioritise 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]. 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. 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.

**Contributions.** EJR and AG defined the research objectives. SL, DS, SS, and EJR jointly designed the methodology and experiments. SL and SS implemented the system infrastructure, including API integrations and formal evaluation procedures. SL led the ontology data acquisition, system optimization, and visualization components. DS designed and implemented the prompt-engineering framework and developed the infrastructure for integrating local language models. EJR and AG contributed to the analysis of the experimental results. All authors contributed to drafting and revising the manuscript and approved the final version.

**Acknowledgement.** This research was supported by the RAI for Ukraine program of the NYU Center for Responsible AI, and by Turing Innovations Limited and The Alan Turing Institute's Defence and Security Programme via the project GUARD. We would also like to acknowledge the support of Dr Dave Herron in the creation of an integrated pipeline for LogMapLLM, which enhances the reproducibility of the results. Finally, we thank the anonymous reviewers for their valuable feedback during the ARR process.

## 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 (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: 6th Conf. on Knowledge Graphs and Semantic Web (KGSWC). LNCS, vol. 15459, pp. 17–31 (2024). https://doi.org/10.1007/978-3-031-81221-7_2, https://doi.org/10.1007/978-3-031-81221-7_2

[3] Bedrock, A.: Meta's Llama in Amazon Bedrock (2025), https://aws.amazon.com/bedrock/meta/ (accessed Sept. 2025)

[4] Bedrock, A.: Mistral AI in Amazon Bedrock (2025), https://aws.amazon.com/bedrock/mistral/ (accessed Sept. 2025)

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

[6] 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)

[7] Chen, J., Jiménez-Ruiz, E., Horrocks, I., Antonyrajah, D., Hadian, A., Lee, J.: Augmenting Ontology Alignment by Semantic Embedding and Distant Supervision. In: 18th Extended Semantic Web Conference (ESWC). LNCS, vol. 12731, pp. 392–408 (2021). https://doi.org/10.1007/978-3-030-77385-4_23, https://doi.org/10.1007/978-3-030-77385-4_23

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

[9] 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. pp. 82–89. AAAI Press (2024). https://doi.org/10.1609/AAAISS.V3I1.31186, https://doi.org/10.1609/aaaiss.v3i1.31186

[10] 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, https://doi.org/10.1186/s13326-017-0166-5

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

[12] Faria, D., Santos, E., Balasubramani, B.S., Silva, M.C., Couto, F.M., Pesquita, C.: AgreementMakerLight. Semantic Web **16**(2), SW–233304 (2025). https://doi.org/10.3233/SW-233304

[13] 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. CEUR Workshop Proceedings, vol. 3591, pp. 164–169 (2023), https://ceur-ws.org/Vol-3591/oaei23_paper6.pdf

[14] Gemini: Gemini API rate limits (2025), https://ai.google.dev/gemini-api/docs/rate-limits (accessed Aug. 2025)

[15] Gemini: Gemini APIs pricing (2025), https://ai.google.dev/gemini-api/docs/pricing (accessed Aug. 2025)

[16] Gemini: OpenAI's SDK endpoint (2025), https://ai.google.dev/gemini-api/docs/openai (accessed Aug. 2025)

[17] 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, https://doi.org/10.48550/arXiv.2404.10317

[18] 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, https://doi.org/10.1016/j.websem.2023.100798

[19] He, Y., Chen, J., Antonyrajah, D., Horrocks, I.: BERTMap: A BERT-Based Ontology Alignment System. In: 36th AAAI Conference on Artificial Intelligence. pp. 5684–5691 (2022). https://doi.org/10.1609/AAAI.V36I5.20510, https://doi.org/10.1609/aaai.v36i5.20510

[20] He, Y., Chen, J., Dong, H., Horrocks, I.: Exploring large language models for ontology alignment. In: Proceedings of the Posters, Demos and Industry Tracks co-located with 22nd International Semantic Web Conference (ISWC). CEUR Workshop Proceedings, vol. 3632 (2023), https://ceur-ws.org/Vol-3632/ISWC2023_paper_427.pdf

[21] 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 (ISWC). LNCS, vol. 13489, pp. 575–591 (2022). https://doi.org/10.1007/978-3-031-19433-7_33, https://doi.org/10.1007/978-3-031-19433-7_33

