Towards Complex Ontology Alignment using Large Language Models

Reihaneh Amini[1], Sanaz Saki Norouzi[1], Pascal Hitzler[1], and Reza Amini[2]

[1] Kansas State University, Manhattan, KS, USA
{reihanea,sanazsn,hitzler}@ksu.edu
[2] Wright State University, Dayton, OH, USA amini.4@wright.edu

Abstract. Ontology alignment, a critical process in the Semantic Web for detecting relationships between different ontologies, has traditionally focused on identifying so-called "simple" 1-to-1 relationships through class labels and properties comparison. The more practically useful exploration of more complex alignments remains a hard problem to automate, and as such is largely underexplored, i.e. in application practice it is usually done manually by ontology and domain experts. Recently, the surge in Natural Language Processing (NLP) capabilities, driven by advancements in Large Language Models (LLMs), presents new opportunities for enhancing ontology engineering practices, including ontology alignment tasks. This paper investigates the application of LLM technologies to tackle the complex ontology alignment challenge. Leveraging a prompt-based approach and integrating rich ontology content – so-called *modules* – our work constitutes a significant advance towards automating the complex alignment task.

Keywords: Complex Ontology Alignment · Ontology · Large Language Model · Knowledge Graph · Modular Ontology Modeling

1 Introduction

Ontology alignment (sometimes called ontology matching) [8] is the task of establishing mappings between different ontologies, and as a research field it is concerned with ways to automate or at least semi-automate this task. For those not familiar with the field: Ontologies, which are usually knowledge bases expressed using Description Logics [12] (including the W3C Web Ontology Language – OWL – standard [22]) in this case act as a type of data schema for data expressed as knowledge graphs [11], i.e. the establishing of mappings between ontologies is central for schema-based data integration purposes.

Ontology alignment has been studied for over two decades, resulting in the development of many alignment approaches and systems. The majority of these systems are designed to detect only so-called "simple" 1-to-1 mappings between ontologies, primarily by establishing equivalence relationships between classes (unary predicates), or between properties (binary relationships); for example, one ontology may have a class called "Person" while another may have a class

called "Human", and an ontology alignment mapping may state that these two classes are in fact equivalent. It has long been recognized in the Semantic Web and Ontologies community that such simple mappings are helpful but ultimately insufficient for data integration task, for which mappings would need to be in the form of complex mapping rules[3], that can be expressed, e.g., as Datalog rules. However, detecting complex alignments between ontologies remains a very challenging and thus largely unexplored area with only few contributions that made progress in restricted settings (see Section 2). In current practice, establishing complex alignments between two or more ontologies requires domain experts to collaborate and manually generate the alignments, and this is usually a very work-intensive and thus expensive task. Any automation or semi-automation would have significant added value.

The Ontology Alignment Evaluation Initiative (OAEI)[4] is a long-standing coordinated international effort aimed at improving and evaluating ontology alignment and coreference resolution technologies.[5] It organizes annual evaluation campaigns [20] that provide a controlled environment where participants can test their ontology alignment systems using various benchmark tests. The benchmarks cover a range of complexity levels and real-world scenarios, aiming to simulate different aspects of the ontology alignment process.

With significant advancements in the natural language processing (NLP) and natural language understanding (NLU) fields, spurred by Large Language Models (LLMs), it has become possible to extract meanings from text and reason about it more effectively. OpenAI[6] has been at the forefront of this research, developing the Generative Pre-trained Transformer (GPT) series of models, which have attracted considerable attention from researchers, developers, and users. One of the most notable models, introduced in March 2022, was GPT-4 [1]. This transformer-based model is designed to predict the next token in generating text and has shown improvements in producing results that more closely align with user intent compared to its predecessor, GPT-3.5, on 70.2% of the prompts.[7]

Recent advancements in applying Large Language Models (LLMs) to Semantic Web and ontology engineering tasks have shown promising results, due to the importance of NLP for such tasks. A notable study [10] demonstrated the effectiveness of zero-shot and few-shot prompting with LLMs on various tasks within the Ontology Alignment Evaluation Initiative (OAEI), highlighting their potential in this area. The study was restricted to simple alignments. Indeed, as we will see later, a straightforward tasking of LLMs with the production of complex alignments does not quite work.

---

[3] See Section 4 for an example.

[4] http://oaei.ontologymatching.org/

[5] Co-reference resolution is about establishing equivalence and non-equivalence mappings between individuals, or constants, a related but different task of similar practical importance.

[6] https://openai.com

[7] See OpenAI: Introducing ChatGPT, 2022, https://openai.com/blog/chatgpt and Greg Brockman, Peter Welinder, Mira Murati, and OpenAI, 2020, https://openai.com/blog/openai-api.

One of the difficulties with ontology alignment is that ontologies often tend to be underspecified and with little internal structure that may add some self-explainability. This can be seen for example by the considerable disagreement between humans as to "correct" alignments, even for the simple alignment task [6]. It has been posited that additional internal structure, e.g. in the form of conceptual "ontology modules" should aid with ontology engineering tasks that are hard to automate [23]. Following this hitherto merely conceptual argument, we made use of ontology modules in our approach to generate complex alignments, and as we will report below, our prompt-based approach for discovering complex alignments between ontologies yields significantly better results when richer content in the form of ontology modules is available.

Since ontologies are knowledge bases expressed using formal logic, and mapping rules are also expressed using formal logic and processed as such, ontology alignment is a key symbolic task that we are here addressing usine "neural" means (i.e. LLMs as artificial neural networks), as such contributing to the body of approaches and methods for neural-symbolic learning and reasoning [13].

The structure of this paper is as follows: Section 2 reviews past research and the current state-of-the-art methods in complex alignment. Section 3 outlines the methodologies and approaches we utilized to experiment with an effective prompt-based method for discovering complex alignments. The evaluation of our approach and the specifics of prompt tailoring are discussed in Sections 4 and 5, respectively.

2   Related Work

In the semantic web domain, ontology alignment is crucial to discerning correspondences between concepts and properties across various ontologies. It involves two primary types of matching: simple and complex alignment [24]. Simple alignment focuses on identifying 1-to-1 equivalence relationships between concepts in two different ontologies. Complex alignment aims to match 1-to-n, n-to-1, or m-to-n equivalence or subsumption relationships between complex expressions in two ontologies (see e.g. [28]); the alignments are usually expressed in form of (first-order logic) rules, description logic axioms, or similar logical or quasi-logical notation.

