Ontology Population Using LLMs: Which Factors Matter?

No Author Given

No Institute Given

Abstract. While LLMs have proven their performance in many ontology-related tasks, it is not yet known how different factors, such as prompting strategies, ontology structure and entity label semantics influence their performance. This paper is aimed to address this gap by exploring effects of a wide range of ontology and LLM characteristics on the accuracy of LLM-driven ontology population. We use three lightweight ontologies and four LLMs to study these effects. Our findings suggest that LLMs are capable of performing ontology population with sufficient accuracy but may struggle to infer concept hierarchies, particularly when their embeddings of concept and individual labels are well-separated. Few-shot prompting is effective in improving performance of both larger and smaller LLMs. Investigating response variation and consistency, we observe larger LLMs exhibit less variability over repetitions, and performance reduces with increasing temperature. We find that the depth and dispersion of concepts does not influence an LLM's ability to predict correct hierarchies. In addition, the size of an ontology is not an influential factor for ontology population.

Keywords: Ontology Population · Ontology Learning · Ontology Engineering · Large Language Model · LLM Prompting · Evaluation

1 Introduction

Manual curation of ontologies is a time and effort-intensive process requiring significant expertise in a target domain and knowledge engineering. Automated ontology learning aims to expedite this process by extracting knowledge components (terms, concepts, relations between them) from unstructured and semi-structured data. The wealth of topical information that Large Language Models (LLMs) can produce on demand has led to the growing interest in the Semantic Web Community to adopt LLMs for various ontology enhancement [16, 19, 25, 30, 36] and ontology learning [12, 23, 24, 26] tasks. However, recent studies highlight that such affordance of information does not automatically translate into strong performance of LLMs on all ontology learning tasks [6, 48].

Understanding the specific conditions under which an LLM performs well on an ontology learning task is vital to obtain the most reliable performance from these models. Ontology population is one such task that requires LLMs to exhibit their understanding of individual-to-concept assertions and concept hierarchies. To provide a better understanding of the underlying factors, we investigate the

effect of several ontology and LLM factors on ontology population: the task of adding new individuals to an ontology through concept assertions. We investigate the effect of the ontology factors: structure, taxonomy and entity labels and the LLM factors: choice of LLM choice, modelling objective, prompting approach, domain context and response variation over 3 ontologies using 4 state-of-the-art LLMs.

The rest of the paper is structured as follows. In the next section, we highlight related work on ontology population and LLM-specific ontology evaluation. Section 3 provides an overview of the different factors that influence ontology learning. Section 4 outlines the factors we investigate in the present work and our experimental setup. Section 5 discusses the observed results and Section 7 concludes the paper. We provide all the necessary materials required to reproduce our work in an Anonymous GitHub Repo.

2 Related Work

The field of ontology learning has primarily focused on (semi-)automatic creation of ontologies. Several works outline the various ontology learning tasks [2, 8, 17, 28, 52, 53] with ontology population being one of them [43]. Contemporary approaches like the LLMs4OL paradigm [22] utilize LLMs instead of conventional linguistic and statistical approaches for ontology learning due to their growing popularity.

Ontology population is the task of adding new instances to an ontology [43]. State-of-the-art approaches often involve applying natural language processing algorithms like name entity recognition to extract concepts and related instances from a provided set of documents and an ontology [33, 43]. Redundancy and entity disambiguation pose several challenges for ontology population. LLM-based approaches include SPIRES [11] where a zero-shot method is used for knowledge base population, and a preliminary work using a combination of text summarization, retrieval-augmented-generation and prompt engineering [38]. Sahbi et al. [45] compare a traditionally semantic approach and an LLM-based approach for ontology population in French. The results illustrate that the semantic approach is consistent and logically coherent, but creates redundancy, while the LLM creates no redundancies but is not always consistent. Bhattacharya et al. [5] finds that LLMs are unable to infer concept hierarchies in zero-shot experiments for concept assertion.

There has been growing interest in LLM-specific ontology learning evaluation based on conventional evaluation measures [6, 7, 25, 51]. Giglou et al. [21] evaluate LLM performance on term typing, taxonomy discovery and relation extraction in a zero-shot setting. Testing on nine different datasets, it provides a comprehensive overview of the effects of size and complexity of ontologies on learning tasks. In their work on how well LLMs actually learn reasoning through concept relations, Mai et al. [34] note inconsistencies suggesting that LLMs tend to fall back to their pre-learnt lexical senses as opposed to using the provided semantic meanings of concepts in ontologies.

Despite growing interest in such methods of evaluation, insufficient attention is paid to the underlying ontology and LLM factors In the next section, we outline

the relevant factors that may contribute to performance variation on ontology learning and enrichment tasks. The influence of these underlying factors can provide strong evidence about the competence of LLMs on various ontology learning and enrichment tasks.

3 Factors

The factors governing LLM-driven ontology learning can be categorized into two groups: Ontology factors and LLM factors. We provide an overview of the relevant factors for each group, highlighting their relevance for various learning tasks. Figure 1 provides an overview of the factors of each group.