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

[23] Iyer, V., Agarwal, A., Kumar, H.: VeeAlign: Multifaceted Context Representation Using Dual Attention for Ontology Alignment. In: Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 10780–10792 (2021). https://doi.org/10.18653/V1/2021.EMNLP-MAIN.842, https://doi.org/10.18653/v1/2021.emnlp-main.842

[24] Jiménez-Ruiz, E., Cuenca Grau, B., Zhou, Y., Horrocks, I.: Large-scale interactive ontology matching: Algorithms and implementation. In: 20th European Conference on Artificial Intelligence (ECAI). Front. Artif. Intell. Appl., vol. 242, pp. 444–449. IOS Press (2012). https://doi.org/10.3233/978-1-61499-098-7-444, https://doi.org/10.3233/978-1-61499-098-7-444

[25] Jiménez-Ruiz, E., Grau, B.C.: Logmap: Logic-based and scalable ontology matching. In: 10th Int'l Semantic Web Conference (ISWC). LNCS, vol. 7031, pp. 273–288 (2011). https://doi.org/10.1007/978-3-642-25073-6_18, https://doi.org/10.1007/978-3-642-25073-6_18

[26] 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. CEUR Workshop Proceedings, vol. 882 (2012), https://ceur-ws.org/Vol-882/elkr_atsf_2012_paper9.pdf

[27] 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

[28] Jiménez-Ruiz, E., Lushnei, S., Shumskyi, D., Shykula, S., d'Avila Garcez, A.: LogMap Family welcomes LogMapLLM in the OAEI 2025. In: Proceedings of the 20th International Workshop on Ontology Matching (OM 2025). pp. 178–181 (2025)

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

[30] 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, https://doi.org/10.1017/S0269888919000080

[31] Liu, S., Tian, Q., Liu, Y., Li, P.: Joint statistical inference for the area under the roc curve and youden index under a density ratio model. Mathematics **12**(13), 2118 (2024). https://doi.org/10.3390/math12132118, https://doi.org/10.3390/math12132118

[32] Nguyen, L., Barcelos, E.I., French, R.H., Wu, Y.: KROMA: Ontology Matching with Knowledge Retrieval and Large Language Models. In: 24th Int'l Semantic Web Conference (ISWC). LNCS, vol. 16140, pp. 629–649 (2025). https://doi.org/10.1007/978-3-032-09527-5_34, https://doi.org/10.1007/978-3-032-09527-5_34

[33] 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)

[34] Norouzi, S.S., Mahdavinejad, M.S., Hitzler, P.: Conversational ontology alignment with ChatGPT. In: 18th Int'l Workshop on Ontology Matching co-located with the 22nd International Semantic Web Conference. CEUR Workshop Proceedings, vol. 3591, pp. 61–66 (2023), https://ceur-ws.org/Vol-3591/om2023_STpaper1.pdf

[35] 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, https://doi.org/10.1093/nar/gkp440

[36] OpenAI: Chat Completions API (2025), https://platform.openai.com/docs/guides/text?api-mode=chat (accessed Aug. 2025)

[37] OpenAI: OpenAI's GPT-4o mini API limits and pricing (2025), https://platform.openai.com/docs/models/gpt-4o-mini (accessed Aug. 2025)

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

[39] Pour, M.A.N., Algergawy, A., Blomqvist, E., et al.: 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. CEUR Workshop Proceedings, vol. 3897, pp. 64–97 (2024), https://ceur-ws.org/Vol-3897/oaei2024_paper0.pdf

[40] Pour, M.A.N., Algergawy, A., et al.: 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). vol. 3063, pp. 62–108 (2021), https://ceur-ws.org/Vol-3063/oaei21_paper0.pdf

[41] Pour, M.A.N., Blomqvist, E., Cotovio, P.G., et al.: Results of the Ontology Alignment Evaluation Initiative 2025. In: Proceedings of the 20th International Workshop on Ontology Matching (OM 2025). pp. 105–139 (2025)

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

[43] 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)

[44] Taboada, M., Martínez, D., Arideh, M., Mosquera, R.: Ontology matching with Large Language Models and prioritized depth-first search. Inf. Fusion **123**, 103254 (2025). https://doi.org/10.1016/J.INFFUS.2025.103254, https://doi.org/10.1016/j.inffus.2025.103254

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

[46] Yang, A., Li, A., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025)

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

[48] 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, https://doi.org/10.1109/TKDE.2023.3325484

# 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 [25, 24].[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. If an Oracle is not available, LogMap can also operate in an automatic (non-interactive) mode.