Most of the work on ontology alignment is evaluated by the OAEI, an annual event focusing on ontology-matching systems' performance, starting in 2004, with focus mostly on simple alignment. A complex alignment track was started in 2018 because simple matching is often not sufficient to capture the rich semantics needed for advanced applications using realistic ontologies. In the context of this track, the datasets utilized throughout these years encompass Complex Conference, Populated Complex Conference, Hydrography, GeoLink, Populated GeoLink, Populated Enslaved, and Taxon. It should be noted, however, that some datasets have ceased to be evaluated in recent years [2,28,20] as each track depends on volunteers to run it.

The GeoLink complex alignment benchmark stands out as a real-world dataset for testing complex ontology alignment; using real-world datasets like GeoLink for testing tasks such as ontology alignment is crucial as it reflects the complexities of practical applications, helps validate the robustness and versatility of ontology alignment algorithms, and ensures their effectiveness beyond synthetic ones. It contains two ontologies, the GeoLink Base Ontology (GBO) and the GeoLink Modular Ontology (GMO). These are paired with a reference alignment created with help from domain specialists [28,4]. During these years, only two or three systems registered for evaluation in this track. Evaluation results show [3,19,18] that the systems that make use of instance data that is shared between the two ontologies perform relatively well, with high precision but low recall, i.e. many matches are missed. Furthermore, the availability of shared instance data, as for this benchmarking competition, while helpful for advancing the state of the art, is unrealistic for most practical application scenarios for ontology alignment: Usually, such shared data would not be availble. The new results which we present in this paper have been obtained without taking shared instance data into consideration, which, in our opinion, constitutes a major advance over the state of the art.

Currently, there is a plethora of reserach on applications of LLMs in a wide variety of fields, often with significant success. In the field of ontology alignment, [17] presents a zero-shot evaluation of ChatGPT-4 using the OAEI conference track ontologies (consisting of small and medium sized ontologies), showing a high recall and F1 of 0.52. Also, [9] provides a zero-shot evaluation of GPT 3.5 and Flan-T5-XXL on the OAEI Bio-ML track, and they also achieved low precision and high recall. [10] employs a comparable approach implemented in MELT, and the candidate generation and matching processes work in collaboration with Sentence-BERT, their evaluation has been done on some OAEI track datasets which almost achieve high F1 scores. In [21], they present a methodology utilizing LLM-based agents for retrieval and matching processes, which they enhance by incorporating Retrieval Augmented Generation (RAG) within their agents. The assessment was conducted on several tracks of OAEI. The results from the conference track indicate a high recall but low precision. The performance of the proposed method on other tracks is relatively good. However all these results do not go beyond simple alignment. In particular, no good results for complex alignment benchmarks have been reported yet.

3   Complex Alignment by Large Language Model

3.1   Ontology Modules

Ontologies that are conceptually clear are inherently more reusable, primarily because they are straightforward to comprehend from the outset. Consequently, a primary objective in the field of ontology research is to develop ontology modeling methodologies for creating ontologies with high conceptual clarity. It is a
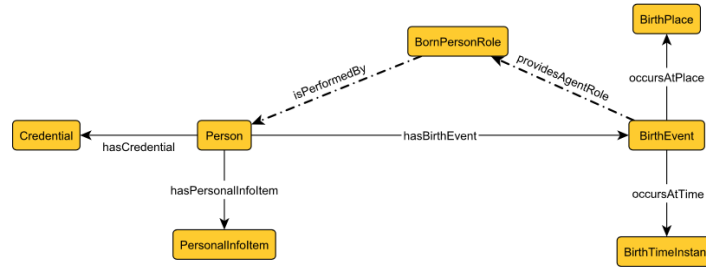
Fig. 1. A diagram for a Person module

fundamental assumption that adhering to well-established modeling principles enhances the understandability and reusability of an ontology [23].

It is known that repurposing or adjusting of ontologies for new applications often presents challenges. These challenges stem from various factors, including: (i) mismatches in the level of detail between ontologies and the intended use-cases, (ii) unclear concepts within the ontologies intended for reuse, (iii) the absence and the complexity of ontology alignment practices, and (iv) the scarcity of tools designed to facilitate ontology reuse and support the process. To tackle these issues, the Modular Ontology Modeling (MOMo) approach was proposed [23]. In the MOMo framework, a key feature is the creation and connection of compact, independent modules, which offers numerous advantages. For instance, it streamlines maintenance because changes to a single module have little to no effect on the entire ontology. Additionally, tracking the origin of every ontology segment back to the initial requirements becomes straightforward through the use of module documentation or metadata.

In the Semantic Web community, the term **"module"** can mean many things. In our case, it refers to a specific part of an ontology that encapsulates a principal concept and its main features, such as a Person module capturing details about "BirthEvent", "PersonalInfoItem", and "Credential" as depicted in Figure 1. Modules serve dual roles: they are technical constructs that, on the one hand, demarcate parts of ontologies that group related classes and their interactions and, on the other hand, they do this in a way that aligns with domain experts' understanding. Despite potential overlaps and hierarchical structures within them, modules organize the ontology into a network of interrelated pieces, each mirroring the domain's conceptual framework as understood by experts.

Modules enable a strategic approach to ontology modeling by allowing the work to be broken down into manageable segments; initially focusing on individual modules before linking them together. This approach provides a clear way to manage the complexities of large, cohesive ontologies by breaking down the process into understanding individual modules and their interconnections. This modular approach also aligns with the way domain experts conceptualize their fields, making both the ontology and its documentation more accessible and comprehensible. Since each module can be easily swapped out for another—perhaps

one that offers a different level of detail—modifications are localized, making the entire system more adaptable [15].

In the research we present herein, we explore the impact of integrating module information on the effectiveness of identifying complex alignments. Specifically, we include descriptions, Core Axioms (where applicable), and Alignment information as outlined in the GeoLink Core Ontology Design Patterns [14] that constitute the GeoLink Modular Ontology (GMO), that had been developed as an integrated schema for combining several large-scale ocean science data repositories [7]. As we will see further below, the utilization of additional module information is of core importance in order to solve the complex alignment problem, in our setting.

