Interpretable Context-Aware Models Improve Expert Validation in Ontology Matching

Pedro Giesteira Cotovio[1,2 [0000−0001−6724−899X]], Susana Nunes[1 [0000−0002−0160−5875]], Ernesto Jiménez-Ruiz[2 [0000−0002−9083−4599]], and Catia Pesquita[1 [0000−0002−1847−9393]]

[1] LASIGE, Faculdade de Ciências, Universidade de Lisboa, Portugal
{pgcotovio,scnunes,clpesquita}@ciencias.ulisboa.pt
[2] City St George's, University of London, London, UK
ernesto.jimenez-ruiz@citystgeorges.ac.uk

Abstract. Recent advances in ontology matching have increasingly relied on language models to capture the lexical information of entities. However, this focus on surface terminology often overlooks the formal semantics encoded in the ontological structure and limits the interpretability of alignment decisions. We introduce EXACT-OM, a context-aware model that integrates ontology-derived semantics directly into the alignment scoring process. It integrates lexical similarity, contextual similarity based on informative relation-specific subgraphs, and language model-based signals within an adaptive scoring framework. The system is intrinsically interpretable: it decomposes every decision into component contributions, provides per-triple importance via perturbation analysis, and renders graph-based visual explanations with concise natural-language summaries. On the OAEI Bio-ML benchmark, EXACT-OM achieves competitive performance while delivering fine-grained, auditable explanations. In a user study (n=12), the majority of users preferred EXACT-OM explanations to more traditional tools, and they were found to substantially improve validation efficiency for high performing users, enabling faster decision-making without compromising accuracy. Importantly, both performance and preferences differed across users, suggesting that explanation effectiveness depends more on individual strategies than on domain knowledge or technical skill. Overall, this shows that the principled integration of semantic structure with language models can enhance explainability and expert validation workflows without sacrificing performance.

Keywords: Ontology Matching · Ontology Alignment · Language Models · User Validation

1 Introduction

Ontology matching is a critical task for establishing semantic interoperability [5], yet traditional matching algorithms and systems often struggle with limited contextual understanding. This is essential not only for injecting the latent domain knowledge needed to distinguish subtly different concepts in complex areas like the life sciences and biomedicine [25,7], but also for supporting explainability: rich contextual cues help

human experts understand why two classes should align, grounding their decisions in deeper, interpretable semantic evidence rather than surface-level similarity alone.

Classical systems such as LogMap [17] and AgreementMakerLight (AML) [8] rely mainly on lexical similarity, string processing, and structural heuristics. While still strong baselines, these methods increasingly struggled with the semantic complexity of biomedical ontologies, where many classes exhibit overlapping labels but differ subtly in definition or contextual meaning. This has motivated a shift toward approaches that capture richer semantics through textual embeddings and graph-based representations [24,16,27,14].

Language models (LM) have become increasingly popular in ontology matching, particularly in the biomedical domain.

While the majority of classical systems rely heavily on sophisticated lexical techniques [7], LMs, both in encoder and generative flavours, are especially well-suited to enrich these aspects with the semantic context they capture, exposing latent synonyms and contextual signals that purely structural or lexical approaches may fail to capture.

The Bio-ML track of the OAEI [15] has been a primary driver of this evolution, providing a benchmark focused explicitly on machine-learning-based ontology matching in the biomedical domain. Since the track's introduction in 2022, it has fostered a diverse ecosystem of approaches that explore neural, classical, and hybrid modelling [14,9,10,16,24,22,13,21,27,31,33,17].

One of the first successful ontology matching systems that integrated language models was BERTMap [14]. It employs a fine-tuned BERT encoder to score candidate mappings based on class labels and textual definitions, followed by a structural refinement step. It supports both unsupervised and semi-supervised modes, achieving state-of-the-art performance. Matcha [9,10] also explores text encoders, integrating semantic similarity based on Sentence-BERT with AML's suite of classical matching algorithms [8]. It preserves AML's robust lexical and structural heuristics while adding contextualised embeddings that capture semantics beyond surface-level label overlap. BioGITOM [24] introduces a hybrid architecture based on a Graph Isomorphism Transformer that combines structural context with transformer-derived semantic embeddings. This fusion allows the system to align heterogeneous biomedical concepts effectively.

Recent works have explored the incorporation of generative language models into the alignment pipeline. GenOM [30] uses GPT-style models to generate textual definitions for ontology terms and then aligns them through language-model querying. Ferraz et al. [11] demonstrated how a simple Retrieval Augmented Generation approach, which extracts relevant contextual information from ontologies and injects it into LM prompts, supports hard-to-find mappings.

Despite substantial advances in automated ontology matching, it is widely acknowledged that fully automatic systems inevitably make mistakes when applied to large, heterogeneous ontologies [5]. Subtle modelling differences, incomplete lexical information, and divergent design patterns often lead matchers to miss valid correspondences or include incorrect ones [26]. Consequently, explainability and human validation have become recognised as essential components of the alignment process [5,4,18]. Rather than merely outputting a ranked list of mappings, a robust matcher should be able to justify each proposed correspondence and support users during the validation process. In

fact, results from the OAEI Interactive Track demonstrate that validating only a handful of uncertain mappings can yield large performance improvements over fully automatic settings [18].

While early work such as CogZ [6] tackled this issue by generating simple natural-language statements describing the rationale behind each suggestion, most state-of-the-art ontology matching systems offer little to no explanatory support; typically, they expose only a numeric similarity score or a confidence value. Li et al. [18] noted that such minimal feedback is often insufficient for users to trust a mapping. In practice, human validators often rely on ontology visualisation tools such as Protégé [23] to assess proposed mappings. However, this approach poses significant challenges in complex ontologies, where domain experts may lack training in knowledge engineering and familiarity with its tools.

More recently, the emergence of LMs has inspired new explanation mechanisms. For instance, the GenOM system [30] generates definitions for ontology terms using generative models, which can serve as a lightweight semantic explanation of each entity prior to matching. Still, all state-of-the-art systems in the Bio-ML track remain primarily optimised for performance metrics, treating alignment as a black-box prediction task. They deliver ranked candidate mappings, but the burden of interpretation falls entirely on the user. This exposes a critical gap in the current landscape: high-performing matchers typically lack transparency, user guidance, and actionable justifications.

Our work addresses this gap by combining competitive matching performance with inherent interpretability. Instead of post-hoc explanations, our model decomposes each score into meaningful lexical, contextual, and generative components, exposes per-triple contextual importance, and produces concise language model summaries. The goal is not only to achieve strong alignment accuracy but also to generate results that a human expert, especially one without expertise in knowledge engineering, can understand, audit, and trust.

In this work, we introduce a novel ontology matching algorithm designed specifically for the ranking task, enabling fine-grained, explanation-oriented assessment of candidate mappings. We evaluated our approach within the OAEI Bio-ML, demonstrating that it achieves competitive performance while providing inherently interpretable scoring signals. To understand how these explanations support real users, we also conducted a small user study involving participants with varying levels of expertise in knowledge graphs, ontology engineering tools, and the biomedical domain. Our analysis revealed that the effectiveness of explanations is highly user-dependent, with formats and supporting tools influencing task efficiency, engagement, and accuracy differently, likely based on individual cognitive strategies rather than prior domain knowledge or technical expertise.

This work contributes: *(i)* a novel method for ontology alignment that achieves state-of-the-art performance in the alignment of biomedical ontologies and supports explainability; *(ii)* an explanation method that generates explanations for mappings based on method scores and context subgraphs; *(iii)* a user study to evaluate the impact of explanations on mapping validation performance.

