Phase 1 Report (GUARD / G2049a)

Ensuring Interoperable and Trustworthy Knowledge Graphs for Defence and National Security AI

PI: Ernesto Jiménez-Ruiz

March 31, 2026

Phase 1 Report: Ensuring Interoperable and Trustworthy Knowledge Graphs for Defence and National Security AI

Project Acronym / Id:	GUARD / G2049a
Principal Investigator:	Dr Ernesto Jiménez Ruiz (Senior Lecturer/Associate Professor in AI)
Named staff:	Dr Dave Herron (PDRA) and Jon Dilworth (Software Engineer)
Collaborators:	Sviatoslav Lushnei, Dmytro Shumskyi, and Severyn Shykula
Institution:	Dept. of Computer Science, City St George's, University of London
Web:	https://ernestojimenezruiz.github.io/projects/guard/

Contents

1 Background and scope

The development of hybrid learning and reasoning systems in general, and Neurosymbolic (NeSy) AI systems [16] in particular, is gaining increasing attention to overcome the challenges of purely data-driven models (*e.g.*, LLMs - Large Language Models [58]) with respect to fairness, privacy, correctness, data and energy efficiency, identifiability, and eXplainability (XAI). In a domain like *defence and national security*, the design and development of robust, reliable and semantically sound AI models is paramount. NeSy systems base their prediction on both available data and knowledge, enabling well-informed decision-making.

Knowledge Graphs (KGs) [33] play a crucial role in orchestrating diverse and heterogeneous data sources, while providing a semantic and mathematical (*i.e.*, logic-based) representation of the domain. KGs are becoming essential components of NeSy systems [31] to *(i)* increase the coverage, validity and quality of the available data, *(ii)* impose constraints during the learning process, and *(iii)* reason about what has been learned (*i.e.*, validity of the predictions).

Note that, although there is no fundamental difference between a KG and an ontology from the knowledge representation point of view, in practice, KGs typically emphasise the encoding of a large amount of graph data, while ontologies focus on the semantic and abstract description of a domain. We use both terms interchangeably, as in this project, it is understood that knowledge graphs are semantically rich and described according to the vocabulary of an ontology [33]. The ontological part will be a critical enabler to data validation and enhancement.

The main challenge when working with KGs is the definition of an ontology to model the domain. The domain for a *defence and national security* application can be, however, quite broad as it may require the interplay of several subdomains to cover, among others, threats (*e.g.*, physical and cyber), relevant actors (*e.g.*, government, terrorist groups, intelligence services), operations (*e.g.*, military, emergency response), infrastructure (*e.g.*, airport, bases), policies (*e.g.*, defence strategy, international law), and geolocations.

There are several contributions from academia, industry, and government to create ontologies and knowledge graphs for security and defence. Prominent examples are the general-purpose models IES[1] and DoDAF.[2] IES (Information Exchange Standard) has been developed by the UK Government; while the DoDAF Formal Ontology has been implemented by the Department of War/Defense Chief Information Officer and serves as a NATO/multi-national model for defence. The IES models embrace semantic web technologies to represent general domain concepts such as locations, legal events, states, and measures. MITRE has also launched a family of cybersecurity ontologies to formalise offensive and defensive techniques [39].[3] These efforts by MITRE have been funded by the National Security Agency. The Department of Homeland Security has also designed a core Ontology for the maritime domain.[4] Efforts from academia include ontologies for military (*e.g.*, [19,45,55,88]), national security (*e.g.*, [12,14,56]), and cybersecurity (*e.g.*, [63,64,84,87]).

There are also prominent examples of collaborations among government, academia and industry. For example, the Intelligence Community (IC) Ontology Working Group (DIOWG), from the US Department of Defense, is leading on the development and implementation of a National Security Ontology Foundry, which will rely on the Basic Formal Ontology (BFO),[5] and the Common Core Ontology (CCO)[6] as their baseline standards.[7] BFO is an academia-driven upper-level ontology, widely adopted in life sciences ontologies (*e.g.*, OBO Foundry[8]), but with an increasing presence

[1] https://github.com/IES-Org/ont-ies

[2] https://dodcio.defense.gov/Library/DoD-Architecture-Framework/

[3] https://d3fend.mitre.org/

[4] https://www.dhs.gov/science-and-technology/publication/department-homeland-security-core-ontology-maritime-domain-awareness-fact-sheet

[5] https://basic-formal-ontology.org/

[6] https://github.com/CommonCoreOntology/CommonCoreOntologies

[7] https://www.buffalo.edu/news/releases/2024/02/department-of-defense-ontology.html

[8] https://obofoundry.org/

in government-related ontologies. CCO, developed by CUBRC,[9] provides a modular suite of BFO-grounded mid-level ontologies covering agents, organisations, events, artefacts, locations, and time.

The interoperability among ontologies at the upper- and mid-level has been facilitated by efforts like BFO and CCO, however, achieving interoperability across application ontologies may be particularly challenging in domains that require reconciling both the wide-ranging scope of concepts and the highly-specialised representations in the subfields. *Defence and national security* is an example of such domains. An ontology foundry for *defence and national security*, as the one proposed by the DIOWG, should ensure that its application ontologies achieve interoperability, comprehensive coverage, and high quality. Without these guarantees, the deployment of ontologies and KGs risks limiting their effectiveness in supporting downstream *defence and national security* tasks.

Challenge 1 (CH1): interoperability and coverage. State-of-the-art application ontologies and knowledge graphs relevant to *defence and national security* model intersecting domains. Although some efforts exist to provide a unified view of some of these ontologies (*e.g.*, [1,75,81]), their interoperability is still limited. Alignment with available linked data resources[10] is also key in a domain like *defence and national security* to extend the coverage with, for example, geolocations, organisations, infectious diseases, and environmental hazards.

Challenge 2 (CH2): quality. The quality of a KG is paramount in domains such as *defence and national security*. Defining the semantics of the KG with expressive languages like the Web Ontology Language (OWL) [15] is key to identifying inconsistencies. Meanwhile, languages like the Shapes Constraint Language (SHACL) [41] or Shape Expressions (ShEx) [71] facilitate the definition of integrity constraints to validate and count violations within the KG. However, SHACL, ShEx and the reasoning capabilities of OWL are not always adopted in practice due to scalability challenges when reasoning with large KGs [66]. Integrating multiple KGs and analysing their compatibility further aggravate scalability issues [79].

The GUARD project aims to lay the foundations to enable the integration and validation of knowledge graphs, making them ready for use in *defence and national security* downstream tasks. The GUARD project was split into two phases. **Phase 1** (November 2025 to March 2026) has dealt with **CH1** through a novel cost-effective utilisation of LLMs for knowledge graph integration. The **proposed Phase 2** (if funded) will focus on addressing **CH2**, leveraging the use of efficient logic-based reasoning for knowledge graph validation and the use of LLMs to provide (potential) solutions for the detected logical errors. LLMs will also be used to further enhance the integration of KGs.

This report presents the main outcomes achieved during **Phase 1 of the project**.

2 Phase 1 delivery

In Phase 1 of the project, we have delivered *(i)* `LogMapLLM`, an open-source LLM-enhanced ontology alignment system built on top of `LogMap`; *(ii)* a comprehensive evaluation of `LogMapLLM` over the OAEI datasets using different prompts and large language models, *(iii)* a comprehensive study of how state-of-the-art ontologies in the *defence and national security* domain semantically overlap; and *(iv)* a new evaluation track in the Ontology Alignment Evaluation Initiative (OAEI) 2026 based on the analysed *defence and national security* domain ontologies.

`LogMapLLM` and its extensive evaluation have led to a publication in one of the main venues in the NLP community: EACL 2026. We are also currently working on a journal paper with an extended zero-shot evaluation, and including the few-shot experiments. These experiments are described in the subsequent sections.

[9] https://www.cubrc.org/data-science-information-fusion/specialized-data-ontology-development/
[10] https://guides.lib.calpoly.edu/c.php?g=261997&p=1749320

Main contributions. 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 the conducted research in Phase 1. *(i)* We have experimented with both commercial LLMs and open-weight models with zero-shot and few-shot prompts. *(ii)* We have investigated 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. *(iii)* 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. *(iv)* We have provided 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. *(v)* The combination of `LogMap` with an LLM-based Oracle (`LogMapLLM` [38, 50]) achieved top-2 overall results in the OAEI 2025 *bio-ml* track.

Publications:

- Sviatoslav Lushnei, Dmytro Shumskyi, Severyn Shykula, Ernesto Jiménez-Ruiz, and Artur Garcez. **Large Language Models as Oracles for Ontology Alignment**. In 19th Conference of the European Chapter of the Association for Computational Linguistics (EACL), 2026. Pre-print: https://openaccess.city.ac.uk/id/eprint/36838/

GitHub repositories:

- `LogMapLLM` (integrated pipeline): https://github.com/city-artificial-intelligence/logmap-llm
- Experiments with different LLMs (and prompts) as diagnostic tools (e.g., Oracles): https://github.com/city-artificial-intelligence/rai-ukraine-kga-llm
- DISO: Defence, Intelligence and Security Ontologies: https://github.com/city-artificial-intelligence/diso
- DISO-OAEI (evaluation track): https://city-artificial-intelligence.github.io/diso-oaei/

Work Plan Status. Phase 1 of the project started on November 11 and has succesfully complete the planned work packages and tasks. We have also applied for a non-cost extension until April 30 to work on the journal paper and bridge with Phase 2 (if funded).

- **WP 1. Tool design and development.**
  - Task 1.1. LLM selection and interaction. (COMPLETED)
  - Task 1.2. Zero-shot prompt design and evaluation. (COMPLETED)
  - Task 1.3. Few-shot prompt design and evaluation. (COMPLETED)
  - Task 1.4. Integration with `LogMap`. (COMPLETED)
- **WP 2. Tool evaluation.**
  - Task 2.1. Domain-specific dataset selection and integration. (COMPLETED)
  - Task 2.2. Evaluation with OAEI datasets. (COMPLETED)
  - Task 2.3. Creation of a new OAEI track. (COMPLETED)

3 Tool design and evaluation

Ontology alignment [21] plays a crucial role in integrating diverse data sources across domains. While numerous ontology matching systems exist (*e.g.*, [68]), 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.*, [72]). Nevertheless, LLMs are computationally or financially costly, and unlimited use is not feasible.

3.1 Preliminaries

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` [35] 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* [46]. 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 $\mathcal{M}_{ask}$. Figure 1 shows the workflow followed by `LogMap` when allowing interaction.

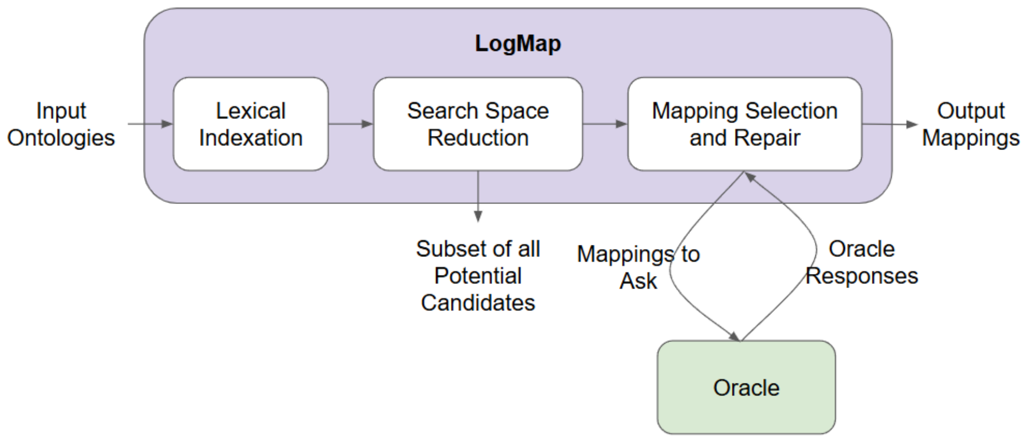

Figure 1: Workflow of the ontology alignment system `LogMap` with calls to an Oracle.

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 [46].

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) [89], 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 or Yes/No)" – possibly enriched with ontology context (*e.g.*, parent classes or synonyms). This approach allows the LLM to be used as a lightweight semantic Oracle.

3.2 Related work

The Ontology Alignment Evaluation Initiative (OAEI) has driven progress since 2004 by providing standardised benchmarks and evaluation protocols for matching systems [68]. Widely-used traditional matchers include `LogMap` [35] and AgreementMakerLight (AML) [22], 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 [46].

In recent years, a new generation of systems leveraging Machine Learning (ML) and (large) language models has emerged. The OAEI Bio-ML track [30] 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.*, [34, 43, 60]). Knowledge graph embedding systems like OWL2Vec* [10] were also deployed in combination with ML to learn and validate ontology alignment (*e.g.*, [11,27]). 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 [28], BioGITOM [65], and the Matcha family [23]. Recent developments in the field are increasingly driven by approaches based on LLMs. Saki Norouzi et al. [61] and He et al. [29] performed exploratory studies about the potential of LLMs at ontology alignment. Amini et al. [5] extended the exploration to discover complex alignments beyond equivalence or subsumption. Systems like OLaLa [32], LLMs4OM [25], MILA [82], Agent-OM [72], KROMA [59] and HybridOM [83] 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 (*e.g.*, [25, 32, 82]). 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.*, [17, 90]). 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 [72].

Our approach builds upon `LogMap` [35] 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 [82] and KROMA [59] have also focused on the limitation of the number of queries, leading to a reduction in the required computation times.

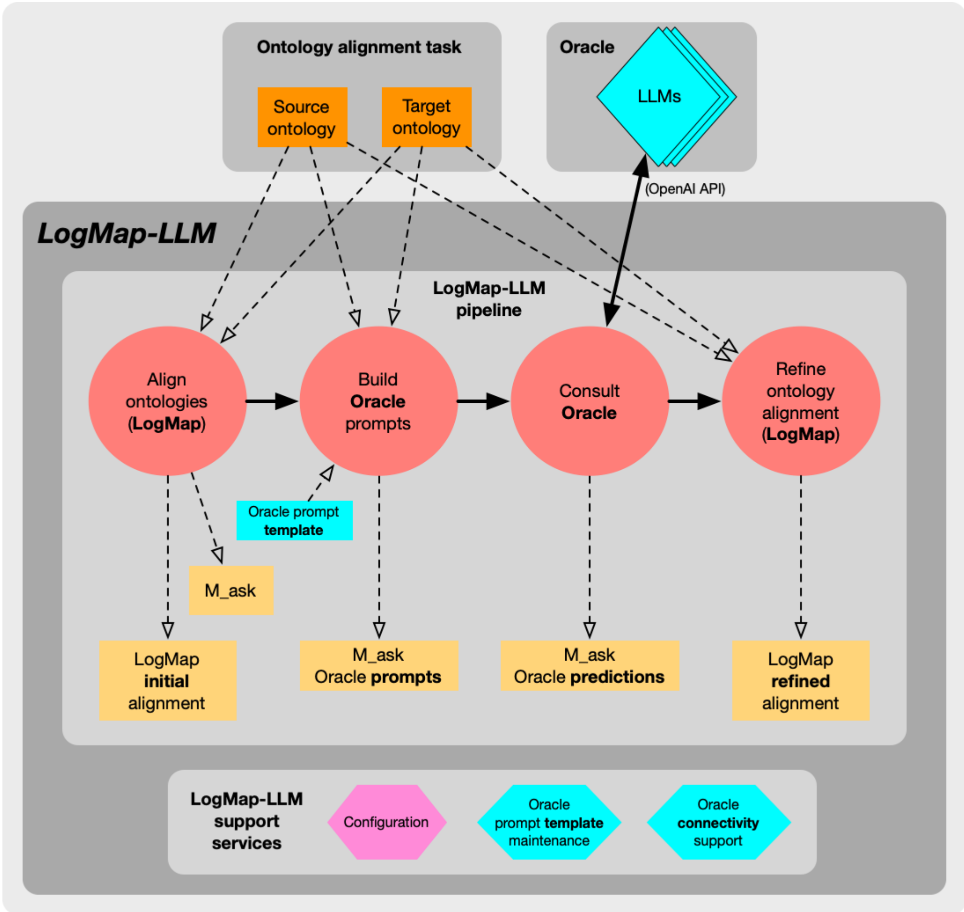

Figure 2: `LogMapLLM` integrated pipeline.