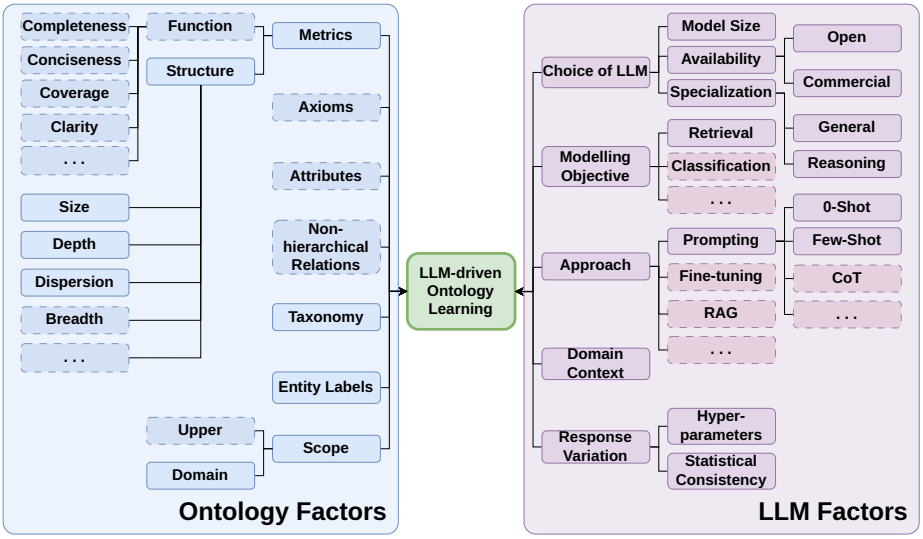


Fig. 1: Overview of Ontology and LLM factors influencing LLM-driven Ontology Population. Dashed boxes are not investigated in the present study

3.1 Ontology Factors

Ontologies vary greatly depending on the complexities of the modeled domain and the design choices taken by ontology engineers and domain experts [39]. Our categorization of ontology factors draws inspiration from the Ontology Learning Layer Cake [20] with additional included factors that are not captured by it.
Scope: The scope of an ontology defines the level of specificity and abstraction of its entities and its structural design. Ontologies can be **upper-level** ontologies that define abstractions enabling integration of heterogeneous knowledge across different domains [35] or of a particular **domain** itself e.g. the Gene Ontology (GO) [1] designed using the principles provided by an upper ontology. The scope of ontologies test the semantic abilities of LLMs with upper ontologies requiring

understanding abstractions and ontological semantics while domain ontologies require direct understanding of the entities themselves.

Metrics: Structural metrics [27](e.g. size, breadth, depth and dispersion) evaluate the structural complexity of an ontology. Structural nuances test the ability of LLMs to identify and populate domains of varying relational complexity.

Functional metrics [27] (e.g. completeness, conciseness, etc.) evaluate the intended use of an ontology [20]. They evaluate the logical consistency and comprehensivenss of an ontology and provides insight into the awareness LLMs possess of an ontology's functional adequacy and how that influences ontology learning.

Entity labels: Entity labels are largely governed by conventions followed in the domain of interest and by any upper ontologies used as reference. Through their pre-training, LLMs exhibit impressive natural language understanding and generation capabilities. Variation in entity label naming highlights the parity between the ontological semantics of the entities and the learnt lexical semantics of LLMs for their labels. This parity is crucial for effective ontology learning. Entity labels highlight whether LLMs are able to effectively utilize lexical semantics or resort to more conventional string edit-distance-based senses for performing ontology learning.

Taxonomy: The taxonomy of an ontology is the backbone structure that defines the hierarchy of concepts in a domain. Understanding taxonomy is pivotal for any ontology learning objective as it showcases the understanding LLMs possess of concept relations.

Non-hierarchical Relations: Parthood and associative relations comprise a large portion of ontological relations. Such relations add complexity to ontology learning tasks, e.g. requiring care to avoid asserting new individuals with closely related concepts like asserting 'Red' as an instance of the concept 'Wine' in the Wines Ontology [39] instead of an attribute value of 'WineColor' during ontology population. Analyzing whether LLMs are susceptible to such pitfalls can provide insight into their ability to distinguish different relations in ontologies.

Attributes: Data and object properties of entities in an ontology provide contextual information that introduces further complexity to entities and different relations in an ontology. The additional information incorporated by such attributes challenges LLMs to identify distinguishing information of ontological entities and separate them from other entity types such as concepts and individuals.

Axioms: The underlying rules and theory of an ontology and its capabilities are defined by its axioms. Axioms represent logical abstractions of the behaviour exhibited by the entities of an ontology in the other layers of the Ontology Learning Layer Cake. As the arguably most difficult ontology learning task, investigation into the awareness LLMs possess of ontology axioms exhibits their comprehensive understanding of ontologies.

3.2 LLM Factors

The capabilities of LLMs in natural language generation and understanding has led to immense growth in research about their abilities and limitations. As an

ever-growing field, factors regularly change with new research. We categorize LLM factors based on already well-established factors in LLM research.

Choice of LLM: Choosing an appropriate LLM from an ever-growing number of LLMs is governed by several factors. **Model size** for in-context learning capabilities, **availability** of model weights enabling task-specific fine-tuning and further analysis of performance (commercial black-boxes or open-source open-weight models) and; **specialization** (general-purpose LLMs versus reasoning-specific LLMs) for more reasoning ability based on the defined task and objective are primary contributors when selecting an LLM.

