Can Language Models Align Biomedical Ontologies?: Evaluating Retrieval-Augmented Prompt Strategies in Bio-ML.

Lucas Ferraz[1], Pedro Giesteira Cotovio[1] and Catia Pesquita[1]

[1]LASIGE, Faculdade de Ciências, Universidade de Lisboa, Portugal

Abstract

Aligning biomedical ontologies presents a significant challenge due to their complexity and the highly domain-specific nature of their vocabulary. Recent advancements in Language Models (LMs) have led to their increasing application in ontology alignment tasks, offering promising results. However, a systematic evaluation of semantics-based prompting strategies for leveraging LMs in this context remains unexplored. This study investigates the effectiveness of different prompting techniques to enhance biomedical ontology alignment performance. We have developed a framework to support the design of LM-based queries to assess the semantic similarity between ontology classes. The framework interrogates the ontologies to align to extract relevant contextual information to inject into the LM prompts allowing the use of Retrieval Augmented Generation (RAG). We conduct preliminary experiments on selected hard cases from biomedical ontologies that compose the Ontology Alignment Evaluation Initiative Bio-ML track and provide some insights into the effectiveness, reliability, and limitations of prompt-based approaches in ontology matching.

Keywords

Language Models, Ontology Alignment, Knowledge Representation

1. Introduction

Ontologies have become increasingly popular in various fields due to their ability to provide structured, formal representations of knowledge. These knowledge structures are particularly valuable in areas such as Artificial Intelligence (AI), Natural Language Processing (NLP), and Semantic Web technologies. An ontology represents a set of concepts within a domain and the relationships between them, allowing for more effective data sharing, discovery, and reasoning across different systems and applications.

As individual ontologies grow and evolve independently from each other, any given concept will inevitably display conceptual, linguistic, and structural differences when modelled in different ontologies, in different contexts and by different creators. These differences often arise from varying domain perspectives, terminologies and modelling choices across the maintainers and communities that develop and use the ontologies.

Ontology alignment addresses this issue through the generation of a set of mappings (correspondences) between entities in different ontologies to establish semantic interoperability [1]. However, automatically identifying these correspondences is a highly complex task. In general, ontologies are typically designed within a specific context, relying on implicit background knowledge that is not explicitly captured in their schema definitions [2].

Most ontology alignment techniques perform entity mapping based on leveraging lexical, structural, semantical, and external information of the entities being matched [1, 3]. Lexical information has proved to be the more successful source for biomedical ontology alignment [4, 5], with algorithms based on exploring the lexical component of ontologies outperforming other approaches by a good margin [6]. Lexical information can be extracted from entity labels, facilitating the exploration of

SeWebMeDa-2025: 8th International Workshop on Semantic Web solutions for large-scale biomedical data analytics, June 1, 2025, Portorož, Slovenia

jlferraz@lasige.di.fc.ul.pt (L. Ferraz); pgcotovio@ciencias.ulisboa.pt (P. G. Cotovio); clpesquita@ciencias.ulisboa.pt (C. Pesquita)

0009-0009-4289-1781 (L. Ferraz); 0000-0001-6724-899X (P. G. Cotovio); 0000-0002-1847-9393 (C. Pesquita)

word sense disambiguation and the inference of lexical relationships between entities. However, the lexical component of biomedical ontologies is typically restricted to the labels of concepts, which results in limitations in capturing mappings that require more contextual information beyond the simple similarity of labels.

The dawn of Large Language Models (LLMs) marked a turning point in our ability to capture and understand deep semantic relationships between terms. Traditional NLP techniques were insufficient at extracting contextual meaning, relying on simpler models that could not fully grasp language nuances. However, with the advent of LLMs such as GPT [7], BERT [8], and other transformer-based architectures [9], we currently have the ability to process and model complex relationships between words, phrases, and even complete documents.

LLMs are trained on massive datasets containing billions of tokens and are capable of understanding / representing not just the meaning of individual terms but how they interact in context. This allows us to capture subtle semantic relationships, such as synonyms, antonyms, hyponyms, and hypernyms, which are extremely useful for tasks such as translation, summarization, and question-answering. These capabilities are likely to translate to the ontology alignment scenario provided a suitable formulation of the problem is achieved. The success of prompt-based strategies in ontology alignment [10] has motivated us to explore whether LLMs are able to tackle the mapping of classes from biomedical ontologies and how well they are able to handle the more difficult cases. However, recent works have highlighted the difficulties in applying prompt-based strategies to real-world ontologies in other domains [11].