3.3 `LogMapLLM` architecture overview

In GUARD's Phase 1, we have extended the state-of-the-art ontology matching system `LogMap` [35] to perform calls to an LLM-based Oracle (`LogMapLLM` [50]). The LLM-based Oracle is used to validate a subset of 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. Figure 2 shows `LogMapLLM`'s integrated pipeline. The GitHub repository https://github.com/city-artificial-intelligence/logmap-llm contains the integrated pipeline to execute `LogMapLLM` (*i.e.*, calling `LogMap` and interacting with an LLM). `LogMapLLM` currently supports the OpenAI API for interacting with LLMs. This permits `LogMapLLM` users to connect with any LLM provider that supports the OpenAI API. For example, the OpenRouter LLM aggregation platform[11] supports the OpenAI API and enables access to 500+ LLM models. Additionally, `LogMapLLM` can seamlessly interact with a locally deployed LLM via the same OpenAI API.

The `LogMapLLM` pipeline, shown in Figure 2, begins with `LogMap` performing an initial alignment between two ontologies. The references to $\mathcal{M}_{ask}$ refer to **mappings to ask** to an Oracle. These are mappings (potential candidate mappings between pairs of ontology entities) of which `LogMap` is uncertain, and for which it invites feedback (opinions or predictions) from an external Oracle of some kind. In the case of `LogMapLLM`, the Oracle is an LLM rather than a human domain expert. For each candidate mapping (pair of ontology entities) in the set $\mathcal{M}_{ask}$, `LogMapLLM` builds a unique Oracle (LLM) **user** prompt to put to an LLM. Different **user** prompt templates can be used that incorporate

[11] https://openrouter.ai/

different types and amounts of ontological context from the respective ontologies. The predictions of an LLM Oracle in relation to these $\mathcal{M}_{ask}$ **user** prompts are collected and fed into `LogMap` so that it can refine its initial alignment, by taking account of the Oracle's feedback (*e.g.*, the predictions of an LLM).

3.4 Ontology-driven prompt builder

The first step in the interaction with the LLM-based Oracle is the creation of a call to the Oracle 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. Each oracle call consists of several components. These components construct the context visible to the LLM and include: *(i)* a system prompt that frames the task for the LLM, *(ii)* a series of (optional) k-shot question-answer pairs, and *(iii)* the ontology-driven prompt that asks for a response from the LLM for a given uncertain mapping.

The construction sequence is as follows: the system prompt appears first in the conversation history, with instructions tailored to the entity type being evaluated (class, property, or instance). Next, each k-shot question-answer pair is retrieved and added in a fixed order: a positive example followed by a negative one. We ensure each example question matches the ontology-driven prompt format and its answer matches the chosen answer format (boolean or natural-language-friendly). Finally, an ontology-driven prompt based on the uncertain mapping from $\mathcal{M}_{ask}$ is constructed and appended to the conversation history.

In our EACL paper [50], we originally designed six different prompt templates (see Section 3.4.1), which have been extended to include additional characteristics and to consider property and instance mappings (see Section 3.4.2).

3.4.1 Original ontology-driven prompts (class mappings)

The original set of prompt templates combined 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**). According to the locality principle [37], 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. 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 $\mathbf{P}^{\mathrm{NLF}}_{\mathrm{EC+S}}$ 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 2 shows the structured prompt with extended context $\mathbf{P}_{\mathrm{EC}}$.

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

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

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

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

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

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

Natural-language friendly prompts. These prompts are based on the assumption that, given 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 3 shows this type of prompt ($\mathbf{P}^{\mathrm{NLF}}$), while Listing 4 includes the version with extended context $\mathbf{P}^{\mathrm{NLF}}_{\mathrm{EC}}$.

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

```
We have two entities from different ontologies.

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

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

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

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

Prompts with synonyms. Although LLMs may inherently encode synonyms and lexical variations related to the ontology entities, the $\mathbf{P}^{\mathrm{NLF}}_{\mathrm{S}}$ and $\mathbf{P}^{\mathrm{NLF}}_{\mathrm{EC+S}}$ prompts are designed to analyse the impact of explicitly including synonyms for both the entities in a given correspondence and their associated context. A $\mathbf{P}^{\mathrm{NLF}}_{\mathrm{S}}$ prompt is shown in Listing 5, while Listing 6 provides its variant with extended context ($\mathbf{P}^{\mathrm{NLF}}_{\mathrm{EC+S}}$).

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

```
We have two entities from different ontologies.

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

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

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

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

3.4.2 Extended ontology-driven prompts (class, property and instance mappings)

The extended ontology-driven prompts were designed to evaluate additional inferencing capabilities of LLMs, as well as to include the necessary context and questions for assessing property and instance mappings. The new prompts build upon the previous experience where natural-language friendly prompts with context and sysnonyms leds to the best results [50]. The new ontology-driven prompt construction automatically detects which track is being evaluated. Each prompt will reference either *'biomedical ontologies'*, *'conference ontologies'*, or *'knowledge graphs'*. In addition, the construction process also allows for the ablation of answer formatting. In particular, we vary the binary response in the model to allow either *"True" or "False"* or *"Yes" or "No"*. The formatting variability is shown by square braces in each example prompt.

System prompts. The LLM system prompt is chosen based on both the entity type and the mapping relation. For class mappings, which are equivalence or subsumption-based, one of two class-level prompts is used. For property and instance mappings, the same property or instance-specific system prompt is always used. The system prompts used are listed below.

Classes: direct equivalence

You are assisting in an OWL ontology alignment exercise. You will be presented with a pair of entities representing classes, one class entity from each of the two different ontologies. Additional ontological context for each of the two class entities may be provided. You will be asked to decide whether the two class entities are semantically equivalent. Adopt the dual personas of ontology alignment expert and domain expert whilst coming to your decision. Give only a one-word binary response regarding the posited equivalence relation: ["True" or "False"]["Yes" or "No"].

Classes: mutual subsumption

You are assisting in an OWL ontology alignment exercise. You will be presented with a pair of entities representing classes, one class entity from each of the two different ontologies. Additional ontological context for each of the two class entities may be provided, such as parent categories. You will be asked to decide whether one class is subsumed by (is a kind of) the other. Adopt the dual personas of ontology alignment expert and domain expert. Give only a one-word binary response regarding the posited subsumption relation: ["True" or "False"]["Yes" or "No"].

Properties

You are assisting in an OWL ontology alignment exercise. You will be presented with a pair of properties (relationships) from different ontologies. Context about what types of entities each property connects (domain and range) may be provided. You will be asked to decide whether the two properties represent the same relationship. Give only a one-word binary response: ["True" or "False"]["Yes" or "No"].

Instances

You are assisting in an entity resolution exercise across knowledge graphs. You will be presented with a pair of entities, one from each of two different knowledge graphs. Additional context such as entity types, attributes, and relationships may be provided. You will be asked to decide whether the two entities refer to the same real-world entity. Give only a one-word binary response: ["True" or "False"]["Yes" or "No"].

Class-mapping prompts. All tasks require at least some class-level correspondences. We organise class-level prompts into equivalence-based and mutual subsumption-based. Note that all operate on the same set of uncertain mappings provided by $\mathcal{M}_{ask}$. They may vary in presentation and may include different structural contexts. Varying the structural context within each prompt tests how well each context guides the model to semantically disambiguate domain-specific terminology. Specifically, we include:

1. **Synonym-based prompts:** include label and synonym annotations.
2. **Parent-based prompts:** include direct antecedent nodes as structural context.
3. **Sibling-based prompts:** include the two most semantically similar siblings as structural context.

Class equivalence: synonyms only

We have two entities from different ["biomedical", "conference", ""] ["ontologies", "knowledge graphs"]. The first one is "Trapezoid", also known as "Trapezoid bone", "Trapezoid", "Os trapezoideum". The second one is "Trapezoid", also known as "Trapezium", "Trapezoid". Do they mean the same thing? Respond with ["True" or "False"]["Yes" or "No"].

Class equivalence: synonyms and parents

We have two entities from different ["biomedical", "conference", ""] ["ontologies", "knowledge graphs"]. The first one is "Trapezoid", also known as "Trapezoid bone", "Trapezoid", "Os trapezoideum", which falls under the category "Short Bone". The second one is "Trapezoid", also known as "Trapezium", "Trapezoid", which falls under the category "Quadrilateral". Do they mean the same thing? Respond with ["True" or "False"]["Yes" or "No"].

Class equivalence: synonyms, parents and siblings

We have two entities from different ["biomedical", "conference", ""] ["ontologies", "knowledge graphs"]. The first one is "Trapezoid", also known as "Trapezoid bone", "Trapezoid", "Os trapezoideum", which falls under the category "Short Bone". Other "Short Bone" concepts include "Scaphoid" and "Lunate". The second one is "Trapezoid", also known as "Trapezium", "Trapezoid", which falls under the category "Quadrilateral". Other "Quadrilateral" concepts include "Rectangle" and "Parallelogram". Do they mean the same thing? Respond with ["True" or "False"]["Yes" or "No"].

Class equivalence: deductive style

We have two entities from different ["biomedical", "conference", ""] ["ontologies", "knowledge graphs"]. Consider the following facts:

- "Trapezoid" is a kind of "Short Bone" (Ontology 1)
- "Trapezoid" is also known as "Trapezoid bone" and "Trapezoid" and "Os trapezoideum" (Ontology 1)
- "Trapezoid" is a kind of "Quadrilateral" (Ontology 2)
- "Trapezoid" is also known as "Trapezium" and "Trapezoid" (Ontology 2)

Given these facts, can you conclude that "Trapezoid" (Ontology 1) and "Trapezoid" (Ontology 2) refer to the same real-world concept? Respond with ["True" or "False"]["Yes" or "No"].

Class equivalence by mutual subsumption. We also included a set of prompts which are used to infer equivalence by means of mutual subsumption. Specifically, two prompts are used to test for subsumption in both directions. Equivalence is determined as the logical conjunction of both answers: $A \sqsubseteq B \land B \sqsubseteq A \iff A \equiv B$. This doubles the number of required prompts when testing for equivalence in uncertain mappings, but is included as an ablation since it may improve the reliability of LLM responses, and can be tested relatively cheaply against open-weight models within a local evaluation environment.

Class equivalence by mutual subsumption: labels and synonyms

We have two entities from different ["biomedical", "conference", ""] ["ontologies", "knowledge graph"]. The first one is "Trapezoid", also known as "Trapezoid bone", "Trapezoid", "Os trapezoideum". The second one is "Trapezoid", also known as "Trapezium", "Trapezoid". If something is a "Trapezoid", is it also a "Trapezoid"? Respond with ["True" or "False"]["Yes" or "No"].

Class equivalence by mutual subsumption: labels, synonyms and parents

We have two entities from different ["biomedical", "conference", ""] ["ontologies", "knowledge graph"]. The first one is "Trapezoid", also known as "Trapezoid bone", "Trapezoid", "Os trapezoideum", which falls under the category "Short Bone". The second one is "Trapezoid", also known as "Trapezium", "Trapezoid", which falls under the category "Quadrilateral". If something is a "Trapezoid" and a "Short Bone", is it also a "Trapezoid" and a "Quadrilateral"? Respond with ["True" or "False"]["Yes" or "No"].

Class equivalence by mutual subsumption: ancestral disjointness

We have two entities from different ["biomedical", "conference", ""] ["ontologies", "knowledge graph"]. The first one is "Trapezoid", also known as "Trapezoid bone", "Trapezoid", "Os trapezoideum". The second one is "Trapezoid", also known as "Trapezium", "Trapezoid"". If something is a "Trapezoid" and a "Short Bone", is it also a "Trapezoid" and a "Quadrilateral"? Respond with ["True" or "False"]["Yes" or "No"].

Property Prompts. Properties have a distinct context compared to classes, as their primary source of contextual information derives from the classes they connect (*e.g.*, domain and range). We explored various natural-language-friendly prompts.

Property labels only

We have two properties from different ["conference", ""] ontologies. The first property is "writePaper". The second property is "is_author_of". Do these properties represent the same relationship? Respond with ["True" or "False"]["Yes" or "No"].

Property labels with domain and range

We have two properties from different ["conference", ""] ontologies. The first property is "writePaper", connecting elements of type "Author" to elements of type "Paper". The second property is "is_author_of", connecting elements of type "Person" to elements of type "Contribution". Do these properties represent the same relationship? Respond with ["True" or

"False"]["Yes" or "No"].

Property labels with domain, range, and synonyms

We have two properties from different ["conference", ""] ontologies. The first property is "writePaper" (also known as "writes"), connecting elements of type "Author" (also known as "Writer") to elements of type "Paper" (also known as "Manuscript", "Written Work"). The second property is "is_author_of", connecting elements of type "Person" (also known as "Individual") to elements of type "Contribution" (also known as "Article", "Publication"). Do these properties represent the same relationship? Respond with ["True" or "False"]["Yes" or "No"].

Instance prompts. Prompt templates for instance mappings consider instance labels, types, attributes, and their entire set of predicates (their 'full context'). Since instances may have multiple attributes, we include two additional prompt variants. One selects context using intersecting predicates, and the other uses entropy-ranked intersecting predicates. For the former, we compute the lexical intersection (exact matches) between predicate names across both instances and include these predicates in the prompt. For the latter (the entropy-ranked variant), we order predicates by maximal Shannon entropy, thereby providing the most distinctive characteristics of each instance first. In this case, we apply a cut-off of 3, meaning only the top-3 ranked predicates are included as context. This avoids including redundant or meaningless context in the prompt, thereby reducing the overall occurrence of noisy prompts.

Instance labels only

We have two entities from different knowledge graphs. The first is "Anakin Skywalker". The second is "Anakin Solo". Do these refer to the same entity? Respond with ["True" or "False"]["Yes" or "No"].

Instance labels with types

We have two entities from different knowledge graphs. The first is "Anakin Skywalker", which is a "Characters". The second is "Anakin Solo", which is a "Characters". Do these refer to the same entity? Respond with ["True" or "False"]["Yes" or "No"].

Instance labels with types and attributes

We have two entities from different knowledge graphs. The first is "Anakin Skywalker", which is a "Characters" and is "gender" "Male" and is "species" "Human" and is "homeworld" "Tatooine". The second is "Anakin Solo", which is a "Characters" and is "gender" "Male" and is "species" "Human" and is "homeworld" "New Republic". Do these refer to the same entity? Respond with ["True" or "False"]["Yes" or "No"].

Instance labels with full context

We have two entities from different knowledge graphs. The first is "Luke Skywalker", which is a "Characters" and has attributes: is "gender" "Male" and is "era" "Rebellion era" and is "height" "1.72" and has relationships: is "species" "Human" and is "affiliation" "Rebel Alliance" and is "homeworld" "Tatooine". The second is "Luke Skywalker", which is a "Characters" and has attributes: is "gender" "Male" and has relationships: is "species" "Human" and is "profession" "Jedi" and is "homeworld" "Tatooine". Do these refer to the same entity? Respond with ["True" or "False"]["Yes" or "No"].

Instance labels with intersecting context

We have two entities from different knowledge graphs. The first is "Luke Skywalker", which is a "Characters" and has attributes: is "gender" "Male" and has relationships: is "species" "Human" and is "homeworld" "Tatooine". The second is "Luke Skywalker", which is a "Characters" and has attributes: is "gender" "Male" and has relationships: is "species" "Human" and is "homeworld" "Tatooine". Do these refer to the same entity? Respond with ["True" or "False"]["Yes" or "No"].

Instance labels with intersecting and entropy ranked context

We have two entities from different knowledge graphs. The first is "Luke Skywalker", which is a "Characters" and has attributes: is "gender" "Male" and has relationships: is "homeworld" "Tatooine" and is "species" "Human". The second is "Luke Skywalker", which is a "Characters" and has attributes: is "gender" "Male" and has relationships: is "homeworld" "Tatooine" and is "species" "Human". Do these refer to the same entity? Respond with ["True" or "False"]["Yes" or "No"].

3.4.3 Few-shot prompts

Our ontology-driven prompt construction framework supports few-shot in-context learning by injecting example question-answer pairs alongside model responses into the message history, separating the system prompt from the actual query. Each example question is formatted using the same template function as the query itself. Each example answer is provided according to the selected answer format. This ensures structural consistency between in-context examples and the task of resolving an uncertain mapping in $\mathcal{M}_{ask}$.

For matching tasks that provide a dedicated training split, positive examples are sampled from that split. For tasks without training data, positive examples are sampled from `LogMap`'s high-confidence anchors. Specifically, this relates to the subset of mappings from the initial alignment that `LogMap` accepted without consulting the oracle. These anchors are filtered by entity type to ensure compatibility with the corresponding prompt template. In any case, any question-answer pairs that happen to overlap with $\mathcal{M}_{ask}$ are excluded to prevent data leakage.

In constructing negative examples, we support both random and hard negatives. Random negatives are constructed by sampling two unrelated positive pairs and crossing their targets. Whereas, hard negatives are sampled by obtaining a positive pair $\langle e_1, e_2 \rangle$ and replacing e_1 with e'_1, where e'_1 is the sibling entity of e_1 whose embedding is most semantically similar to e_1 by cosine similarity. This should (in principle) produce a challenging near-miss $\langle e'_1, e_2 \rangle$. We compute sibling embeddings using SapBERT [47] for biomedical tasks and *all-MiniLM-L12-V2* [86] for other domains.

