LLM Driven Justified Ontology Alignment

Diego Conde-Herreros[1,*,†], George Hannah[2,†], Terry R. Payne[2], Jacopo de Berardinis[2], Valentina Tamma[2], David Chaves-Fraga[3] and Oscar Corcho[1]

[1]*Facultad de Informática, Universidad Politécnica de Madrid, Campus de Montegancedo s/n, Boadilla del Monte, Madrid 28660, Spain*

[2]*School of Computer Science and Informatics, University of Liverpool, Brownlow Hill, Liverpool, L69 7ZX, United Kingdom*

[3]*CiTIUS, Universidade de Santiago de Compostela, 15705 Santiago de Compostela, A Coruña, Galicia, Spain*

Abstract

Ontology alignment is a key task for achieving semantic interoperability across heterogeneous knowledge graphs; yet, it remains a time-consuming and expert-driven process. Recent advances in Large Language Models (LLMs) offer new opportunities to automate this task, particularly in scenarios involving the parallel development of multiple ontologies by automatic means that follow different approaches. In this paper, we propose an approach for the automatic generation of ontology alignments using LLMs, producing mappings that are compliant with the SSSOM standard and enriched with explicit justifications and provenance metadata that can support ontology developers and domain experts in developing ontologies from multiple automatically generated versions. We design and evaluate a set of progressively refined prompts to guide LLMs in generating structured and explainable alignments. The approach is assessed using multiple state-of-the-art models (GPT-5.4, GPT-5 Mini, Gemini Flash, and Gemini Pro) against a selection of ontology pairings from the OAEI Conference dataset. The evaluation combines structural validation, standard alignment metrics (precision, recall, and F1-score), and expert qualitative analysis. The results observed indicate that LLMs possess the potential to generate high-quality candidate mappings, particularly for lexically similar entities. However, limitations persist in semantic discrimination, predicate selection, and the exploitation of ontology structure. These findings indicate that LLMs are best suited as assistive tools for knowledge engineers and domain experts in managing the parallel evolution of ontologies.

Keywords

Ontologies, Semantic Web, Ontology Alignment, Ontology Matching, Large Language Models, Explainable AI

1. Introduction

An ontology is the specification of a shared conceptualization that describes structural and domain-specific knowledge for a variety of tasks, such as data integration, transformation, and homogenization processes [1]. It traditionally relies on knowledge elicitation from domain experts, and the manual formalization of domain knowledge using formal representation languages such as RDFs [2] and OWL [3]. Although well-established methodologies exist to guide this process, ontology development remains a time-consuming and knowledge-intensive task. that relies heavily on manual effort by ontology engineers and domain experts.

AI-based approaches have been applied to various stages of the ontology lifecycle, including concept extraction, ontology learning from text, ontology population, and ontology alignment. More recently, the emergence of Large Language Models (LLMs) has opened new opportunities for supporting ontology engineering tasks. LLMs can assist in generating ontologies, documenting ontologies, and identifying correspondences between ontological entities. Recent studies have explored the potential of conversational AI systems such as GPT models for tasks including ontology generation, ontology alignment,

ELMKE: Evaluation of Language Models in Knowledge Engineering, 3rd Workshop co-located with ESWC 2026, Dubrovnik, Croatia

*Corresponding author.

†These authors contributed equally.

diego.conde.herreros@upm.es (D. Conde-Herreros); g.t.hannah@liverpool.ac.uk (G. Hannah); trp@liverpool.ac.uk (T. R. Payne); Jacopo.De-Berardinis@liverpool.ac.uk (J. d. Berardinis); valli@liverpool.ac.uk (V. Tamma); david.chaves@usc.es (D. Chaves-Fraga); oscar.corcho@upm.es (O. Corcho)

0000-0002-4788-1509 (D. Conde-Herreros); 0000-0002-3218-4559 (G. Hannah); 0000-0002-0106-8731 (T. R. Payne); 0000-0001-6770-1969 (J. d. Berardinis); 0000-0002-1320-610X (V. Tamma); 0000-0003-3236-2789 (D. Chaves-Fraga); 0000-0002-9260-0753 (O. Corcho)

and semantic annotation [4, 5].

As these approaches reach a higher level of maturity, they are being further incorporated into real-world production environments. However, since a single language model can't be reliably used for all ontology engineering tasks due to biases, and because models are prone to hallucinations, several models used in conjunction with oversight from knowledge engineers, domain experts, and stakeholders can improve results [6].

One example of an industrial integration of semantic technologies is Unilever[1], where there is desire to implement semantic technologies, such as ontologies and knowledge graphs, into their product development and autonomous formulation workflows. To maintain clear 1:1 mappings with their source data, stored in the XML derived format AnIML (Analytical Information Markup language)[2], the creation of an ontology that achieves 100% coverage of the AnIML schema is required. In the development of this ontology, several approaches have been employed concurrently: manually curated ontologies and automatically generated ontologies created using different models, prompting strategies, and input data [7]. The resulting ontologies are then compared to each other, corrected, integrated, and iterated upon after being manually evaluated. This particular scenario is a unique case of parallel ontology evolution, where several ontologies are created simultaneously through automated means that are developed at the same time as new versions are created and fed into one another.

Despite saving time in ontology engineering tasks, this approach also presents its own set of challenges. It requires knowledge engineers and domain expert intervention to compare the different approaches and make informed decisions on what aspects must be addressed. The number of comparisons that must be made increases exponentially as new versions of the ontology are created.