In this paper, we present a preliminary study that focuses on investigating the impact of including hierarchical relations in the prompt, exploring different design patterns for its verbalisation. We performed an evaluation of prompt-design strategies using a carefully selected set of challenging mappings extracted from Bio-ML [12], highlighting the pitfalls and strengths of each strategy.

2. Related Work

Ontology alignment is the process of identifying correspondences between entities in two distinct ontologies, typically referred to as the source ontology (O_s) and the target ontology (O_t). The goal of ontology alignment is to establish meaningful mappings between entities, ensuring interoperability between heterogeneous data sources. The resulting alignment consists of a set of mappings, often represented as tuples $<e_s, e_t, r, c>$, where e_s and e_t are entities from O_s and O_t respectively, r denotes a semantic relation (e.g., equivalence, subsumption), and c represents the confidence score of the mapping. These mappings are crucial for tasks such as data integration, Knowledge Graph fusion, and semantic interoperability in domains like healthcare, biology, and the Semantic Web.

Ontology matching systems are predominantly unsupervised, relying on heuristics and rules instead of deriving a mapping function through learning. These systems typically include three stages: pre-processing (the identification and retrieval of entities from the ontologies based on specific criteria); candidate generation (the use of diverse matching techniques to generate possible correspondences based on ontology features); and filtering (the refinement of initial matches through discarding of unlikely mappings).

In recent years, with the advent of language models, more attention has been devoted to Machine Learning-based ontology alignment, with several systems incorporating it [13, 14, 15, 16] and the creation of the Bio-ML track at the Ontology Alignment Evaluation Initiative. ML introduces a data-focused approach to ontology alignment, shifting away from heuristic and rule-based approaches. Unlike traditional OM systems, ML-based methods aim to learn a mapping function using labelled reference alignments, enabling more adaptive and scalable matching solutions. In principle, this allows for improved candidate generation, better matcher combination strategies, and more effective filtering techniques. Most approaches employ BERT-like methods [8] and typically sacrifice recall in favour of precision [17].

3. Methodology

3.1. Overview

To investigate the impact of semantic context (in the form of verbalized hierarchical relations) in prompt-based ontology matching tasks we developed a simple framework to design prompts based on combinations of relevant elements into meaningful patterns. Prompts built using this framework were evaluated using different language models with different model sizes. Our approach takes as input two candidate entities (typically classes) from each ontology to align, designs a prompt to evaluate the validity of a mapping between them based on different parameters, interrogates the language model using the prompt, and evaluates its output. In our study, we assume candidates are already selected, and focus only on prompt design and evaluation.

3.2. Prompt design for matching

We present a two-stage framework for generating context-aware prompts designed for tasks such as ontology alignment. This framework decomposes prompt generation into a **static stage** — where invariant templates (static skeletons) are constructed from a base template using task-specific configuration parameters — and a **dynamic stage** — where these skeletons are enriched with instance-specific data. This modular design allows a small number of configurable templates to be efficiently adapted to large datasets.

3.2.1. Static Stage: Template Construction

The static stage begins with a base template, denoted by $\mathcal{S}$, which contains symbolic placeholders indicating the different elements that compose a prompt, where specific types of information will be inserted. The elements are listed in Table 1.

Element Tag	Description
$TC (Task Context)	A description of the task the model should perform.
$I (Instruction)	A description of the nature of the question that will be asked and the expected answer format.
$CONF (Confidence)	The type of confidence that the model should output (if any).
$S (Source)	The main label(s) of the source ontology entity.
$CTX_S (Source Context)	Labels of potentially meaningful entities to the source entity.
$T (Target)	The main label(s) of the target ontology entity.
$CTX_T (Target Context)	Labels of potentially meaningful entities to the target entity.
$TYPE (Equivalence Type)	The type of equivalence to be assessed.

Table 1
Description of each base template prompt element

A brief description of each of these elements is presented in Table 1 and their possible values are presented in Table 2.

Category	Type	String Value
Task Context	—	"You are doing an ontology alignment task,"
Instruction	—	"I am going to ask you a question and you should answer 'yes' or 'no'."
Confidence	float	"Followed by confidence as a score from 0 to 1 (e.g., 'yes:0.8')"
	categorical	"Followed by confidence as 'Not Confident', 'Confident', or 'Very Confident' (e.g., 'yes:Confident')"
Context	subclass_of	"a subclass of $SC", with '$SC' being a superclass of either $S or $T
	kind_of	"a kind of $SC", with '$SC' being a superclass of either $S or $T
Equivalence Type	Equivalent	"equivalent"

Table 2
The possible types of each category of prompt elements and their respective values

You are doing an ontology alignment task. I am going to ask you a question and you should answer 'yes' or 'no', followed by your confidence in your answer as a score from 0 to 1, like this: 'yes:0.8'.
Question: Are 'Neuraminidase Deficiency' and 'glycoproteinosis' equivalent?