4 Experimental evaluation

We have performed an extensive evaluation with both commercial and open weight models, using local infrastructure and external resources like OpenRouter. When using external resources, all the experiments reported here were obtained with a budget of less than £100.

We used the datasets provided by the OAEI evaluation initiative [67, 68]. We used the *anatomy* [18], *largebio* [36], *bio-ml* [30], *conference*, and *knowledge graph* tasks. As shown in Table 1, the covered matching tasks involved ontologies of diverse sizes . The reference alignments ($\mathcal{M}^{RA}$) of

OAEI track	Matching task	$\|\mathcal{O}_1\|$	$\|\mathcal{O}_2\|$	$\|\mathcal{M}^{RA}\|$
Anatomy	**Mouse-Human**	2,755	3,313	1,516
Bio-ML	**NCIT-DOID**	15,991	8,516	4,686
	OMIM-ORDO	9,662	9,320	3,721
	SNOMED-FMA.body	34,562	89,180	7,256
	SNOMED-NCIT.neoplas	23,116	20,497	3,804
	SNOMED-NCIT.pharm	29,646	22,387	5,803
Largebio	**FMA-NCI**	79,049	66,919	3,024
	FMA-SNOMED	79,049	122,521	9,008
	SNOMED-NCI	122,521	66,919	18,844
Conference	$*$-$*$ ***(Avg.)***	109	120	15
Knowledge Graph	**marvelcinematicuniverse-marvel**	137,784	1,134,039	1,667
	memoryalpha-memorybeta	160,712	169,069	9,365
	memoryalpha-stexpanded	160,712	41,252	1,779
	starwars-swg	364,829	29,714	1,121
	starwars-swtor	364,829	15,735	1,429

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.

these matching tasks have different sources. For example, in *anatomy* and *conference*, the reference alignment has been manually curated, while in *bio-ml* and *largebio* the reference alignment relies on public resources like MONDO [77] and UMLS [8].

4.1 Evaluation with commercial models (zero-shot)

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). 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.

Diagnostic capability. We tested over the 9 bio matching tasks a total of 30 LLM-based Oracles ($\texttt{Or}^{LLM}$), combining the six prompt templates introduced in Section 3.4.1 Figure 3 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 Figure 4.

Oracles relying on the Gemini 2.5 Flash model led to the best results on average, as summarised in Figure 3. 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 2 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 for low, moderate and high effectiveness, respectively (*e.g.*, [48]). Due to the complexity of the mappings in $\mathcal{M}_{ask}$, moderate YI values can be expected of an Oracle.

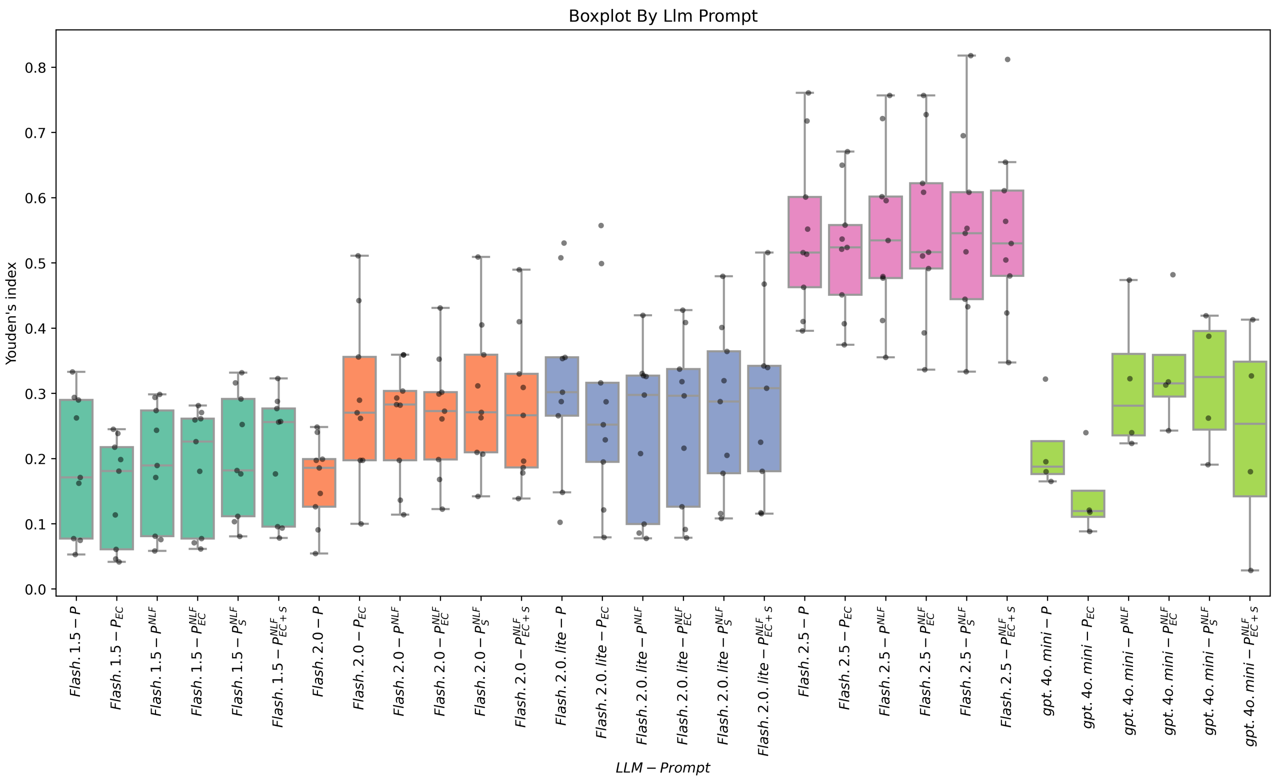

Figure 3: Summary of the diagnostic results (Youden's index) for the LLM-based Oracles.

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

Impact of the prompt template. Figure 3 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 $\mathbf{P}^{\text{NLF}}_{\text{S}}$ (natural-language friendly with synonyms). We emphasise that different ontologies and matching tasks pose varying challenges due to lexical and structural differences. Our tested prompts aim to capture this by leveraging both structural and lexical information in the input ontologies. There is also a dependency on

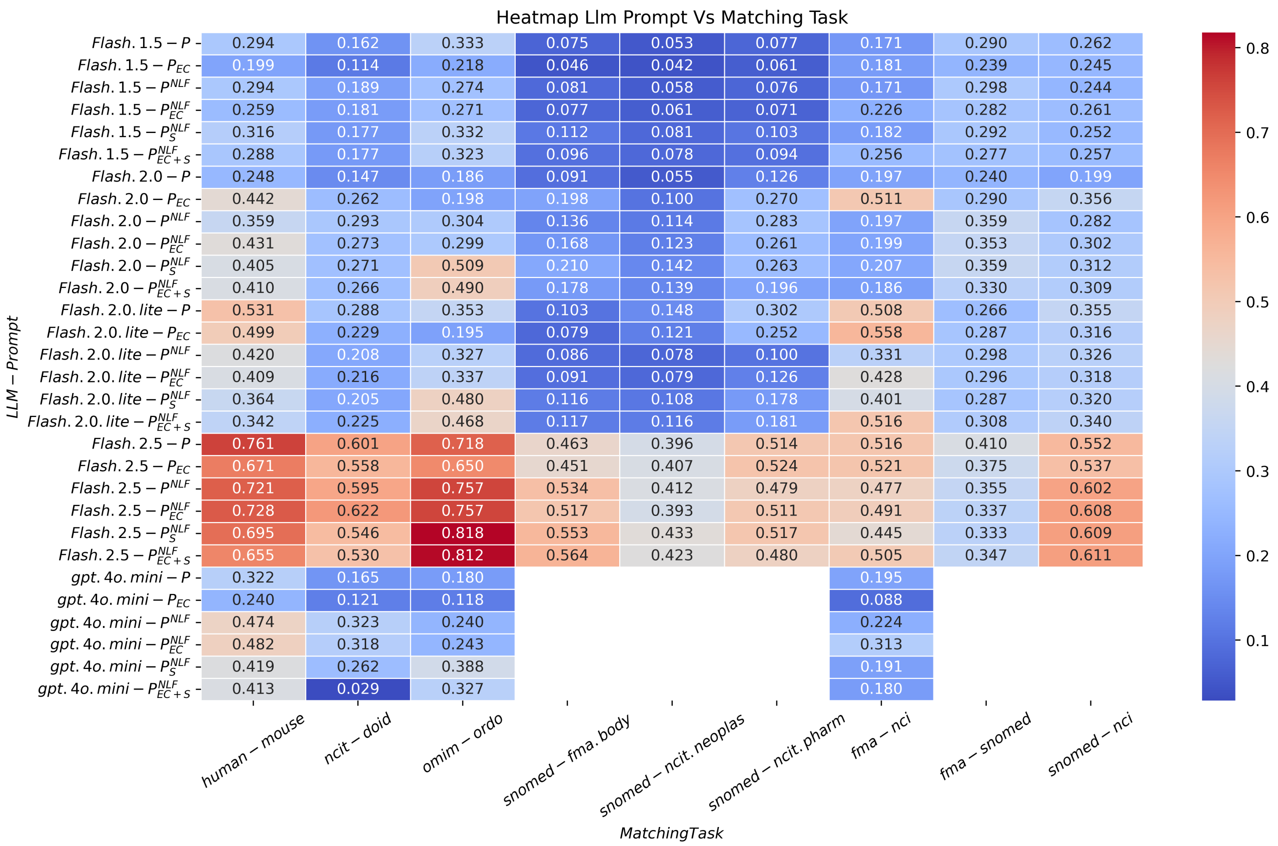

Figure 4: Diagnostic results (Youden's index) by the LLM-based Oracles over the selected ontology matching tasks. For example, *Flash 2.5*-$\mathrm{P}^{\mathrm{NLF}}_{\mathrm{EC+S}}$ represents the LLM-based Oracle relying on the Gemini Flash 2.5 model and evaluated with the natural-language friendly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

Matching task	**LogMap**			**LogMap - Or$^{LLM}_{GF2.0}$**			**LogMap - 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 ($\mathtt{Or}^{LLM}_{GF2.0}$ and $\mathtt{Or}^{LLM}_{GF2.5}$) on all matching tasks. Pr denotes Precision, Re Recall and F is the F-score.

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).

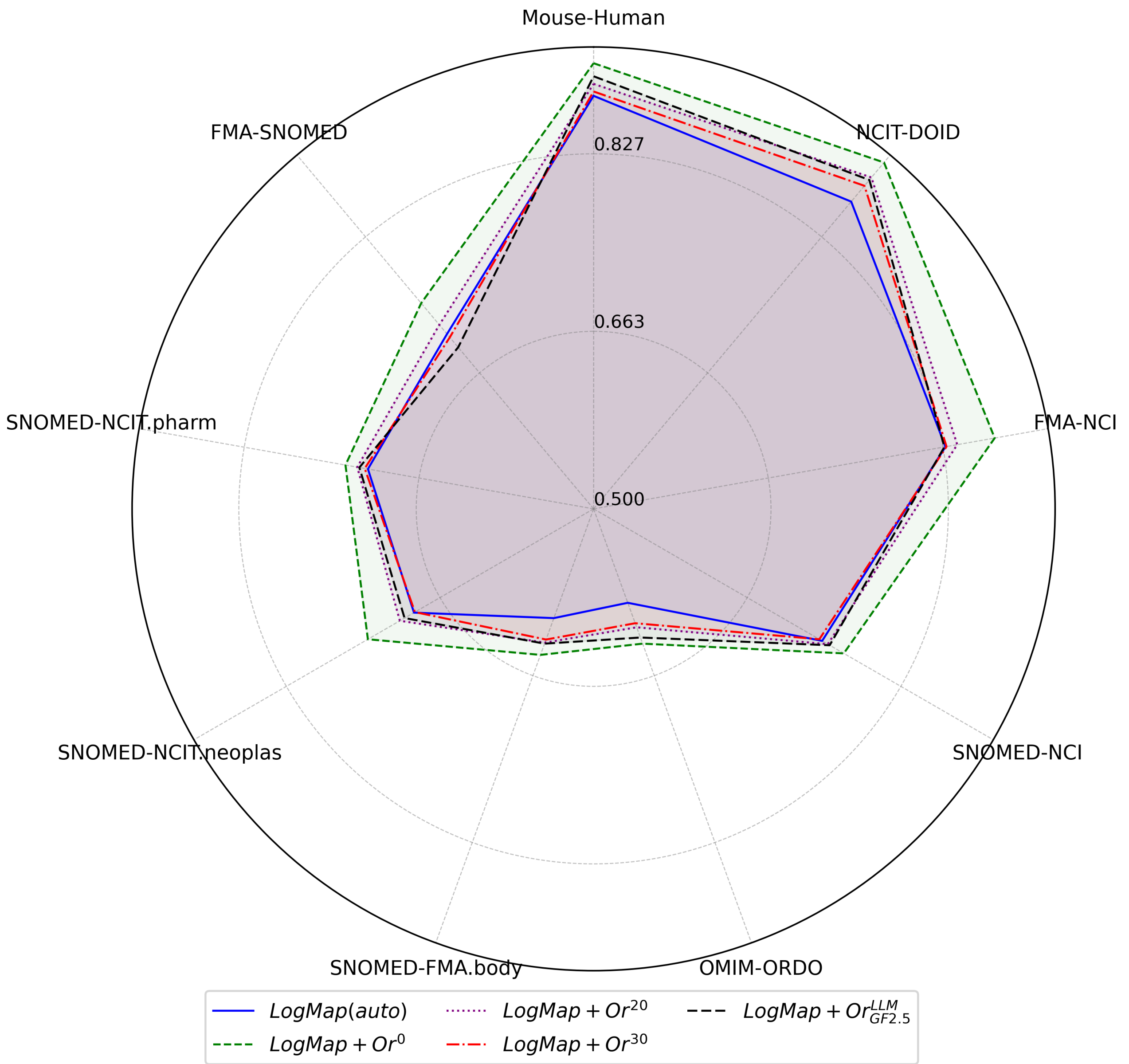

Figure 5: Comparison of LogMap, LogMap with $\texttt{Or}^{LLM}_{GF2.5}$, and LogMap in combination with Oracles with different error rates ($\texttt{Or}^{0}$, $\texttt{Or}^{20}$, and $\texttt{Or}^{30}$).

Overall matching task. Table 3 shows the results obtained using LogMap (automatic mode) compared to LogMap integrated with an LLM-based Oracle. We selected the top-performing LLM-based Oracles using $\mathbf{P}^{\mathrm{NLF}}_{\mathrm{S}}$ 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. LogMap+$\texttt{Or}^{LLM}_{GF2.5}$ dominates the *anatomy* and *bio-ml* tasks, while LogMap+$\texttt{Or}^{LLM}_{GF2.0}$ 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 [46]. In this work, we have evaluated 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 [46]. Figure 5 compares performance across all nine ontology matching tasks for: LogMap,

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 4: 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.

LogMap+$\texttt{Or}^{LLM}_{GF2.5}$, and LogMap with the simulated Oracles $\texttt{Or}^{0}$, $\texttt{Or}^{20}$, and $\texttt{Or}^{30}$ (corresponding to error rates of 0%, 20%, and 30%, respectively). We can observe that $\texttt{Or}^{LLM}_{GF2.5}$ performs similarly to $\texttt{Or}^{20}$, except in the FMA-SNOMED task (lower F-score) and OMIM-ORDO task (higher F-score). In line with previous studies [35, 46], LogMap+$\texttt{Or}^{30}$ still outperforms LogMap (without Oracle).

Comparison with OAEI systems. The results of LogMap+$\texttt{Or}^{LLM}_{GF2.5}$ are highly competitive when compared with the state-of-the-art systems participating in the OAEI campaign (see results in the OAEI 2021 [67] for *largebio*, and in the OAEI 2024 [68], and OAEI 2025 [69] for *anatomy* and *bio-ml*). For example, LogMap+$\texttt{Or}^{LLM}_{GF2.5}$ would have ranked top-3 in the 2024 *anatomy* track, top-2 in the 2021 *largebio* track, achieving performance comparable to leading systems such as BertMap [28], Matcha [23], and LogMap-Bio [13] in the 2024 *bio-ml* track. In the OAEI 2025, LogMap+$\texttt{Or}^{LLM}_{GF2.5}$ participated under the name LogMapLLM, achieving top-4 results in the *anatomy* track and top-2 results in the *bio-ml* track [38, 69]. Table 4 reports the official OAEI results and ranks achieved by LogMap and LogMapLLM in the *bio-ml* track.[12] LogMap-Bio[13] 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. 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).