Objective Modelling: LLMs are used for a broad set of tasks that can be modelled as various NLP objectives. Modelling an ontology learning task as **classification**, **retrieval** or another objective results in varied responses generated by LLMs and, as a result, varied performance.

Approach: LLMs can be utilized as instruction-following agents leveraging their in-context learning capabilities through **prompting** strategies [32, 37, 44, 46, 50] like zero-shot, few-shot and Chain-of-Thought [49], **fine-tuned** [31] for a specific objective or coupled with a data store to facilitate data-substantiated text generation [29](**Retrieval Augmented Generation (RAG)**) Comparison of performance across different approaches highlights the relevance of pre-learnt information, task-specific augmentation and data availability for ontology learning.

Domain Context: LLMs can be prompted to assume certain roles/personas e.g. 'helpful assistant' while performing a certain task. These roles define a domain space for the LLM and is a powerful tool that contextualizes the task and user input data for the LLM. As a factor, domain context analyzes the amount of data domain context necessary for an LLM to perform an ontology learning task.

Response Variation: LLMs are non-deterministic based on their probabilistic generation strategy. The determinism, length and repetitiveness of LLM responses can be controlled using a wide variety of **hyperparameters**. `top_k` and `top_p` influence the choice of tokens selected by confining the sampling space to a fixed number of tokens based or based on a cumulative frequency, `frequency penalty` to penalize repetition and reward response variation and `temperature` performs simulated annealing of the token probabilities and controls randomness of responses (higher is more creative, lower is more deterministic). In conjunction with the various hyperparameters, **statistical consistency** of LLM responses display how consistent the results are across repetitions.

4 Experimentation

Ontology and LLM factors influencing LLM-driven ontology learning are not independent of each other. Ontology factors like structure are intrinsically linked to entities in an ontology and therefore to the OL Layer Cake of ontology factors, whereas functional measures of an ontology are driven by the taxonomical and non-hierarchical modeling of the underlying domain. The complex inter-factor variability influencing LLM-driven ontology learning adds several levels of complexity to accurately understand how each factor acts in isolation. Effective marginalization of each factor and comparing performance by varying

inter-related variables provides insight into the exact manner of influence each factor introduces.

We investigate the influence of the ontology factors: structure, taxonomy and entity labels; and the LLM factors: choice of LLM, modelling objective, prompting approach, domain context and response variation on the ontology learning task of population. Our selection of ontology factors is based on them being the most relevant for ontology population. The choice of LLM factors is based on areas of growing interest in LLM research.

4.1 Ontology Population

We select ontology population as the first learning task to investigate as it is the simplest of all ontology learning tasks. Despite being simple, ontology population requires an LLM to "understand" concepts and instances in an ontology to make new assertions. LLM-driven ontology population helps us investigate the degree of awareness LLMs posses of ontology assertion relations and of the entities themselves.
Following Hlomani and Stacey [27], we define an ontology O as a 4-tuple:

$$O = \langle C, H, R, A \rangle \tag{1}$$

where, C is the set of Concepts $\{c\}$; H is the set of taxonomical/hierarchical ('is a') relations over C; R is the set of non-taxonomical relations over C and; A is a set of axioms

Let t be a term to be mapped as an individual to a concept in C. Then the task of ontology population is:

$$f(t;O)^{Ontology}_{Population} = \{c_i \mid c_i \in C \ ; \ c_i \subseteq c_{i+1} \ ; \ t \in \Sigma_{c_1} \ ; \ 1 \leq i \leq D - d + 1\} \tag{2}$$

where, D is the depth of the ontology O; d is the depth of the concept c_1 in O; $\subseteq$ denotes the 'subclass of' or 'is a' relation (taxonomical relations of O i.e. H); c_1 is the directly asserted concept of the term t and; Σ_{c_1} is the extension [27] of c_1 (i.e. set of asserted individuals)

4.2 Ontology Factors

We select three domain-specific lightweight ontologies characterized by varying structural complexity and, (by virtue of their differing domains) distinct entity labels. We view these ontologies as lightweight as they do not include relational loops, any complex composite entity and relation types and, are of single parentage. Despite being lightweight, these ontologies possess sufficient domain, functional and structural variability to effectively assess the influence of various factors.

The **Wines** Ontology [39] is a structurally simple and well-known ontology regarding wines. It represents information about the colour, flavour and origin of various wines. The Wines Ontology provides a simple ontology learning use-case that we expect most LLMs to have good knowledge of either directly or through the topic of wines.

The **CASE** Ontology [10] is a larger, more complex and newer ontology focused on accurately capturing the life-cycle of digital evidence. We incorporate individuals from the Owl Trafficking example available on the CASE website[1] to investigate ontology population with the ontology. As the CASE Ontology is built on top of the UCO Ontology [9], we include concepts from it in our experimentation and refer to this composite constructed ontology as the CASE Ontology henceforth. The ontology poses a complex and lesser known scenario that LLMs may struggle with.