Figure 1: Example of a prompt without hierarchical context.

A set of configuration parameters $i \in \{1, \ldots, N\}$ governs the substitution of these placeholders with specific verbalizations. We define four vectors:

$$\boldsymbol{\tau} = (\tau_1, \tau_2, \ldots, \tau_N), \quad \boldsymbol{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_N),$$

$$\boldsymbol{\sigma} = (\sigma_1, \sigma_2, \ldots, \sigma_N), \quad \boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_N),$$

where:

- $\tau_i \in \{\text{True}, \text{False}\}$ indicates whether to include a task context.
- $\gamma_i \in \Gamma$ specifies the comparison prompt.
- $\sigma_i \in \Sigma$ represents the semantic context prompt.
- $\lambda_i \in \mathcal{L}$ denotes the confidence type (e.g., `float` or `cat`).

Then, for each configuration $i \in \{1, \ldots, N\}$, the static base template $\mathcal{S}_i$ corresponds to a unique pattern combining each of the four elements $< \tau_i, \gamma_i, \sigma_i, \lambda_i >$ by replacing each placeholder in $\mathcal{S}$ by the appropriate string. The output of this stage is the set $\{\mathcal{S}_i\}_{i=1}^{N}$ of static templates that capture the invariant, configuration-specific aspects of the prompt.

3.2.2. Dynamic Stage: Instance-Specific Enrichment

Let $\{(s_j, t_j)\}_{j=1}^{M}$ be the set of source–target entity pairs in the dataset. In the dynamic stage, each static template $\mathcal{S}_i$ is enriched with instance-specific information to produce a dynamically-built prompt.

For each entity pair (s_j, t_j) and each static template $\mathcal{S}_i$, a dynamic prompt P_{ij} is generated according to

$$P_{ij} = f_{\text{dynamic}}(\mathcal{S}_i; s_j, t_j, p_i, n_i),$$

where:

1. **Label Formatting:** The entities s_j and t_j provide label sets, which are formatted (e.g., truncated to a specified cardinality and concatenated with a given delimiter) to yield $L(s_j)$ and $L(t_j)$. These formatted labels replace the placeholders $S and $T, respectively.
2. **Contextual Enrichment:** Additional contextual information is extracted from the ontology and formatted as $C(s_j)$ and $C(t_j)$, replacing the placeholders $CTX_S and $CTX_T. In cases of absent context, extraneous semantic tokens may be removed. In this work, we focused on subsumption relations to include hierarchical context.

Thus, the dynamic prompt for each static skeleton and entity pair is obtained via the function f_{dynamic}, and for each j, the complete set of dynamic prompts is given by

$$\mathcal{P}_j = \{P_{ij}\}_{i=1}^{N}.$$

These dynamically enriched prompts form the final output features for the dataset.

In summary, the **static stage** produces a family of invariant templates and the **dynamic stage** adapts these templates to each instance (s_j, t_j). Figures 1 and 2 illustrate two prompt examples with and without hierarchical context.

You are doing an ontology alignment task. I am going to ask you a question and you should answer 'yes' or 'no', followed by your confidence in your answer as a score from 0 to 1, like this: 'yes:0.8'. Question: Are 'Neuraminidase Deficiency' (a kind of Mucolipidosis) and 'glycoproteinosis' (a kind of lysosomal storage disease) equivalent?

Figure 2: Example of a prompt with hierarchical context.

3.3. Models

Our experiments evaluated the prompts in five different language models with varying numbers of parameters and reasoning capabilities. The Flan-T5-Base model [18] (with 220 million parameters), is a lightweight transformer model developed by Google, tailored for instruction-based tasks and without any reasoning capabilities. The Claude 3.7 Sonnet model [19] was developed by Anthropic and is a significantly larger model than lightweight models such as Flan-T5-base, possessing 137 billion parameters but also lacking reasoning capabilities. Our experiments also incorporate GPT4 [20], a large-scale model comprising 1.76 trillion parameters, which represents a significant milestone in enhancing linguistic fluency and contextual comprehension within generative language models. Additionally, we also analyse the performance of two state-of-the-art reasoning models: GPT4o [21], a multimodal model with 200 billion parameters and OpenAIo1 [22], another multimodal model with 175 billion parameters.