Our proposed approach to address this issue is the automatic generation of ontology alignments, compliant with the Simple Standard for Sharing Ontological Mappings (SSSOM) [8] standard, to document the conceptual similarities and differences between the modelled concepts using LLMs. The SSSOM standard provides terms to document the curation rules that have been used to obtain a match in ontology alignment and mapping justifications that inform about the conceptual relationships that have been identified. This approach is expected to ease the development of a tabular format that uses a clear and understandable structure, allowing domain experts to make informed decisions on the development of the ontology, which can be obtained automatically.

Thus, the research questions that will be answered in this study are as follows:

1. **RQ1**: Can LLMs be used for the automatic generation of ontology alignments that are SSSOM compliant and include explicit justifications?
2. **RQ2**: How do different models with different capabilities fare in the ontology alignment task in terms of precision, recall, and F1-score?
3. **RQ3**: Are LLMs better suited as autonomous tools or as assistive tools in a human-in-the-loop workflow for ontology alignment?

The paper follows this structure: Section 2 highlights related work in the field of automatic ontology alignment extraction, Section 3 details the prompt design and evaluation methodology, Section 4 presents the results of said evaluation, and Section 5 discusses them. Finally, the conclusions, limitations, and future work will be reflected upon in Section 6.

2. Related Work

Ontology Alignment is a key technique for achieving semantic interoperability between heterogeneous knowledge representations. It consists of identifying the correspondences between OWL entities, such as classes, properties, or instances, that are defined in different ontologies. Ontology alignment enables data integration, knowledge reuse,and interoperability across different systems [9]. The early work on ontology alignment focused on identifying the categories of matching techniques. In [10] the ontology

[1] https://www.unilever.com/

[2] https://www.animl.org/current-schema

alignment process was formalized, and a framework for matching systems based on similarity measures and alignment generation strategies was proposed. Ontology alignment methods combine different measures of similarity, such as lexical similarity, structural similarity, semantic similarity, and logical reasoning [9].

The approaches for ontology alignment can be grouped into the following categories

1. **Linguistic approaches**: They rely on string similarity measures or linguistic resources such as WordNet to identify correspondences between ontology entities. They also exploit labels, comments, and other textual metadata.
2. **Structural approaches**: They analyze the graph structure of the ontology. They exploit relationships between entities, such as `rdfs:subClassOf` hierarchies or property restrictions, to infer correspondences between semantically similar structures.
3. **Semantic approaches**: They incorporate logical reasoning and formal semantics, ensuring that the generated correspondences are consistent with the logical structure of the ontologies.
4. **Hybrid approaches**: They combine multiple strategies to improve alignment quality; these approaches are common in practice because the task requires combining different sources to produce accurate mappings [11].
5. **Complex ontology matching**: It includes those approaches where correspondences involve complex expressions rather than simple entity-to-entity mappings [12]

The evaluation of the ontology alignment systems has been driven by the Ontology Alignment Evaluation Initiative (OAEI), an international benchmarking campaign organized annually since 2004 [13]. They provide standardized datasets, evaluation tracks, and metrics that enable systematic comparison between ontology matching systems. It is currently the primary experimental benchmark in ontology alignment and has led research in this field.

Recent surveys emphasize the increasing use of machine learning and embedding-based techniques to capture deeper semantic relations between ontology entities [14]. Embedding-based methods represent entities as vectors in a continuous space, enabling similarity computation through vector operations. These approaches rely on graph embeddings or neural networks to encode structural and semantic information from ontologies [15]. Other work has explored the use of embedding techniques that adapt random walk strategies to alignment tasks [16].

Large Language Models (LLMs) have also emerged as a promising approach for ontology alignment. These models leverage large-scale pretraining on natural language corpora to capture semantic relationships between concepts that can be exploited for matching ontology entities. Norouzi et al. [17] evaluates the performance of conversational LLMs such as ChatGPT for ontology alignment tasks and compares the results with systems that participate in the OAEI benchmarks. Hertling and Paulheim [18] propose OLaLa, which applies zero-shot and few-shot prompting with multiple open LLMs to OAEI tasks, showing that with a handful of examples and a well-designed prompt, results comparable to supervised matching systems can be achieved; OLaLa obtained an F1 score of 0.91 in the OAEI 2023 Anatomy track [19]. There are other approaches that combine LLM reasoning with retrieval and candidate generation pipelines. For instance, Taboada et al. [20] propose a hybrid method that integrates embeddings, heuristic search strategies, and LLM prompting to improve alignment accuracy while reducing overall computational costs. More recently, Qiang et al. [21] introduced the OAEI-LLM benchmark to specifically study LLM hallucinations in ontology matching, classifying them into categories such as missing mappings, false mappings, and hierarchical misalignments. This dataset was later extended to cover ten LLMs across seven TBox datasets [22], confirming that hallucinations remain pervasive across models. The LLMs4OL 2025 Challenge [23] further showed that hybrid pipelines combining commercial LLMs with domain-tuned embeddings achieve the strongest performance for ontology learning tasks.

Explainable ontology alignments have been an unexplored topic in research compared to the ontology matching task itself. One form of explainability comes from the mapping justification: why a correspondence holds. Alignment systems such as Logmap [24] use techniques to compute logical explanations for ontology mappings to detect and repair inconsistencies produced by alignments. It introduces

methods to extract minimal sets of axioms responsible for a mapping or a logical conflict. This approach frames the explanation in terms of a logical justification rather than as a natural language explanation that can be easily understood by domain experts. There has also been work on making the matching process itself interpretable; AgreementMaker [25] presents detailed similarity scores and supporting evidence for each of the candidate alignments. In [10], the need for explanations of similarity measures and matching decisions is also emphasized to support validation by domain experts. As for interactive alignment processes that integrate users in validating and refining alignments, [26] studied interactive alignment systems where explanations are used to guide user feedback during alignment validation and repair. Their work shows that providing understandable justification significantly improves the acceptance and correction of mappings.