3.2   Design of the Prompting Process

Tailoring an LLM for specific tasks can be achieved either by fine-tuning the model with select data or by using prompts for in-context learning. Fine-tuning requires significant computational resources and expertise, which may not be feasible for all users or organizations. Moreover, it risks over-fitting the model to the training data, potentially diminishing its ability to generalize to other tasks. In contrast, employing various prompting techniques on a fixed LLM is more resource-efficient, requiring less computational power, time, and expertise. Additionally, prompts allow for the flexible and on-the-fly adaptation of the model to a wide range of tasks by applying different prompting techniques [16].

Prompt engineering is increasingly vital for enhancing the performance of large language models (LLMs) across a wide range of tasks. These models have shown impressive capabilities using zero-shot prompts, where they generate responses without prior specific training examples. However, for more complex tasks requiring deeper reasoning, advanced prompting techniques are necessary. Incorporating context into prompts significantly improves model performance. Techniques such as few-shot prompting, where the model is given a few examples to learn from, or chain-of-thought prompting, which guides the model through a series of reasoning steps, are particularly effective in enhancing the model's understanding and response accuracy [26].

There is a range of task-agnostic prompting techniques available for use with large language models.[8] These include:

- Zero-shot prompting: This technique involves providing a single, natural language description of the task at inference time, without any prior examples.
- Few-shot prompting: This approach includes giving the model a few examples of the task, complete with context and successful outcomes, to guide its understanding and responses [5].
- Chain-of-thought prompting: This method involves supplying the model with a series of thought process examples, helping it to navigate through reasoning steps to arrive at a conclusion [25].

---

[8] https://www.promptingguide.ai/techniques

These techniques enable large language models to adapt and respond to a wide variety of tasks with varying levels of guidance and specificity.

Figure 2 illustrates our workflow, which begins by uploading the entire GMO file as an initial prompt. Subsequently, we include specific entities from the second ontology, the GBO, to inquire about their alignment with the GMO. If the initial prompt successfully provides the relevant segments, further prompts in the chain-of-thought process are unnecessary. Otherwise, we can provide a list of all GMO module names and ask GPT to identify the most related modules for the GBO entities. Then, in a subsequent prompt, we supplement the inquiry with module information and request the related segments again.

Fig. 2. Prompting workflow

In the following section, we examine the performance of GPT-4 in response to various entities, modules, and prompts it processes.

4   Evaluation

For the evaluation of our methodology, we employ the GeoLink Complex Alignment dataset previously detailed in Zhou et al., 2018 [27]. This dataset comprises 109 complex alignment rules between GMO and GBO, each of which has been thoroughly analyzed to test our hypothesis. Detailed findings from this study, including both the prompts used and the responses obtained, are available online.[9]

### 4.1   Prompting and Detailed Example

OpenAI's functionality includes the ability to load a prompt from a file, triggering backend processing to parse the uploaded data. We uploaded the entire GMO ontology RDF file in turtle (.ttl) format. OpenAI GPT-4 parsed the data, albeit with some latency, prompting a request for more specific tasks to be directed to it (Figure 3).

The complex alignment rule we are targeting in this example is

$$
\begin{aligned}
&\mathtt{Award}(x) \wedge \mathtt{hasCoPrincipalInvestigator}(x,z) \leftrightarrow \\
&\mathtt{FundingAward}(x) \wedge \mathtt{providesAgentRole}(x,y) \\
&\wedge \mathtt{CoPrincipalInvestigatorRole}(y) \wedge \mathtt{performedBy}(y,z)
\end{aligned}
\tag{1}
$$

where GBO entities `hasCoPrincipalInvestigator` (a relation or binary predicate or so-called object property) and `Award` (a class or unary predicate)

---

[9] https://daselab.cs.ksu.edu/publications/alignment-rules-gbo-gmo
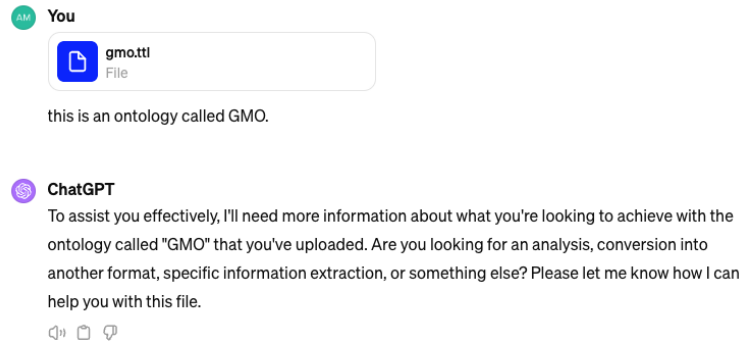
Fig. 3. Uploading GMO.ttl file as prompt.

can be found on the left-hand side, and the GMO entities on the right-hand side. This is an example of a complex alignment rule expressed in first-order predicate logic and our objective is to assess the effectiveness of detecting complex alignments by prompting GPT-4.

After uploading the GMO, we extract the GBO entities found on the left-hand side of the complex alignment rule (i.e., **"Award"** and **"hasCoPrincipal-Investigator"**) from the GBO RDF file in its original format. For our example, this looks as follows.

```
###  http://gbo#Award
main:Award rdf:type owl:Class ;
          rdfs:comment "Funding provided by an Organization
          enabling Participation.";
          rdfs:label "Award" .
###  http://gbo#hasCoPrincipalInvestigator
main:hasCoPrincipalInvestigator rdf:type owl:ObjectProperty ;
                  owl:inverseOf main:isCoPrincipalInvestigatorOf ;
                  rdfs:domain [ rdf:type owl:Class ;
                                owl:unionOf ( main:Award
                                              main:Program
                                            )
                              ] ;
                  rdfs:range main:Person ;
                  rdfs:label "hasCoPrincipalInvestigator" .
```

Next, our prompt instructs GPT-4 to examine the components in GMO that are associated with these elements in GBO, as depicted in Figure 4.