Model Source	Number of Parameters	Reasoning
Flan-t5-base	220 million	No
Claude 3.7 Sonnet	137 billion	No
OpenAIo1	175 billion	Yes
GPT4o	200 billion	Yes
GPT4	1.76 trillion	No

Table 3
Number of parameters in the language models employed in the experiments, ordered by ascending size.

3.4. Parsing Model Responses

Let F denote a parsing function that maps a textual response y, generated by a predictive model, into a numerical confidence score $c \in [-1, 1]$. This confidence score quantifies the certainty associated with a binary classification decision, indicating either a positive or negative outcome.

Parsing Process: The function F is defined through the sequential application of the following procedures:

1. **Text Normalization:**
   - Transform the textual response y into lowercase form: $y \leftarrow \text{lowercase}(y)$.
   - Remove leading and trailing white space: $y \leftarrow \text{trim}(y)$.
2. **Numeric Confidence Extraction:**
   - Use regular expressions to search for numeric confidence values within y.
   - If a numeric value is found, convert it to a float and clip its value to the range $[0, 1]$. For instance, a response including "0.85" would yield $c = 0.85$.
3. **Default Uncertainty Handling:**
   - In the absence of numeric values, assign a default confidence $c = 1.0$. This default ensures reliance exclusively on the binary signal derived from keyword polarity.

4. **Solution Polarity Determination:**
   - Adjust the polarity of the mapping based on explicit binary indicators:

$$polarity = \begin{cases} 0, & \text{if "no" (negative) is detected,} \\ 1, & \text{if "yes" (positive) is detected,} \\ 0.0, & \text{if neither or both indicators ("yes" and "no") are detected.} \end{cases}$$

This parsing approach enables consistent extraction of numerical confidence scores from multiple textual responses generated for each query.

3.5. Evaluation

Our preliminary experiments focused on a subset of the mappings for the NCIT-DOID task of Bio-ML[12]. This track includes a special dataset, Bio-ML LLM, which contains 50 randomly selected matched class pairs from ground truth mappings, excluding pairs that can be aligned with direct string matching (i.e., having at least one shared label). This restricts the efficacy of conventional lexical matching. Of these 50 pairs, we selected the six which were considered as particularly hard to detect. For each source class in these "very hard" mappings, we created an additional "hard" negative (i.e., a target class with some lexical similarity to the source). The mappings are listed in Table 4.

Source	Target	Status
Esophageal Verrucous Carcinoma	esophagus verrucous carcinoma	1
Esophageal Verrucous Carcinoma	esophageal varix	0
Diabetic Vascular Disorder	diabetic angiopathy	1
Diabetic Vascular Disorder	diabetic encephalopathy	0
Malignant Hypopharyngeal Neoplasm	hypopharynx cancer	1
Malignant Hypopharyngeal Neoplasm	malignant granular cell skin tumor	0
Neuraminidase Deficiency	glycoproteinosis	1
Neuraminidase Deficiency	biotinidase	0
Bone Necrosis	ischemic bone disease	1
Bone Necrosis	dysbaric osteonecrosis	0
Microcystic Adnexal Carcinoma	malignant syringoma	1
Microcystic Adnexal Carcinoma	nasopharynx carcinoma	0

Table 4
Source and target classes with corresponding mapping status (1: correct mapping; 0: incorrect mapping).

4. Results

Table 5 presents the confusion matrix for the preliminary experiments. When the prompt does not include hierarchical contextual information, the best-performing models are OpenAIo1 and GTP4o, which despite being smaller than GPT4 have improved reasoning capabilities. These reasoning capabilities may help the models perform better when there is less information available. In fact, GPT4 ranks fourth despite being the largest model.

When semantic contextual information is given to the models, we observed very different behaviours between the *"kind_of"* prompt and the *"subclass_of"* prompt. While the *"kind_of"* resulted in improved results for the non-reasoning models, for the reasoning models, it had either no impact or a small negative impact. The *"subclass_of"* prompt, however, did not perform as well, having a negative impact in most models. These results demonstrate that hierarchical contextual information should be considered when designing prompts for biomedical ontology alignment. It is well worth noting that the second best performing approach was the pairing between Claude-3.7-Sonnet and the *"kind_of"* prompt, which achieved nearly identical results with GPT4, while being 10% of its size.