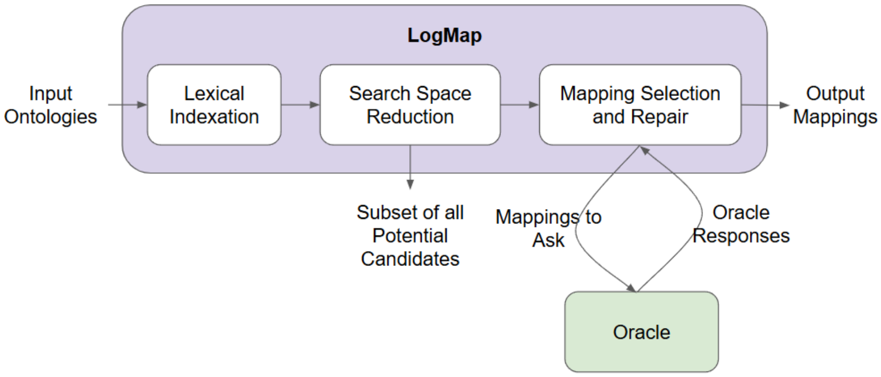

Figure 4: Workflow of the ontology alignment system LogMap with calls to an Oracle.

# 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.

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

```
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).

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

```
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: $\mathbf{P}^{\mathrm{NLF}}_{\mathrm{EC+S}}$ 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 5 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 6 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 7 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 8 illustrates the effect of different system prompts on the average YI index, revealing performance differences across prompts.

# E Experiments on statistical analysis

We conducted t-test and Wilcoxon statistical tests to analyse whether the performance differences reported in Table 4 and Figure 3 were significant ($p$-value $< 0.05$). Table 7 confirms that $\texttt{LogMap+Or}^{0}$ and $\texttt{LogMap+Or}^{10}$ lead to significantly better results than both $\texttt{LogMap+Or}^{LLM}_{GF2.0}$ and $\texttt{LogMap+Or}^{LLM}_{GF2.5}$ (*i.e.*, $p < 0.01$ in the "less" and "two-sided" settings with both t-test and Wilcoxon). The comparison of $\texttt{LogMap+Or}^{LLM}_{GF2.0}$ and $\texttt{LogMap+Or}^{LLM}_{GF2.5}$ with $\texttt{LogMap+Or}^{20}$ yields no significant differences ($p > 0.1$ in the two-sided setting). $\texttt{LogMap+Or}^{LLM}_{GF2.0}$ and $\texttt{LogMap+Or}^{LLM}_{GF2.5}$ also significantly improve $\texttt{LogMap+Or}^{30}$ 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.