What we observed from GPT-4's responses in this step is often not a comprehensive answer. Many times, it explicitly states that *It appears there were no results found in the ontology file "gmo.ttl" directly related to the terms you mentioned from the "GBO" ontology, such as "Award", "hasCoPrincipalInvestigator", "Person", "Program", or "isCoPrincipalInvestigatorOf".* GPT-4 typically continues its response with further inquiries such as *To proceed, I can*
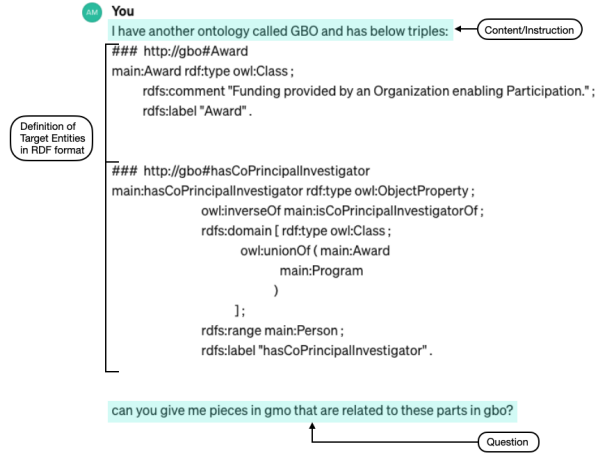
Fig. 4. GBO-related instruction and question in prompt

manually examine the "gmo.ttl" file to identify any related concepts or proper-
ties that might align with those in the "GBO" ontology. This process involves
looking for classes or properties that serve similar purposes or are conceptually
related, even if they are not directly named the same. Would you like me to per-
form this manual examination, or is there another way I can assist you with
these ontologies?

Our consistent response to this question has always been "Yes", and the
typical output we received from GPT is illustrated in Fig 5. As depicted, the
LLM successfully identifies a similar entity to **gbo#Award**, which in GMO
is **gmo#AwardAmount** (highlighted in green). However, it struggles to con-
duct further analysis to identify alignments in GMO related to the relation-
ship between "Award" and "hasCoPrincipalInvestigator" in GBO. Rather, it
attempts to generate random information about recently detected entities such
as "FundingAward", which hypothetically arises from its next-word-prediction
logic (highlighted in yellow). However, the remaining generated information, such
as "gmo#Program" (highlighted in red), is not relevant to this component in
GBO. Ultimately, it states that *There were no direct matches found for hasCo-*
PrincipalInvestigator or isCoPrincipalInvestigatorOf in the GMO ontology.

We also attempted zero-shot prompting at this stage, combining both the
GMO file and the second prompt into a single input to GPT-4. This approach
resulted in increased latency and a confused response.

Our conclusion from this analysis is that GPT-4 failed to detect the com-
plex alignment between the two ontologies. Instead, it provided partially related
objects, including random information about different unrelated entity classes.
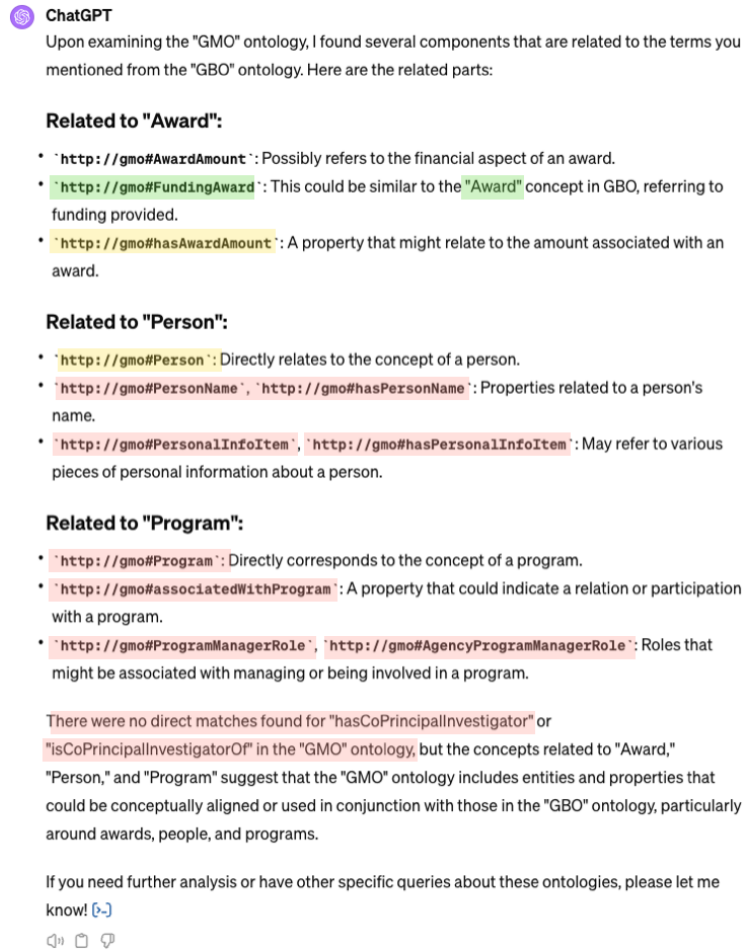
ChatGPT

Upon examining the "GMO" ontology, I found several components that are related to the terms you mentioned from the "GBO" ontology. Here are the related parts:

Related to "Award":

* `http://gmo#AwardAmount`: Possibly refers to the financial aspect of an award.
* `http://gmo#FundingAward`: This could be similar to the "Award" concept in GBO, referring to funding provided.
* `http://gmo#hasAwardAmount`: A property that might relate to the amount associated with an award.

Related to "Person":

* `http://gmo#Person`: Directly relates to the concept of a person.
* `http://gmo#PersonName`, `http://gmo#hasPersonName`: Properties related to a person's name.
* `http://gmo#PersonalInfoItem`, `http://gmo#hasPersonalInfoItem`: May refer to various pieces of personal information about a person.

Related to "Program":

* `http://gmo#Program`: Directly corresponds to the concept of a program.
* `http://gmo#associatedWithProgram`: A property that could indicate a relation or participation with a program.
* `http://gmo#ProgramManagerRole`, `http://gmo#AgencyProgramManagerRole`: Roles that might be associated with managing or being involved in a program.

There were no direct matches found for "hasCoPrincipalInvestigator" or "isCoPrincipalInvestigatorOf" in the "GMO" ontology, but the concepts related to "Award," "Person," and "Program" suggest that the "GMO" ontology includes entities and properties that could be conceptually aligned or used in conjunction with those in the "GBO" ontology, particularly around awards, people, and programs.

If you need further analysis or have other specific queries about these ontologies, please let me know! (>_-)

Fig. 5. GPT-4 response to our initial question

This is a list of all the modules and patterns that are defined in GMO ontology:

Agent , Agent Role, Event, Information Object, Identifier, Person, Personal Info Item pattern, Person Name, Organization, Funding Award, Program, Place, Cruise, Platform Pattern Stub, Vessel pattern, Physical Sample pattern, Property Value pattern.