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,

[12]Note that the F-scores in Table 3 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.

[13]LogMap-Bio [13] uses BioPortal as a source of mediating (biomedical) ontologies.

Table 5: Open-weight models evaluated locally.

Model	Total	Active	Architecture	Quantisation	Backend
Qwen3-32B	32B	32B	Dense	—	vLLM
Qwen3-235B-A22B	235B	22B	MoE	GPTQ-Int4	vLLM
Qwen3.5-35B-A3B	35B	3B	MoE	—	SGLang
Qwen3.5-122B-A10B	122B	10B	MoE	GPTQ-Int4	SGLang
Mistral-Small-3.2-24B	24B	24B	Dense	—	vLLM

indicating very limited evidence of leakage of the OAEI *bio-ml* ground truths.

4.2 Evaluation with open-weight models (zero-shot)

We have deployed locally and evaluated the open-weight models in Table 5. The requests are performed via vLLM [44] or SGLang [91] backend. The local environment runs on a workstation with the following specifications:

- **CPU:** AMD Ryzen 9975WX (24 cores, 48 threads, base clock: 4.2 GHz, boost clock: 5.4 GHz).
- **GPU:** 2 $\times$ NVIDIA RTX PRO 6000 Max-Q Edition 96GB (192GB available VRAM).
- **RAM:** 768GB DDR5 ECC RDIMM at 5600mhz.
- **OS:** Ubuntu 24.04.3 LTS; CUDA version 13.1.

By leveraging open-weight models, we can achieve competitive results, comparable to those of commercial models as reported in Section 4.1. For example, we achieve a global F1 score of 0.897 on Anatomy, and remain competitive (within 1% of the global F1 score compared to the best-performing commercial model). If configured properly, local models are often as efficient as commercial models when run on the specified hardware, since we avoid network latency, complex routing, and request queues. For instance, we conducted 277 oracle consultations in less than 3 seconds (wall-clock time). Moreover, running locally maintains data sovereignty. All local experiments can be reproduced on an air-gapped machine, ensuring privacy.

Evaluation on Bio-tracks. We evaluated the new prompts on class mappings introduced in Section 3.4.2 on the *anatomy* and *bio-ml* tracks. We compared the results against the original prompts *natural-language friendly with synonyms* and *natural-language friendly with synonyms and extended context* introduced in Section 3.4.1 (baseline prompts).

We evaluated the new prompts on the class mappings presented in Section 3.4.2 across the *anatomy* and *bioml* tracks. We compared the results with the original prompts-*natural-language friendly with synonyms* and *natural-language friendly with synonyms and extended context*-introduced in Section 3.4.1 (baseline prompts). Table 6 summarises the average results in terms of Youden's Index, where the new prompts show an important improvement for the evaluated open-weight models.

Evaluation on the OAEI Conference track. This dataset includes seven small ontologies from the conference organisation domain: Cmt, ConfTool, Edas, Ekaw, Iasted, Sigkdd, and Sofsem. Reference alignments are available for only 21 ontology pairs. We used the public reference alignments ra1-M3.

To test the *conference* track, we created several prompt combinations by taking the product between the set of class prompts and the set of property prompts. This results in the following prompt combinations:

- **PC1:** cls-syns+pars+sibs/P:D+R
- **PC2:** cls-syns+pars+sibs/P:D+R+Syn
- **PC3:** cls-syns+pars+sibs/P:Labels

Table 6: Oracle diagnostic quality on *anatomy* and *bioml* (macro-averaged Youden's J across tasks, yes/no answer format).

Model	Baseline Prompts		New Prompts				
	syns (only)	syns+pars	syns+pars+sibs	subs (syns)	subs (syns+pars)	subs (ans-disj)	deductive
Qwen3-32B	0.466	0.489	0.493	**0.543**	0.477	0.472	0.536
Qwen3-235B	0.412	0.453	0.493	0.537	**0.567**	0.549	0.448
Qwen3.5-35B	0.494	0.495	0.504	**0.593**	0.568	0.568	0.544
Qwen3.5-122B	0.532	0.552	0.551	**0.621**	0.597	0.596	0.589
Mistral-3.2-24B	0.447	0.427	0.430	**0.529**	0.522	0.469	0.443

- **PC4:** subs-AncDis/P:D+R
- **PC5:** subs-AncDis/P:D+R+Syn
- **PC6:** subs-AncDis/P:Labels
- **PC7:** cls-syns+pars/P:D+R
- **PC8:** cls-syns+pars/P:D+R+Syn
- **PC9:** cls-syns+pars/P:Labels
- **PC10:** subs-syns/P:D+R
- **PC11:** subs-syns/P:D+R+Syn
- **PC12:** subs-syns/P:Labels
- **PC13:** cls-syns/P:D+R
- **PC14:** cls-syns/P:D+R+Syn
- **PC15:** cls-syns/P:Labels

The class prompts are shown on the left-hand side of the / symbol; whereas the property prompts are provided on the right-hand side of the symbol (see Section 3.4.2). We obtain the following results for Youden's Index J on the $\mathcal{M}_{ask}$ set, shown in Table 7, and results for global F_1 shown in Table 8. `LogMap`'s global F_1 at the OAEI 2025 was 0.680. By carefully selecting an appropriate prompt combination, we can slightly improve on this score by +1%. This improvement is significant as: *(i)* we are using open-weight models, and *(ii)* the upper-bound performance with a perfect Oracle for `LogMap` in the *conference* track is 0.72 [46].

Table 7: Conference Track: Micro-averaged Youden's J on the $\mathcal{M}_{ask}$ set.

Model	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13	PC14	PC15
Qwen3-32B	0.539	0.539	0.240	0.333	0.333	**0.667**	0.468	0.468	0.169	**0.667**	**0.667**	**0.667**	0.539	0.539	0.169
Qwen3-235B	0.403	0.403	0.221	**0.667**	**0.667**	**0.667**	0.403	0.403	0.221	**0.667**	**0.667**	**0.667**	0.474	0.474	0.292
Qwen3.5-35B	0.695	**0.766**	0.208	0.333	0.333	0.000	0.623	0.552	0.065	0.667	0.667	0.667	**0.766**	**0.766**	0.377
Qwen3.5-122B	0.429	0.429	0.292	**0.667**	**0.667**	**0.667**	0.403	0.474	0.266	**0.667**	**0.667**	**0.667**	0.474	0.429	0.292
Mistral-3.2-24B	0.558	0.604	0.123	0.667	**1.000**	0.667	0.532	0.604	0.097	0.667	0.667	0.667	0.721	0.675	0.240

Evaluation on the OAEI Knowledge Graph track. This track comprises nine independent knowledge graphs with both instance- and schema-level data, aiming to align entities at both levels. The track involves 5 large alignment tasks. Table 9 shows the Youden's Index J on the $\mathcal{M}_{ask}$ set. As expected, the best prompt template was the one using the information content entropy to select the most meaningful attributes for each instance to be used in the prompt (see Section 3.4.2). It is worth mentioning that adding context to the instance prompts is only beneficial when selecting the right attributes (*e.g.*, entropy-based); otherwise, the performance degrades in some tasks.

Table 8: Conference Track: Micro-averaged global F_1 on the $\mathcal{M}_{ask}$ set.

Model	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13	PC14	PC15
Qwen3-32B	**0.683**	**0.683**	0.676	0.674	0.674	0.677	0.681	0.681	0.673	0.677	0.677	0.677	**0.683**	**0.683**	0.673
Qwen3-235B	0.679	0.679	0.675	0.677	0.677	0.677	0.679	0.679	0.675	0.677	0.677	0.677	**0.682**	**0.682**	0.678
Qwen3.5-35B	0.687	**0.690**	0.676	0.674	0.674	0.672	0.685	0.682	0.673	0.677	0.677	0.677	**0.690**	**0.690**	0.679
Qwen3.5-122B	0.681	0.681	0.678	0.677	0.677	0.677	0.679	**0.682**	0.677	0.677	0.677	0.677	**0.682**	0.681	0.678
Mistral-3.2-24B	0.681	0.682	0.672	0.677	0.680	0.677	0.679	0.682	0.670	0.677	0.677	0.677	**0.686**	0.685	0.676

Table 9: KG Track: Oracle discrimination (Youden's $J = Se + Sp - 1$) on $\mathcal{M}_{ask}$, macro-averaged across 5 KG tasks.

Model	I1	I2	I3	I4	I4-int	I4-ent
Qwen3.5-122B	0.3925	0.4749	0.4701	0.5790	0.5692	0.5895
Qwen3.5-35B	0.4443	0.4478	0.3149	0.3299	0.4232	0.3237
Qwen3-32B	0.3896	0.2796	0.2954	0.3105	0.2941	0.3281
Qwen3-235B	0.3099	0.1967	0.1290	0.3344	0.3571	0.5248
Mistral-3.2-24B	0.6126	0.5110	0.4812	0.4739	0.4886	0.4561
Macro avg.	0.4298	0.3820	0.3381	0.4055	0.4264	0.4444

Prompts: I1=Labels only; I2=Labels+types; I3=Types+attributes; I4=Full context; I4-int=Full context (shared predicates); I4-ent=Full context (entropy-ranked).

4.3 Evaluation with few-shot prompts

We focused the few-shot evaluation on small models to better understand the effects on model performance. We tested from 1 to 75 shots, using dense intervals for low shot counts (1–10) and sparse intervals for higher counts. For example, Figure 6 shows the (positive) impact of the number of few-shot samples for Gemini 2.0 Flash Lite on the *bio-ml* omim-ordo task.

Performance trends. Across the majority of tasks and models, we observed a consistent non-linear performance curve summarised as follows:

1. **Immediate Gain:** The most significant marginal gain occurs between 0 and 5 shots.
2. **Intermediate Drop:** As the context grows (5–15 shots), we often observed a plateau or a slight

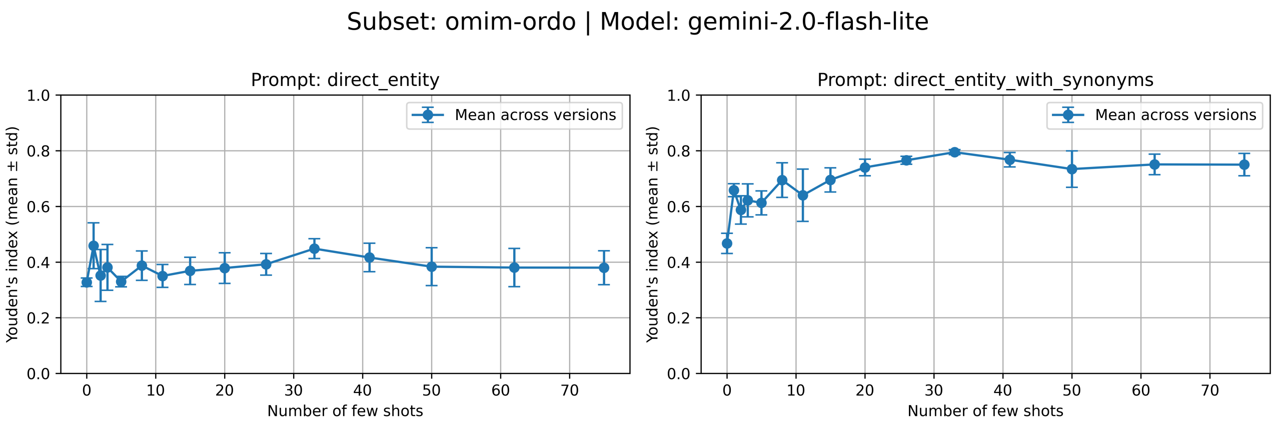

Figure 6: Impact of the Number of few-shot samples for Gemini 2.0 Flash Lite on the *bioml* omim-ordo task.

dip in performance, likely due to the model struggling to attend to a growing list of examples without a strong pattern.

3. **Global Peak (25–35 shots):** Performance typically recovers and reaches a global maximum between 25 and 35 examples. Beyond this point, adding more shots yields diminishing returns or degrades performance due to context saturation.

The "Cold Start" in small models. An interesting finding in this study concerns the behaviour of smaller open-weight models (e.g., `ministral-3b-2512`). In zero-shot settings, these models frequently failed completely, yielding Youden's Index scores of 0. However, providing just a few-shot samples was often sufficient to "jump-start" the model, raising performance from near-zero to a competitive baseline (see Figure 7). This suggests that for small models, few-shot examples serve as critical instruction guidelines.

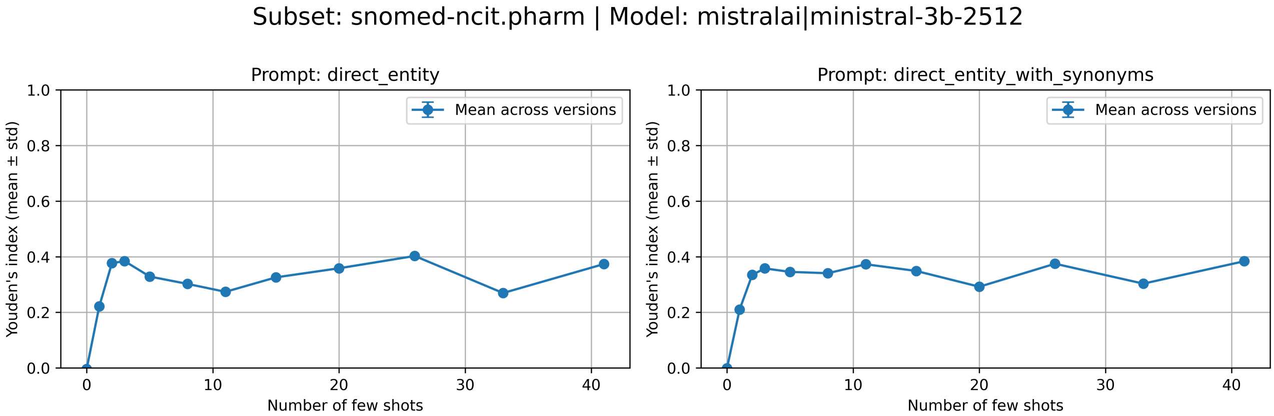

Figure 7: Cold Start behavior of `ministral-3b-2512` on snomed-fma.body task.

Stability and variance. By averaging results across 10 different permutations of few-shot samples, we analysed the stability of the models. While most tasks showed deterministic behaviour with low standard deviation, certain complex tasks in *largebio* exhibited high variance, where specific samples drastically improved output while others did not (see Figure 8). Thus, the right selection of shots seems to be critical for some tasks. Despite this noise, the general trend lines (peaking at 25–35 shots) remained consistent.

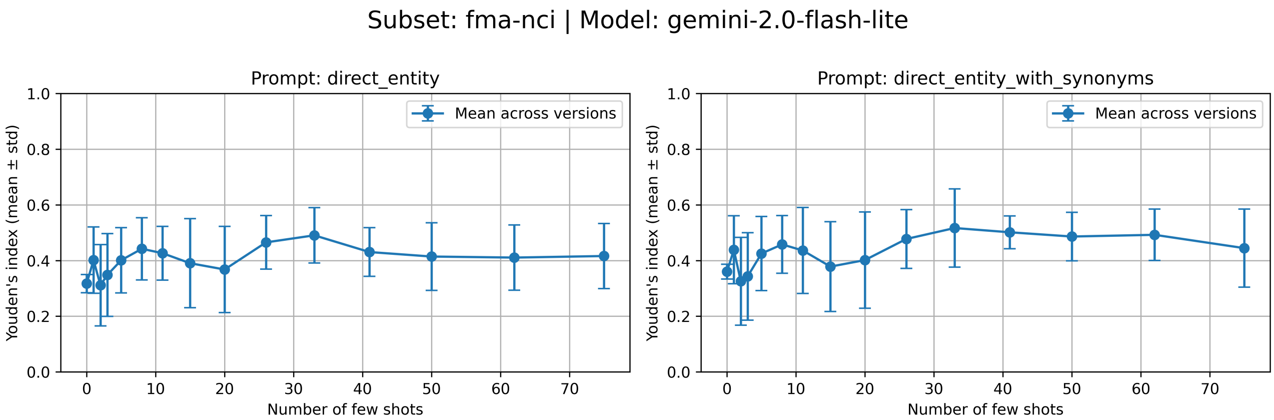

Figure 8: High results variance with preserved common trend on fma-nci task.

Context Window Limitations. We found that "more is not always better" for some models (See Figure9). Pushing the shot count towards 75 often degraded performance, particularly for models with smaller context windows or weaker attention mechanisms. Consequently, we identify the range of **25–35 shots** as the optimal trade-off between maximising performance and maintaining reasonable inference costs and latency.