2 Problem Definition

Ontology matching aims to identify semantic correspondences between entities belonging to two heterogeneous ontologies that describe overlapping domains of knowledge. Formally, let

$$\mathcal{O}_s = (C_s, P_s, I_s, A_s), \qquad \mathcal{O}_t = (C_t, P_t, I_t, A_t)$$

denote the *source* and *target* ontologies, respectively, where C is the set of classes, P is the set of properties, I is the set of individuals, and A is the set of axioms, including both logical assertions and lexical annotations (e.g., labels, synonyms, textual descriptions). We write $V_s = C_s \cup P_s \cup I_s$ and $V_t = C_t \cup P_t \cup I_t$ for the sets of ontology entities.

Alignment Definition. An *alignment* $\mathcal{A}$ is a finite set of mappings of the form

$$\mathcal{A} = \{(s, t, rel, c) \mid s \in V_s,\ t \in V_t,\ rel \in \{\equiv, \sqsubseteq, \sqsupseteq, \perp\},\ c \in [0, 1]\},$$

where rel denotes the semantic relation inferred between entities s and t, and c is a confidence score reflecting the model's belief that the relation holds. The relation space includes equivalence ($\equiv$), subsumption ($\sqsubseteq$), supersumption ($\sqsupseteq$), and disjointness ($\perp$) relations.

In this work, we focus on equivalence relations ($\equiv$), since they constitute the standard setting in most benchmarks. Each mapping (s, t, c) is assigned a scalar score $c = S_{\text{final}}(s, t) \in [0, 1]$ representing the estimated likelihood that the semantic relation rel (equivalence) holds.

Within the OAEI Bio-ML benchmark [15], ontology matching systems are evaluated under two complementary tasks: (i) the *full alignment task*, corresponding to the classical ontology matching objective, and (ii) the *ranking task*, a simplified formulation designed to facilitate the optimisation and comparative testing of machine-learning models on biomedical ontologies.

In this paper, we focus on the ranking task, as it provides a more direct and fine-grained setting to analyse how different model formulations and parameters influence alignment quality. Ranking also offers a natural bridge to user evaluation; by asking participants to reorder the top candidate mappings according to perceived equivalence, we can compare their judgments to the model's ordering and quantify preference, agreement, and divergence beyond simple top-1 accuracy. This finer structure allows us to capture partial agreement, measure how explanations affect user confidence, and examine how decision time relates to performance, offering a richer perspective on interpretability than a binary classification setup.

Ranking Task. The *ranking* task isolates the scoring component of ontology matching. For each source entity $s \in V_s$, a predefined set of target candidates $T_s \subseteq V_t$ is supplied. The model assigns a similarity score to each candidate and produces a ranking

$$\pi_s = \text{argsort}_{t \in T_s}\ S_{\text{final}}(s, t).$$

Performance is evaluated using common ranking-based metrics: Mean Reciprocal Rank (MRR) and Hits@k.

$$\text{MRR} = \frac{1}{|V_s|} \sum_{s \in V_s} \frac{1}{\text{rank}_s(t^*)}, \qquad \text{Hits@k} = \frac{1}{|V_s|} \sum_{s \in V_s} \mathbf{1}(\text{rank}_s(t^*) \leq k),$$

where t^* is the true corresponding target of s in the reference alignment. MRR captures how early in the ranking the correct match appears on average (higher values indicate that correct matches tend to be near the top), while Hits@k measures the proportion of source entities for which the correct target appears within the top-k predictions. Together, these metrics quantify a system's ability to prioritise correct correspondences within a fixed candidate set, making the ranking task particularly well-suited for rapid model development and hyperparameter optimisation in embedding-based or neural approaches.

Interpretability Objective. Beyond predictive accuracy, the proposed framework explicitly targets *explainable alignment*. For each mapping (s, t), the model outputs: *(i)* weighted lexical and contextual evidence; *(ii)* a subgraph-based visual explanation highlighting informative triples; and *(iii)* LLM-generated textual summaries for ambiguous or conflicting cases. These artefacts provide transparent reasoning traces for each correspondence, supporting human validation and improving trust in the resulting alignments.

3 Methods

Our method is designed to address the key challenges that make biomedical ontology matching challenging: ambiguous or overlapping labels, modelling heterogeneity across ontologies, large and noisy neighbourhoods, and the need for interpretable, user-oriented explanations [7].

The alignment process (Figure 1) begins by resolving high-confidence correspondences through exact lexical matching (step 1), reducing the search space and eliminating trivial cases. For more ambiguous entities, we calculate lexical similarity (step 2.1) and construct context subgraphs that capture the most relevant neighbouring classes while filtering noise arising from ontology size and heterogeneity. These subgraphs are then verbalised and summarised into natural-language descriptions (step 2.2), bridging symbolic structure with LLMs and supporting downstream interpretability. In step 3, we integrate lexical and contextual signals within an adaptive scoring framework that selectively invokes a language model when uncertainty is high (step 4). Finally, for candidate mappings in the low-confidence review band, the system generates structured explanations that combine quantitative signals (confidence and importance of lexical, contextual, and LLM-based evidence) with qualitative components (class labels, key context classes, and, when applicable, a short LM rationale summarising the decision) to support human expert validation.

3.1 Label Normalisation and Exact Matching.

The alignment process begins with an exact matching stage to retrieve trivial matches and reduce the search space that our model needs to explore to obtain a correct ranking. All entity labels and synonyms in the source and target ontologies are first normalised by lowercasing and removing spaces, punctuation, and special characters. Considering the *ranking* setting (where candidate lists are predefined), exact matches are similarly

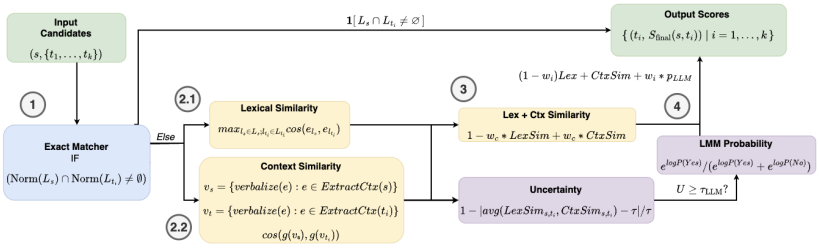


Fig. 1: Overview of the alignment process using Exact-OM for a source and a set of candidate pairs.

assigned a score of 1, while all other candidates for that source entity receive a score of 0. This step captures high-precision correspondences that can be resolved without further analysis.

3.2 Context Subgraph Extraction

Ontology entities s are embedded within assertions that collectively define their meaning. To enable scalable yet semantically faithful inference over these assertions, we first project each ontology into a typed, directed graph representation that captures its relational semantics; we then extract, for each source or target entity s, a compact, high-information subset of triples (its *context subgraph*) that balances informativeness, compactness, and relevance.

Graph Projection. Let $\mathcal{O} = (C, P, I, A)$ denote an OWL ontology. We employ an implementation of the projector proposed in OWL2Vec* [2], which systematically translates TBox and ABox axioms into typed edges between ontology entities to construct a labelled, directed multigraph

$$G = (V, E), \qquad V = C \cup P \cup I, \qquad E = \mathsf{Proj}_{\text{OWL2Vec}*}(A),$$

edges encode syntactic transformations of TBox and ABox axioms (e.g., `subClassOf`, `equivalentClass`, property assertions, domain/range, and annotations). This projection yields a semantically enriched graph in which nodes are ontology entities and edge labels are relation types r. In what follows, we write each edge as a triple $(u, r, v) \in E$ and bound the neighbourhood of s by a hop radius N.

Objective. For each entity $s \in V$, we seek a localised semantic representation $H_s = (V_s, E_s) \subseteq (V, E)$ that balances three design goals: **informativeness** (favouring rare or surprising patterns), **compactness** (respecting a token budget $C_{\max}$ for downstream language models), and **relevance** (favouring triples close to s to preserve local coherence).