The **Astronomy** Ontology [47] is the largest ontology of the three ontologies concerning astronomical phenomena including planets, stars and the relevant laws of physics. Despite its large size, the ontology concerns a well-known topic that one might expect LLMs to have 'seen' abundantly in its pre-training corpus.

Table 1 provides an overview of the structural metrics of the three ontologies.

Table 1: Ontology Structure Metrics for Wines, CASE and Astronomy ontologies

Metric	**Wines**	**CASE**	**Astronomy**
Classes (no.)	76	434	1663
Individuals (no.)	161	131	68
Depth [20]	4	8	10
Breadth [20]	62	228	989
Dispersion (max.) [20]	3	118	44

Structure: To investigate the influence of structure, we conduct a correlation study between the ability of LLMs to place concepts at the correct position in their hierarchy during retrieval and ontology population performance.

For any concept c of depth d, we compute the Correct Retrieval (CR) as the percentage of retrievals at the correct hierarchy position across all individual terms $t \in T$ where the concept was retrieved.

$$CR(c) = \frac{\left|\{t \in T; c \in \{y_i\}_t; y_{D-d+1} = c\}\right|}{\left|\{t \in T; c \in \{y_i\}_t\}\right|} \tag{3}$$

CR highlights the ability of LLMs to place concepts at the correct taxonomical position during ontology population. We compute the Pearson correlation between CR and two structural metrics: depth and dispersion of concepts. It highlights whether LLMs struggle with correctly retrieving concepts of greater depth or dispersion. Using the Pearson correlation values, we compute the average correlation [13] across all 96 experiments conducted for each ontology.

Taxonomy: The designed modelling objective emphasizes predicting hierarchies over only directly-asserted concepts. Doing so allows us to investigate the taxonomical understanding LLMs possess when performing ontology population. We evaluate the ability of LLMs to predict correct hierarchies using the mAP@D

[1] `https://caseontology.org/examples/`

metric. mAP@1 helps understand an LLM's ability to identify directly-asserted concepts. Comparing mAP@D and mAP@1 highlights whether LLMs perform ontology population with an ontological reliable sense of entities or simply perform entity matching using lexical similarity.

Entity Labels: We investigate the parity between ontological label semantics and LLM lexical understanding using similarity measurement between LLM embeddings of individuals and concepts. We generate 3072-dimensional embeddings of each individual and concept label for all three ontologies using OpenAI's `text-embedding-3-large` [42] model. Utilizing the embeddings of concepts and individual labels, we measure inter-entity semantic separation through centroid distance of the embedding entity clusters and overall entity semantic homegeneity using the Davies-Bouldin Index (DBI) [14]. We hypothesize that the cluster separability between concepts and individuals should be low for LLMs to perform ontology population (i.e a lower centroid distance and a higher DBI is better).

We supplement the analysis of entity labels by investigating the lexical matching between concepts and individuals, assessing whether LLMs resort to simpler string similarity-based ontology population. The naming conventions of entities in ontologies can lead to similar but distinct concepts possessing lexically similar labels. This can cause confusion for LLMs if they resort to string matching for ontology population. We evaluate the string edit-distance between individual labels and two groups: G1) the directly-asserted concept of the individual and, G2) all other ancestor concepts of the individual (averaged). The combination of the two highlights the ability of an LLM to identify a concept assertion from the directly-asserted concept and individual labels versus its ability to additionally predict other ancestors using lexical matching. Based on our formulation of the modelling objective (Section 4.3), the second distance provides an indicator of the semantics-based ontology population that LLMs are required to employ to be robust.

4.3 LLM Factors

Our experimentation involves several LLMs, a modelling objective that expounds an LLM's ontological taxonomical understanding, multiple prompting approaches, domain contexts and an investigation of response consistency over temperature variation.

Modelling Objective: Following a similar investigation of term typing in Giglou et al. [22], we model the task of ontology population as a retrieval problem that requires LLMs to generate a ranked list of concepts of length up to the depth [20] of the ontology. We provide minimal context about the ontology from which individuals and concepts are taken to force LLMs to utilize their own knowledge. This choice highlights the relevance and utility of an LLM's "world knowledge" for a basic ontology learning task. Requiring an LLM to retrieve concepts up to the depth of the ontology provides insight into their ability to infer concept hierarchies when only provided with concept labels. It illustrates better ontological understanding.

We evaluate using the standard information retrieval metric Mean Average Precision (mAP) [4] with mAP computed at 1 (mAP@1) highlighting the ability

of LLMs to identify the directly-asserted concept for an individual and at the depth D (mAP@D) of each ontology to better understand an LLM's ability to infer the correct concept hierarchies. The mathematical formulation can be found in Appendix A.

Choice of LLM: We conduct experiments using four instruction-tuned LLMs: OpenAI's GPT-4o [40], OpenAI's o1-preview [41], Meta's Llama3-8B [18] and DeepSeek's R1-Distil-Llama-8B [15]. Our selection involves two larger commerical models: GPT-4o and o1-preview and two smaller open-source models: Llama3-8B and R1-Distil-Llama-8B. GPT-4o and Llama3-8B are general LLMs while o1-preview and R1-Distil-Llama-8B are reasoning-specific LLMs.