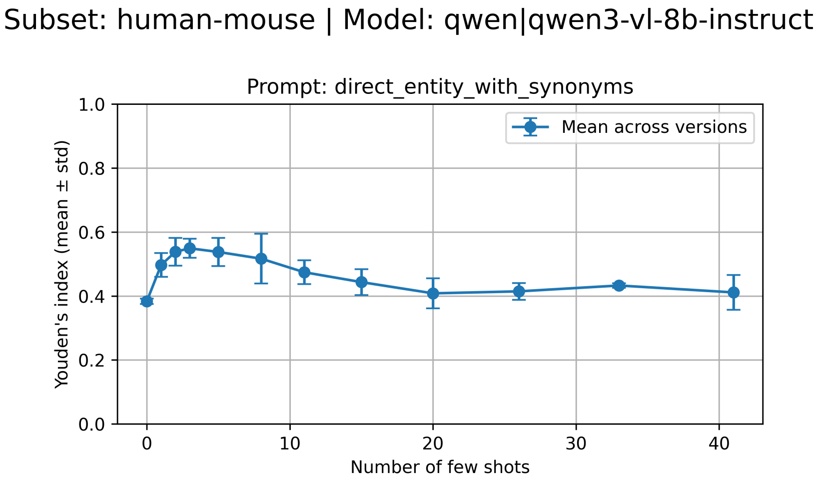

Figure 9: Performance deprecation with more samples added for Qwen 3 8B model.

5 Defence, Intelligence and Security Ontologies (DISO)

DISO (Defence, Intelligence and Security Ontologies)[14] is a collection of 60+ publicly available Web Ontology Language (OWL) [15] ontologies relating to the domains of defence, intelligence and national security. The ontologies in DISO were identified and obtained during an ontology search exercise undertaken primarily during Nov/Dec of 2025.

5.1 DISO clusters (subdomains)

Figure 10 shows an overview of the identified clusters (or subdomains) and the conceptual overlaps that exist among them. The structure of the DISO repository reflects this clustering. If an ontology fits into more than one cluster, it only appears in the cluster where the fit is deemed strongest. In the DISO repository, each ontology is documented by providing information such as a full name, short description, relevant context, web presence, and the srce from which the ontology was obtained.

Table 10 shows the 11 clusters (or subdomains) of the DISO collection in alphabetic order, along with the number of ontologies assigned to each cluster.

Three factors make providing precise counts of the number of ontologies in the DISO collection, and in its component clusters (subdomains), somewhat awkward. First, many of the ontologies obtained for DISO are physically represented as networks of multiple component ontology files. SOUPA [9], for example, an ontology for ubiquitous and pervasive computing whose development was partly funded by DARPA, and which is assigned to the DISO cluster 'situation awareness', is presented as a network of 18 component ontology files. For DISO ontology counting purposes, however, we count SOUPA as a single ontology. Hence, in this respect, the ontology counts reported in Table 10 significantly understate the number of physically distinct ontology files present within the DISO repository. Second, to make it easy for researchers to reuse such networked DISO ontologies,

[14] https://github.com/city-artificial-intelligence/diso/

Table 10: A view of the conceptual structure of the DISO collection of defence, intelligence and security ontologies. The clusters (or subdomains) of the DISO collection are listed together with the number of ontologies assigned to each. An ontology assigned to more than one DISO cluster is physically stored in a primary cluster, and is present only virtually in the alternate cluster.

DISO Cluster Name	Ontologies Present Physically	Ontologies Present Virtually	Ontology Present Virtually	Virtually Present Ontology Primary DISO Cluster	DISO Cluster Size	DISO Subcluster Size
agentic	1	0			1	
ambient intelligence	0	1	BOnSAI	smart buildings	1	
context awareness	2	1	SOUPA	situation awareness	3	
cyber-security	15	0			15	
digital twins	1	0			1	
information exchange	2	0			2	
information security	5	0			5	
mid-level	12[1]	0			12	
risk management	1	0			1	
robotics	1	1	OntoSecRPA	information security	2	
situation awareness	3	1	JC3IEDM	information exchange	4	
smart environments	8	0			8	
» smart buildings						4
» smart cities						2
» smart homes						2
upper-level	8[2]	0			8	
totals	59	4			63	

[1] The mid-level ontology IES-core is physically stored with the rest of IES in the 'information exchange' DISO cluster, but it is counted as a mid-level ontology, not as an 'information exchange' ontology.
[2] The upper-level ontology IES-top is physically stored with the rest of IES in the 'information exchange' DISO cluster, but it is counted as an upper-level ontology, not as an 'information exchange' ontology.

we have frequently merged such networks into single, all-inclusive ontology files. This serves to increase the number of ontology files physically present in the DISO repository, but this should not be allowed to change the number of meaningfully distinct ontologies that we regard as comprising DISO. Third, some ontologies within DISO fit naturally into more than one DISO cluster. For example, the SOUPA ontology is regarded as being a member of both the 'situation awareness' cluster and the 'context awareness' cluster. This complicates the counting of DISO ontologies because, as seen in Table 10, a sum of the cluster sizes somewhat overstates the number of unique ontologies distributed across the DISO clusters.

Situation awareness The left half of Figure 10 is anchored around the notion of situation awareness [20, 53]. Briefly, situation awareness has to do with perceiving and understanding the elements of a dynamic environment, and the relations between them, whilst considering both space and time, and the prediction of the environment's future state(s). The generality of the notion of situation awareness is reflected in the number of other DISO clusters with which it has close conceptual associations.

Context awareness The DISO cluster 'context awareness', for instance, has primarily to do with individuals and their immediate surroundings: their nearby environmental context. The common thread (and motivation) shared by the ontologies in this cluster is the idea of pervasive computing, where the objective is for mobile application services to be able to adapt their behaviour dynamically based on an application user's perceived current environmental context. Thus, the notion of context awareness can be regarded as a specialised form of situation awareness, where the situation of interest is the surroundings of a mobile, software application user.

Smart environments It is similarly reasonable to regard smart environments (whether they be homes, or buildings, or cities) as being specialised forms of situations: built environment situations of various size and scale. The notion of 'smart' links to the notion of 'awareness' because the ontologies

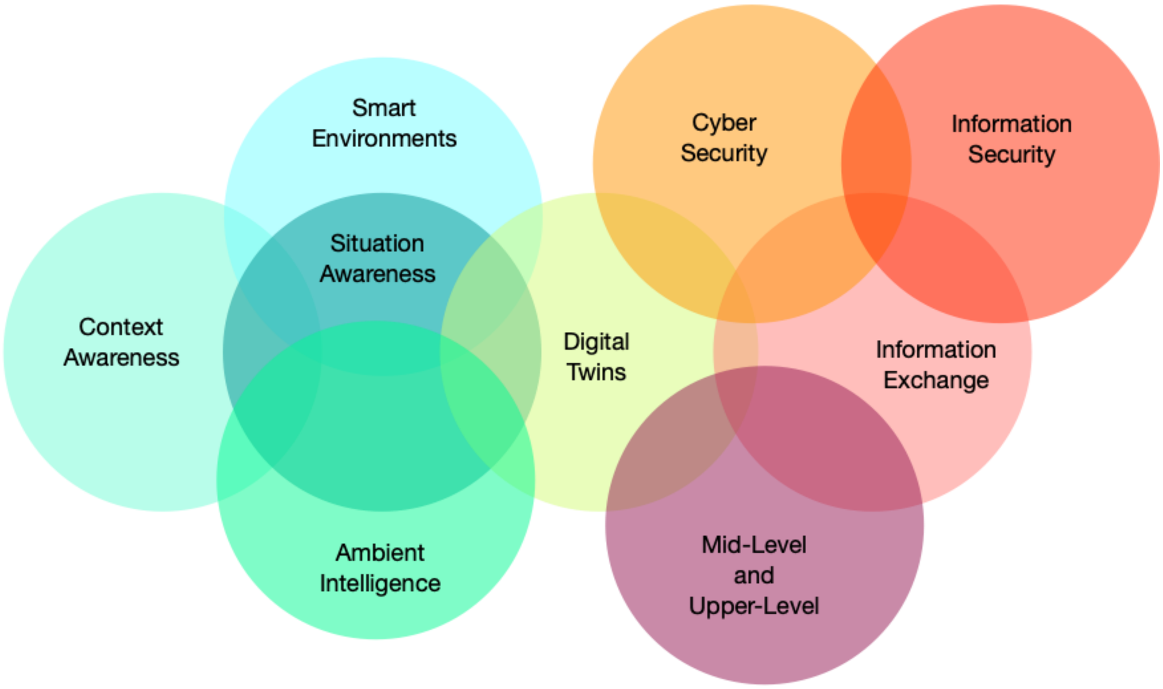

Figure 10: A view of the conceptual overlaps that exist between some of the clusters (or subdomains) of the ontologies within the DISO (Defence, Intelligence and Security Ontology) collection.

of the DISO 'smart environments' cluster generally presume some amount of sensor-based monitoring of the environment in question. For example, the DISO 'smart homes' ontology ThinkHome [73] has an emphasis on the monitoring and control of home energy use.

Ambient intelligence Ambient intelligence [6] has to do with environments featuring devices that seamlessly interconnect and collaborate with each other as well as with human users. Like context awareness, ambient intelligence supports the idea of pervasive computing, but with more attention to supporting ideas like the Internet of Things (IoT). BOnSAI [80] is a smart building ontology for ambient intelligence, and is assigned to both the 'smart buildings' and 'ambient intelligence' DISO clusters.

Digital twins The UK's National Digital Twin Programme (NDTP)[15] defines a *digital twin*[16] as a digital representation of a physical thing, environment or process that exchanges data bi-directionally with the physical world, including with its physical counterpart. Further, a digital twin must replicate its physical counterpart without statistical bias, and it must be able to predict how the physical counterpart will respond to a given stimulus. Ontologies are well suited to supporting the realisation of digital twins [51], and their use in connection with digital twins has been examined in a recent survey [40]. Ontologies facilitate data exchange between internal models and digital twin service programs, and between the digital twin and external systems, agents and users. They can support distinguishing between domain knowledge and operational knowledge; they can support digital twin knowledge bases; and they can support the monitoring of digital twin data and state integrity and consistency. The notions of situation awareness, smart environments and ambient intelligence relate strongly to the notion of digital twins. Indeed, what DISO refers to as smart environments (smart homes, buildings and cities) are prime examples of the types of physical things that the UK's NDTP envisages having digital twins. SAREF (a Smart Applications REFerence ontology)[17], created by

[15] https://ndtp.co.uk
[16] https://ndtp.co.uk/digital-twin-definition/
[17] https://saref.etsi.org/

the European Telecommunications Standards Institute (ETSI), is a suite of ontologies (a core with extensions for vertical domains) for enabling semantic interoperability between disparate providers. It is one of the main ontologies discussed in connection with creating digital twins for buildings in [51]. Digital twins deserve to feature in the DISO collection. The NDTP's notion of a UK National Digital Twin (NDT) that integrates disparate individual digital twins as nodes in a national digital fabric ranks as critical national infrastructure with clear national security implications. And one recent paper explores the integration of digital twin technologies directly within a variety of defence applications [24].

Information security Information security has to do with protecting information generally, in all its forms, whether physical or digital, tangible (*e.g.*, paperwork) or intangible (*e.g.*, knowledge), from things like unauthorised access or inappropriate use, modification, disclosure or corruption. Its guiding principles are known as the CIA triad: Confidentiality, Integrity and Availability [76]. The DISO 'information security' cluster contains five ontologies covering this domain. One of these, MDISOnt [54], a multi-dimensional information security ontology, decomposes information security into several perspectives using dimensional views and modules.

Cyber-security Cyber-security is readily understood as being a specialised form of information security whose focus is on the digital domain and the securing of assets against digital threats. Three recent surveys of cyber-security ontologies [49, 74, 78] attest to the fact that the cyber-security domain appears to appeal to ontologists. The DISO collection includes 15 such ontologies that feature prominently in the literature.

Information exchange Ontologies have long been linked with the notion of knowledge sharing [26]. And the notion of information exchange clearly links with knowledge sharing. So it should not be surprising that ontologies can also play a role in information exchange. Such ontologies seek to standardise data representations so as to facilitate the exchange (and aggregation) of data across multiple, diverse domains, sectors, organisations and systems. The DISO collection contains two authoritative information exchange ontologies. One, IES (Information Exchange Standard)[18], is a standard for information exchange developed within the UK Government. DISO contains the latest version of IES, version 5, released in November 2025. The IES plays a critical role in enabling the UK's vision of a National Digital Twin (NDT), since the IES is fundamental to enabling the NDTP's Information Management Framework (IMF) that is responsible for securely and reliably interconnecting all of the individual digital twin nodes comprising the UK's NDT. The other information exchange ontology in DISO is JC3IEDM (Joint Consultation, Command and Control Information Exchange Data Model), jointly developed by the European Defence Agency and NATO.

Mid-level and Upper-level All of the DISO clusters described to this point contain domain-level ontologies only. But DISO contains clusters for mid-level and upper-level ontologies as well. It does so because mid-level and upper-level ontologies featured so prominently in *(i)* the ontology search exercise for defence, intelligence and security ontologies from which DISO emerged, and in *(ii)* the UK's vision of a National Digital Twin (NDT). One explanation for this is the greater generality of the concepts described in mid-level and upper-level ontologies. Upper-level ontologies seek to describe concepts general enough to embrace (subsume) all domain-level ontologies. Mid-level ontologies specialise upper-level ontologies for particular (vertical) domains, and thereby play a bridging role between (general) upper-level ontologies and (specialised) domain-level ontologies. The value of the greater generality of mid-level and upper-level ontologies lies in the prospect they offer for enabling the integration of more diverse, heterogeneous knowledge bases than might otherwise have been

[18] https://informationexchangestandard.org/

Table 11: A comparison of the (partial) distributions of alignment size (measured in entity mapping counts) for two different sets of DISO ontology pairs, as generated by one ontology matching system.

	1653 DISO ontology pairs			105 DISO cyber-security ontology pairs		
Alignment Mapping Count	Ontology Pair Count	Ontology Pair Count Cumulative	Cumulative Pair Count As Proportion of Total Pairs	Ontology Pair Count	Ontology Pair Count Cumulative	Cumulative Pair Count As Proportion of Total Pairs
0	1006	1006	0.61	39	39	0.37
1	213	1219	0.74	17	56	0.53
2	107	1326	0.80	13	69	0.66
3	90	1416	0.86	11	80	0.76
4	48	1464	0.89	3	83	0.79
5	29	1493	0.90	3	86	0.82
6	31	1524	0.92	6	92	0.88
7	17	1541	0.93	2	94	0.90
8	14	1555	0.94	1	95	0.90
9	13	1568	0.95	0	95	0.90
10	12	1580	0.96	3	98	0.93
11	5	1585	0.96	1	99	0.94
12	5	1590	0.96	0	99	0.94
13	9	1599	0.97	0	99	0.94
14	4	1603	0.97	0	99	0.94
15	5	1608	0.97	0	99	0.94
...	...	...	...	...	...	...

conceivable [52]. Evidence of the criticality of mid-level and upper-level ontologies for enabling integration of diverse, heterogeneous knowledge sources is evidenced by the fact that the version 5 of the UK Government's IES information exchange standard, which underpins the UK's vision for a National Digital Twin, includes 3 levels of ontologies: domain-level ('ies-common'), mid-level ('ies-core'), and upper-level ('ies-top').

5.1.1 DISO documentation

To help researchers comprehend the DISO collection, and to thereby facilitate its reuse, the DISO repository contains comprehensive documentation in the form of markdown README files. Each ontology is accompanied by its own README file that provides information such as: a full name, short description, relevant context, web presence, and the source from which the ontology was obtained. If an ontology has been assigned to more than one DISO cluster, the ontology's README makes this clear. Each DISO cluster and subcluster folder contains a README file that introduces the cluster or subcluster, describing something of the character/nature of the subdomain, to help provide conceptual context for DISO users. If we merged a network of ontology files into a single, all-inclusive ontology file, the README file declares this and clearly specifies the name the merged ontology file. All such merging was undertaken using the Protege ontology editor. In some cases, supplementary material (such as .pdf ontology guides or papers) is provided. Where applicable, cluster README files also highlight other DISO clusters regarded as having strong conceptual overlaps.

As is perhaps apparent from the range of cluster (subdomain) names appearing in Table 10, the umbrella domains of defence, intelligence and national security have been interpreted broadly. In cases where the connection between a DISO cluster and the umbrella domains might appear tenuous, the cluster README files provide supplementary information and cite sources to help explain and justify our decision to associate a cluster with the themes of defence, intelligence and security.