Informativeness. We quantify informativeness *relative to each relation type*. For a relation r, we define its induced subgraph as:

$$G_r = (V_r, E_r), \quad E_r = \{(u, r, v) \in E\}, \quad V_r = \{u, v \mid (u, r, v) \in E_r\}.$$

Let $\deg_r(x)$ be the number of occurrences of node x in triples of relation r. The relation-local information content of $x \in V_r$ is

$$IC_r(x) = -\log\left(\frac{\deg_r(x)}{|E_r|}\right).$$

This captures the intuition that a node can be globally frequent yet *relation-specifically* distinctive. To score triples, we average the information content within relation r. Edges that connect relation-distinctive nodes receive higher scores.

Compactness. Each triple (u, r, v) is associated with a token cost $w(u, r, v)$ (estimated encoding length using the LMs). The context must satisfy a maximum cost:

$$\sum_{(u,r,v)\in E_s} w(u, r, v) \leq C_{\max}.$$

Relevance. To enforce locality and focus, we penalise by distance from the root s. Let $d(u, v)$ be the shortest distance from s to either endpoint; with $\alpha > 0$ as a tunable coefficient, the penalty $\alpha \cdot d(u, v)$ discourages distant edges.

Optimisation Problem. We formalise the selection of triples to include in the context subgraph as a constrained maximisation (more details in Appendix C):

$$\begin{aligned} \max_{E_s \subseteq E} \quad & \sum_{(u,r,v)\in E_s} \Big[IC_{\text{edge}}(u, r, v) \; - \; \alpha \cdot d(u, v) \Big] \\ \text{s.t.} \quad & \sum_{(u,r,v)\in E_s} w(u, r, v) \; \leq \; C_{\max}, \\ & E_s \text{ consists of simple paths of length } \leq N \text{ rooted at } s. \end{aligned}$$

We adopt a benefit–cost greedy procedure with a $(1 - 1/e)$ guarantee under standard assumptions. At each step, an oracle `BestPath` returns a bounded-hop path from s that maximises semantic gain per token among yet-unselected edges:

$$\max_{p:\, \ell(p)\leq N} \quad \frac{\sum_{(u,r,v)\in p} \big[IC_{\text{edge}}(u, r, v) - \alpha \cdot d(u, v)\big]}{\sum_{(u,r,v)\in p \setminus E_{\text{used}}} w(u, r, v)}.$$

We use a label-setting dynamic programming approach over $(\text{node}, c, \text{reward}, h)$ as our `BestPath` oracle.

3.3 Verbalisation and Summarisation

To support contextual similarity computation and natural-language inference, each extracted context subgraph is verbalised into coherent sentences that describe its triples.

Triple Verbalisation. For each relation type r, a language model is prompted to generate a reusable sentence template that expresses the relationship between the head and tail entities while preserving their exact lexical forms. A fixed prompt ensures grammatical completeness and lexical fidelity. The first valid sentence produced for a given relation type is stored as a verbalisation template and reused for all triples of that type. When a generated template is found to be invalid or missing replacement keys, a secondary corrective prompt is applied. All verbalised triples are concatenated using a fixed delimiter to produce a coherent natural-language representation of the entity's local context, denoted $v_C(s)$. This verbalised representation is subsequently encoded to compute contextual similarity during scoring.

Entity Summarisation. Summarisation is performed only for entities involved in candidate pairs that will be evaluated by the language model-based oracle, thereby avoiding unnecessary computation for confidently resolved mappings. For each selected entity, the verbalised context is provided to a language model together with the entity label to produce a concise textual summary of its meaning. The resulting summaries supply high-level semantic abstractions that complement the lexical and contextual embeddings, forming a third textual representation used during language model-based verification (Section 3.4). [3]

3.4 Scoring Formulation

Following the extraction of lexical features and context subgraphs, each candidate pair (s, t) is assigned a confidence score that integrates three complementary sources of evidence: *(i) lexical evidence*, capturing surface-form similarity between entity labels; *(ii) contextual evidence*, reflecting structural similarity between their local subgraphs; and *(iii)* an uncertainty-sensitive *language-model signal*, used only when the evidence from the first two sources is insufficiently decisive. Rather than treating these components as independent predictors, our scoring function integrates them through adaptive weighting rules that reflect their relative reliability for each pair, yielding a final confidence value together with a transparent decomposition of its contributing factors.

Lexical and Contextual Similarities. Entities may contain multiple lexical variants (preferred label, synonyms, and abbreviations). The lexical similarity (s_{label}) is computed as the maximum cosine similarity across all label pairs. Contextual similarity (s_{ctx}) is calculated as the cosine similarity between the embeddings obtained by encoding each entity's extracted subgraph verbalisation.

Adaptive Context Weighting. To dynamically balance lexical and contextual evidence, we measure each modality's confidence relative to a neutrality threshold τ

$$\sigma_{\text{label}} = |s_{\text{label}} - \tau|, \qquad \sigma_{\text{ctx}} = |s_{\text{ctx}} - \tau|.$$

[3] Further implementation details and prompts used for the language model components are described in Appendix D.

The contextual weight is then

$$w_c = \frac{\sigma_{\text{ctx}}^{\gamma}}{\sigma_{\text{ctx}}^{\gamma} + \sigma_{\text{label}}^{\gamma}},$$

where the γ acts as a temperature parameter, controlling how strongly the model favours the more confident modality. Values $\gamma > 1$ sharpen the contrast, amplifying differences between σ_{ctx} and σ_{label}, whereas $0 < \gamma < 1$ produces a smoother trade-off that preserves contributions from both modalities, even when their confidence diverges. The joint lexical–contextual similarity becomes

$$S_{lctx}(s,t) = (1 - w_c)\, s_{\text{label}} + w_c\, s_{\text{ctx}}.$$

Uncertainty Measure. Let $m = \frac{1}{2}\big(s_{\text{label}} + s_{\text{ctx}}\big)$. We quantify the overall ambiguity of the lexical and semantic evidence (uncertainty) as

$$U(s,t) = 2\big(\tau - |m - \tau|\big) \in [0,1],$$

maximal at $m = \tau$ and decreasing toward 0 as evidence coalesces toward match ($m \to 1$) or non-match ($m \to 0$).

Language Model Weighting, Verification, and Calibration. The contribution of the language model-based oracle is modulated by the estimated uncertainty

$$w_i = \beta\, U(s,t),$$

where $\beta \in [0,1]$ controls sensitivity to ambiguous evidence.

We invoke the language model if $U(s,t) \geq \tau_{\text{LLM}}$. For selected pairs, the model receives the verbalised summaries of both entities and produces a binary (*Yes*/*No*) prediction. From its logits, we compute the contrastive probability

$$p_{\text{LLM}} = \frac{e^{\log P(\text{Yes})}}{e^{\log P(\text{Yes})} + e^{\log P(\text{No})}} \qquad p_{\text{LLM}}^{\text{cal}} = f(p_{\text{LLM}}),$$

where $f(\cdot)$ is a post-hoc logistic calibration fitted to validation data.

Final Alignment Score. The final interpretable confidence score integrates lexical–contextual similarity with the calibrated LLM decision

$$S_{\text{final}}(s,t) = \begin{cases} S_{lctx}(s,t), & U < \tau_{\text{LLM}} \\ (1 - w_i)\, S_{lctx}(s,t) + w_i\, p_{\text{LLM}}^{\text{cal}}(s,t), & U \geq \tau_{\text{LLM}} \end{cases}$$

When ambiguity is low ($U \approx 0$), $S_{\text{final}} \approx S_{lctx}$; when it is high ($U \approx 1$), the language model's judgement predominates.