Prompting approach: To investigate the effect of different prompting approaches, we perform experiments on all ontologies with zero-shot and few-shot prompts. For few-shot prompts, we experiment with providing different numbers of examples ranging from 1 to 10. The provided examples are selected randomly from the concepts with the highest number of individuals, ensuring that each example is taken from a different concept.

Domain Context: We experiment with four types of domain context: 1) *Generic* The LLM is only asked to perform a generic task e.g. ranked retrieval. 2) *Ontology:* The LLM is defined as an expert in ontologies. 3) *Topic:* The LLM is defined as an expert in the topic of an ontology e.g. `'You are a wine expert'` for the Wines Ontology [39]. 4) *Ontology and Topic:* The LLM is defined as an expert in ontologies and an expert in the topic. Examples of the templates for different prompting approaches and domain context variations can be found at our repository.

Response Variation: We evaluate response variation in terms of statistical consistency at different temperature settings under specific values of different factors. Our approach involves statistical consistency evaluation of LLM responses in terms of mAP@D at several temperature values: 0.0, 0.2, 0.4, 0.6, 0.8, 1.0. We investigate response variation using GPT-4o and Llama3-8B on the Wines and CASE ontologies with a 3-shot prompting approach utilizing ontology domain contextualiztion. Statistical consistency is measured across 10 repetitions at each temperature value. Lower temperature values (< 0.5) highlight performance variation at low response variability while higher values (≥ 0.5) illustrate performance variation when response variation is encouraged.

5 Results and Discussion

This section presents the results and observations across the 288 conducted experiments (zero-shot: 4 LLMs $\times$ 3 ontologies $\times$ 4 domain contexts = 48; few-shot: 2 LLMs $\times$ 3 ontologies $\times$ 4 domain contexts $\times$ 10 few-shot variations = 240). Due to space limitations, we are unable to report all the computed scores here. A complete table with all the results and figures can be found online [2]. Table 2 reports a selection of scores representative of the primary trends observed across all considered factors. When analyzing any factor, we draw our conclusions from all experiments performed. When analyzing the effects of different factors, we draw

[2] https://anonymous.4open.science/r/llm_ontology_awareness-91C4/

our conclusions with an emphasis on mAP@D as it better highlights ontological awareness with the inclusion of taxonomies.

A macroscopic overview of performance shows that LLMs are quite capable of performing ontology population when provided with simple labels. LLMs struggle with zero-shot hierarchy retrieval-based ontology population (mAP@D) for the CASE Ontology but few-shot context examples largely help alleviate the issue. The performance differences between the three ontologies highlight the importance of the domain of the ontologies even on a simple ontology learning task such as ontology population. In the subsequent sections, we explore the effects of the outlined Ontology and LLM Factors in our experimentation.

Table 2: Ontology domain context prediction results at zero-shot, 1-shot, 3-shot and 10-shot for all ontologies and LLMs (DeepSeek-R1-Distil-Llama-8B abbreviated to DeepSeek-R1* to save space)

N-Shot	Ontology	LLM	mAP@1	mAP@D
0	Wines	o1-preview	0.826	0.871
		GPT-4o	0.733	0.785
		Llama3-8B	0.484	0.571
		DeepSeek-R1*	0.472	0.542
	CASE	o1-preview	0.878	0.578
		GPT-4o	0.779	0.316
		Llama3-8B	0.626	0.270
		DeepSeek-R1*	0.557	0.196
	Astronomy	o1-preview	0.956	0.870
		GPT-4o	0.926	0.750
		Llama3-8B	0.544	0.420
		DeepSeek-R1*	0.618	0.387
1	Wines	GPT-4o	0.800	0.868
		Llama3-8B	0.500	0.536
	CASE	GPT-4o	0.823	0.608
		Llama3-8B	0.700	0.372
	Astronomy	GPT-4o	0.940	0.830
		Llama3-8B	0.776	0.627
3	Wines	GPT-4o	0.873	0.889
		Llama3-8B	0.690	0.720
	CASE	GPT-4o	0.836	0.772
		Llama3-8B	0.711	0.552
	Astronomy	GPT-4o	0.954	0.807
		Llama3-8B	0.769	0.648
10	Wines	GPT-4o	0.894	0.901
		Llama3-8B	0.603	0.630
	CASE	GPT-4o	0.851	0.829
		Llama3-8B	0.785	0.673
	Astronomy	GPT-4o	0.914	0.778
		Llama3-8B	0.810	0.670

5.1 Analysis of Ontology Factors

Structure: For all three ontologies, we did not observe a statistically significant correlation between either the depth or dispersion of a concept with their CR. In terms of size, mAP@D is the highest for the smallest Wines Ontology. However, LLMs perform better on the largest Astronomy Ontology compared to the CASE Ontology. The variable performance based on size and the lack of statistically significant correlation between correct retrieval and concept structural metrics highlight that structural metrics do not influence the ability of LLMs to perform ontology population.