From this list, give me relevant module names that can help us find the most related GMO pieces.
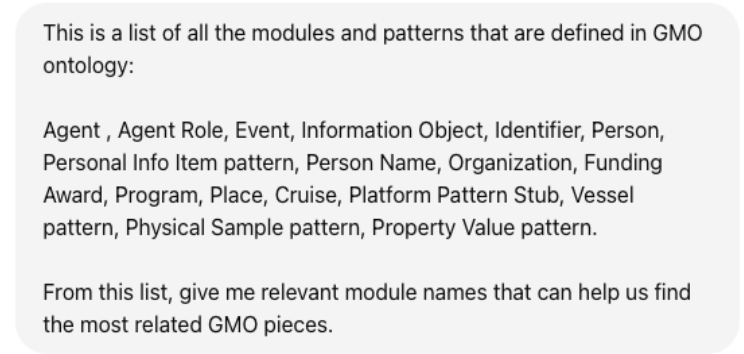
Fig. 6. Prompt for module name suggestion

To proceed, let us revisit the GMO modules discussed in Section 3.1. The documentation [14] contains an informal description of the modules (in the documentation called *patterns*) in the GMO ontology, accompanied by visual depictions of these patterns. In the ongoing chain-of-thoughts prompting processes, we included the module names and asked for the most relevant module names related to the GBO pieces, as shown in Figure 6. The result of this prompt is usually a single module or a list of suggested modules based on the information that the GPT has processed through the chain of prompts. In the next prompt, we included the descriptions of the suggested modules along with the question, as illustrated in the prompt below.

```
you couldn't give me all the pieces I need in GMO. here is more info that
can help: The Funding Award pattern describes the funding awards that
fund  all kinds of ocean science research activities. We use the
isFundedBy property to connect anything to a funding award if the funding
award funds it. Each funding award has exactly one starting and ending
date (aligned with time:Instant). It provides at most one award amount,
which is described via a pair of decimal value and currency code. The
currency code is not specified here, but existing standards can be used,
e.g., ISO  4217.  There may be people or organizations that have a role
in a funding award. This is modeled by re-using (and aligning with) the
Agent Role pattern. In this version, we include the following types of
agent-roles, represented as classes: SponsorRole,
AgencyProgramManagerRole, PrincipalInvestigatorRole, and
CoPrincipalInvestigatorRole. Additional roles are possible in the future
versions. Each funding award is described by an InformationObject,
which when aligned to the Information Object pattern, allows one to
represent additional information such as identifier, description, etc.

ObjectProperty: isFundedBy
ObjectProperty: startsOnDate
ObjectProperty: endsOnDate
ObjectProperty: isDescribedBy
ObjectProperty: providesAgentRole
ObjectProperty: isPerformedBy
ObjectProperty: hasAwardAmount
ObjectProperty: hasCurrencyCode
DataProperty: hasCurrencyValue
Class: FundingAward
Class: Agent
Class: TimeInstant
Class: AgentRole
Class: SponsorRole
Class: PrincipalInvestigatorRole
Class: CoPrincipalInvestigatorRole
Class: AgencyProgramManagerRole
Class: InformationObject
Class: AwardAmount
Class: CurrencyCode
```

Updated Analysis for Equivalence in "GMO"

- **Class "FundingAward"** in GMO corresponds to "Award" in GBO.
- **Object Property "providesAgentRole"** in GMO can be used to connect a "FundingAward" to different roles, including "PrincipalInvestigatorRole" and "CoPrincipalInvestigatorRole".

Constructing the Equivalence Rule

Given the "GBO" relationship:

- **"Award" (x)** ∧ **"hasCoPrincipalInvestigator" (x, z)**

The equivalent in "GMO", incorporating the additional details provided, would be:

- **"FundingAward" (x)** ∧ **"providesAgentRole" (x, y)** ∧ **"CoPrincipalInvestigatorRole" (y)** ∧ **"isPerformedBy" (y, z)**

Here's the breakdown:

- **"FundingAward" (x)**: Directly equates to the "Award" concept in GBO, representing the funding award.
- **"providesAgentRole" (x, y)**: Connects a funding award to a role, which can include "CoPrincipalInvestigatorRole".
- **"CoPrincipalInvestigatorRole" (y)**: Specifies the role related to the funding award, equivalent to having a co-principal investigator in GBO.
- **"isPerformedBy" (y, z)**: Links the role (e.g., CoPrincipalInvestigatorRole) to the person or organization performing it, closing the equivalence to "hasCoPrincipalInvestigator" in GBO.

Fig. 7. GPT-4 response to our question with Module information included

By evaluating the outcome (see Fig. 7) of the prompt, it is evident that GPT-4 has a more informed and precise approach to conducting investigations and identifying semantically related components if it has the module information available. GPT-4 identifies all components related to the GMO as outlined in the alignment rule mentioned above. It further elaborates on each component and explains their interconnections. The full evaluation data of our study over all complex alignment rules is available online.[10]

Impact of few-shot vs chain-of-thought: Comparing few-shot and chain-of-thought approaches, we noted the differences between providing information in a single zero-shot prompt versus delivering it in a series of prompts. GPT tends to become confused about the question and the relevant information it needs to process in a zero-shot scenario. In contrast, introducing the prompt

---

[10] `https://github.com/reddraa/complex_ontology_alignment`

as part of a sequential chain of information clarifies the data pieces and their meanings for the model.

Impact of adding modules to the prompt: We observed that while GPT-4 nearly grasps the query regarding the relevant components needed to fulfill the rule, it typically identifies only 10-20% of the required elements if no module information is given. However, if details about the module in the GMO are presented in a single prompt before posing the question again, GPT-4 is significantly more effective, clearly identifying the majority of the targeted components.