A central element of our approach is the use of the SSSOM standard [8] for representing alignments. SSSOM was developed to address the lack of metadata in existing mapping exchange formats: prior standards such as EDOAL, while expressive, were not widely adopted outside the OAEI community due to their limited metadata model [8]. SSSOM introduces a simple table-based format with an extensible vocabulary that makes imprecision and incompleteness in mappings explicit. Of particular relevance to this work is the Semantic Mapping Vocabulary (SEMAPV), introduced in the 2022 update of the standard [27], which provides a controlled vocabulary for describing mapping justifications and matching approaches. To our knowledge, no prior work has explored the automatic generation of SSSOM-compliant mappings using LLMs.

Our prompt engineering strategy builds on established techniques for structured LLM generation. Chain-of-thought prompting [28] has shown that decomposing complex tasks into intermediate reasoning steps improves output quality, and role-based instruction paradigms have been shown to improve domain-specific behavior [4]. These insights inform the step-wise, role-based prompt design described in Section 3.

Despite the existing literature on the automatic generation of ontology alignments and the research on explainable alignments, there is currently no work that bridges the two. Modern LLM-based approaches such as OLaLa [18] and MILA [20] focus on alignment accuracy but do not produce justifications; conversely, systems like LogMap [24] or AgreementMaker [25] provide logical justifications but do not leverage LLM generation capabilities. Furthermore, none of these approaches produce output compliant with the SSSOM standard. Our approach addresses this gap by combining LLM-based alignment generation with SSSOM-compliant, justified, and human-readable correspondences, aimed at supporting domain experts in the parallel evolution of ontologies.

3. Methods

The goal of this approach is the automatic generation of a set of SSSOM compliant alignments that provide both the conceptual relationships between the ontology concepts and the justifications for why these matches were made. Using LLMs leverages the language models for the identification of lexical similarities in the OWL entities and the reasoning capabilities of the model to compute the semantic and structural similarities. They also provide the means to produce a structured output through the usage of JSON Schema[3] . The SSSOM standard also provides a set of mapping justifications that facilitate the classification task. From the SSSOM specification, the subset of terms that has been chosen for the alignments is shown in Table 1.

One of the expected challenges for this approach, which will be shown in the Evaluation (Section 4) and Discussion (Section 5), is the hallucinations that LLMs can exhibit, how well they fit the structured output, the consistency of the results, and how well they infer the mapping relationships. Section 3.1 will show how the models are instructed, and in Section 3.2, it is described how the models are used in the experiments.

[3]https://json-schema.org/

Table 1
Description of SSSOM mapping fields.

Field	Description
subject_id	The ID of the subject of the mapping.
subject_label	The label of subject of the mapping.
predicate_id	The ID of the predicate or relation that relates the subject and object.
object_id	The ID of the object of the mapping.
object_label	The label of object of the mapping.
mapping_justification	An action of showing a mapping to be right or reasonable.
mapping_date	The date the mapping was asserted.
author_id	Persons or groups responsible for asserting the mappings.
subject_source	URI of vocabulary or identifier source for the subject.
subject_source_version	Version IRI or version string of the source of the subject term.
object_source	URI of vocabulary or identifier source for the object.
object_source_version	Version IRI or version string of the source of the object term.
confidence	A score between 0 and 1 denoting the probability the match is correct.
comment	Free text with curator notes or tool-generated information.
mapping_set_id	A globally unique identifier for the mapping set.
mapping_set_version	A version string for the mapping.
mapping_set_description	A description of the mapping set.
license	A URL to the license of the mapping.

3.1. Prompt Engineering

The design of the set of prompts used in the generation of ontology alignments compliant with the SSSOM standard has evolved iteratively across three versions: early, extended, and improved. We have followed a role-based instruction paradigm where the model is assigned the role of a knowledge engineer specialized in ontology alignments. This ensures that the model adopts domain-specific reasoning and behavior. The early version provides the task in a monolithic structure, combining role information, task description, and instructions in a single block. This is a comprehensive guide, but it places a high cognitive load and may lead to inconsistencies in execution. The extended prompt is a modified version of the early prompt with more precise language to direct the model by decomposing the task into several steps (subject extraction, matching, predicate selection, metadata generation) and allowing for several justifications for each mapping, which limits the number of mapping predicates to only one. It also restricts the output to the TSV format, but it is more verbose. This aligns with best practices in prompt engineering, as it provides a clearer structure and reduces ambiguity. The improved prompt focuses on presenting the same instructions to the model in a much more concise manner and reducing verbosity. Contextualization regarding RDF and SSSOM has been removed, an output format section has been added, and a new set of restrictions has been introduced. Each of these prompts shares the goal of increasing output correctness, structural consistency, and reproducibility.

The task steps that are followed by all three prompts are as follows:

1. Extraction of entities from Ontology 1.
2. Identification of candidate matches from Ontology 2.
3. Selection of appropriate semantic predicates.
4. Assignment of justification and confidence.
5. Enrichment with provenance metadata.

The prompt includes, as part of the classification task for the mapping predicate, explicit matching strategies. The mapping justifications from the Semantic Mapping Vocabulary (SEMAPV)[4] ensure that the model considers multiple alignment sources rather than relying solely on surface-level similarities.