Taxonomy: Comparing mAP@1 and mAP@D in Table 2, we observe that LLMs are adept at inferring direct assertions during ontology population but can struggle with predicting hierarchies. This is particularly evident for the zero-shot experiments with CASE where mAP@1 is comparable to that of the other two ontologies but mAP@D is significantly lower. Addition of examples through few-shot prompting improves performance and inclusion of even a few examples significantly improves ontology population taxonomy retrieval (Figure 2). Providing examples leads to significant improvements in taxonomy retrieval (mAP@D) compared to direct assertion (mAP@1).

We observed that 83 out of the 161 individuals (51.5%) for Wines, 94 of the 131 individuals (71.7%) for CASE and, 17 out of the 68 individuals (25%) for the Astronomy ontology utilize the name of their directly-asserted concept in their labels. For these individuals, lexical matching is sufficient for direct assertion. Analysis using mAP@1 under such circumstances provides an inflated measure of the understanding LLMs possess of ontologies. In contrast, mAP@D provides a more ontologically grounded evaluation approach highlighting that LLMs on their own may not possess good taxonomical sense but in-context learning with few-shot prompting can largely address this problem.

Entity Labels: Table 3 highlights the entity label measures to evaluate the influence of entity labels. We observe that the normalized centroid distance between concept and individual clusters for CASE is twice as large as the other two ontologies (larger values indicate greater inter cluster separation). The DBI for CASE is much lower than the other two ontologies. These two entity label metrics (Table 3) together suggest that the entity labels of CASE might make ontology population difficult for LLMs. The performance in Table 2 supports this observation with performance on CASE being significantly lower than the other two ontologies in the zero-shot scenario and when using one or two examples. Beyond that, the provision of more examples in few-shot prompting counters the effects of this entity semantic separation.

Looking at the string edit-distance metrics, G1 disparity is highest for Wines and lowest for CASE. This highlighs that LLMs cannot resort to lexical matching to find the directly-asserted concept for Wines. A large G2 disparity across all ontologies highlights that lexical matching is insufficient for LLMs to infer concept hierarchies. Combining insights from the two disparities and the fact that several individual labels utilize names of the directly-asserted concepts in their label suggests that LLMs might be able to infer the directly-asserted concepts from

Table 3: Entity label metrics for Wines, CASE and Astronomy ontologies (Levenshtein shortened to L. for readability)

Metric	Wines	CASE	Astronomy
Centroid Distance	0.259	0.540	0.261
DBI [14]	5.995	2.891	6.362
Direct L. Distance (G1)	10.285	4.137	5.167
Ancestor L. Distance (G2)	12.191	14.092	15.381

lexical matching but cannot infer hierarchies. Looking at Table 2, we observe that under zero-shot conditions, LLMs are able to predict the directly-asserted concepts for all ontologies (mAP@1) but struggle with predicting complete hierarchies (mAP@D), particularly on the CASE ontology. Manually inspecting the nature of predicted hierarchies in the zero-shot setting, we observed a large tendency to predict hierarchies based on string similarity. The influence of string similarity is reduced in the few-shot experiments. An example of the preference for string similarity when predicting hierarchies is shown in the Appendix. Entity labels greatly influence the ability of LLMs to perform ontology population. When the semantic separation between the two types of entities is pronounced, LLMs may resort to using string matching to predict hierarchies.

5.2 Analysis of LLM Factors

Choice of LLM: We find that larger (commercial) LLMs outperform their smaller (open) competitors. o1-preview outperforms all other models in our zero-shot experiments indicating the potential of reasoning LLMs for ontology-related tasks. DeepSeek's R1-Distil-Llama-8B has the worse performance despite being a reasoning LLM. Our experimentation suggests that reasoning LLMs do not always outperform general LLMs. For larger models, we observe that GPT-4o, when provided with a few examples (Table 2), performs as well as o1-preview under zero-shot conditions. The performance gap between GPT-4o and Llama3-8B is very pronounced in the zero-shot setting, but is significantly reduced in the few-shot experiments. Figures 2a and 2b highlight that both models benefit considerably from providing examples to the CASE ontology with the primary performance improvement observed on mAP@D. The performance improvement exhibited by Llama3-8B highlights how even smaller models have the potential to successfully support ontology population.

Prompting Approach: Including a single example leads to significant performance improvements, particularly on CASE. Significant improvements in mAP@D highlight how LLMs can adapt with little context. Improvements in mAP@1 are less pronounced in comparison. Adding examples primarily improves an LLM's ability to infer the correct ontology concept hierarchy. Figure 2 highlights that adding more examples yields continued performance improvements that plateaus at 3 - 4 examples. Performance gains from adding additional examples is more pronounced for Llama3-8B compared to GPT-4o. This suggests that smaller LLMs benefit more from seeing more examples. Experimentation with