4.2 Quantitative Evaluation

To assess the effectiveness of our approach, we identified key entities within our complex alignment rules to serve as the basis for our metrics. As illustrated in Figure 7, the entities detected as relevant by GPT-4 include FundingAwards(z), providesAgentRole(x,y), CoPrinciplaInvestigatorRole(y), isPerformedBy(y,z). It is important to note that response formats may vary, hence each response might be unique. We employed two key performance metrics: recall and precision. Recall, defined as the ratio of detected GMO-related instances to the total number of expected instances, evaluates our study's ability to identify all relevant GMO pieces comprehensively. Precision, on the other hand, measures the accuracy of our detection process by calculating the proportion of correctly identified GMO pieces out of all the instances flagged in our findings. Together, these metrics provide a holistic view of our study's identification capability, striking a balance between thoroughness and accuracy in detecting GMO-related instances of a complex alignment rule. The well-known definitions are as follows.

$$\text{Recall} = \frac{\text{Number of Correctly Detected Pieces}}{\text{Total Number of GMO Pieces in Complex Alignment}}$$

$$\text{Precision} = \frac{\text{Number of Correctly Detected Pieces}}{\text{Total Number of Detected Pieces}}$$

According to Table 1, in the study presented in the paper, an analysis of 109 complex alignment rules was conducted. The findings revealed that in only 4.5% of these alignments, GMO components were identified solely through the use of GBO entities (i.e., not using module information). Both identified complex alignment pieces, without the module, are actually simple $1 : 1$ mappings, e.g., Program(x) $\leftrightarrow$ Program(x), where 'Program' is an existing class in both the GMO and GBO ontology. However, in over 95% of the cases, the identification of GMO components was achieved through the application of information from the GMO module information, highlighting its significant role in the detection process.

In assessing GPT-4's performance on entity detection in the complex alignment, we analyzed the example alignment mentioned earlier in this paper. We compared the alignment pieces returned by GPT-4, shown in Figure 7 for our

Table 1. Distribution of successful approaches for detecting complex alignment rules

detected pieces without module Information	5
detected pieces with module Information	104
total number of complex alignment rules	109

running example based on rule (1), with four corresponding entities defined in the GMO ontology's alignment rule, i.e., FundingAward(x), providesAgent-Role(x,y), CoPrincipalInvestigatorRole(y), performedBy(y,z). Note that we did not assess the actual return of alignment rules, but rather whether the relevant pieces (predicates) were detected. While the detection (without actual composition of the pieces into a rule) constitutes a simpler task than producing the rules, it nevertheless captures the core difficulty for complex ontology alignment. In fact, if the pieces are correct, the actual rule can easily be assembled by a human (or, in many cases, by a symbolic algorithm based on the ontology and example data).

In evaluating recall (coverage) in this setting, we found that all four expected GMO entities were accurately identified by GPT-4 for our running example, yielding a recall of 1.0 in this case. Additionally, for precision, we examined the entities returned by GPT-4 and found that aside from the four correct GMO entities, no irrelevant entities were detected, also resulting in a precision of 1.0. This indicates perfect alignment detection by GPT-4 in this instance. To further clarify the evaluation of recall, consider that if the prompt response from GPT-4 had included additional entities such as "Event" or "Place" which were not part of the expected entities, it would have negatively impacted the precision. These extraneous entities, not being included in the set of expected results, would reduce the precision as they represent incorrect identifications according to the specified alignment rule. In practical terms, the return of (in particular, many) superfluous entities by the system would make assembly of the actual rule by a human more difficult.

In Table 2 we see that for 73.3% of the complex alignment rules evaluated, the recall value exceeded 0.5. This indicates that more than half of the GMO entities involved in a complex alignment rule were successfully detected for this ratio in our population. Furthermore, the recall value surpasses 0.75 for approximately 62.3% of these rules, signifying a higher accuracy in detection. It's also noteworthy that *for 45% of the alignment rules recall is a perfect 1.0, and for 45.8% precision is also 1.0*, because of the integration of module information in the analysis process.

The precision metric indicates the accuracy of the responses in directing us toward expected entities within the GMO ontology. Our analysis found that responses achieved a precision higher than 0.75 for 59.6% of the evaluated records. Additionally, when the precision threshold is lowered to 0.5—implying that half of the entities suggested by the language model are the expected ones, while the other half may include relevant or irrelevant entities—the coverage of alignment rules increases to 69.7%.

Table 2. Recall and Precision for detected GMO entities using module information

	Recall			Precision		
	≥ 0.5	≥ 0.75	$= 1$	≥ 0.5	≥ 0.75	$= 1$
with Module Information	73.3%	62.3%	45.0%	69.7%	59.6%	45.8%
mean	0.67			0.67		
median	0.75			0.80		
standard deviation	0.37			0.37		

Note that detection (recall) of half or more of the correct body predicates, paired with high precision, is already very helpful for human assembly of a rule.

The recall data exhibits a mean of 0.67 and a median of 0.75, indicating that half of the recall values exceed 0.75, while the other half fall below this threshold. This distribution highlights that the majority of our data points demonstrate recall values above 0.75, offering insight into the central tendency of our dataset. Similarly, for precision, both the median and mean values suggest promising results, with at least 50% of the records achieving a precision greater than 0.8. This underscores a generally reliable level of accuracy in the data.

We close the discussion with a number of additional more detailed observations.

– Type and Class Type alignments: An intriguing insight from the analysis of complex alignment rules is the difficulty encountered with Type or Class alignment. This process frequently falls short in identifying pertinent alignments. A case in point is the alignment where GeoFeatureType(x) in GBO transitions to rdfs:subClassOf(x, Place) in GMO. In such instances, the GMO ontology attempts to consolidate individual classes into types or subclasses. Nonetheless, the absence of this specific information in the GMO's module data impedes the language model's (LLM) ability to accurately discern these sophisticated alignments, resulting in the provision of erroneous information.
– Rich Modules, Improved Discovery: Another observation is that modules enriched with detailed descriptions, core axioms or alignments, as discussed in the referenced paper [15], depict significantly higher accuracy in discovery. Specifically, modules that offer comprehensive information facilitate better recall and precision rates. For example, the Cruise() entity, which involves approximately 18 complex alignment rules, demonstrates an impressive average recall rate of 0.93. This indicates that the depth of information provided directly influences the effectiveness of discovery processes.

5   Discussion and Future Work

The quantitative results we have just presented indicate that our setting—i.e., under inclusion of module information—produces high precision and recall values in many cases. Our results demonstrate a significant improvement compared

to this baseline. While, absolutely speaking, the quantitative results are still moderate, we are in fact presenting the *very first* reasonably working approach for generating complex alignments that does not require shared individuals; as noted, shared individuals are far from a realistic setting in practice. In the latest evaluation of the participating systems in the OAEI complex alignment track, which was in 2021, most systems failed to detect any m:n complex alignments.[11] Only two systems procuded complex alignments,[12] but they required instance data. As such our contribution shows a path forward towards complex alignment in realistic settings, a challenge that had so far eluded researchers.