[4]https://mapping-commons.github.io/semantic-mapping-vocabulary/

Table 2
List of LLMs used in the evaluation of the approach.

Model	Reason for Inclusion	Expected Strength
GPT-5.4	Reasoning	High accuracy, strong constraint adherence
GPT-5 Mini	Efficiency	Cost-effective, moderate reasoning
Gemini Flash Lite Preview	Speed	Low latency, weaker reasoning
Gemini Pro Preview	Balanced	Trade-off between performance and efficiency

The context consists of two input ontologies (Ontology 1 and Ontology 2) that are provided as external context during execution. The prompts clearly distinguish between the source ontology (subject) and target ontology (object) roles and instruct the model to extract the entities and labels from each ontology. Since the test ontologies used during the evaluation are not large, no chunking strategy has been implemented. One will be implemented when an extended evaluation is performed in future work.

Another critical component of the prompt engineering strategy is the enforcement of a strict output structure. All prompts require the generation of a TSV table compliant with the SSSOM standard, including the aforementioned set of terms. The early prompt introduces the full schema and column definitions but does not strongly enforce output constraints. The extended prompt clearly separates output instructions, requires the inclusion of all metadata fields, and emphasizes completeness. The improved prompt fixes the column order and prohibits additional text outside of the TSV output. The set of output constraints has additionally been implemented using the pydantic library and is publicly available at the resource's repository[5].

Due to the preliminary nature of this work, prompt engineering was restricted to purely the augmentation of the lexical content prompt. In future work, we aim to expand prompt engineering to cover more complex prompting techniques, such as few-shot prompting and chain-of-thought prompting.

3.2. Models and Evaluation Methods

To evaluate the use of LLMs for ontology alignment generation, four state of the art models have been chosen: GPT-5.4, GPT-5 Mini, Gemini Flash Lite Preview, and Gemini Pro Preview, with specific snapshot dates of 2026-03-05 and 2025-07-08 for GPT-5.4 and GPT-5 Mini, respectively. The selection aims to cover a range of reasoning performance, efficiency, and latency, which are known to impact structured generation tasks [29]. These LLMs have shown good results in instruction-following, structured output generation, and multi-step reasoning, making them suitable for automatic ontology alignment [28, 30].

From the OpenAI models, there is GPT-5.4, which is a high-capacity model optimized for complex reasoning and structured generation; it is the reference model. GPT-5 Mini is a smaller, efficiency-oriented variant that reduces the computational cost and latency while maintaining competitive performance in structured tasks [31]. It allows for evaluating the trade-off between efficiency and alignment quality. Gemini Flash is optimized for low-latency and high throughput-inference, prioritizing speed over deep reasoning [29, 32]. It allows for assessing whether strict prompt constraints and structured outputs can compensate for limited reasoning capability. Gemini Pro is the higher-capacity model that balances reasoning ability and efficiency. In this study, it serves as a baseline, allowing for the comparison between GPT-based and Gemini-based models and assessing the generalizability of the prompts used. The summary of the differences between the different models can be seen in Table 2.

For the evaluation of this approach, the following procedure has been followed. All three of the prompts have been used with the four models contemplated for this evaluation. The data used for the evaluation is the Conference ontologies alignments from the Ontology Alignment Evaluation Initiative[6].As for the selection of hyperparameters in the generation of the alignments, the temperature

[5]https://github.com/DiegoCondeHerreros/Evol_Align/blob/master/structured_outputs.py
[6]https://oaei.ontologymatching.org/2025/conference/index.html

has been kept at a consistent value of 0.7[7] to avoid compromising the determinism and the randomness. Further experimentation with the hyperparameters will be addressed in a future extended evaluation. The OAEI conference dataset is made up of alignments between 16 different ontologies dealing with conference organization. This dataset was originally in the OAEI format, but it has been manually converted into the SSSOM format. As the prompt engineering survey presented in this work is a preliminary survey, we only tested the *ekaw-iasted*, *edas-ekaw*, and *cmt-confOf* ontology pairs, consisting of 10, 23, and 16 alignments, respectively. Structural data on the ontologies included can be seen in Table 3.

Table 3
Structural statistics of the OAEI Conference ontologies used in the evaluation.

Metric	cmt	confOf	edas	ekaw	iasted
Schema elements					
Named classes	29	38	103	73	140
Object properties	38	12	28	33	38
Datatype properties	9	16	18	0	3
Total properties	47	28	46	33	41
Named individuals	0	0	0	0	0
Class axioms					
`rdfs:subClassOf`	33	64	92	85	247
`owl:disjointWith`	54	86	814	148	2
`owl:equivalentClass`	1	0	7	0	16
OWL Restrictions	8	31	14	14	134
Property axioms					
`rdfs:domain`	59	36	50	24	41
`rdfs:range`	59	36	50	24	41
`owl:inverseOf`	40	3	28	30	16
`rdfs:subPropertyOf`	0	0	0	8	0
Functional properties	8	8	4	0	0
Annotations					
`rdfs:label`	0	0	0	0	0
`rdfs:comment`	4	65	11	2	0
Structural complexity					
Max hierarchy depth	3	2	3	5	5
Approx. logical axioms	254	256	1055	333	497
File size (KB)	27	36	100	35	70

First, the generated output is validated against the expected structural constraints: TSV format compliance, column completeness and ordering validation, and CURIE syntax. Then, the presence of the required metadata is ensured, and finally, it is checked that the output is machine-readable and conforms to the SSSOM specification, removing invalid generations before further evaluation.