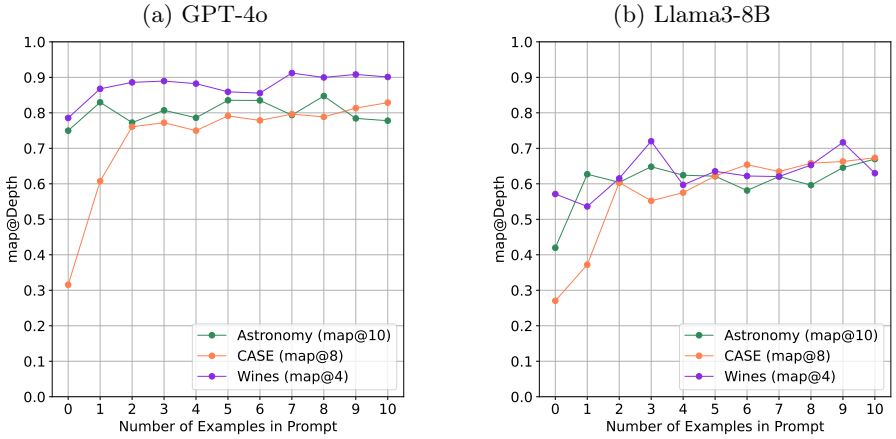


Fig. 2: mAP@D variation with number of examples for GPT-4o (a) and Llama3-8B (b) using ontology domain context

zero-shot and few-shot strategies highlights that few-shot prompting can provide significant performance improvements for ontology population, particularly when an LLM's entity label semantics differs from the ontological entity properties. Largest performance benefits are observed when using 3 or 4 examples beyond which smaller LLMs may benefit more.

Domain Context: Of the 72 (288/4) domain context experimentation groups (each combination of LLM, ontology and prompting strategy, over four types of domain context together comprise one group), ontology context yielded the best performance for 31 groups (43.1%) followed by ontology and topic (double) context for 21 groups (29.2%). From these 52 groups with ontology or double context as the best performer, in 27 groups, the second-best performer was the other context method. Of the remaining 20 groups, ontology or double context was second-best in 12 groups. This clearly highlights the preference for the taxonomical context variants. Topic context provided the best performance in 16 groups (22.2%). Table 2 reports the ontology domain context scores for each ontology and LLM under different prompting strategies. We observed the difference between the best and second-best strategies to be marginal (< 0.017) in several groups.

Response Variation: Figure 3 presents mAP@D variability exhibited by GPT-4o and Llama3-8B models run 10 times over Wines and CASE ontologies under different temperature settings. GPT-4o is fairly consistent across all temperature settings on both ontologies with variability increasing with temperature although it is marginal. Despite the fact that setting a temperature value of 0.0 should result in a deterministic output, both LLMs exhibit response variation at that temperature although it is quite small for GPT-4o. In contrast, Llama3-8B exhibits a great deal of variation at each temperature value. There is an observed degradation in performance with increase in temperature in GPT-4o, with the

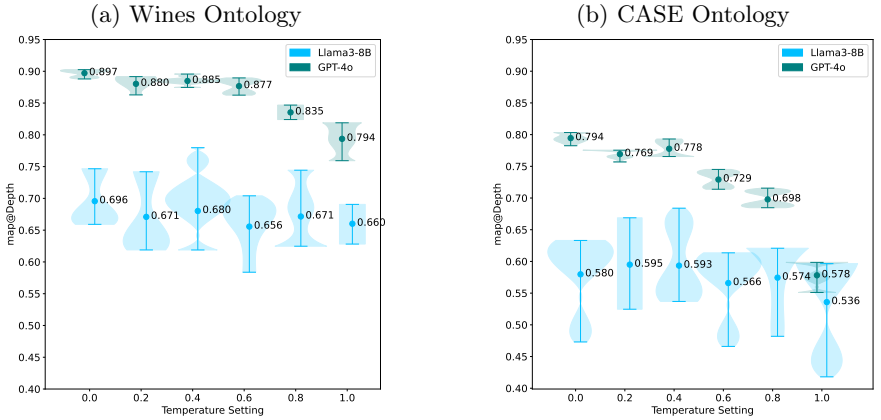


Fig. 3: Temperature variation for (a) Wines and (b) CASE ontologies for GPT-4o and Llama3-8B with 3-shot, ontology domain context. Central points are the average value over 10 runs. Lighter shaded regions are approximated densities over the range of values

highest performance observed at the lowest temperature. This decline is slightly more pronounced in the CASE Ontology. Similar performance trends are not observed for Llama3-8B. Performing Welch's T-test for statistically significant performance difference across both LLMs and ontologies against the default value of 0.6, we find performance differences observed by GPT-4o to be statistically significant for Wines at temperatures of 0.0, 0.8 and 1.0 and, for CASE at all temperatures. The same tests were found to be statistically insignificant in all scenarios for Llama3-8B except on Wines at temperature 0.0. For both ontologies, we observe a reduction in average performance with increase in temperature with GPT-4o. A similar trend is not observed for experiments with Llama3-8B.

6 Limitations and Future Directions

The present work does not explore the influence of all the outlined factors for ontology population. Our experiments are confined to domain-specific lightweight ontologies and do not explore the effects of using large-scale ontologies such as DBPedia [3]. The embeddings utilized for the analysis of entity label semantics are primarily representative of OpenAI models. Further analysis with more embeddings would better highlight the robustness of the observations made. The present experimentation does not include any few-shot experimentation with the reasoning models. Performance observed with the smaller reasoning models suggests more nuanced prompting is possibly required.

In future work, we wish to address the outlined limitations and extend our analysis to other ontology learning and enrichment tasks with additional emphasis on inter-factor variability.