For our approach to work, we provided the LLM with module information, following the previously presented arguments that (1) ontologies without additional internal structure or meaningful additional information are often too ambiguous for automated complex alignment tasks, (2) module identification during the ontology design process can easily be provided by the ontology modelers at least as part of the documentation (while doing so post-hoc, by others, requires major efforts), and (3) providing such module information as part of ontologies (and/or their documentations) would likely significantly decrease the effort and cost of many ontology engineering tasks, including complex alignment [23]. As such, this is also a (repeated) call for improving ontology modeling methods to additionally provide module structure.

While our results, as presented, are very encouraging, substantial future investigations will be required to cast them into an ontology alignment system that can work autonomously at high precision. Intermediate steps could constitute human-in-the-loop approaches where a human ontology engineer receives suggestions from an LLM, e.g., as to the relevant modules for a question, post-processes the LLM responses by manually checking the small number of suggestions, and feeding the correct suggestions back to the LLM for obtaining more complete responses. This is in line with the idea of an assisting system that limits the number of checks for the human, as opposed to the vast number of potential checks that would have to be done manually without such a limiting system.

In the future, we also intend to extend out approach to additional datasets featuring complex alignments for both evaluation and analytical objectives. Furthermore, we aim to explore alternative representations of modules to LLMs and evaluate the model's performance with these variations. Fine-tuning existing LLMs, but also the integration of additional symbolic data or algorithms, e.g. pertaining to logical axioms that come with well-designed ontologies, and also the integration of traditional simple alignment algorithms to further assist with the complex alignment task are all on our path forward.

6   Conclusion

We have presented the very first approach that is able to achieve good accuracy for complex ontology alignment without relying on shared individuals. The sys-

---

[11] https://oaei.ontologymatching.org/2021/results/complex/geolink/index.html
[12] https://oaei.ontologymatching.org/2021/results/complex/popgeolink/index.html

tem is neural-symbolic in its nature as it addresses a symbolic task (complex ontology alignment), the output of which are alignment rules expressed in some logic, it furthermore makes decisive use of additional symbolic input in the form of ontology modules, and it uses an LLM as core processing engine. Our results suggest that further work bear the promise to result in strong complex ontology alignment systems, for ontologies that carry sufficient internal structure.

References

1. Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
2. Alsayed Algergawy, Michelle Cheatham, Daniel Faria, Alfio Ferrara, Irini Fundulaki, Ian Harrow, Sven Hertling, Ernesto Jiménez-Ruiz, Naouel Karam, Abderrahmane Khiat, et al. Results of the ontology alignment evaluation initiative 2018. In *13th International Workshop on Ontology Matching co-located with the 17th ISWC (OM 2018)*, volume 2288, pages 76–116, 2018.
3. Alsayed Algergawy, Daniel Faria, Alfio Ferrara, Irini Fundulaki, Ian Harrow, Sven Hertling, Ernesto Jiménez-Ruiz, Naouel Karam, Abderrahmane Khiat, Patrick Lambrix, Huanyu Li, Stefano Montanelli, Heiko Paulheim, Catia Pesquita, Tzanina Saveta, Pavel Shvaiko, Andrea Splendiani, Élodie Thiéblin, Cássia Trojahn, Jana Vatascinová, Ondrej Zamazal, and Lu Zhou. Results of the ontology alignment evaluation initiative 2019. In Pavel Shvaiko, Jérôme Euzenat, Ernesto Jiménez-Ruiz, Oktie Hassanzadeh, and Cássia Trojahn, editors, *Proceedings of the 14th International Workshop on Ontology Matching co-located with the 18th International Semantic Web Conference (ISWC 2019), Auckland, New Zealand, October 26, 2019*, volume 2536 of *CEUR Workshop Proceedings*, pages 46–85. CEUR-WS.org, 2019.
4. Reihaneh Amini, Lu Zhou, and Pascal Hitzler. Geolink cruises: A non-synthetic benchmark for co-reference resolution on knowledge graphs. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 2959–2966, 2020.
5. Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
6. Michelle Cheatham and Pascal Hitzler. Conference v2. 0: An uncertain version of the oaei conference benchmark. In *The Semantic Web–ISWC 2014: 13th International Semantic Web Conference, Riva del Garda, Italy, October 19-23, 2014. Proceedings, Part II 13*, pages 33–48. Springer, 2014.
7. Michelle Cheatham, Adila Krisnadhi, Reihaneh Amini, Pascal Hitzler, Krzysztof Janowicz, Adam Shepherd, Tom Narock, Matt Jones, and Peng Ji. The geolink knowledge graph. *Big Earth Data*, 2(2):131–143, 2018.
8. Jérôme Euzenat and Pavel Shvaiko. *Ontology Matching, Second Edition*. Springer, 2013.

9. Yuan He, Jiaoyan Chen, Hang Dong, and Ian Horrocks. Exploring large language models for ontology alignment. In Irini Fundulaki, Kouji Kozaki, Daniel Garijo, and José Manuél Gómez-Pérez, editors, *Proceedings of the ISWC 2023 Posters, Demos and Industry Tracks: From Novel Ideas to Industrial Practice co-located with 22nd International Semantic Web Conference (ISWC 2023), Athens, Greece, November 6-10, 2023*, volume 3632 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2023.

10. Sven Hertling and Heiko Paulheim. Olala: Ontology matching with large language models. In *Proceedings of the 12th Knowledge Capture Conference 2023*, pages 131–139, 2023.

11. Pascal Hitzler. A review of the semantic web field. *Commun. ACM*, 64(2):76–83, 2021.

12. Pascal Hitzler, Markus Krötzsch, and Sebastian Rudolph. *Foundations of Semantic Web Technologies*. Chapman and Hall/CRC Press, 2010.

13. Pascal Hitzler, Md. Kamruzzaman Sarker, and Aaron Eberhart, editors. *Compendium of Neurosymbolic Artificial Intelligence*, volume 369 of *Frontiers in Artificial Intelligence and Applications*. IOS Press, 2023.

14. Adila Krisnadhi, Yingjie Hu, Robert Arko, Suzanne Carbotte, Cynthia Chandler, Michelle Cheatham, Douglas Fils, Timothy Finin, Pascal Hitzler, Krzysztof Janowicz, et al. Geolink core ontology design patterns. Technical report. Available from `https://people.cs.ksu.edu/~hitzler/pub2/gmo-tr.pdf`.