5.2 A network of DISO ontologies

We conducted a bulk ontology alignment exercise using `LogMap` over 58 DISO ontologies, resulting in the (potential) generation of alignments for 1653 ontology pairs. Table 11 compares the distribu-

	task id	source ontology	target ontology	alignment mappings	m_ask mappings	coverage source	coverage target	simplicity	labels_1 source	labels_1 target	labels_1b source	labels_1b target	labels_2 source	labels_2 target	comments source	comments target	logmap anchors	logmap discards	logmap hard discards	logmap logical conflicts	logmap over-estimations	score sum minus over-estimations	logmap anchors as proportion of alignment mappings
0	199	JC3IEDM	mlOmerged	219	40	0.057	0.112	0.023	1.0	1.023	0.973		0.0	0.0	1.0	0.005	10	9	914	0	1143	-1	0.046
1	267	uco1-merge	stix-spec-merged	131	43	0.075	0.237	0.496	1.0	0.992	0.969		0.0	0.0	0.634	0.992	32	0	113	24	269	-1	0.244
2	1441	ies-core	ies-common	105	29	0.474	0.145	0.8	0.971	0.99			0.0	0.0	0.79	0.99	64	0	254	0	359	0	0.610
3	1145	CurrencyUnitOntology	Brick+imports	74	36	0.569	0.033	0.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0	0	0	0	74	0	0.000
4	210	JC3IEDM	Brick+imports	49	69	0.013	0.022	0.082	1.0	2.265	0.837		0.0	0.0	1.0	0.061	3	26	1789	0	1866	-2	0.061
5	172	JC3IEDM	FacilityOntology	44	23	0.011	0.467	0.0	1.0	1.0	0.25	0.0	0.0	0.0	1.0	0.0	20	0	272	0	316	0	0.455
6	1217	SOmerged	SecAOnto_V4	44	23	0.041	0.145	0.727	0.023	0.0			0.0	0.0	0.182	0.023	32	5	200	14	264	-1	0.727
7	1619	stix-spec-merged	d3fend	37	31	0.053	0.006	0.514	0.973	1.0		0.405	0.0	0.0	0.973	0.0	5	0	603	0	641	-1	0.135
8	1252	bonsaiMerged	mlOmerged	35	29	0.143	0.015	0.943	0.029	1.486			0.0	0.0	0.2	0.171	21	3	119	0	158	-1	0.600
9	269	uco1-merge	d3fend	34	57	0.019	0.006	0.706	1.0	1.0	1.0	0.559	0.0	0.0	0.971	0.029	12	12	1669	57	1773	-1	0.353
10	1563	ThinkHomeMerged	Brick+imports	33	42	0.011	0.014	0.515	0.333	1.697		0.515	2.667	0.0	0.303	0.152	18	34	1385	195	1648	-1	0.545
11	1582	Brick+imports	SmartEnvMerged	32	45	0.01	0.049	0.594	1.625	1.312			0.0	2.75	0.312	0.656	21	3	269	13	319	-2	0.656
12	466	cwe	capec	32	5	0.444	0.525	0.875	1.0	1.031	0.938	0.984	0.0	0.0	0.0	0.0	3	0	1	0	33	0	0.094
13	1395	mlOmerged	SmartEnvMerged	31	31	0.014	0.057	0.903	0.645	1.419			0.0	0.0	0.258	1.0	18	3	189	8	233	-2	0.581
14	1612	saref4cityMerge	SmartEnvMerged	31	25	0.127	0.053	0.806	0.935	1.129			0.0	0.032	0.774	0.903	18	0	44	11	88	-2	0.581
15	1388	mlOmerged	Brick+imports	30	91	0.015	0.013	0.733	0.833	0.833			0.0	0.0	0.367	0.167	18	10	927	5	972	0	0.600
16	1581	Brick+imports	d3fend	30	22	0.013	0.005	0.533	0.933	1.0		0.867	0.0	0.0	0.167	0.0	10	8	1954	0	1994	-2	0.333
17	1543	ops-merged	SmartEnvMerged	29	29	0.035	0.053	0.586	0.897	1.414			0.0	0.0	0.414	0.931	16	0	151	16	196	0	0.552
18	273	uco1-merge	ies-common	28	10	0.015	0.039	0.786	1.0	1.0	1.0	0.679	0.0	0.0	1.0	1.0	24	6	270	35	340	-1	0.857
19	1372	saref4bldgMerge	SmartEnvMerged	26	24	0.045	0.035	0.692	0.923	1.038			0.0	0.077	0.923	0.923	12	0	50	7	84	-1	0.462
20	1521	CityOWLmerged	Brick+imports	24	27	0.016	0.01	0.583	0.708	1.917			3.708	0.0	0.0	0.167	13	3	613	8	648	0	0.542
21	1385	mlOmerged	ops-merged	24	18	0.012	0.032	0.542	0.875	0.708			0.0	0.0	0.333	0.417	11	0	433	2	460	-1	0.458
22	1578	Brick+imports	saref4cityMerge	24	16	0.009	0.107	0.625	0.792	0.958			0.0	0.0	0.167	0.792	10	13	144	0	182	-1	0.417
23	964	soupa-merged	mlOmerged	23	38	0.062	0.009	0.696	0.652	0.348			0.0	0.0	0.0	0.304	13	6	118	3	152	-2	0.565
24	1536	ops-merged	Brick+imports	22	21	0.028	0.009	0.545	0.955	1.909			4.0	0.0	0.227	0.182	11	4	506	6	540	-2	0.500
25	1212	SOmerged	d3fend	21	40	0.02	0.003	0.524	0.048	1.0		0.429	0.0	0.0	0.238	0.0	19	11	1172	2	1207	-1	0.905
26	1633	d3fend	SmartEnvMerged	21	25	0.002	0.043	0.857	1.0	1.286	0.905		0.0	0.0	0.0	0.857	15	3	349	139	514	-2	0.714
27	1585	Brick+imports	ies-common	21	21	0.008	0.024	0.476	0.905	0.952			0.0	0.0	0.286	0.952	15	8	602	1	632	0	0.714
28	1384	mlOmerged	CityOWLmerged	21	19	0.01	0.011	0.714	0.857	0.571			0.0	0.0	0.238	0.095	6	0	252	4	278	-1	0.286
29	874	SMOF	ArtifactOntology	21	11	0.037	0.036	0.0	0.0	1.0		0.0	0.0	0.0	0.048	0.048	10	0	230	22	271	2	0.476
30	1046	UnitsOfMeasureOntology	Brick+imports	21	8	0.101	0.009	0.0	1.0	2.524	0.0	0.048	0.0	0.0	0.0	0.048	0	0	26	0	47	0	0.000
31	982	soupa-merged	SmartEnvMerged	20	45	0.046	0.033	0.9	0.55	1.55			0.0	0.0	0.0	1.1	14	11	72	4	106	1	0.700
32	1251	bonsaiMerged	saref4bldgMerge	20	15	0.072	0.038	0.75	0.05	1.0		0.7	0.0	0.0	0.55	1.0	8	0	23	0	43	0	0.400
33	900	SMOF	Brick+imports	19	24	0.033	0.009	0.789	0.0	1.0		0.737	0.0	0.0	0.0	0.0	9	7	597	23	642	4	0.474

Figure 11: LogMap bulk alignment summary and analysis of ontology pair metric scores.

tions of alignment sizes with respect to two different sets of DISO ontologies: *(i)* the full set of 1653 DISO ontology pairs derived from 58 DISO ontologies, and *(ii)* the subset of 105 DISO cyber-security ontology pairs derived from the 15 cyber-security ontologies. We observe that LogMap identifies an intersection between at least 647 ontology pairs (66 about cyber-security), indicating that the DISO ontology network has an important number of connections (*i.e.*, ontology mappings) across its ontologies.

5.3 Towards a DISO-OAEI evaluation track

The ontology pairs (with their accompanying metric scores) were sorted in descending order by *(i)* the LogMap alignment mapping count and *(ii)* the LogMap $\mathcal{M}_{ask}$ mapping count. This decision reflects a preference for larger, more substantial alignments over smaller, thinner ones. A traffic light (red, amber, green) colour-coding scheme was then applied to the metric scores for the top 41 ontology pairs (ranked in descending order by mapping count). As is evident in Figure 11, the metric 'simplicity' ended up playing a pivotal role in distinguishing between candidate and unattractive ontology pairs. The metric 'simplicity' measures the proportion of the entity mappings in LogMap's alignment where the entity names in the URIs of the two entities are identical. Large values (near 1) were interpreted as indicating that the alignment task is too simple — too simple to be of enduring interest to the OAEI community. Small values, near zero, are preferred. As Figure 11 suggests, high simplicity scores disqualified a high proportion of the ontology pairs. Fifty-nine per cent of the top 41 ontology pairs (ranked in descending order by mapping count) were disqualified for this reason.

Amongst the remaining 17 of the top 41 ontology pairs, a small set of candidate ontology pairs was identified. Table 12 shows the top 8 candidate ontology pairs that emerged from the analytic process. It indicates that this set of ontology pairs has a coherent structure. First, the eight ontology pairs cluster into three distinct subdomains: cybersecurity (2 pairs), situation awareness (3 pairs), and smart environments (3 pairs). Second, each subdomain has what we call an 'anchor ontology' — an ontology that appears in every ontology pair in the subdomain. These ontology pairs will form the basis for the new DISO-OAEI 2026 track: https://github.com/city-artificial-intelligence/diso-oaei.

Table 12: The top 8 ontology pairs that emerged from the LogMap-driven analytic process. 'SA' refers to 'situation awareness; 'CA' refers to 'context awareness'.

Source Ontology	Target Ontology	Source DISO Cluster	Target DISO Cluster	Alignment Mappings	$\mathcal{M}_{ask}$ Mappings	OAEI Track Subdomain	OAEI Track Subdomain Anchor Ontology
UCO	STIX	cybersecurity	cybersecurity	131	43	cybersecurity	
STIX	D3FEND	cybersecurity	cybersecurity	37	31	cybersecurity	STIX
JC3IEDM	mIO!	SA	CA	219	40	SA	
JC3IEDM	Brick	SA	smart buildings	49	69	SA	JC3IEDM
JC3IEDM	FacilityOntology	SA	cco-modules	44	23	SA	
ThinkHome	Brick	smart homes	smart buildings	33	42	smart envir	
Brick	SmartEnv	smart buildings	smart homes	32	45	smart envir	Brick
CityOWL	Brick	smart cities	smart buildings	24	27	smart envir	

5.3.1 Ontology provenance and quantitative analysis

Table 13 shows a quantitative analysis and comparison of the 10 ontologies involved in the top 8 candidate ontology pairs that emerged from the LogMap-driven analytic process. The largest ontologies with respect to class count are D3FEND (3,495), JC3IEDM (2,922) and Brick (2,305), whereas the smallest is STIX (88). We can also see that the three ontologies with the largest class counts also have the largest counts of individuals. Overall, one gets the positive impression that the ontologies have substance without being awkwardly large, and that the entities of each ontology are distributed in relatively balanced ways across the four entity types: classes, object properties, data properties and individuals. Importantly, there is nothing in these quantitative profiles that leads one to question the suitability of one or more of these ontologies for inclusion in a prospective OAEI track.

Next, for each ontology, we briefly describe its name and its domain/nature, and we discuss its provenance in order to establish its credibility.

UCO: Unified Cyber Ontology UCO[19] is a domain-level ontology assigned to the DISO 'cyber-security' cluster. It focuses on concepts such as cyber investigations, computer / network defence, threat intelligence, malware analysis, vulnerability research, and offensive operations. UCO is an open, community-developed ontology, partly funded by the Linux Foundation, whose ambition is to support the cybersecurity domain in its entirety. (Note: There is an unfortunate name clash in the literature whereby two distinct cyber-security ontologies both share the acronym UCO. To distinguish between them, DISO refers to the two ontologies using the acronyms UCO1 and UCO2. Here, references to UCO correspond to DISO ontology UCO1 — the Unified Cyber Ontology. DISO ontology UCO2 refers to the Unified Cybersecurity Ontology [81], the product of a project undertaken by individual researchers.)

STIX: Structured Threat Information eXpression STIX [7] is a domain-level ontology assigned to the DISO 'cyber-security' cluster. Initially developed by MITRE Corporation, STIX was designed as a language and serialisation format for exchanging cyber threat intelligence. The official STIX specification is managed by the OASIS Cyber Threat Intelligence (CTI) Technical Committee[20]. The OASIS Threat Actor Context (TAC) technical committee[21] created the STIX OWL ontology for STIX v2.1.

[19] https://unifiedcyberontology.org/

[20] https://oasis-open.github.io/cti-documentation/

[21] https://github.com/oasis-tcs/tac-ontology

Table 13: Quantitative descriptions of the 10 ontologies appearing in the top 8 ontology pairs that emerged from the LogMap-driven analytic process. The metric scores are as reported by the Protege ontology editor.

Metric	UCO	STIX	D3FEND	JC3IEDM	mIO!	Brick	Facility	ThinkHome	SmartEnv	CityOWL
Axioms	11,552	3,903	32,832	26,622	7,727	117,811	7,083	8,122	4,146	7,387
Class count	429	88	3,495	2,922	624	2,305	859	1,109	161	516
Object property count	192	68	205	619	364	131	146	406	242	522
Data property count	581	386	42	313	310	80	12	303	50	162
Individual count	507	4	2,340	4,088	520	8,556	81	54	17	17
Annotation property count	43	6	35	2	24	165	40	22	44	42
Class axioms										
SubClassOf	444	1270	5,391	2,567	627	2,494	925	2,344	401	1,250
EquivalentClasses	0	0	0	272	36	105	25	282	6	2
DisjointClasses	8	18	8	0	62	15	23	123	21	36
Object property axioms										
SubObjectPropertyOf	11	0	250	0	92	19	84	61	164	48
EquivalentObjectProperties	0	0	0	0	0	4	0	0	13	13
InverseObjectProperties	9	0	41	116	70	9	63	48	80	7
ObjectPropertyDomain	10	89	2	503	344	0	127	350	177	517
ObjectPropertyRange	181	71	4	503	337	2	128	340	173	511
Data property axioms										
SubDataPropertyOf	40	1	42	0	92	6	0	12	13	9
EquivalentDataProperties	0	0	0	0	0	0	0	0	0	0
DataPropertyDomain	4	448	2	313	304	2	8	82	48	162
DataPropertyRange	579	385	17	313	284	2	8	277	43	159
Individual axioms										
ClassAssertion	558	0	1,194	4,088	904	18,882	83	53	29	25

D3FEND D3FEND [39] is a domain-level ontology assigned to the DISO 'cyber-security' cluster. MITRE Corporation[22], its developer, refers to it as a knowledge graph of cybersecurity countermeasures, and as a framework for cybersecurity operations and strategic decision-making. Development of D3FEND was funded by the National Security Agency (NSA), the U.S. Cyber Warfare Directorate, and the U.S. Office of the Under Secretary of Defense for Research and Engineering.

JC3IEDM: Joint Consultation, Command and Control Information Exchange Data Model JC3IEDM [57] is a domain-level ontology whose primary DISO cluster is 'information exchange' and whose alternate (equally applicable) DISO cluster is 'situation awareness' [85]. The aim of the JC3IEDM is to support the exchange of consultation, command and control (C3) information, where heterogeneous command and control information systems (C2IS) are sharing military situational awareness data. The Multilateral Interoperability Programme (MIP)[23] is a military standardisation body comprising 24 member nations, the European Defence Agency (EDA), and the North Atlantic Treaty Organisation (NATO). The MIP, in conjunction with NATO, is the configuration manager of the JC3IEDM (a set of UML specifications). In 2009, employees of Vistology[24] created an OWL ontology version of JC3IEDM based on the specifications for JC3IEDM v3.0 from 2006.

mIO!: A Context Ontology for Mobile Environments mIO! [70] is a domain-level ontology assigned to DISO cluster 'context awareness'. It is a context ontology network whose aim is to model context-related knowledge that allows mobile applications to adapt their behaviour based on user context (per the idea of pervasive computing from some years ago). Context awareness can be regarded as a specialised form of situation awareness, where the situation of interest is what is

[22] https://d3fend.mitre.org/

[23] https://en.wikipedia.org/wiki/Multilateral_Interoperability_Programme

[24] https://vistology.com/

happening nearby an individual who is interacting with a mobile application of some kind. mIO! was developed by the Ontology Engineering Group at University of Madrid[25].

Brick: A Uniform Metadata Schema for Buildings Brick[26] is a domain-level ontology assigned to DISO cluster 'smart buildings', within cluster 'smart environments'. It is an open-source project whose aim is to standardise descriptions of physical, logical and virtual aspects of buildings, and the relationships between them. The Brick GitHub repository[27] indicates that the Brick development team consists of academics spanning IBM and many (primarily U.S.-based) universities.