Afterward, for the quantitative evaluation of the alignment, the following standard metrics are used, these are calculated using the ground truth from the manually crafted data.

1. Precision: The proportion of generated alignments that are correct.
2. Recall: The proportion of correct alignments that are retrieved.
3. F1-score: The harmonic mean of precision and recall.

Due to the subjective nature of ontology alignment tasks and the open world assumption, it cannot be assumed that the set of alignments we are considering a "ground truth" is comprehensive. It can only be assumed that all alignments contained within the ground truth dataset are correct. To account for these assumptions, we define the components of the confusion matrix as follows.

1. True Positive: An alignment that the domain expert has assessed as valid.

[7], with the exception of GPT-5 Mini, which was run at a temperature of 1.0, as this is the only value supported.

Table 4
The number of alignments generated by each model for each ontology pair across each tested prompt.

	ekaw-iasted			edas-ekaw			cmt-confOf		
Model	Early	Extended	Improved	Early	Extended	Improved	Early	Extended	Improved
GPT 5.4	30	24	20	30	32	30	21	15	15
GPT 5 Mini	14	5	12	13	14	14	11	15	13
Gemini Pro	15	9	13	23	16	18	5	7	9
Gemini Flash	5	6	4	9	8	10	4	5	6

2. True Negative: An alignment that is not expected and not generated by the LLM[8]
3. False Positive: An alignment that the domain expert has assessed as invalid.
4. False Negative: An alignment that exists in the ground truth dataset but is not captured by the LLM.

Finally, for the qualitative evaluation of the alignments, a domain expert from the academic field has performed an expert review of the mappings between the conference ontologies, providing comments on the validity of the alignments, the equivalent entity in the ground truth, the validity of the comments added by the LLM, and the validity of the justification provided. The outcomes of said results are shown in Section 4. The combination of structural validation, quantitative metrics, and qualitative analysis ensures a comprehensive evaluation of both the syntactic correctness and semantic quality of the generated alignments. The full workflow of the proposed approach can be seen in Figure 1.

4. Results

4.1. Quantitative Assessment

When comparing the two model families tested, Table 4 shows that each family shares a "preference" for the type of response generated. OpenAI's GPT models overall produced a larger number of alignments, whilst Google's Gemini models produced fewer. This gap is further widened when comparing the larger and smaller models against their counterparts from the other family.

Across all the tested completions, the behavior of each model remains relatively consistent. The larger *"pro"* models outperform their smaller counterparts. However, this higher performance comes at the price of higher token costs[9][10]. Whilst the overall performance of the larger models, as shown in Table 7, is higher than that of the smaller models, in cases where precision is more important, as shown in Table 5, Gemini Flash is the most performant model with a precision of 1 across all prompts and ontology pairs. However, the same model is notably weak in performance regarding recall, as shown in Table 6. This behavior is also exhibited, albeit to a lesser extent, by Gemini Pro. For example, in the case of the Early prompt for the ekaw-iasted case, Gemini pro achieves a high precision of 1 compared to GPT 5.4's precision of 0.8. However, Gemini Pro only generated 15 alignments as opposed to the 30 generated by GPT 5.4. Whilst both "pro" models performed well, each being the most performant for at least one ontology pairing, in these experiments, GPT 5.4 was the model best suited to the task with a maximum F1 score of 0.918 and a minimum F1 score of 0.820. Gemini pro, whilst having a high maximum F1 score of 0.968, achieved a much lower minimum F1 score of 0.313. This performance is comparable to, if not worse than, the performance of the smaller models in the worst case scenarios.

Across all models and ontology pairs compared, there is a limited difference in response quality between the three different prompts tested. This indicates that there may not be a significant difference

[8]Due to the experimental setup, it is not possible for true negatives to exist. However, introducing negative, or "false" alignments into the ground truth dataset would alleviate this limitation.
[9]https://developers.openai.com/api/docs/pricing
[10]https://ai.google.dev/gemini-api/docs/pricing

Table 5
Precision of the generated alignments by each model for all ontology pairs across all prompts.

	ekaw-iasted			edas-ekaw			cmt-confOf		
Model	Early	Extended	Improved	Early	Extended	Improved	Early	Extended	Improved
GPT 5.4	0.800	0.750	0.750	0.867	0.844	0.933	0.762	0.933	0.933
GPT 5 Mini	0.929	**1.000**	0.750	**1.000**	**1.000**	0.929	**1.000**	0.867	0.692
Gemini Pro	**1.000**	0.889	0.923	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
Gemini Flash	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**

Table 6
Recall of the generated alignments by each model for all ontology pairs across all prompts.

	ekaw-iasted			edas-ekaw			cmt-confOf		
Model	Early	Extended	Improved	Early	Extended	Improved	Early	Extended	Improved
GPT 5.4	**1.000**	**1.000**	0.938	0.813	0.844	**0.903**	**0.888**	0.824	0.824
GPT 5 Mini	0.722	0.500	0.600	0.464	0.500	0.481	0.611	0.684	0.500
Gemini Pro	0.938	0.800	0.857	0.793	0.593	0.667	0.313	0.438	0.563
Gemini Flash	0.500	0.600	0.400	0.375	0.333	0.400	0.250	0.313	0.375

Table 7
F1 Score of generated alignments by each model for all ontology pairs across all prompts.

	ekaw-iasted			edas-ekaw			cmt-confOf		
Model	Early	Extended	Improved	Early	Extended	Improved	Early	Extended	Improved
GPT 5.4	0.889	0.857	0.834	0.839	0.844	**0.918**	0.820	**0.875**	**0.875**
GPT 5 Mini	0.813	0.667	0.667	0.634	0.667	0.634	0.759	0.765	0.581
Gemini Pro	**0.968**	0.842	0.889	0.885	0.745	0.800	0.477	0.609	0.720
Gemini Flash	0.667	0.750	0.571	0.545	0.500	0.571	0.400	0.477	0.545

in the content of the prompt tests. Further augmenting these prompts in future work may lead to more significant differences in response quality; however, there is no evidence to support this in these results.