This expression unifies lexical, semantic, and language model signals within a continuous, bounded scoring function. The approach queries the language model only when the embedding-based evidence is uncertain, thus trying to maximise scalability while obtaining accurate predictions.[4]

[4] We rely on BERT and BGE-style encoders for lexical and contextual embeddings, and an instruction-tuned Qwen model for generative tasks; complete descriptions are provided in Appendix D.

3.5 Explanation Generation

We focus our explanations on low-confidence, ambiguous mappings; for this reason, we define a review band:

$$\mathcal{C}_{\text{review}} = \{(s,t) \mid \theta_{\text{review,low}} \leq S_{\text{final}}(s,t) \leq \theta_{\text{review,high}}\}$$

For mappings within it, the system produces structured explanations that integrate quantitative and qualitative components.

Confidence and Importance. Each evidence source contributes both a *confidence* (its intrinsic similarity signal) and an *importance* (its normalised contribution in S_{final}):

- **Confidence** (C): $s_{\text{label}} = \max \cos(e_{l_s}, e_{l_t})$ for lexical; $s_{\text{ctx}} = \cos(e_{C_s}, e_{C_t})$ for context; and $p^{\text{cal}}_{\text{LLM}}$ for the language model.
- **Importance** (I): derived from the weights in S_{final}:

  $$I_{\text{label}} = (1-w_c)(1-w_i), \qquad I_{\text{ctx}} = w_c(1-w_i), \qquad I_{\text{LLM}} = w_i,$$

  with $I_{\text{label}} + I_{\text{ctx}} + I_{\text{LLM}} = 1$ ensuring that the importance values form a convex decomposition of the overall score:

  $$S_{\text{final}}(s,t) = I_{\text{label}}\, s^*_{\text{label}} + I_{\text{ctx}}\, s_{\text{ctx}} + I_{\text{LLM}}\, p^{\text{cal}}_{\text{LLM}}(s,t).$$

Hence, *confidence* quantifies the quality of each signal in isolation, while *importance* reflects its actual influence on the final decision.

Label Evidence. We report the selected label pair (l^*_s, l^*_t) used in the max-pooling step, its cosine similarity s_{label}, and its importance I_{label}.

Contextual Importance. For each triple (u, r, v) in the source/target context subgraphs, we estimate its contribution to s_{ctx} via iterative perturbation:

$$I_i = \frac{s_{\text{ctx}} - s^{(-i)}_{\text{ctx}}}{\sum_j |s_{\text{ctx}} - s^{(-j)}_{\text{ctx}}|}.$$

Triples inducing larger drops carry a higher I_i. The aggregate contextual importance I_{ctx} shows how strongly contextual evidence influenced S_{final} relative to other components.

Visualisation. Figure 2 presents an example explanation for a candidate alignment, showing the entity's context subgraph with edges weighted by triple importance I_i. On the right, we present the language model entity summary together with $(I_{\text{label}}, I_{\text{ctx}}, I_{\text{LLM}})$ and $(s_{\text{label}}, s_{\text{ctx}}, p^{\text{cal}}_{\text{LLM}})$, yielding a coherent quantitative–qualitative explanation. In this example, the mapping between "Congenital disorders of glycosylations, iia iim" and "MOGS-CDG" requires understanding that the source class is a subclass of "Congenital disorders of glycosylations, type II", whereas the target is modelled as *part_of* "Congenital disorders of glycosylation with epilepsy as a major feature". The explanation highlights that these two classes share a high degree of semantic similarity, a relationship that is not readily apparent from simple lexical approaches alone.

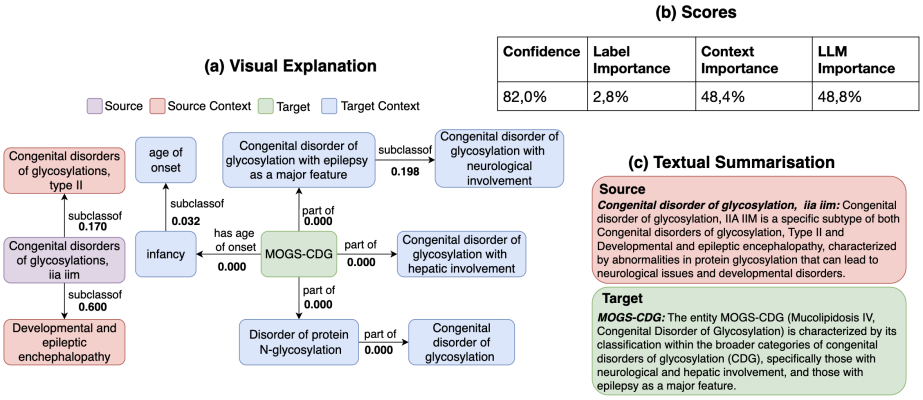


Fig. 2: Example of EXACT-OM explanations for a candidate mapping between *Congenital disorders of glycosylations, iia iim* and *MOGS-CDG*.

Table 1: Equivalence Matching Results on OAEI 2025 Bio-ML benchmark.

	NCIT-DOID			OMIM-ORDO			SNOMED-NCIT†			SNOMED-FMA‡			SNOMED-NCIT§		
System	MRR	Hits@1	MRank	MRR	Hits@1	MRank	MRR	Hits@1	MRank	MRR	Hits@1	MRank	MRR	Hits@1	MRank
BERTMap	.959	.937	1	.880	.830	1	.954	.928	1	.944	.920	2	.969	.951	2
BERTMapLt	.890	.861	5	.766	.716	4	.891	.859	4	.892	.865	5	.849	.773	8
BioGITOM	.913	.891	3	.834	.768	2	.929	.884	2	.909	.854	4	.913	.855	4
BioSTransMatch	.856	.812	8	.693	.620	7	.779	.698	8	.633	.724	6	.908	.877	5
Logmap+OWL2Vec4OA	.873	.834	7	.692	.618	8	.808	.739	7	–	–	–	.852	.777	7
Matcha	.902	.873	4	.815	.782	3	.889	.936	5	.950	.935	1	.936	.921	3
OWL2Vec4OA	.879	.841	6	.707	.635	6	.828	.770	6	–	–	–	.864	.796	6
EXACT-OM	.935	.910	2	.745	.688	5	.911	.876	3	.930	.899	3	.973	.966	1

MRR: Mean Reciprocal Rank; MRank: ranking of system by MRR;‡Body subset; †Neoplasm subset; §Pharm subset.

4 Results and Discussion

4.1 Performance

Our method achieves competitive performance across the OAEI 2025 Bio-ML benchmark while remaining fully unsupervised and inherently interpretable (Table 1). On NCIT[12]–DOID[29], it reaches an MRR of .935, ranking second and closely trailing BERTMap, while outperforming hybrid and GNN-based systems such as Matcha and BioGITOM. On SNOMED[3]–NCIT, it ranks third on the neoplasm subset (MRR .911) and first on the pharm subset (MRR .973, Hits@1 .966), surpassing BERTMap and Matcha on the latter. On SNOMED–FMA[28], it ranks third with an MRR of .930, remaining close to the top lexical-semantic systems. The main performance drop occurs on OMIM[1]–ORDO[32], where lexical ambiguity is high and rich textual definitions provide a stronger advantage, here, it ranks fifth but remains competitive with the OWL2Vec-based baselines. Overall, the results indicate that integrating structured context with interpretable scoring can match or exceed the performance of several state-of-the-art systems, particularly in settings where lexical similarity alone is insufficient.

Table 2: Local conference-track equivalence matching results for EXACT-OM.