FacilityOntology FacilityOntology is a CCO (Common Core Ontology)[28] mid-level ontology assigned to DISO cluster 'cco-modules', within cluster 'mid-level'. This mid-level ontology was designed to represent facilities (such as buildings, campuses, etc.) that serve some specific purpose and which are common in multiple domains.

ThinkHome: Smart Home Ontology for Human Activity Recognition ThinkHome [73] is a domain-level ontology assigned to DISO cluster 'smart homes', within cluster 'smart environments'. It is a smart home ontology intended to support smart home systems, especially those with a focus on home control, and particularly on energy efficiency. It consists of five component ontologies that address individual aspects of the target application domain: actors, buildings, energy resources, processes and weather. ThinkHome was developed at Technical University Vienna[29].

SmartEnv: Smart Home Environments SmartEnv [2–4, 42] is a domain-level ontology assigned to DISO cluster 'smart homes', within cluster 'smart environments'. SmartEnv, initially referred to as 'E-care@home' in the literature, is a network of 8 component ontologies modelling various aspects of smart homes: temporal, spatial, objects, events, agents, observations, etc.. Its notion of 'smart' refers to environments (such as homes) that are monitored by sensors of various kinds—sensors that can gather and feed environment data.

CityOWL CityOWL is a domain-level ontology assigned to DISO cluster 'smart cities', within cluster 'smart environments'. It is an OWL rendering of a standard schema defined by the Open Geospatial Consortium called CityGML[30]. The standard defines a conceptual model for representing virtual 3D models (aka digital twins) of cities. The CityOWL ontology obtained for DISO was acquired from a URL associated with the Computer Science Laboratory for Image and Information Systems at the University of Lyon[31].

5.3.2 `LogMap` ontology alignment summarisation

Table 14 shows the alignments of `LogMap`'s top 8 candidate ontology pairs summarised by relation type vs. entity type. Relation types are signified as follows: '=' (equivalence), '<' (subClassOf), '>' (superClassOf). Entity types include classes, object properties, data properties, and individuals. We can see that all four entity types are reasonably represented within the mappings for the alignments and LogMap's m_ask mappings (mappings of which LogMap is uncertain and invites feedback from an oracle of some kind).

[25] https://oeg.fi.upm.es/index.php/en/ontologies/index.html
[26] https://brickschema.org/
[27] https://github.com/brickschema/Brick
[28] https://www.commoncoreontologies.org
[29] https://www.auto.tuwien.ac.at/index.php/projectsites/155-thinkhome
[30] https://www.ogc.org/standards/citygml/
[31] https://liris.cnrs.fr/

Table 14: The LogMap alignment and $\mathcal{M}_{ask}$ mappings for the top 8 candidate ontology pairs summarised by relation type vs. entity type.

Source Ontology	Target Ontology	Alignment Mappings	M_ask Mappings	Alignment Mappings Rel	Cls	Oprop	Dprop	Ind	M_ask Mappings Rel	Cls	Oprop	Dprop	Ind
UCO	STIX	131	43	'='	43	4	84	0	'='	12	22	9	0
STIX	D3FEND	37	31	'='	31	0	6	0	'='	26	2	3	0
JC3IEDM	mIO!	219	40	'='	12	0	0	207	'='	26	3	0	11
JC3IEDM	Brick	49	69	'='	6	0	0	43	'='	26	1	0	42
JC3IEDM	FacilityOntology	44	23	'='	43	0	0	0	'='	23	0	0	0
ThinkHome	Brick	33	42	'='	24	0	1	7	'='	34	6	0	1
				'<'	1	0	0	0	'<'	1	0	0	0
Brick	SmartEnv	32	45	'='	25	0	0	7	'='	26	18	0	1
CityOWL	Brick	24	27	'='	14	1	0	7	'='	14	12	0	1
				'>'	2	0	0	0	'>'	0	0	0	0

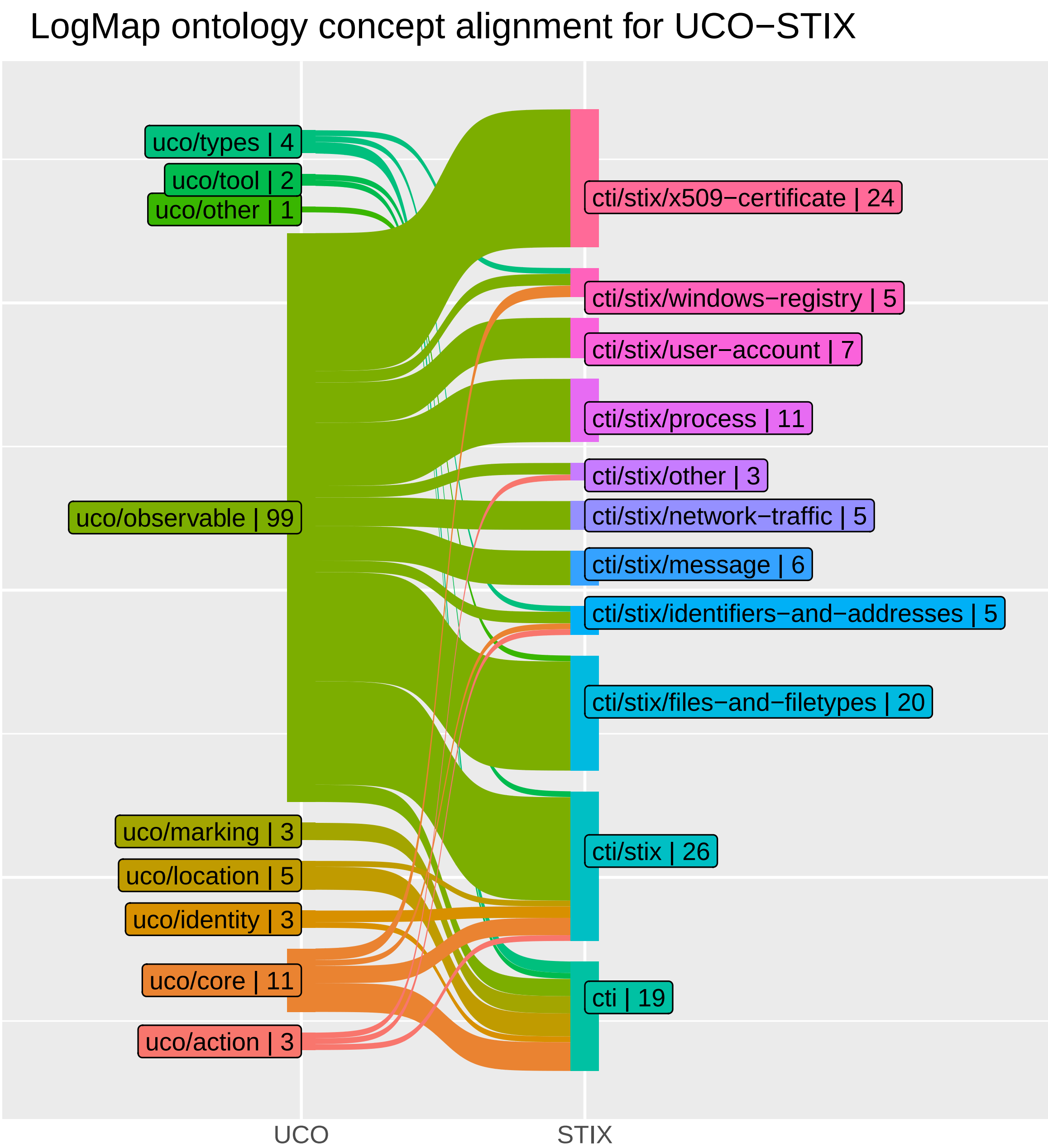

Figure 12: An ontology concept alignment derived from the ontology entity alignment generated by LogMap for the ontology pair UCO-STIX.

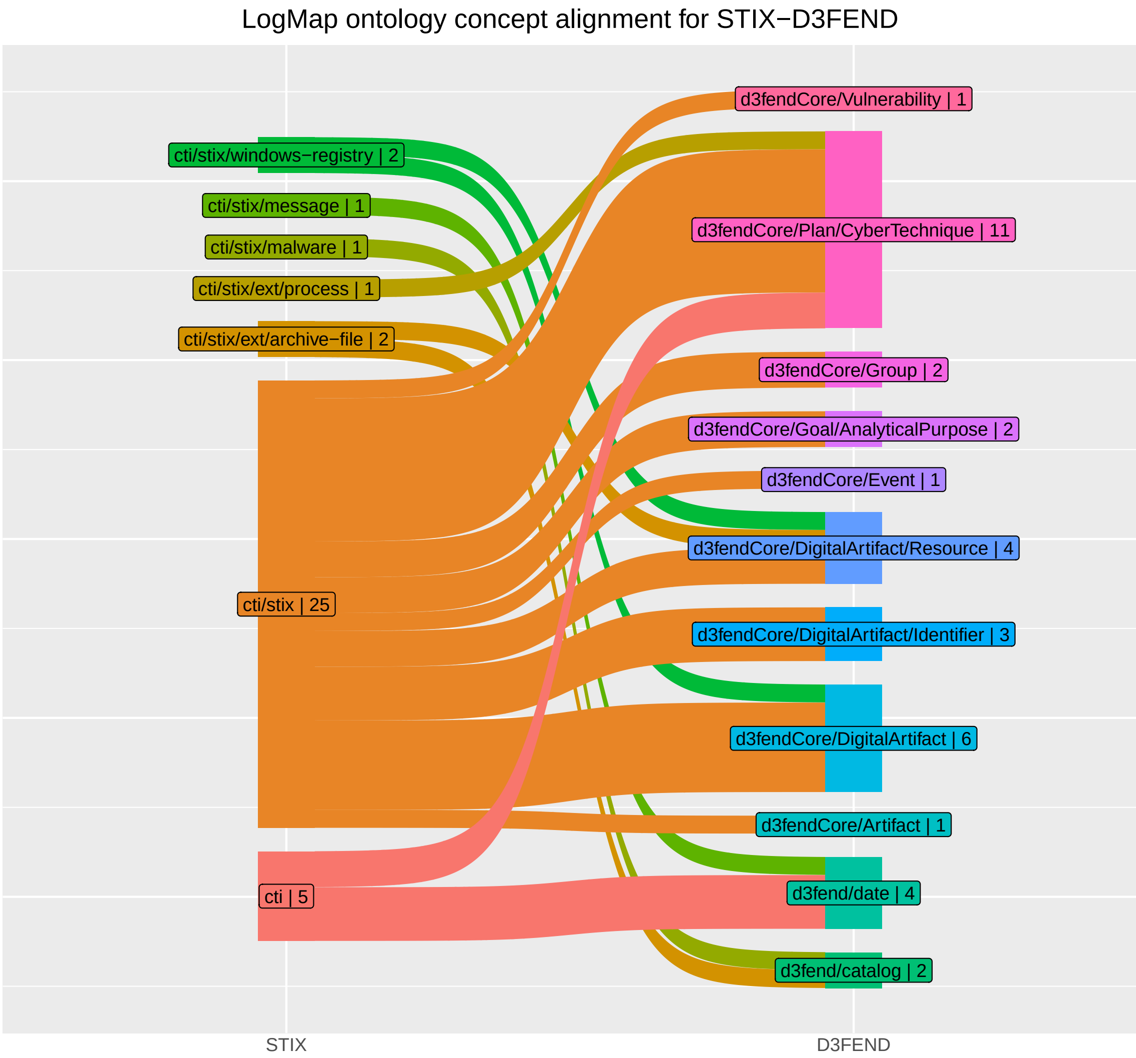

Figure 13: An ontology concept alignment derived from the ontology entity alignment generated by LogMap for the ontology pair STIX-D3FEND.

Figures 12 and 13 visually summarise the ontology alignment (clustered by topic) produced by `LogMap` for the ontology pairs UCO-STIX and STIX-D3FEND, respectively.

6 Conclusions

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 the GUARD project (Phase 1), 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.

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, 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 [62], leading to more informed prompts and accurate diagnostic capabilities.

We have also delivered DISO (Defence, Intelligence and Security Ontologies), a curated collection of over 60 publicly available OWL ontologies covering key subdomains of defence, intelligence, and national security. The repository organises these ontologies into 11 thematic clusters (subdomains), capturing both their primary focus and conceptual overlaps, while documenting each resource with metadata such as description, context, provenance, and web references. We have selected a subset of the DISO ontologies to establish a new OAEI track. This track will support a realistic and meaningful evaluation of ontology alignment systems in a domain like *defence and national security*.

Acknowledgements

This research was supported by Turing Innovations Limited and The Alan Turing Institute's Defence and Security Programme via the project GUARD. Sviatoslav Lushnei, Dmytro Shumskyi, and Severyn Shykula were able to collaborate with the GUARD team thanks to the RAI for Ukraine program of the NYU Center for Responsible AI.

References

[1] K. A. Akbar, F. I. Rahman, A. Singhal, L. Khan, and B. Thuraisingham. The Design and Application of a Unified Ontology for Cyber Security. In *Information Systems Security - 19th International Conference, ICISS*, volume 14424 of *Lecture Notes in Computer Science*, pages 23–41. Springer, 2023.

[2] M. Alirezaie, K. Hammar, and E. Blomqvist. SmartEnv as a network of ontology patterns. *Semantic Web*, 9(6):903–918, 2018.

[3] M. Alirezaie, K. Hammar, E. Blomqvist, M. Nyström, and V. Ivanova. SmartEnv Ontology in E-care@home. In *Proceedings of the 9th International Semantic Sensor Networks Workshop co-located with 17th International Semantic Web Conference*, CEUR Workshop Proceedings, pages 72–79. CEUR-WS.org, 2018.

[4] M. Alirezaie, J. Renoux, U. Köckemann, A. Kristoffersson, L. Karlsson, E. Blomqvist, N. Tsiftes, T. Voigt, and A. Loutfi. An Ontology-based Context-aware System for Smart Homes: E-care@home. *Sensors*, 17(7), 2017.

[5] R. Amini, S. S. Norouzi, P. Hitzler, and R. Amini. Towards Complex Ontology Alignment Using Large Language Models. In *Knowledge Graphs and Semantic Web - 6th International Conference, KGSWC 2024, Paris, France, December 11-13, 2024, Proceedings*, volume 15459 of *Lecture Notes in Computer Science*, pages 17–31. Springer, 2024.

[6] J. C. Augusto and P. J. McCullagh. Ambient Intelligence: Concepts and applications. *Comput. Sci. Inf. Syst.*, 4(1):1–27, 2007.

[7] S. Barnum. Standardizing Cyber Threat Intelligence Information with the Structured Threat Information eXpression (STIX). Technical report, MITRE Corporation, 2014.

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

[9] H. Chen, F. Perich, T. W. Finin, and A. Joshi. SOUPA: Standard Ontology for Ubiquitous and Pervasive Applications. In *1st Annual International Conference on Mobile and Ubiquitous Systems (MobiQuitous)*, pages 258–267. IEEE Computer Society, 2004.

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

[11] J. Chen, E. Jiménez-Ruiz, I. Horrocks, D. Antonyrajah, A. Hadian, and J. Lee. Augmenting Ontology Alignment by Semantic Embedding and Distant Supervision. In *The Semantic Web - 18th International*

Conference, ESWC 2021, Virtual Event, June 6-10, 2021, Proceedings, volume 12731 of *Lecture Notes in Computer Science*, pages 392–408. Springer, 2021.

[12] J. Chen, Y. Lu, Y. Zhang, F. Huang, and J. Qin. A management knowledge graph approach for critical infrastructure protection: Ontology design, information extraction and relation prediction. *Int. J. Crit. Infrastructure Prot.*, 43:100634, 2023.

[13] X. Chen, W. Xia, E. Jiménez-Ruiz, and V. V. Cross. 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, 2014.

[14] D. Costa, M. Collins, S. J. Perl, M. Albrethsen, G. Silowash, and D. Spooner. An ontology for insider threat indicators: Development and application. In *Proceedings of the Ninth Conference on Semantic Technology for Intelligence, Defense, and Security*, volume 1304 of *CEUR Workshop Proceedings*, pages 48–53. CEUR-WS.org, 2014.

[15] B. Cuenca Grau, I. Horrocks, B. Motik, B. Parsia, P. F. Patel-Schneider, and U. Sattler. OWL 2: The next step for OWL. *J. Web Semant.*, 6(4):309–322, 2008.

[16] A. d'Avila Garcez and L. C. Lamb. Neurosymbolic AI: the 3rd wave. *Artif. Intell. Rev.*, 56(11):12387–12406, 2023.