MODEL		w/o HC		w/ HC (kind of)		w/ HC (subclass of)	
		Pred: 1	Pred: 0	Pred: 1	Pred: 0	Pred: 1	Pred: 0
flan-t5-base	Actual: 1	1	5	3	3	1	5
	Actual: 0	0	6	1	5	0	6
Claude 3.7 Sonnet	Actual: 1	3	3	5	1	4	2
	Actual: 0	0	6	1	5	0	6
OpenAIo1	Actual: 1	4	2	4	2	3	3
	Actual: 0	0	6	0	6	0	6
GPT4o	Actual: 1	5	1	4	2	3	3
	Actual: 0	1	5	0	6	0	6
GPT4	Actual: 1	3	3	5	1	3	3
	Actual: 0	1	5	0	6	0	6

Table 5
Confusion Matrix Results for Different Models

We also investigated in more depth some false negative cases, depicted in Table 6. Some mappings, such as "Neuraminidase Deficiency - glycoproteinosis" are missed by all models, regardless of the context that is imparted in the prompt. Curiously, some sources indicate that this may not actually be an equivalence but rather a subsumption, with the corresponding diseases being modelled as such in ICD-10 (categories E77 and E77.1). However, including the hierarchical context in the form of *"kind_of"* prompts mitigates these issues, with most models, especially the mid to large sized, improving their recall of hard-to-find positive mappings.

Source	Target	HC	Fail to find
Neuraminidase Deficiency	glycoproteinosis	all	all fail
Microcystic Adnexal Carcinoma	malignant syringoma	No	Claude-3.7-Sonnet, flan-t5-base, GPT4
Microcystic Adnexal Carcinoma	malignant syringoma	*subclass_of*	, flan-t5-base, GPT4
Microcystic Adnexal Carcinoma	malignant syringoma	*kind_of*	none fail
Bone Necrosis	ischemic bone disease	No	Claude-3.7-Sonnet, flan-t5-base, GPT4
Bone Necrosis	ischemic bone disease	*subclass_of*	flan-t5-base, GPT4, GPT4o
Bone Necrosis	ischemic bone disease	*kind_of*	flan-t5-base

Table 6
Examples of mappings that the models failed to find when combined with different hierarchical context prompts.

5. Conclusion

This study explored the effectiveness of semantic prompting strategies, particularly the use of hierarchical contextual information, in enhancing biomedical ontology alignment with language models. Our experiments revealed that the impact of the inclusion of hierarchical context depended on the prompt wording. While the "kind_of" prompt — which more closely aligns with everyday language — improved the performance for non-reasoning models, the "subclass_of" prompt generally led to decreased performance. These findings highlight that the value of adding semantic context is heavily influenced by the verbalisation used when designing prompts.

We also found that smaller models like OpenAIo1 and GPT4o outperformed larger models like GPT4 when no hierarchical context was included in the prompt. This suggests that smaller models with better reasoning capabilities may perform more effectively when limited information is provided. Interestingly, the pairing of Claude-3.7-Sonnet with the "kind_of" prompt delivered nearly identical results to GPT4, despite being only 10% of its size, showing that less resource-intensive models can still achieve strong performance when combined with the right prompting strategies.

Additionally, the inclusion of hierarchical context through the "kind_of" prompt improved the recall of hard-to-find mappings, especially for mid- to large-sized models. However, some mappings remained challenging for all models, indicating that certain biomedical ontology mappings require more advanced approaches.

Future work will focus on extending the prompt design framework to include in-context learning based on positive and negative examples and developing additional strategies to extract semantic context by exploring common biomedical ontology features such as partonomy, rich synonyms and logical definitions.

Acknowledgments

This work was supported by FCT through the fellowships 2022.10557.BD (Pedro Cotovio), and the LASIGE Research Unit, ref. UID/00408/2025. It was also partially supported by the KATY project which has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 101017453. This work was also supported partially by project 41, HfPT: Health from Portugal, funded by the Portuguese Plano de Recuperação e Resiliência.

Declaration on Generative AI

During the preparation of this work, the author(s) used GPT-4o in order to: Grammar and spelling check. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] J. Euzenat, P. Shvaiko, Ontology matching, 2nd ed., Springer-Verlag, Heidelberg (DE), 2013.

[2] J. Portisch, M. Hladik, H. Paulheim, Background knowledge in ontology matching: A survey, Semantic Web 15 (2024) 2639–2693. doi:10.3233/SW-223085.

[3] X. Liu, Q. Tong, X. Liu, Z. Qin, Ontology matching: State of the art, future challenges, and thinking based on utilized information, in: Proceedings of the 19th International Workshop on Ontology Matching., volume 9, 2021, pp. 91235–91243. doi:10.1109/ACCESS.2021.3057081.