Pair	MRR	Hits@1	**Pair**	MRR	Hits@1	**Pair**	MRR	Hits@1
CMT-CONFERENCE	.664	.600	CONFERENCE-EDAS	.785	.647	CONFOF-SIGKDD	.810	.714
CMT-CONFOF	.646	.562	CONFERENCE-EKAW	.772	.720	EDAS-EKAW	.816	.739
CMT-EDAS	.714	.692	CONFERENCE-IASTED	.542	.429	EDAS-IASTED	.693	.579
CMT-EKAW	.699	.636	CONFERENCE-SIGKDD	.856	.733	EDAS-SIGKDD	.758	.667
CMT-IASTED	1.000	1.000	CONFOF-EDAS	.862	.842	EKAW-IASTED	.910	.900
CMT-SIGKDD	.903	.833	CONFOF-EKAW	.926	.900	EKAW-SIGKDD	.894	.818
CONFERENCE-CONFOF	.835	.733	CONFOF-IASTED	.807	.778	IASTED-SIGKDD	.868	.800

MRR: Mean Reciprocal Rank

To demonstrate the generalisability of our method beyond the biomedical domain, we ran 21 additional experiments on the conference track (Table 2). Because the ontologies in this track are small, we considered every possible source-target mapping as a candidate rather than relying on candidate pruning, maintaining a local equivalent for comparison with our other tasks. Even under this exhaustive setup, EXACT-OM achieved a mean MRR of .798 and a mean Hits@1 of .730. Performance varies across ontology pairs, but the results overall suggest that the method transfers well beyond Bio-ML and remains effective even when ranking over the full cross-product candidate space. [5]

4.2 User Study

Our user study focuses on the OMIM-ORDO task, where EXACT-OM achieves its lowest performance. This choice represents a scenario where expert validation would be crucial to improving alignment quality.

Study Design We conducted a counterbalanced within-subjects study with 12 participants from computer science and health/life sciences backgrounds to assess whether EXACT-OM's explanations improve alignment validation.

Task. Participants evaluated 10 case studies, each presenting a source entity along with its top-5 candidate mappings from ORDO to OMIM alignment task (Figure 2). For each case, participants re-ranked the candidates by their confidence that each represents a true equivalence (rank 1 = most likely equivalent).

Conditions. Each participant completed the task under two conditions:

- *NotExplain:* access to the model confidence score and Protégé with the ontologies loaded.
- *Explain:* access to EXACT-OM's score decomposition, visual subgraph explanations, and textual summaries.

Setup. The study was conducted using LimeSurvey [19]. Participants were randomly assigned to one of two groups: Group A evaluated cases 1–5 under *NotExplain*

[5] Detailed dataset statistics are provided in Appendix A, while a comprehensive description of the experimental setup—including hyperparameter settings, running times, and computational resources—is presented in Appendix B.

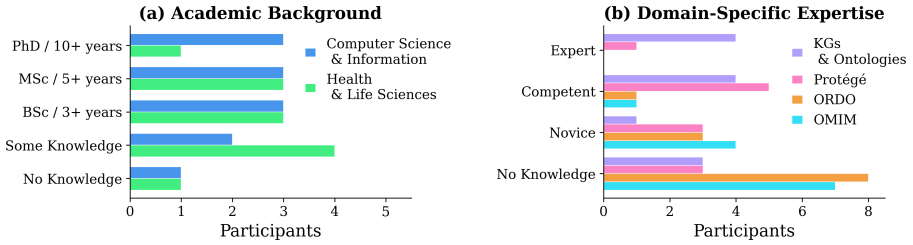


Fig. 3: Participant demographics: (a) academic background by domain and experience level, and (b) expertise with domain-specific resources.

and cases 6–10 under *Explain*; Group B received the reverse assignment. The case presentation order was randomised within each block.

Procedure. Participants completed a consent form, provided background information, and evaluated the 10 cases. At the end, participants selected which approach contributed most to their validation: *(i)* The scores table, visual aids, and text explanations helped me more; *(ii)* Protégé and confidence score helped me more; *(iii)* I liked both approaches; or *(iv)* Neither approach helped me validate. Participants justified their selection and could provide optional comments. The study was conducted remotely via video call, with a researcher available to answer questions, and typically lasted around 60 minutes.

Users' Background and Expertise We recruited 12 participants with diverse backgrounds. The sample was predominantly early-career, with 75% aged under 30 (n=9), while the remainder represented more experienced professionals (ages 30–60+). Participants self-reported their academic background and expertise with domain-specific resources. As shown in Figure 3(a), most participants (n=9) held at least BSc-level qualifications in Computer Science and Information, while 7 had equivalent training in Health and Life Sciences. Regarding expertise (Figure 3(b)), most participants rated themselves as competent or expert in Knowledge Graphs and Ontologies (n=8), while Protégé proficiency varied widely. Notably, familiarity with the specific ontologies used in the evaluation was low, with most participants reporting no prior knowledge.

User Validation Performance and Preferences We analysed participant performance using Mean Reciprocal Rank (MRR) of the correct mapping, grouped by self-reported expertise levels (Table 3). A consistent pattern emerged: participants with no prior knowledge of domain-specific tools or knowledge graphs achieved notably higher MRR in the *Explain* condition (.774) compared to *NotExplain* (.628), representing a 23% relative improvement. Surprisingly, when aggregated by these expertises, the remaining participants achieved better validation performance in the *NotExplain* setting.

To better understand these mixed effects, we regrouped participants based on their observed validation performance: those achieving high validation accuracy (MRR > .7) in at least one condition and those who did not. For participants with strong performance in at least one setting, we examined their *explanation gain*—that is, how much

Table 3: Mean Reciprocal Rank (MRR) of validation performance by expertise level and condition. Best results in bold.

(a) Protégé expertise

Expertise	NotExplain	Explain
No Knowledge	.628	**.774**
Novice	**.883**	.529
Competent	**.720**	.522
Expert	**1.00**	.910

(b) KG expertise

Expertise	NotExplain	Explain
No Knowledge	.628	**.774**
Novice	**.750**	.297
Competent	**.717**	.528
Expert	**.908**	.673

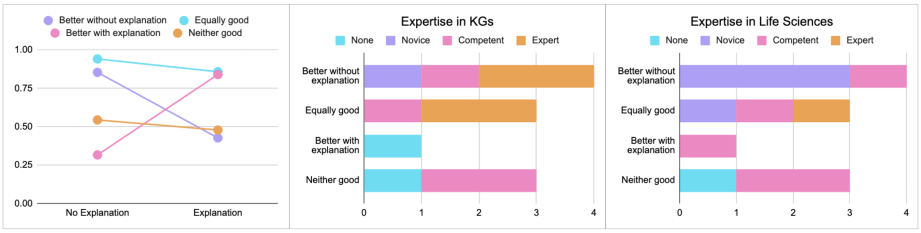


Fig. 4: Left: Validation performance (MRR) change by participant cluster with and without explanations. Centre: distribution of KG expertise by cluster. Right: distribution of Life Sciences expertise by cluster. BSc and MSc have been grouped into competencies.

their validation accuracy changed when shown EXACT-OM explanations instead of Protégé. This analysis yielded four meaningful clusters of users (see Figure 4): participants who performed equally well in both settings, those who performed better without explanations, a single participant who performed markedly better with explanations, and those who performed poorly overall. [6]

Users who performed equally well across conditions showed no clear preference between Protégé and EXACT-OM explanations; however, they completed tasks roughly 1.5x faster when explanations were provided compared to using Protégé. Participants who performed better without explanations also showed no clear preference patterns and completed tasks 1.3x faster with explanations than when using Protégé.[7] Conversely, users with overall low validation performance preferred receiving explanations, but showed no clear trend in task time. Finally, a single participant exhibited a clear performance benefit from explanations, but required more time to complete tasks when using them. Importantly, neither prior experience with knowledge graphs nor domain expertise appeared to determine validation success, explanation gain, or even explanation preference at the individual level.

[6] A complete summary of the cluster analysis results is provided in Appendix E.

[7] Some participants elected to skip Protégé altogether for some tasks, deciding based on labels and scores alone. These times were not considered.