4.2. Qualitative Assessment

The use of the SSSOM vocabulary allows for the justification of a given alignment to be represented [8]. Whilst the primary role of this feature is to facilitate the sharing of mappings between developers, these justifications also provide grounds to qualitatively evaluate the alignments generated by LLMs. Across prompts and models, whilst there are edge cases where a given LLM cannot understand the nature of the relationship between two concepts, generally, the natural language justification of the alignments and the justification following SEMAPV remain coherent and consistent. An exception to this is GPT 5 Mini responses, where in certain cases no natural language justifications are generated, but even in these cases SEMAPV justifications are still generated and are of high quality.

The consistency and quality of the SEMAPV justifications expose a new avenue for future experimentation, as removing the natural language justifications from the response would reduce the number of output tokens and, therefore, reduce the cost of the prompts. However, it may be the case that the natural language justification *"points the LLM in the right direction"* and that the quality of the SEMAPV justifications is tied to the existence of the natural language justifications.

5. Discussion

The generated alignments show that LLMs are capable of identifying high-confidence correspondences for lexically similar entities , such as direct matches, that are mapped to the `skos:exactMatch` term. This, as expected, indicates that LLMs perform reliably when surface-level similarity aligns with semantic equivalence. However, the models tend to over-rely on lexical similarity, defaulting to strong equivalence

relations even in cases where a weak semantic relation (`skos:closeMatch, skos:relatedMatch`) may be more appropriate.

Despite the relative size of the dataset, some patterns can be observed that are significant for the entire conference dataset. The higher capacity models (GPT-5.4 and Gemini Pro) produce more complete and consistent outputs in metadata generation and justification assignment. The lightweight models (GPT-5 Mini, Gemini Flash) omit some of the metadata fields, are less precise in candidate selection, and have a greater dependence on lexical similarity. This is not unexpected since model scale correlates with reasoning capability.

There are limitations to the alignments. First, there is an overgeneralization of the equivalence relationship, where strong equivalence relations are added even when the semantic relationship is weaker or hierarchical. For instance, `ConferencePaper` is classified as the exact match of a `Paper` , where a `rdfs:subClassOf` would be more appropriate. Also, despite the models including structural and logical matching strategies, these rarely exploit ontology hierarchies or relations, relying mostly on label similarity without inferring richer mappings based on hierarchical alignment or contextual classification.

The results suggest that LLMs are a promising tool for semi-automatic ontology alignment for the generation of candidate mappings, the acceleration of the alignment workflows, and the production of structured outputs that are compliant with standards such as SSSOM. However, their use in a fully automatic setting is limited by the lack of robust semantic discrimination, sensitivity to prompt design, and incomplete exploitation of ontology structure. Therefore, the LLM-generated alignments are best positioned as assistive tools within a human-in-the-loop or hybrid pipeline, making it a functional working solution for tackling the parallel evolution of ontologies and informing knowledge engineers and domain experts about which decisions to make during development.

6. Conclusion

This paper presents an approach for the automatic generation of SSSOM-compliant ontology alignments using Large Language Models, with a particular focus on producing not only mappings but also justified and explainable correspondences. Through a preliminary assessment of prompts of varying lexical complexity and state-of-the-art LLMs, the results demonstrate that LLMs are capable of generating structurally valid and semantically meaningful candidate alignments, especially in cases of strong lexical similarity.

The study highlights that whilst model capacity significantly influences semantic accuracy and completeness, different model families have differing “preferences” when generating alignments, which can have a large impact on the number and quality of generated alignments. However, limitations remain in terms of over-reliance on lexical similarity, imperfect predicate selection, and limited exploitation of ontology structure, indicating that LLMs are not yet suitable for fully autonomous alignment generation.

Overall, the findings suggest that LLMs are best positioned as assistive tools within human-in-the-loop workflows, supporting ontology engineers and domain experts in the alignment process. Future work will focus on scaling the evaluation to larger datasets, incorporating retrieval and hybrid alignment strategies, and improving the semantic grounding and explainability of the generated mappings.

This preliminary investigation and its results present many avenues for future work. There is scope to significantly expand the prompt engineering survey and incorporate more complex prompting strategies. The datasets tested will also be expanded to cover a wider range of domains and ontology sizes, which will include the exploration of ontology chunking methods. Additionally, including open weight models in the evaluation will be a valuable contribution.

Acknowledgments

George Hannah has been funded by an EPSRC ICASE studentship, 201146 with Unilevel PLC. David Chaves-Fraga is funded by the Agencia Estatal de Investigación (Spain) (PID2023-149549NB-I00 &

CPP2024-011786), the Xunta de Galicia — Consellería de Cultura, Educación, Formación Profesional e Universidades (Centro de investigación de Galicia accreditation 2024–2027 ED431G-2023/04) and the European Union (European Regional Development Fund–ERDF).