15. Adila Krisnadhi, Yingjie Hu, Krzysztof Janowicz, Pascal Hitzler, Robert Arko, Suzanne Carbotte, Cynthia Chandler, Michelle Cheatham, Douglas Fils, Timothy Finin, et al. The geolink modular oceanography ontology. In *The Semantic Web-ISWC 2015: 14th International Semantic Web Conference, Bethlehem, PA, USA, October 11-15, 2015, Proceedings, Part II 14*, pages 301–309. Springer, 2015.

16. Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35, 2023.

17. Sanaz Saki Norouzi, Mohammad Saeid Mahdavinejad, and Pascal Hitzler. Conversational ontology alignment with chatgpt. In Pavel Shvaiko, Jérôme Euzenat, Ernesto Jiménez-Ruiz, Oktie Hassanzadeh, and Cássia Trojahn, editors, *Proceedings of the 18th International Workshop on Ontology Matching co-located with the 22nd International Semantic Web Conference (ISWC 2023), Athens, Greece, November 7, 2023*, volume 3591 of *CEUR Workshop Proceedings*, pages 61–66. CEUR-WS.org, 2023.

18. Mina Abd Nikooie Pour, Alsayed Algergawy, Florence Amardeilh, Reihaneh Amini, Omaima Fallatah, Daniel Faria, Irini Fundulaki, Ian Harrow, Sven Hertling, Pascal Hitzler, Martin Huschka, Liliana Ibanescu, Ernesto Jiménez-Ruiz, Naouel Karam, Amir Laadhar, Patrick Lambrix, Huanyu Li, Ying Li, Franck Michel, Engy Nasr, Heiko Paulheim, Catia Pesquita, Jan Portisch, Catherine Roussey, Tzanina Saveta, Pavel Shvaiko, Andrea Splendiani, Cássia Trojahn, Jana Vatascinová, Beyza Yaman, Ondrej Zamazal, and Lu Zhou. Results of the ontology alignment evaluation initiative 2021. In Pavel Shvaiko, Jérôme Euzenat, Ernesto Jiménez-Ruiz, Oktie Hassanzadeh, and Cássia Trojahn, editors, *Proceedings of the 16th International Workshop on Ontology Matching co-located with the 20th International Semantic Web Conference (ISWC 2021), Virtual conference, October 25, 2021*, volume 3063 of *CEUR Workshop Proceedings*, pages 62–108. CEUR-WS.org, 2021.

19. Mina Abd Nikooie Pour, Alsayed Algergawy, Reihaneh Amini, Daniel Faria, Irini Fundulaki, Ian Harrow, Sven Hertling, Ernesto Jiménez-Ruiz, Clément Jonquet,

Naouel Karam, Abderrahmane Khiat, Amir Laadhar, Patrick Lambrix, Huanyu Li, Ying Li, Pascal Hitzler, Heiko Paulheim, Catia Pesquita, Tzanina Saveta, Pavel Shvaiko, Andrea Splendiani, Élodie Thiéblin, Cássia Trojahn, Jana Vatascinová, Beyza Yaman, Ondrej Zamazal, and Lu Zhou. Results of the ontology alignment evaluation initiative 2020. In Pavel Shvaiko, Jérôme Euzenat, Ernesto Jiménez-Ruiz, Oktie Hassanzadeh, and Cássia Trojahn, editors, *Proceedings of the 15th International Workshop on Ontology Matching co-located with the 19th International Semantic Web Conference (ISWC 2020), Virtual conference (originally planned to be in Athens, Greece), November 2, 2020*, volume 2788 of *CEUR Workshop Proceedings*, pages 92–138. CEUR-WS.org, 2020.

20. Mina Abd Nikooie Pour, Alsayed Algergawy, Patrice Buche, Leyla Jael Castro, Jiaoyan Chen, Adrien Coulet, Julien Cufi, Hang Dong, Omaima Fallatah, Daniel Faria, Irini Fundulaki, Sven Hertling, Yuan He, Ian Horrocks, Martin Huschka, Liliana Ibanescu, Sarika Jain, Ernesto Jiménez-Ruiz, Naouel Karam, Patrick Lambrix, Huanyu Li, Ying Li, Pierre Monnin, Engy Nasr, Heiko Paulheim, Catia Pesquita, Tzanina Saveta, Pavel Shvaiko, Guilherme Sousa, Cássia Trojahn, Jana Vatascinová, Mingfang Wu, Beyza Yaman, Ondrej Zamazal, and Lu Zhou. Results of the ontology alignment evaluation initiative 2023. In Pavel Shvaiko, Jérôme Euzenat, Ernesto Jiménez-Ruiz, Oktie Hassanzadeh, and Cássia Trojahn, editors, *Proceedings of the 18th International Workshop on Ontology Matching co-located with the 22nd International Semantic Web Conference (ISWC 2023), Athens, Greece, November 7, 2023*, volume 3591 of *CEUR Workshop Proceedings*, pages 97–139. CEUR-WS.org, 2023.

21. Zhangcheng Qiang, Weiqing Wang, and Kerry Taylor. Agent-om: Leveraging large language models for ontology matching. *CoRR*, abs/2312.00326, 2023.

22. Sebastian Rudolph, Markus Krötzsch, Peter Patel-Schneider, Pascal Hitzler, and Bijan Parsia. OWL 2 web ontology language primer (second edition). W3C recommendation, W3C, December 2012. https://www.w3.org/TR/2012/REC-owl2-primer-20121211/.

23. Cogan Shimizu, Karl Hammar, and Pascal Hitzler. Modular ontology modeling. *Semantic Web*, 14(3):459–489, 2023.

24. Pavel Shvaiko and Jérôme Euzenat. Ontology matching: state of the art and future challenges. *IEEE Transactions on knowledge and data engineering*, 25(1):158–176, 2011.

25. Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.

26. Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C Schmidt. A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*, 2023.

27. Lu Zhou, Michelle Cheatham, Adila Krisnadhi, and Pascal Hitzler. A complex alignment benchmark: Geolink dataset. In *The Semantic Web–ISWC 2018: 17th International Semantic Web Conference, Monterey, CA, USA, October 8–12, 2018, Proceedings, Part II 17*, pages 273–288. Springer, 2018.

28. Lu Zhou, Michelle Cheatham, Adila Krisnadhi, and Pascal Hitzler. Geolink data set: A complex alignment benchmark from real-world ontology. *Data Intelligence*, 2(3):353–378, 2020.