More than half of the participants (55%, n=6) preferred EXACT-OM's score tables, visual, and textual explanations, noting their accessibility and interpretability, particularly for users unfamiliar with the domain ontologies or Protégé. One noted that this approach was "much easier to handle and interpret, which also helped to ground the predictions more, rather than blindly trusting the confidence scores." A smaller group (18%, n=2) with higher domain expertise preferred Protégé alongside confidence scores, valuing ontology annotations (e.g., definitions, cross-references) not fully captured by EXACT-OM. The remaining participants (27%, n=3) preferred combining both approaches, recognising their complementary strengths: Protégé for detailed ontological information and EXACT-OM for intuitive overviews.

Taken together, these patterns highlight the complex interplay between explanation style, user performance, and cognitive load. List-based tools like Protégé, that present more detailed information about each class in the mapping separately, and graph-based explanations (EXACT-OM) appear to support different interpretive strategies. High-performing users succeeded regardless of explanation style, suggesting a robust understanding of the task, yet they benefited from the efficiency gains of graph explanations, completing tasks roughly 1.5x faster. Users whose performance decreased with explanations may have experienced additional cognitive overhead or distraction from the information presented, indicating that graph explanations are not universally beneficial. Those with lower overall validation accuracy favoured receiving explanations, although this did not translate into improved correctness, implying that explanations can influence task engagement without guaranteeing better outcomes. Finally, the single participant who showed clear performance improvement with EXACT-OM explanations demonstrates that, for some users, graph-based explanations can provide a substantial accuracy advantage despite a small increase in completion time. Overall, these findings suggest that the effectiveness of explanation formats is highly user-dependent, likely shaped more by cognitive processing style and the specific strategies users employ when interacting with domain-specific knowledge representations than by domain knowledge and technical expertise.

5 Conclusion

We introduced EXACT-OM, an explainable, context-aware ontology matching model that integrates lexical similarity, informative context, and language model signals into a unified scoring function. The system achieves competitive performance across the OAEI 2025 Bio-ML benchmark while remaining fully unsupervised and inherently interpretable.

Our user study reveals that explanations can substantially reduce validation time for participants who excel at alignment, with many completing tasks 1.3–1.5 $\times$ faster when explanations were available. At the same time, performance and preferences varied across users, indicating that the effectiveness of explanations depends strongly on individual strategies rather than on domain familiarity or technical expertise.

These findings point to several avenues for future work: tailoring explanations to user profiles, enriching them with additional ontological context, incorporating more expressive generative rationales, and extending evaluation to larger user cohorts. In-

tegrating explanations into interactive, user-in-the-loop workflows that allow for the fine-tuning of the underlying models also represents a promising direction. Overall, EXACT-OM demonstrates that high-quality ontology matching can be achieved without sacrificing transparency, supporting both accuracy and expert trust.

Our context-aware methodology also offers a potentially valuable approach for capturing alignment relations beyond equivalence, especially relevant in the context of the new OAEI benchmark [20]. This would require an adaptation of the context subgraph extraction to account for the different semantics.

Acknowledgments. This work was supported by the LASIGE Research Unit, ref. UID/00408/2025 and the FCT through the fellowships `https://doi.org/10.54499/2022.10557.BD` (Pedro Cotovio) and `https://doi.org/10.54499/2023.00653.BD` (Susana Nunes). It was also partially supported by the KATY project, which has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 101017453, and by the CancerScan project, which received funding from the European Union's Horizon Europe Research and Innovation Action (EIC Pathfinder Open) under grant agreement No. 101186829. Ernesto Jiménez-Ruiz was supported by Turing Innovations Limited and The Alan Turing Institute's Defence and Security Programme via the project GUARD. Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or the European Innovation Council and SMEs Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

Disclosure of Interests. The authors declare no competing interests.

Resources The implementation of the proposed system is publicly available at `https://github.com/liseda-lab/Exact-OM` under the Apache-2.0 license. The datasets used in the experiments can be obtained from the OAEI repository at `https://oaei.ontologymatching.org/2025/`.

A Datasets

Table 4: Statistics of the OAEI datasets used in the experiments.

OAEI Track	Matching Task	$\|O_1\|$	$\|O_2\|$	$\|RA\|$
Bio-ML	NCIT–DOID	15,762	8,465	4,686
	OMIM–ORDO	9,648	9,275	3,721
	SNOMED–NCIT.neoplas	22,971	20,247	3,804
	SNOMED–NCIT.pharm	29,500	22,136	5,803
	SNOMED–FMA.body	34,418	88,955	7,256
Conference	CMT–Conference	36	60	15
	CMT–ConfOf	36	38	16
	CMT–Edas	36	104	13
	CMT–Ekaw	36	74	11
	CMT–Iasted	36	140	4
	CMT–Sigkdd	36	49	12
	Conference–ConfOf	60	38	15
	Conference–Edas	60	104	17
	Conference–Ekaw	60	74	25
	Conference–Iasted	60	140	14
	Conference–Sigkdd	60	49	15
	ConfOf–Edas	38	104	19
	ConfOf–Ekaw	38	74	20
	ConfOf–Iasted	38	140	9
	ConfOf–Sigkdd	38	49	7
	Edas–Ekaw	104	74	23
	Edas–Iasted	104	140	19
	Edas–Sigkdd	104	49	15
	Ekaw–Iasted	74	140	10
	Ekaw–Sigkdd	74	49	11
	Iasted–Sigkdd	140	49	15

$|O_1|$ and $|O_2|$ denote source and target ontology sizes, respectively, and $|RA|$ denotes the size of the reference alignment.

B Experimental Setup

B.1 Experiments Hyperparameters

Table 5 reports the hyperparameters used in the experiments.

Table 5: Hyperparameters used in the experiments.

Category	Parameter	NCIT–DOID	OMIM–ORDO	SNOMED–NCIT	Explanation
Adaptive	τ	0.5	0.5	0.5	Neutral similarity pivot for adaptive weighting.
Adaptive	γ	0.73	1.42	1.00	Balances label and context contributions.
LLM Gate	β	0.8	0.8	0.8	Scales LLM score contribution.
LLM Gate	τ_{LLM}	0.8	0.8	0.8	LLM activated when $U \geq \tau_{\text{LLM}}$.
Context	n_{hops}	2	2	2	Depth of extracted context graph.
Context	`BestPath`	dp	dp	dp	Greedy path scoring strategy.
Context	α	0.1	0.1	0.1	Hop penalty in path extraction.
Context	$C_{\max}$	512	128	128	Token limit for context expansion.

B.2 Running Setup and Times

All experiments were executed on a single machine equipped with one RTX 5090 GPU with 32 GB of memory. No distributed execution or multi-GPU setup was used. We report total runtime, dataset preparation time, and alignment time. As expected, alignment dominates the overall runtime, whereas dataset construction remains comparatively small.

Table 6: Recorded runtimes for the main experiments.

Experiment	**Total (min)**	**Dataset (min)**	**Alignment (min)**
NCIT–DOID	361.4	0.9	359.2
OMIM–ORDO	30.5	0.3	30.0
SNOMED–NCIT[†]	371.7	0.5	369.3
SNOMED–FMA[‡]	1174.6	2.0	1167.6
SNOMED–NCIT[§]	408.3	0.6	403.3
Conference (avg., 21 runs)	0.9	0.1	0.5
Conference (min.)	0.6	0.1	0.2
Conference (max.)	1.5	0.1	1.1

[†]Neoplasm subset; [‡]Body subset; [§]Pharm subset.

The conference experiments are substantially faster because the ontologies are much smaller. Across the 21 local conference runs, the average total runtime was 0.9 minutes, with alignment taking 0.5 minutes on average. In this setting, every possible source–target pair was considered as a candidate mapping.