[4] D. Faria, E. Santos, B. S. Balasubramani, M. C. Silva, F. M. Couto, C. Pesquita, Agreementmakerlight, Semantic Web (2024) 1–13.

[5] E. Jiménez-Ruiz, B. Cuenca Grau, Logmap: Logic-based and scalable ontology matching, in: The Semantic Web–ISWC 2011: 10th International Semantic Web Conference, Bonn, Germany, October 23-27, 2011, Proceedings, Part I 10, Springer, 2011, pp. 273–288.

[6] D. Faria, C. Pesquita, I. Mott, C. Martins, F. M. Couto, I. F. Cruz, Tackling the challenges of matching biomedical ontologies, Journal of biomedical semantics 9 (2018) 1–19.

[7] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, et al., Improving language understanding by generative pre-training (2018).

[8] J. Devlin, M. W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, volume 1, 2019.

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).

[10] Y. He, J. Chen, H. Dong, I. Horrocks, Exploring large language models for ontology alignment, arXiv preprint arXiv:2309.07172 (2023).

[11] G. Macilenti, A. Stellato, M. Fiorelli, Prompting is not all you need evaluating gpt-4 performance on a real-world ontology alignment use case, Procedia Computer Science 246 (2024) 1289–1298.

[12] Y. He, J. Chen, H. Dong, E. Jiménez-Ruiz, A. Hadian, I. Horrocks, Machine learning-friendly biomedical datasets for equivalence and subsumption ontology matching, in: International Semantic Web Conference, Springer, 2022, pp. 575–591.