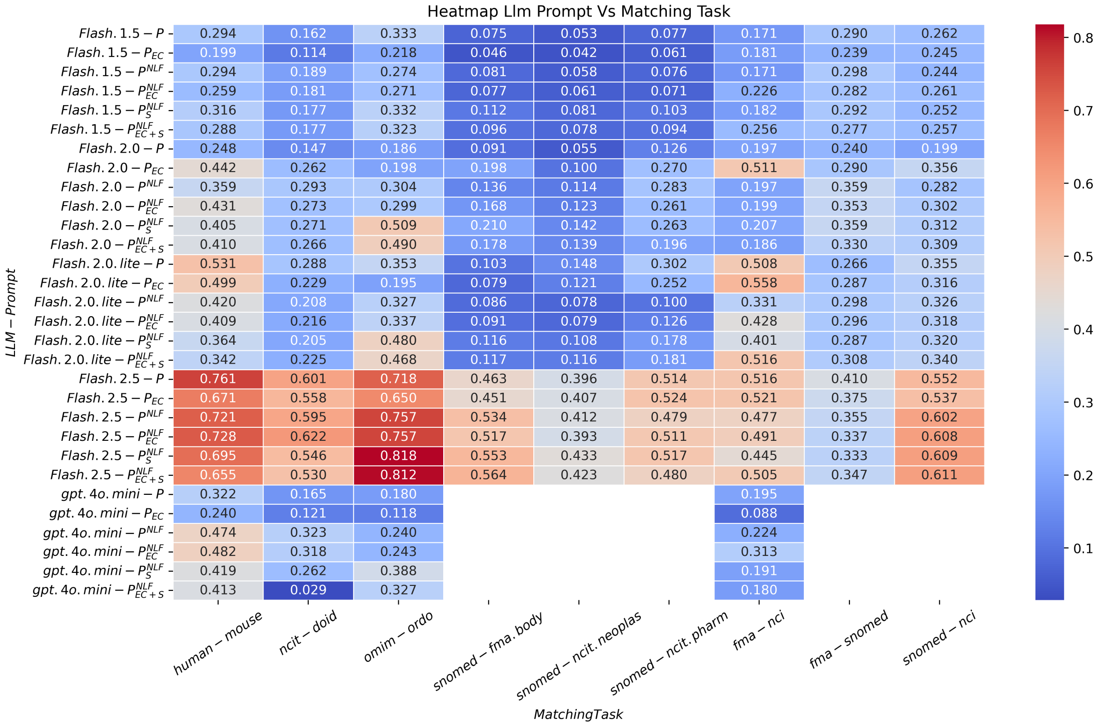

Figure 5: Diagnostic results (Youden's index) by the LLM-based Oracles over the selected ontology matching tasks. For example, *Flash 2.5*-$\mathbf{P}^{\text{NLF}}_{\text{EC+S}}$ represents the LLM-based Oracle relying on the Gemini Flash 2.5 model and evaluated with the natural-language friendly (NLF) prompts with extended context (EC) and synonyms (S). We only completed a subset of experiments with GPT-4o Mini as a reference.

# F Additional experiments with Qwen models

Table 8 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.

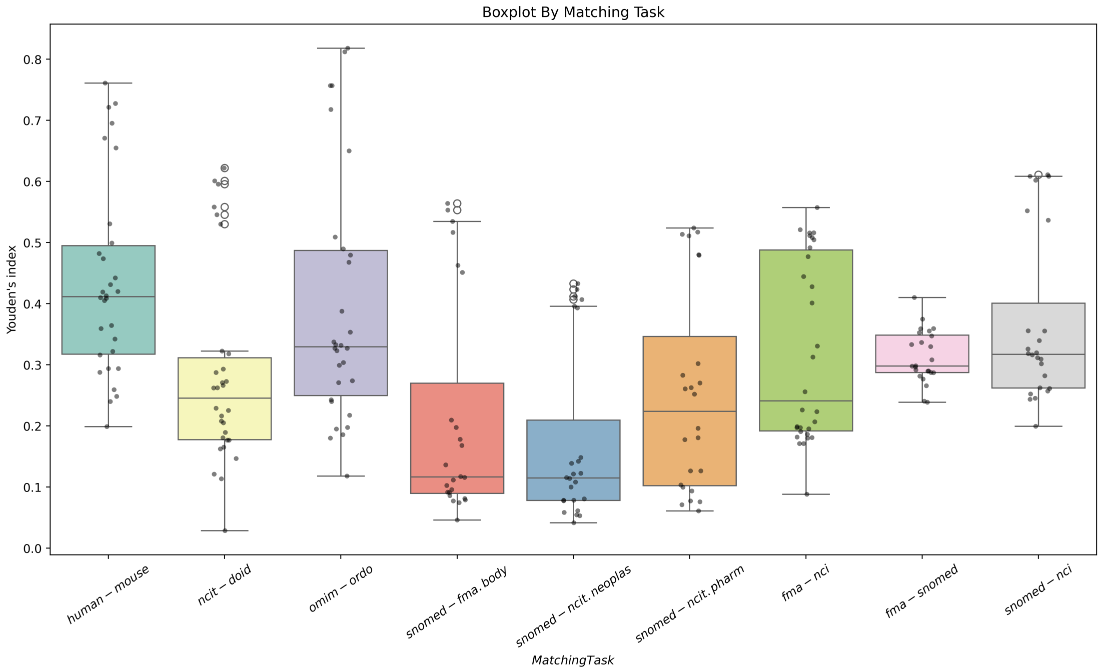

Figure 6: Average YI values per ontology matching task, reflecting the varying complexity of $\mathcal{M}_{ask}$ across tasks.