Declaration on Generative AI

During the development of this work and its corresponding experiments, the author(s) employed AI models from both OpenAI and Google AI. These activities included the generation of responses to be manually evaluated as a core part of the experiments presented, and the generation of generic code for data translation tasks. After using these tool(s)/service(s), the author(s) reviewed and edited the content as required and take(s) full responsibility for the publication's content.

References

[1] R. Studer, V. R. Benjamins, D. Fensel, Knowledge engineering: Principles and methods, Data & knowledge engineering 25 (1998) 161–197.
[2] D. Brickley, R. Guha, Rdf schema, 2014. URL: https://www.w3.org/TR/rdf-schema/.
[3] D. L. McGuinness, F. Van Harmelen, et al., Owl web ontology language overview, W3C recommendation 10 (2004) 2004.
[4] J. Huang, K. C.-C. Chang, Towards reasoning in large language models: A survey, in: Findings of the association for computational linguistics: ACL 2023, 2023, pp. 1049–1065.
[5] S. Pan, L. Luo, Y. Wang, C. Chen, J. Wang, X. Wu, Unifying large language models and knowledge graphs: A roadmap, IEEE Transactions on Knowledge and Data Engineering 36 (2024) 3580–3599.
[6] J. Kampars, G. Mosans, T. Jogi, F. Roters, N. Vajragupta, Llm-supported collaborative ontology design for data and knowledge management platforms, Frontiers in big Data 8 (2025) 1676477.
[7] G. Hannah, J. de Berardinis, T. R. Payne, V. Tamma, A. Mitchell, E. Piercy, E. Johnson, A. Ng, H. Rostron, B. Konev, Relrae: Llm-based relationship extraction, labelling, refinement, and evaluation, 2025. URL: https://arxiv.org/abs/2507.03829. arXiv:2507.03829.
[8] N. Matentzoglu, J. P. Balhoff, S. M. Bello, C. Bizon, M. Brush, T. J. Callahan, C. G. Chute, W. D. Duncan, C. T. Evelo, D. Gabriel, et al., A simple standard for sharing ontological mappings (sssom), Database 2022 (2022) baac035.
[9] F. Ardjani, D. Bouchiha, M. Malki, Ontology-alignment techniques: survey and analysis, International Journal of Modern Education and Computer Science 7 (2015) 67.
[10] J. Euzenat, P. Shvaiko, Ontology matching, Springer, 2007.
[11] P. Ochieng, S. Kyanda, Large-scale ontology matching: State-of-the-art analysis, ACM Computing Surveys (CSUR) 51 (2018) 1–35.
[12] E. Thiéblin, O. Haemmerlé, N. Hernandez, C. Trojahn, Survey on complex ontology matching, Semantic Web 11 (2020) 689–727.
[13] J. Euzenat, C. Meilicke, H. Stuckenschmidt, P. Shvaiko, C. Trojahn, Ontology alignment evaluation initiative: six years of experience, in: Journal on data semantics XV, Springer, 2011, pp. 158–192.
[14] H. Khan, M. Saqib, H. A. Khattak, S. I. Ali, S. Lee, Ontology alignment for accurate ontology matching: A survey, in: International Conference on Smart Homes and Health Telematics, Springer, 2023, pp. 338–349.
[15] Z. Hao, W. Mayer, J. Xia, G. Li, L. Qin, Z. Feng, Ontology alignment with semantic and structural embeddings, Journal of Web Semantics 78 (2023) 100798.
[16] S. Teymurova, E. Jiménez-Ruiz, T. Weyde, J. Chen, Owl2vec4oa: Tailoring knowledge graph embeddings for ontology alignment, in: International Knowledge Graph and Semantic Web Conference, Springer, 2024, pp. 168–182.
[17] S. S. Norouzi, M. S. Mahdavinejad, P. Hitzler, Conversational ontology alignment with chatgpt, arXiv preprint arXiv:2308.09217 (2023).
[18] S. Hertling, H. Paulheim, OLaLa: Ontology Matching with Large Language Models, in: Proceedings

of the 12th Knowledge Capture Conference (K-CAP '23), ACM, Pensacola, FL, USA, 2023, pp. 131–139. doi:10.1145/3587259.3627571.

[19] S. Hertling, H. Paulheim, OLaLa results for OAEI 2023, in: Proceedings of the 18th International Workshop on Ontology Matching (OM 2023), volume 3591 of *CEUR Workshop Proceedings*, 2023, pp. 170–177.

[20] M. Taboada, D. Martinez, M. Arideh, R. Mosquera, Ontology matching with large language models and prioritized depth-first search, Information Fusion 123 (2025) 103254.

[21] Z. Qiang, K. Taylor, W. Wang, J. Jiang, OAEI-LLM: A Benchmark Dataset for Understanding Large Language Model Hallucinations in Ontology Matching, in: Proceedings of HGAIS 2024, volume 3953 of *CEUR Workshop Proceedings*, 2024. ArXiv:2409.14038.

[22] Z. Qiang, K. Taylor, W. Wang, J. Jiang, OAEI-LLM-T: A TBox Benchmark Dataset for Understanding Large Language Model Hallucinations in Ontology Matching, arXiv preprint arXiv:2503.21813 (2025).

[23] H. B. Giglou, J. D'Souza, N. Mihindukulasooriya, S. Auer, LLMs4OL 2025 Overview: The 2nd Large Language Models for Ontology Learning Challenge, in: Open Conference Proceedings, volume 6, 2025. doi:10.52825/ocp.v6i.2913.

[24] E. Jiménez-Ruiz, B. Cuenca Grau, Logmap: Logic-based and scalable ontology matching, in: International Semantic Web Conference, Springer, 2011, pp. 273–288.