[13] Y. He, J. Chen, D. Antonyrajah, I. Horrocks, Bertmap: a bert-based ontology alignment system, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 36, 2022, pp. 5684–5691.
[14] S. Oulefki, L. Berkani, L. Bellatreche, N. Boudjenah, A. Mokhtari, Results for biogitom in oaei 2024, in: Proceedings of the 19th International Workshop on Ontology Matching., 2024.
[15] S. Menad, S. Abdeddaïm, L. F. Soualmia, Biostransmatch results@ oaei 2024, in: Proceedings of the 19th International Workshop on Ontology Matching., 2024.
[16] D. Faria, M. C. Silva, P. Cotovio, L. Ferraz, L. Balbi, C. Pesquita, Results in oaei 2024 for matcha, in: Proceedings of the 19th International Workshop on Ontology Matching., 2024.
[17] E. B. P. B. J. C. P. G. C. A. C. J. C. H. D. D. F. L. F. S. H. Y. H. I. H. L. I. S. J. E. J.-R. N. K. F. K. P. L. H. L. Y. L. P. M. H. P. C. P. A. S. P. S. M. S. G. S. C. T. J. V. B. Y. O. Z. Mina Abd Nikooie Pour, Alsayed Algergawy, L. Zhou, Results of the ontology alignment evaluation initiative 2024, in: Proceedings of the 19th International Workshop on Ontology Matching., 2024.
[18] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma, A. Webson, S. S. Gu, Z. Dai, M. Suzgun, X. Chen, A. Chowdhery, A. Castro-Ros, M. Pellat, K. Robinson, D. Valter, S. Narang, G. Mishra, A. Yu, V. Zhao, Y. Huang, A. Dai, H. Yu, S. Petrov, E. H. Chi, J. Dean, J. Devlin, A. Roberts, D. Zhou, Q. V. Le, J. Wei, Scaling instruction-finetuned language models, 2022. URL: https://arxiv.org/abs/2210.11416. arXiv:2210.11416.
[19] Anthropic, Claude 3.7 sonnet, 2025. URL: https://www.anthropic.com/claude/sonnet.
[20] OpenAI, J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, R. Avila, I. Babuschkin, S. Balaji, V. Balcom, P. Baltescu, H. Bao, M. Bavarian, J. Belgum, I. Bello, J. Berdine, G. Bernadett-Shapiro, C. Berner, L. Bogdonoff, O. Boiko, M. Boyd, A.-L. Brakman, G. Brockman, T. Brooks, M. Brundage, K. Button, T. Cai, R. Campbell, A. Cann, B. Carey, C. Carlson, R. Carmichael, B. Chan, C. Chang, F. Chantzis, D. Chen, S. Chen, R. Chen, J. Chen, M. Chen, B. Chess, C. Cho, C. Chu, H. W. Chung, D. Cummings, J. Currier, Y. Dai, C. Decareaux, T. Degry, N. Deutsch, D. Deville, A. Dhar, D. Dohan, S. Dowling, S. Dunning, A. Ecoffet, A. Eleti, T. Eloundou, D. Farhi, L. Fedus, N. Felix, S. P. Fishman, J. Forte, I. Fulford, L. Gao, E. Georges, C. Gibson, V. Goel, T. Gogineni, G. Goh, R. Gontijo-Lopes, J. Gordon, M. Grafstein, S. Gray, R. Greene, J. Gross, S. S. Gu, Y. Guo, C. Hallacy, J. Han, J. Harris, Y. He, M. Heaton, J. Heidecke, C. Hesse, A. Hickey, W. Hickey, P. Hoeschele, B. Houghton, K. Hsu, S. Hu, X. Hu, J. Huizinga, S. Jain, S. Jain, J. Jang, A. Jiang, R. Jiang, H. Jin, D. Jin, S. Jomoto, B. Jonn, H. Jun, T. Kaftan, Łukasz Kaiser, A. Kamali, I. Kanitscheider, N. S. Keskar, T. Khan, L. Kilpatrick, J. W. Kim, C. Kim, Y. Kim, J. H. Kirchner, J. Kiros, M. Knight, D. Kokotajlo, Łukasz Kondraciuk, A. Kondrich, A. Konstantinidis, K. Kosic, G. Krueger, V. Kuo, M. Lampe, I. Lan, T. Lee, J. Leike, J. Leung, D. Levy, C. M. Li, R. Lim, M. Lin, S. Lin, M. Litwin, T. Lopez, R. Lowe, P. Lue, A. Makanju, K. Malfacini, S. Manning, T. Markov, Y. Markovski, B. Martin, K. Mayer, A. Mayne, B. McGrew, S. M. McKinney, C. McLeavey, P. McMillan, J. McNeil, D. Medina, A. Mehta, J. Menick, L. Metz, A. Mishchenko, P. Mishkin, V. Monaco, E. Morikawa, D. Mossing, T. Mu, M. Murati, O. Murk, D. Mély, A. Nair, R. Nakano, R. Nayak, A. Neelakantan, R. Ngo, H. Noh, L. Ouyang, C. O'Keefe, J. Pachocki, A. Paino, J. Palermo, A. Pantuliano, G. Parascandolo, J. Parish, E. Parparita, A. Passos, M. Pavlov, A. Peng, A. Perelman, F. de Avila Belbute Peres, M. Petrov, H. P. de Oliveira Pinto, Michael, Pokorny, M. Pokrass, V. H. Pong, T. Powell, A. Power, B. Power, E. Proehl, R. Puri, A. Radford, J. Rae, A. Ramesh, C. Raymond, F. Real, K. Rimbach, C. Ross, B. Rotsted, H. Roussez, N. Ryder, M. Saltarelli, T. Sanders, S. Santurkar, G. Sastry, H. Schmidt, D. Schnurr, J. Schulman, D. Selsam, K. Sheppard, T. Sherbakov, J. Shieh, S. Shoker, P. Shyam, S. Sidor, E. Sigler, M. Simens, J. Sitkin, K. Slama, I. Sohl, B. Sokolowsky, Y. Song, N. Staudacher, F. P. Such, N. Summers, I. Sutskever, J. Tang, N. Tezak, M. B. Thompson, P. Tillet, A. Tootoonchian, E. Tseng, P. Tuggle, N. Turley, J. Tworek, J. F. C. Uribe, A. Vallone, A. Vijayvergiya, C. Voss, C. Wainwright, J. J. Wang, A. Wang, B. Wang, J. Ward, J. Wei, C. Weinmann, A. Welihinda, P. Welinder, J. Weng, L. Weng, M. Wiethoff, D. Willner, C. Winter, S. Wolrich, H. Wong, L. Workman, S. Wu, J. Wu, M. Wu, K. Xiao, T. Xu, S. Yoo, K. Yu, Q. Yuan, W. Zaremba, R. Zellers, C. Zhang, M. Zhang, S. Zhao, T. Zheng, J. Zhuang, W. Zhuk, B. Zoph, Gpt-4 technical report, 2024. URL: https://arxiv.org/abs/2303.08774. arXiv:2303.08774.
[21] OpenAI, Gpt-4o system card, arXiv preprint arXiv:2410.21276 (2024). URL: https://arxiv.org/abs/

2410.21276.