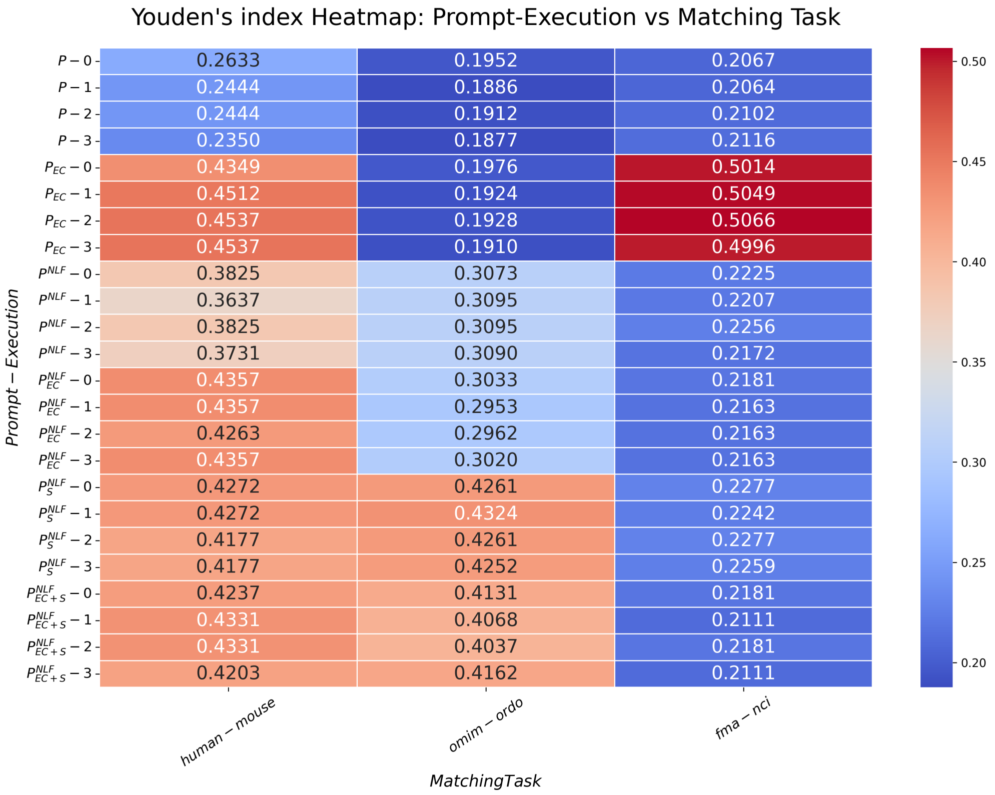

Figure 7: Determinism of Gemini 2.0 Flash across four runs for three matching tasks and all prompt templates.

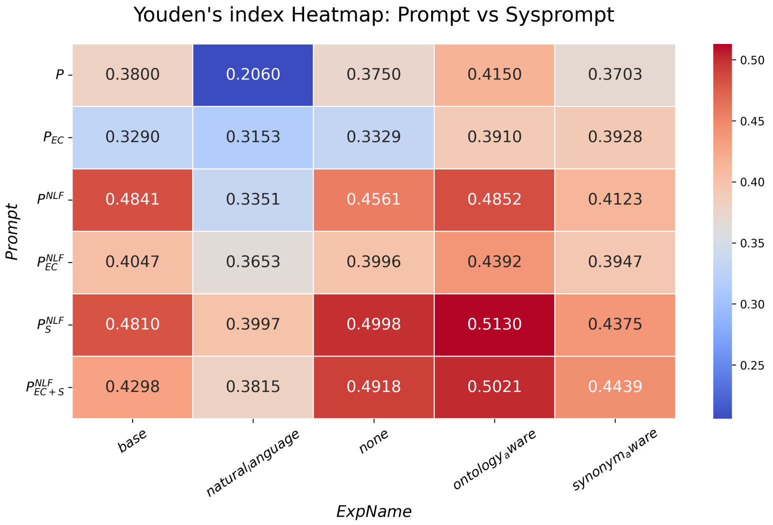

Figure 8: Performance variation of Gemini 2.0 Flash with different system prompts.

| **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 7: 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.

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

Table 8: 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.