[17] S. Dernbach, K. Agarwal, A. Zuniga, M. Henry, and S. Choudhury. GLaM: Fine-Tuning Large Language Models for Domain Knowledge Graph Alignment via Neighborhood Partitioning and Generative Subgraph Encoding. In *Proceedings of the AAAI 2024 Spring Symposium Series, Stanford, CA, USA, March 25-27, 2024*, pages 82–89. AAAI Press, 2024.

[18] Z. Dragisic, V. Ivanova, H. Li, and P. Lambrix. Experiences from the anatomy track in the ontology alignment evaluation initiative. *J. Biomed. Semant.*, 8(1):56:1–56:28, 2017.

[19] V. Dragos. Ontology modeling for intelligence: the ONTO-CIF model. In *Advances in Knowledge-Based and Intelligent Information and Engineering Systems - 16th Annual KES Conference*, volume 243 of *Frontiers in Artificial Intelligence and Applications*, pages 1543–1552. IOS Press, 2012.

[20] M. R. Endsley. Toward a Theory of Situation Awareness in Dynamic Systems. *Human Factors*, 37(1):32–64, 1995.

[21] J. Euzenat and P. Shvaiko. *Ontology Matching, Second Edition*. Springer, 2013.

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

[23] D. Faria, M. C. Silva, P. Cotovio, L. Ferraz, L. Balbi, and C. Pesquita. Results for Matcha and Matcha-DL in OAEI 2023. In *Proceedings of the 18th International Workshop on Ontology Matching co-located with the 22nd International Semantic Web Conference (ISWC 2023), Athens, Greece, November 7, 2023*, volume 3591 of *CEUR Workshop Proceedings*, pages 164–169. CEUR-WS.org, 2023.

[24] M. Giberna, H. Voos, P. Tavares, J. Nunes, T. Sorg, A. Masini, and J. L. Sanchez-Lopez. On Digital Twins in Defence: Overview and Applications. *CoRR*, abs/2508.05717, 2025.

[25] H. B. Giglou, J. D'Souza, F. Engel, and S. Auer. LLMs4OM: Matching Ontologies with Large Language Models. *CoRR*, abs/2404.10317, 2024.

[26] T. R. Gruber. Toward principles for the design of ontologies used for knowledge sharing? *International Journal of Human-Computer Studies*, 43(5):907–928, 1995.

[27] Z. Hao, W. Mayer, J. Xia, G. Li, L. Qin, and Z. Feng. Ontology alignment with semantic and structural embeddings. *J. Web Semant.*, 78:100798, 2023.

[28] Y. He, J. Chen, D. Antonyrajah, and I. Horrocks. BERTMap: A BERT-Based Ontology Alignment System. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 5684–5691. AAAI Press, 2022.

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

[30] Y. He, J. Chen, H. Dong, E. Jiménez-Ruiz, A. Hadian, and I. Horrocks. Machine Learning-Friendly Biomedical Datasets for Equivalence and Subsumption Ontology Matching. In *21st International Se-*

mantic Web Conference, volume 13489 of *Lecture Notes in Computer Science*, pages 575–591. Springer, 2022.

[31] D. Herron, E. Jiménez-Ruiz, and T. Weyde. On the Potential of Logic and Reasoning in Neurosymbolic Systems using OWL-based Knowledge Graphs. *Neurosymbolic Artificial Intelligence*, 1, April 2025.

[32] S. Hertling and H. Paulheim. OLaLa: Ontology Matching with Large Language Models. In *Proceedings of the 12th Knowledge Capture Conference 2023, K-CAP 2023, Pensacola, FL, USA, December 5-7, 2023*, pages 131–139. ACM, 2023.

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

[34] V. Iyer, A. Agarwal, and H. Kumar. VeeAlign: Multifaceted Context Representation Using Dual Attention for Ontology Alignment. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 10780–10792. Association for Computational Linguistics, 2021.

[35] E. Jiménez-Ruiz, B. Cuenca Grau, Y. Zhou, and I. Horrocks. Large-scale interactive ontology matching: Algorithms and implementation. In *ECAI - 20th European Conference on Artificial Intelligence.*, volume 242 of *Frontiers in Artificial Intelligence and Applications*, pages 444–449. IOS Press, 2012.

[36] E. Jiménez-Ruiz, B. C. Grau, and I. Horrocks. Exploiting the UMLS metathesaurus in the ontology alignment evaluation initiative. In *Proceedings of the 2nd International Workshop on Exploiting Large Knowledge Repositories, Castellón de la Plana, Spain, September 7, 2012*, volume 882 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2012.

[37] E. Jiménez-Ruiz, B. C. Grau, I. Horrocks, and R. B. Llavori. Logic-based assessment of the compatibility of UMLS ontology sources. *J. Biomed. Semant.*, 2(S-1):S2, 2011.

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

[39] P. E. Kaloroumakis and M. J. Smith. Toward a Knowledge Graph of Cybersecurity Countermeasures. Technical report, MITRE Corporation, 2021.

[40] E. Karabulut, S. F. Pileggi, P. Groth, and V. Degeler. Ontologies in digital twins: A systematic literature review. *Future Gener. Comput. Syst.*, 153:442–456, 2024.

[41] H. Knublauch and D. Kontokostas. Shapes constraint language (shacl). Technical report, The World Wide Web Consortium (W3C), July 2017.

[42] U. Köckemann, M. Alirezaie, J. Renoux, N. Tsiftes, M. U. Ahmed, D. Morberg, M. Lindén, and A. Loutfi. Open-Source Data Collection and Data Sets for Activity Recognition in Smart Homes. *Sensors*, 20(3), 2020.

[43] P. Kolyvakis, A. Kalousis, and D. Kiritsis. DeepAlignment: Unsupervised ontology matching with refined word vectors. In *Proceedings of NAACL*, pages 787–798, 2018.

[44] W. Kwon, Z. Li, S. Zhuang, Y. Sheng, L. Zheng, C. H. Yu, J. Gonzalez, H. Zhang, and I. Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*, pages 611–626, 2023.

[45] E. Leblanc, D. Nguyen, M. Balduccini, W. Regli, J. Kopena, and T. Wambold. Military ontologies for information dissemination at the tactical edge. In *Joint Ontology Workshops*. CEUR, 2015.

[46] H. Li, Z. Dragisic, D. Faria, V. Ivanova, E. Jiménez-Ruiz, P. Lambrix, and C. Pesquita. User validation in ontology alignment: functional assessment and impact. *Knowl. Eng. Rev.*, 34:e15, 2019.

[47] F. Liu, E. Shareghi, Z. Meng, M. Basaldella, and N. Collier. Self-alignment pretraining for biomedical entity representations. In *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 4228–4238, 2021.

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

[49] B. Lourenço, P. Adão, J. F. Ferreira, M. M. Marques, and C. Vaz. Structuring Security: A Survey of Cybersecurity Ontologies, Semantic Log Processing, and LLMs Application. *CoRR*, abs/2510.16610, 2025.

[50] S. Lushnei, D. Shumskyi, S. Shykula, E. Jiménez-Ruiz, and A. d'Avila Garcez. Large Language Models as Oracles for Ontology Alignment. In *19th Conference of the European Chapter of the Association for*

Computational Linguistics (EACL 2026). Association for Computational Linguists, 2026.

[51] O. Maryasin. Home Automation System Ontology for Digital Building Twin. In *2019 XXI International Conference Complex Systems: Control and Modeling Problems (CSCMP)*, pages 70–74, 2019.

[52] V. Mascardi, V. Cordì, and P. Rosso. A Comparison of Upper Ontologies. In *WOA 2007: Dagli Oggetti agli Agenti. 8th AI*IA/TABOO Joint Workshop "From Objects to Agents": Agents and Industry: Technological Applications of Software Agents, 24-25 September 2007, Genova, Italy*, pages 55–64. Seneca Edizioni Torino, 2007.

[53] M. Matthews, L. Shattuck, S. Grham, J. Weeks, M. Endsley, and L. Strater. Situation Awareness for Military Ground Forces: Current Issues and Perspectives. *Human Factors*, 45(4):351–355, 2001.

[54] I. Meriah, L. B. A. Rabai, and R. Khédri. Building a comprehensive and multi-dimensional information security ontology: elicitation process and OWL implementation. *Knowl. Inf. Syst.*, 67(1):167–195, 2025.

[55] P. Morosoff, R. Rudnicki, J. Bryant, R. Farrell, and B. Smith. Joint Doctrine Ontology: A Benchmark for Military Information Systems Interoperability. In *10th Conference on Semantic Technology for Intelligence, Defense, and Security*, volume 1523 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2015.

[56] D. A. Mundie, R. Ruefle, A. J. Dorofee, S. J. Perl, J. McCloud, and M. Collins. An incident management ontology. In *Proceedings of the Ninth Conference on Semantic Technology for Intelligence, Defense, and Security*, volume 1304 of *CEUR Workshop Proceedings*, pages 62–71. CEUR-WS.org, 2014.

[57] NATO/MIP. THE JOINT C3 INFORMATION EXCHANGE DATA MODEL OVERVIEW. Technical report, NATO/MIP, 2007.

[58] H. Naveed, A. U. Khan, S. Qiu, M. Saqib, S. Anwar, M. Usman, N. Akhtar, N. Barnes, and A. Mian. A comprehensive overview of large language models. *ACM Trans. Intell. Syst. Technol.*, 16(5), Aug. 2025.

[59] L. Nguyen, E. I. Barcelos, R. H. French, and Y. Wu. KROMA: Ontology Matching with Knowledge Retrieval and Large Language Models. In *24th Int'l Semantic Web Conference (ISWC)*, volume 16140 of *LNCS*, pages 629–649, 2025.

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

[61] S. S. Norouzi, M. S. Mahdavinejad, and P. Hitzler. Conversational ontology alignment with ChatGPT. In *Proceedings of the 18th International Workshop on Ontology Matching co-located with the 22nd International Semantic Web Conference (ISWC 2023), Athens, Greece, November 7, 2023*, volume 3591 of *CEUR Workshop Proceedings*, pages 61–66. CEUR-WS.org, 2023.

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

[63] L. Obrst, P. Chase, and R. Markeloff. Developing an Ontology of the Cyber Security Domain. In *Proceedings of the Seventh International Conference on Semantic Technologies for Intelligence, Defense, and Security*, volume 966 of *CEUR Workshop Proceedings*, pages 49–56. CEUR-WS.org, 2012.

[64] A. Oltramari, L. F. Cranor, R. J. Walls, and P. D. McDaniel. Building an Ontology of Cyber Security. In *Proceedings of the Ninth Conference on Semantic Technology for Intelligence, Defense, and Security*, volume 1304 of *CEUR Workshop Proceedings*, pages 54–61. CEUR-WS.org, 2014.

[65] S. Oulefki, L. Berkani, L. Bellatreche, N. Boudjenah, and A. Mokhtari. Results for BioGITOM in OAEI 2024. In *Proceedings of the 19th International Workshop on Ontology Matching co-located with the 23rd International Semantic Web Conference (ISWC 2024), Baltimore, USA, November 11, 2024*, volume 3897 of *CEUR Workshop Proceedings*, pages 104–109. CEUR-WS.org, 2024.

[66] A. Polleres, R. Pernisch, A. Bonifati, D. Dell'Aglio, D. Dobriy, S. Dumbrava, L. Etcheverry, N. Ferranti, K. Hose, E. Jiménez-Ruiz, M. Lissandrini, A. Scherp, R. Tommasini, and J. Wachs. How does knowledge evolve in open knowledge graphs? *TGDK*, 1(1):11:1–11:59, 2023.

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

[68] M. A. N. Pour, A. Algergawy, E. Blomqvist, P. Buche, et al. Results of the Ontology Alignment Evaluation Initiative 2024. In *19th International Workshop on Ontology Matching (OM)*, volume 3897 of *CEUR Workshop Proceedings*, pages 64–97. CEUR-WS.org, 2024.

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

[70] M. Poveda-Villalón, M. C. Suárez-Figueroa, R. Garcia-Castro, and A. Gómez-Pérez. A Context Ontology for Mobile Environments. In *Proceedings of the Second Workshop on Context, Information and Ontologies, CIAO@EKAW*, CEUR Workshop Proceedings. CEUR-WS.org, 2010.

[71] E. Prud'hommeaux, I. Boneva, J. E. Labra Gayo, and G. Kellogg. Shape expressions language 2.1. Technical report, W3C Shape Expressions Community Group, Oct. 2019.

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

[73] C. Reinisch, M. J. Kofler, F. Iglesias, and W. Kastner. ThinkHome Energy Efficiency in Future Smart Homes. *EURASIP J. Embed. Syst.*, 2011, 2011.

[74] W. Rivadeneira and O. S. Gómez. Cybersecurity Ontologies: A Systematic Literature Review. *RECIBE, Revista ELECTRoNICA DE COMPUTACIoN, INFORMATICA, BIOMEDICA Y ELECTRONICA*, 9:C1–C18, 2021.

[75] K. Saint-Hilaire, C. Neal, F. Cuppens, N. Boulahia-Cuppens, and M. Hadji. Matching Knowledge Graphs for Cybersecurity Countermeasures Selection. In *Science of Cyber Security - 6th International Conference, SciSec 2024, Copenhagen, Denmark, August 14-16, 2024, Proceedings*, volume 15441 of *Lecture Notes in Computer Science*, pages 118–137. Springer, 2024.

[76] S. Samonas and D. Coss. The CIA Strikes Back: Redefining Confidentiality, Integrity and Availability in Security. *Journal of Information Systems Security*, 10(3):21–45, 2014.

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

[78] L. F. Sikos. Cybersecurity Knowledge Graphs. *Knowl. Inf. Syst.*, 65(9):3511–3531, 2023.

[79] A. Solimando, E. Jiménez-Ruiz, and G. Guerrini. Pushing the limits of OWL 2 reasoners in ontology alignment repair problems. *Intelligenza Artificiale*, 10(1):1–18, 2016.

[80] T. G. Stavropoulos, D. Vrakas, D. Vlachava, and N. Bassiliades. BOnSAI: A Smart Building Ontology for Ambient Intelligence. In *2nd International Conference on Web Intelligence, Mining and Semantics, WIMS*, pages 30:1–30:12. ACM, 2012.

[81] Z. Syed, A. Padia, T. Finin, M. L. Mathews, and A. Joshi. UCO: A unified cybersecurity ontology. In *Artificial Intelligence for Cyber Security, Papers from the 2016 AAAI Workshop, Phoenix, Arizona, USA, February 12, 2016*, volume WS-16-03 of *AAAI Technical Report*. AAAI Press, 2016.

[82] M. Taboada, D. Martínez, M. Arideh, and R. Mosquera. Ontology Matching with Large Language Models and Prioritized Depth-First Search. *CoRR*, abs/2501.11441, 2025.

[83] M. Totoian, A. Marginean, P. Blohm, and M. N. Hussain. HybridOM: Ontology Matching using Hybrid Search. In *Proceedings of the 19th International Workshop on Ontology Matching co-located with the 23rd International Semantic Web Conference (ISWC 2024), Baltimore, USA, November 11, 2024*, volume 3897 of *CEUR Workshop Proceedings*, pages 138–145. CEUR-WS.org, 2024.

[84] J. M. J. Valero, A. L. Martínez, P. M. S. Sánchez, D. Navarro-Martínez, R. V. López, J. I. R. Lacal, A. L. Vivar, M. A. S. Monge, M. G. Pérez, and G. M. Pérez. Unlocking the potential of knowledge graphs: A cyber defense ontology for a knowledge representation and reasoning system. In *19th International Conference on Availability, Reliability and Security (ARES)*, pages 73:1–73:9. ACM, 2024.

[85] M.-C. Valiente, R. Machín, E. García-Barriocanal, and M.-Á. Sicilia. An Ontology-Based Integrated Approach to Situation Awareness for High-Level Information Fusion in C4ISR. In *Advanced Information Systems Engineering Workshops*, pages 513–527. Springer, 2011.

[86] W. Wang, H. Bao, S. Huang, L. Dong, and F. Wei. MiniLMv2: Multi-head self-attention relation distillation for compressing pretrained transformers. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2140–2151, Online, Aug. 2021. Association for Computational Linguistics.

[87] Z. Wang, H. Zhu, P. Liu, and L. Sun. Social engineering in cybersecurity: a domain ontology and knowledge graph application examples. *Cybersecur.*, 4(1):31, 2021.

[88] H. Yongqin, C. Xushan, Y. Anlian, and P. Shuo. Research on application of knowledge graph in war

archive based on big data. In *Big Data and Security*, pages 234–247. Springer Nature, 2023.

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

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

[91] L. Zheng, L. Yin, Z. Xie, C. Sun, J. Huang, C. H. Yu, S. Cao, C. Kozyrakis, I. Stoica, J. E. Gonzalez, et al. Sglang: Efficient execution of structured language model programs. *Advances in neural information processing systems*, 37:62557–62583, 2024.