[22] OpenAI, :, A. Jaech, A. Kalai, A. Lerer, A. Richardson, A. El-Kishky, A. Low, A. Helyar, A. Madry, A. Beutel, A. Carney, A. Iftimie, A. Karpenko, A. T. Passos, A. Neitz, A. Prokofiev, A. Wei, A. Tam, A. Bennett, A. Kumar, A. Saraiva, A. Vallone, A. Duberstein, A. Kondrich, A. Mishchenko, A. Applebaum, A. Jiang, A. Nair, B. Zoph, B. Ghorbani, B. Rossen, B. Sokolowsky, B. Barak, B. McGrew, B. Minaiev, B. Hao, B. Baker, B. Houghton, B. McKinzie, B. Eastman, C. Lugaresi, C. Bassin, C. Hudson, C. M. Li, C. de Bourcy, C. Voss, C. Shen, C. Zhang, C. Koch, C. Orsinger, C. Hesse, C. Fischer, C. Chan, D. Roberts, D. Kappler, D. Levy, D. Selsam, D. Dohan, D. Farhi, D. Mely, D. Robinson, D. Tsipras, D. Li, D. Oprica, E. Freeman, E. Zhang, E. Wong, E. Proehl, E. Cheung, E. Mitchell, E. Wallace, E. Ritter, E. Mays, F. Wang, F. P. Such, F. Raso, F. Leoni, F. Tsimpourlas, F. Song, F. von Lohmann, F. Sulit, G. Salmon, G. Parascandolo, G. Chabot, G. Zhao, G. Brockman, G. Leclerc, H. Salman, H. Bao, H. Sheng, H. Andrin, H. Bagherinezhad, H. Ren, H. Lightman, H. W. Chung, I. Kivlichan, I. O'Connell, I. Osband, I. C. Gilaberte, I. Akkaya, I. Kostrikov, I. Sutskever, I. Kofman, J. Pachocki, J. Lennon, J. Wei, J. Harb, J. Twore, J. Feng, J. Yu, J. Weng, J. Tang, J. Yu, J. Q. Candela, J. Palermo, J. Parish, J. Heidecke, J. Hallman, J. Rizzo, J. Gordon, J. Uesato, J. Ward, J. Huizinga, J. Wang, K. Chen, K. Xiao, K. Singhal, K. Nguyen, K. Cobbe, K. Shi, K. Wood, K. Rimbach, K. Gu-Lemberg, K. Liu, K. Lu, K. Stone, K. Yu, L. Ahmad, L. Yang, L. Liu, L. Maksin, L. Ho, L. Fedus, L. Weng, L. Li, L. McCallum, L. Held, L. Kuhn, L. Kondraciuk, L. Kaiser, L. Metz, M. Boyd, M. Trebacz, M. Joglekar, M. Chen, M. Tintor, M. Meyer, M. Jones, M. Kaufer, M. Schwarzer, M. Shah, M. Yatbaz, M. Y. Guan, M. Xu, M. Yan, M. Glaese, M. Chen, M. Lampe, M. Malek, M. Wang, M. Fradin, M. McClay, M. Pavlov, M. Wang, M. Wang, M. Murati, M. Bavarian, M. Rohaninejad, N. McAleese, N. Chowdhury, N. Chowdhury, N. Ryder, N. Tezak, N. Brown, O. Nachum, O. Boiko, O. Murk, O. Watkins, P. Chao, P. Ashbourne, P. Izmailov, P. Zhokhov, R. Dias, R. Arora, R. Lin, R. G. Lopes, R. Gaon, R. Miyara, R. Leike, R. Hwang, R. Garg, R. Brown, R. James, R. Shu, R. Cheu, R. Greene, S. Jain, S. Altman, S. Toizer, S. Toyer, S. Miserendino, S. Agarwal, S. Hernandez, S. Baker, S. McKinney, S. Yan, S. Zhao, S. Hu, S. Santurkar, S. R. Chaudhuri, S. Zhang, S. Fu, S. Papay, S. Lin, S. Balaji, S. Sanjeev, S. Sidor, T. Broda, A. Clark, T. Wang, T. Gordon, T. Sanders, T. Patwardhan, T. Sottiaux, T. Degry, T. Dimson, T. Zheng, T. Garipov, T. Stasi, T. Bansal, T. Creech, T. Peterson, T. Eloundou, V. Qi, V. Kosaraju, V. Monaco, V. Pong, V. Fomenko, W. Zheng, W. Zhou, W. McCabe, W. Zaremba, Y. Dubois, Y. Lu, Y. Chen, Y. Cha, Y. Bai, Y. He, Y. Zhang, Y. Wang, Z. Shao, Z. Li, Openai o1 system card, 2024. URL: https://arxiv.org/abs/2412.16720. arXiv:2412.16720.