C Generating Context: Optimisation Problem

We formalise context selection as a constrained maximisation:

$$\max_{E_s \subseteq E} \quad \sum_{(u,r,v)\in E_s} \Big[\, IC_{\text{edge}}(u,r,v) \;-\; \alpha \cdot d(u,v) \,\Big]$$

$$\text{s.t.} \quad \sum_{(u,r,v)\in E_s} w(u,r,v) \;\le\; C_{\max},$$

$$E_s \text{ consists of simple paths of length } \le N \text{ rooted at } s.$$

This is a submodular knapsack–type problem: the objective is modular over edges, the budget is monotone submodular, and the hop bound enforces locality.

Greedy Approximation. We adopt a benefit–cost greedy procedure with a $(1 - 1/e)$ guarantee under standard assumptions. At each step, an oracle `BestPath` returns a bounded-hop path from s that maximises semantic gain per token among yet-unselected edges:

$$\max_{p:\, \ell(p)\le N} \quad \frac{\sum_{(u,r,v)\in p} \big[IC_{\text{edge}}(u,r,v) - \alpha \cdot d(u,v)\big]}{\sum_{(u,r,v)\in p \setminus E_{\text{used}}} w(u,r,v)}.$$

We support three `BestPath` strategies: full dynamic programming, a Lagrangian-relaxed search, and a lightweight greedy heuristic. Each progressively reduces computational cost and allows users to balance context optimality against speed when dealing with large or dense ontologies.

- **Full DP:** label-setting DP over $(\text{node}, c, \text{reward}, h)$; complexity $O(N\,|E|\,C_{\max})$.
- **Lagrangian relaxation:** set $w_\lambda(e) = [IC_{\text{edge}}(e) - \alpha] - \lambda\, w(e)$, solve the hop-limited longest path in $O(N\,|E|)$, and bisect λ (about $T \approx 8$ iterations): $O(T\,N\,|E|)$.
- **Local greedy:** iteratively choose the outgoing edge that maximises $(IC_{\text{edge}} - \alpha)/w$, building a path up to N hops; $O(N\,\bar{d})$ where $\bar{d}$ is the average out-degree.

Edges from the chosen path are added until the budget is exhausted or no positive-gain path remains.

Algorithm 1 Greedy Context Subgraph Extraction

procedure EXTRACTCONTEXT($G, s, N, C_{\max}, \alpha$)
  $E_s \leftarrow \varnothing$ ▷ Selected triples
  $T_{\text{used}} \leftarrow \varnothing$ ▷ Already added
  $C \leftarrow C_{\max}$ ▷ Remaining budget
  while $C > 0$ **do**
    $(p, \Delta C) \leftarrow$ BESTPATH$(s, T_{\text{used}}, C, N, \alpha)$
    if $\Delta C > C$ **then**
      break
    end if
    $E_s \leftarrow E_s \cup p$
    $T_{\text{used}} \leftarrow T_{\text{used}} \cup p$
    $C \leftarrow C - \Delta C$
  end while
  return E_s
end procedure

Implementation Notes.

- *Token cost estimation.* We approximate tokenization by word count with a settable word-to-token ratio and apply a safety margin by targeting $x\%$ of $C_{\max}$ to avoid runtime overflows without per-triple tokenization.
- *Hop penalty scaling.* We optionally rescale α relative to the typical edge gain by setting $\alpha = \lambda \cdot \overline{IC}_{\text{edge}}$, where $\overline{IC}_{\text{edge}}$ is the running mean edge IC and $\lambda \in (0, 1)$ (e.g., $\lambda = 0.1$ means "pay" $\sim$10% of a typical edge's information gain per extra hop).
- *Oracle choice.* In practice, the Lagrangian and local-greedy oracles achieve near-optimal contexts in milliseconds per root on large biomedical graphs; full DP remains available for exactness checks on smaller slices.

D Language Models

D.1 Prompts

Verbalisation

"You are an ontology expert natural language generator that converts structured data into a coherent sentence. Given the triple below, generate one complete, grammatically correct sentence that clearly expresses the relationship between the head and the tail as indicated by the relation. You must use the exact wording of $HEAD$ and $TAIL$. The triple is: head: $HEAD$, relation: REL, tail: $TAIL$. The sentence should be:"

"You are an ontology expert natural language generator that converts structured data into a coherent sentence. Given the triple head: $HEAD$, relation: REL, tail: $TAIL$, you generated the following incorrect sentence: '$SENTENCE$'. Please regenerate this sentence, ensuring you use exactly the words $HEAD$ and $TAIL$. Regenerated sentence:"

Entity Summarisation

> "You are an ontology expert entity summariser. Given the following context subgraph, presented as a verbalised list of triples that describe various aspects of the entity '$ENTITY$', please provide a concise and informative summary that captures its key characteristics. The context subgraph is as follows: $CONTEXT$. Summary:"

Decision Making

> "You are an ontology alignment expert. Determine whether the following two ontology entities refer to the same concept. Provide your answer using a single token: `Yes` or `No`. Source entity: SRC_{LABEL} Summary: $SRC_{SUMMARY}$ Target entity: TGT_{LABEL} Summary: $TGT_{SUMMARY}$ Answer:"

D.2 Language Models Used

EXACT-OM uses four pretrained language models, each with a distinct role in the pipeline. All models are used in a frozen, inference-only setup; there is no fine-tuning. SapBERT provides domain-specialised lexical embeddings for labels and synonyms, while BGE-large encodes verbalised context triples into dense contextual representations. A smaller Qwen model (3B) is used to induce verbalisation templates for relations, which are then reused throughout the system to turn graph triples into short natural-language statements. Finally, a larger Qwen model (7B) is used online both to summarise the verbalised context for each entity and to produce a calibrated `Yes`/`No` decision for ambiguous mappings. The LLM-generated summaries are consumed only within the decision LLM itself and are *not* embedded by BGE; contextual embeddings are always computed directly from the verbalised triples.

Table 7: Pretrained models used in EXACT-OM.

Component	Model Name	Size	Role in the System
Lexical encoder	`cambridgeltl/SapBERT-from-PubMedBERT-fulltext`	∼110M	Encodes labels and synonyms of ontology entities into lexical embeddings. Used to compute lexical similarity $s_{\text{label}}(s,t)$ via max-pooled cosine similarity over label pairs.
Context encoder	`BAAI/bge-large-en-v1.5`	∼335M	Encodes verbalised context triples (produced from ontology subgraphs) into contextual embeddings. Used to compute contextual similarity $s_{\text{ctx}}(s,t)$ by comparing source and target context representations.
Verbalisation LLM (templates)	`Qwen/Qwen2.5-3B-Instruct`	3B	Used to generate natural-language verbalisation templates for relations (e.g., turning a triple (h,r,t) into a short pattern such as "h is a subtype of t"). These templates are then reused to verbalise all subgraph triples; the model is not used during scoring.
Decision LLM (summaries & Yes/No)	`Qwen/Qwen2.5-7B-Instruct`	7B	Used in two steps for ambiguous pairs: (i) to generate concise summaries of the verbalised context for source and target entities, and (ii) to answer whether they refer to the same concept with a single-token `Yes`/`No`. The last-token logits for `Yes`/`No` are converted into a calibrated probability $s_{\text{llm}}(s,t)$, which is mixed with the lexical–contextual score under the LLM gating mechanism.

E User Evaluation

Table 8: Cluster analysis summary: performance metrics, preferences, and expertise distribution.