[25] I. F. Cruz, F. P. Antonelli, C. Stroe, Agreementmaker: efficient matching for large real-world schemas and ontologies, Proceedings of the VLDB Endowment 2 (2009) 1586–1589.

[26] P. Lambrix, R. Kaliyaperumal, A session-based approach for aligning large ontologies, in: Extended Semantic Web Conference, Springer, 2013, pp. 46–60.

[27] N. Matentzoglu, J. Flack, J. Graybeal, N. L. Harris, H. B. Hegde, C. T. Hoyt, H. Kim, S. Toro, N. Vasilevsky, C. J. Mungall, A Simple Standard for Ontological Mappings 2022: Updates of data model and outlook, in: CEUR Workshop Proceedings, volume 3324, 2022, pp. 61–66.

[28] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al., Chain-of-thought prompting elicits reasoning in large language models, Advances in neural information processing systems 35 (2022) 24824–24837.

[29] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang, et al., A survey on evaluation of large language models, ACM transactions on intelligent systems and technology 15 (2024) 1–45.

[30] R. OpenAI, Gpt-4 technical report. arxiv 2303.08774, View in Article 2 (2023) 1.

[31] J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. Casas, L. A. Hendricks, J. Welbl, A. Clark, et al., Training compute-optimal large language models, arXiv preprint arXiv:2203.15556 10 (2022).

[32] G. Team, R. Anil, S. Borgeaud, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, K. Millican, et al., Gemini: a family of highly capable multimodal models, arXiv preprint arXiv:2312.11805 (2023).

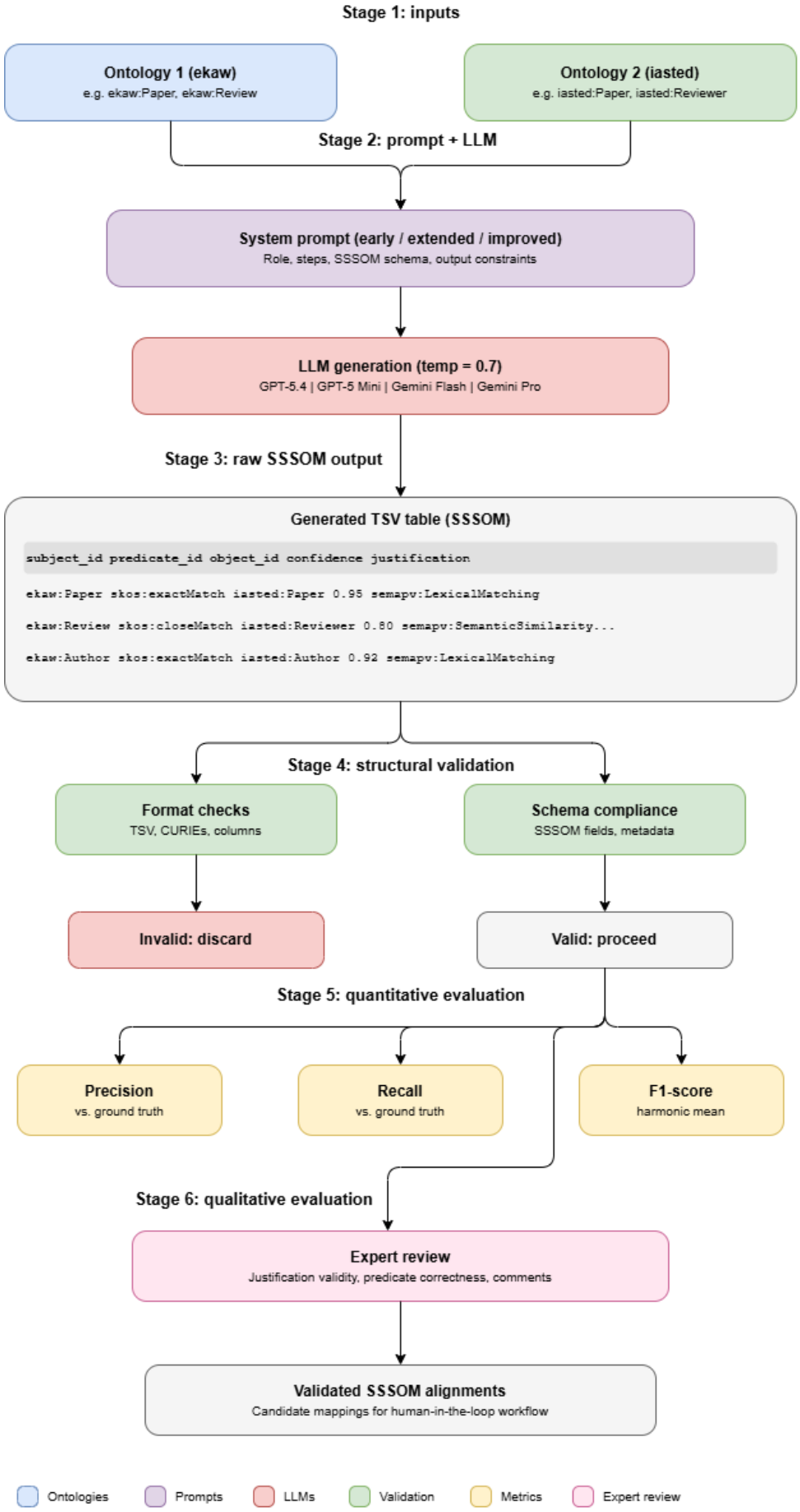

Figure 1: Diagram of the workflow of the proposed approachs