	Avg MRR		Avg Time (s)			Preferences					KG Competence				LS Competence		
Cluster	NotExplain	Explain	NotExplain	Explain	**n**	Explanations	Both	Protégé.	Neither	None	Nov.	Comp.	Exp.	None	Nov.	Comp.	Exp.
Better without explanation	0.85	0.43	200.8	239.7	4	1	1	1	1	0	1	1	2	0	3	1	0
Equally good	0.94	0.86	412.5	268.3	3	1	2	–	–	0	0	1	2	0	1	1	1
Better with Explanation	0.32	0.84	120.8	196.2	1	1	–	–	–	0	0	0	0	1	0	0	1
Neither good	0.54	0.48	352.3	349.9	3	3	–	–	–	0	0	1	2	0	0	2	1

n = number of participants, LS = Life Sciences, Nov. = Novice, Comp. = Competent, Exp. = Expert.

References

1. Amberger, J.S., Bocchini, C.A., Schiettecatte, F., Scott, A.F., Hamosh, A.: OMIM. org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. Nucleic acids research **43**(D1), D789–D798 (2015)
2. Chen, J., Hu, P., Jimenez-Ruiz, E., Holter, O.M., Antonyrajah, D., Horrocks, I.: OWL2Vec*: embedding of OWL ontologies. Machine Learning **110**(7), 1813–1845 (2021)
3. Donnelly, K., et al.: SNOMED-CT: The advanced terminology and coding system for eHealth. Studies in health technology and informatics **121**, 279 (2006)
4. Dragisic, Z., Ivanova, V., Lambrix, P., Faria, D., Jiménez-Ruiz, E., Pesquita, C.: User validation in ontology alignment. In: International Semantic Web Conference. pp. 200–217. Springer (2016)
5. Euzenat, J., Shvaiko, P.: Ontology matching. Springer-Verlag, Heidelberg (DE), 2nd edn. (2013)
6. Falconer, S.M., Storey, M.A.: CogZ: cognitive support and visualization for semi-automatic ontology mapping. In: International Conference on Biomedical Ontology, Software Demonstration (2009)
7. Faria, D., Pesquita, C., Mott, I., Martins, C., Couto, F.M., Cruz, I.F.: Tackling the challenges of matching biomedical ontologies. Journal of biomedical semantics **9**, 1–19 (2018)
8. Faria, D., Santos, E., Balasubramani, B.S., Silva, M.C., Couto, F.M., Pesquita, C.: AgreementMakerLight. Semantic Web **16**(2) (2025)
9. Faria, D., Silva, M., Cotovio, P., Ferraz, L., Balbi, L., Pesquita, C.: Matcha and Matcha-DL results for OAEI 2023. In: Proceedings of the 18th International Workshop on Ontology Matching. (2023)
10. Faria, D., Silva, M.C., Cotovio, P., Ferraz, L., Balbi, L., Pesquita, C.: Results in OAEI 2024 for Matcha. In: Proceedings of the 19th International Workshop on Ontology Matching. (2024)
11. Ferraz, L., Cotovio, P.G., Pesquita, C.: Can language models align biomedical ontologies?: Evaluating retrieval-augmented prompt strategies in bio-ml. In: CEUR Workshop Proceedings. vol. 4001. CEUR-WS (2025)
12. Golbeck, J., Fragoso, G., Hartel, F., Hendler, J., Oberthaler, J., Parsia, B.: The national cancer institute's thesaurus and ontology. Journal of web semantics **1**(1), 75–80 (2003)
13. Gosselin, F., Zouaq, A.: SORBETMatcher Results for OAEI 2023. In: Proceedings of the 18th International Workshop on Ontology Matching. (2023)
14. He, Y., Chen, J., Antonyrajah, D., Horrocks, I.: BERTMap: a BERT-based ontology alignment system. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 5684–5691 (2022)
15. He, Y., Chen, J., Dong, H., Jiménez-Ruiz, E., Hadian, A., Horrocks, I.: Machine learning-friendly biomedical datasets for equivalence and subsumption ontology matching. In: International Semantic Web Conference. pp. 575–591. Springer (2022)
16. Hertling, S., Paulheim, H.: Olala: Ontology matching with large language models. In: Proceedings of the 12th knowledge capture conference 2023. pp. 131–139 (2023)

17. Jiménez-Ruiz, E., Cuenca Grau, B.: LogMap: Logic-based and scalable ontology matching. In: The Semantic Web–ISWC 2011: 10th International Semantic Web Conference, Bonn, Germany, October 23-27, 2011, Proceedings, Part I 10. pp. 273–288. Springer (2011)
18. Li, H., Dragisic, Z., Faria, D., Ivanova, V., Jiménez-Ruiz, E., Lambrix, P., Pesquita, C.: User validation in ontology alignment: functional assessment and impact. The Knowledge Engineering Review **34**, e15 (2019)
19. LimeSurvey GmbH: LimeSurvey: An open source survey tool, http://www.limesurvey.org
20. Liu, X., Hertling, S., Hansen, M.R.: Beyond equivalence: Benchmark datasets for ontology alignment. In: Knowledge Graphs and Semantic Web: 7th International Conference, KGSWC 2025, Leipzig, Germany, November 26–28, 2025, Proceedings. pp. 122–136. Springer-Verlag, Berlin, Heidelberg (2025)
21. Lushnei, S., Shumskyi, D., Shykula, S., Jiménez-Ruiz, E., d'Avila Garcez, A.: 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)
22. Menad, S., Abdeddaïm, S., Soualmia, L.F.: BioSTransMatch Results@ OAEI 2024. In: Proceedings of the 19th International Workshop on Ontology Matching. (2024)
23. Musen, M.A.: The Protégé project: a look back and a look forward. AI Matters **1**(4), 4–12 (2015). https://doi.org/10.1145/2757001.2757003, https://doi.org/10.1145/2757001.2757003
24. Oulefki, S., Berkani, L., Bellatreche, L., Boudjenah, N., Mokhtari, A.: Results for BioG-ITOM in OAEI 2024. In: Proceedings of the 19th International Workshop on Ontology Matching. (2024)
25. Portisch, J., Hladik, M., Paulheim, H.: Background knowledge in ontology matching: A survey. Semantic Web **15**(6), 2639–2693 (2024). https://doi.org/10.3233/SW-223085
26. Portisch, J., Hladik, M., Paulheim, H.: Background knowledge in ontology matching: A survey. Semantic Web **15**(6), 2639–2693 (2024)
27. Qiang, Z., Wang, W., Taylor, K.: Agent-OM: Leveraging LLM Agents for Ontology Matching. Proceedings of the VLDB Endowment **18**(3), 516–529 (2024)
28. Rosse, C., Mejino, J.L.: A reference ontology for biomedical informatics: the foundational model of anatomy. Journal of Biomedical Informatics **36**(6), 478–500 (2003), unified Medical Language System
29. Schriml, L.M., Arze, C., Nadendla, S., Chang, Y.W.W., Mazaitis, M., Felix, V., Feng, G., Kibbe, W.A.: Disease Ontology: a backbone for disease semantic integration. Nucleic acids research **40**(D1), D940–D946 (2012)
30. Song, Y., Chen, J., Schmidt, R.A.: GenOM: Ontology Matching with Description Generation and Large Language Model. arXiv preprint arXiv:2508.10703 (2025)
31. Teymurova, S., Jiménez-Ruiz, E., Weyde, T., Chen, J.: OWL2Vec4OA: Tailoring Knowledge Graph Embeddings for Ontology Alignment. In: International Knowledge Graph and Semantic Web Conference. pp. 168–182. Springer (2024)
32. Vasant, D., Chanas, L., Malone, J., Hanauer, M., Olry, A., Jupp, S., Robinson, P.N., Parkinson, H., Rath, A.: ORDO: an ontology connecting rare disease, epidemiology and genetic data. In: Proceedings of ISMB. vol. 30, pp. 1–4 (2014)
33. Wang, Z.: AMD Results for OAEI 2022. In: Proceedings of the 17th International Workshop on Ontology Matching. vol. 3324 (2022)