7 Conclusion

We present an analysis of the ontology and LLM factors influencing LLM-driven ontology population. We investigate the ontology factors: structure, taxonomy and entity labels and, the LLM factors: choice of LLM, modelling objective, prompting approach, domain context and response variation over three domain ontologies and several LLMs.

Through extensive experimentation (288), we find that LLMs are capable of performing ontology population well, utilizing only their own pre-learnt knowledge and in-context learning. Our findings suggest that size of ontologies and the depth and dispersion of concepts do not influence LLM ontology population. In contrast, entity labels do contribute to performance variation. When an LLM's semantic understanding of entity labels deviates from their ontological nature, they struggle to predict hierarchies and are prone to use string-based matching for prediction. LLMs are therefore, quite capable of identifying directly-asserted concepts but often struggle with predicting correct hierarchies without additional context. Larger LLMs outperform smaller LLMs. Reasoning LLMs, under zero-shot conditions perform well but conventional models with few-shot prompts are equally adept. Inclusion of context through few-shot examples significantly improves performance across all LLMs and ontologies. The improvement in performance stems from better hierarchy prediction. We observe the largest benefits when providing three or four examples. Our experimentation with domain context strategies shows stronger benefits of the ontology domain context in most scenarios. However, the observed differences in performance are fairly marginal. Larger LLMs are more robust with consistent response generation and exhibit a mild decrease in performance with increase in temperature. Significant variation in responses from small LLMs renders temperature-based performance variation inconclusive.

Appendix A: Formalization of Evaluation Metrics

For each individual, the ground truth is defined as the sequence of concepts, starting at the individual's directly asserted concept, along the path to the top concept of that ontology. We provide a mathematical formulation of mAP@K with $K = 1, D$ being applied to obtain mAP@1 and mAP@D respectively.

Let the ground truth hierarchy for an individual t be given by:

$$\{c_i \mid c_i \in C \ ; \ c_i \subseteq c_{i+1} \ ; \ t \in \Sigma_{c_1} \ ; \ 1 \leq i \leq D - d + 1\} \tag{4}$$

where d is the depth of the directly-asserted concept c_1 and $d \leq D$
Let $\{y_i\}_{i=1}^{D-d+1}$ be the predicted hierarchy for t
The mAP at a sequence length of $K(\leq D - d + 1)$ is given by:

$$\text{mAP@K} = \frac{1}{|T|} \sum_{t \in T} \text{AP@K}_t \tag{5}$$

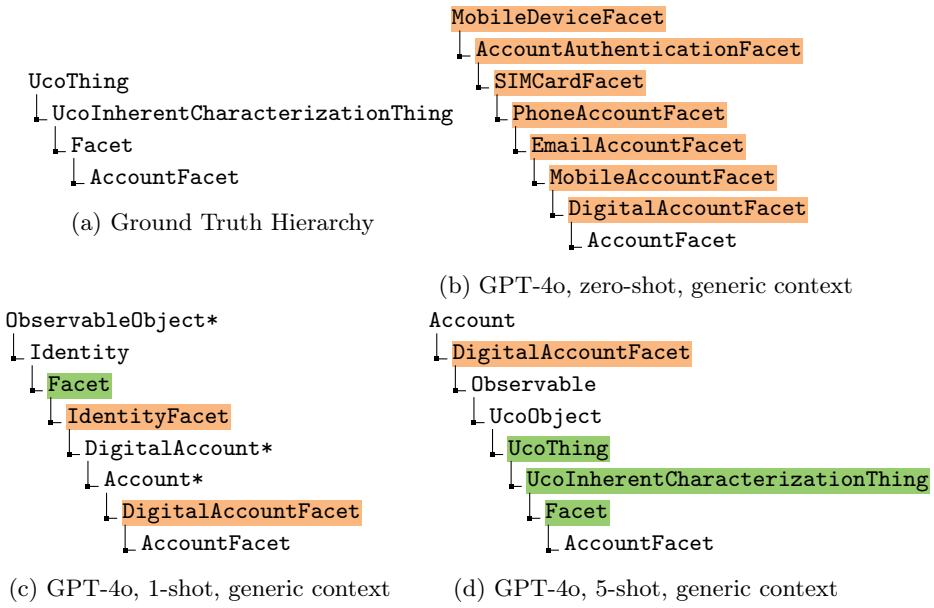
where,
T is the set of individuals;
$|\cdot|$ is set cardinality and;
AP@K_t is the average precision for t at a sequence length K, given by:

$$\text{AP@K}_t = \frac{1}{K} \sum_{x=1}^{K} \frac{\left|\{y_i\}_{i=1}^{x} \cap \{c_j\}_{j=1}^{x}\right|}{\left|\{y_k\}_{k=1}^{x}\right|} \tag{6}$$

The numerator in Equation 6 is the cardinality of the set of *true positives* and the denominator is the cardinality of the set of *predictions* for a single indvidual t.

Appendix B: String Similarity Prediction

Figure 4 highlights the hierarchy predicted by GPT-4o for an individual 'AccountFacet' in the CASE ontology on GPT-4o under different zero and few-shot conditions. In the zero-shot scenario, GPT-4o predicts primarily siblings rather than ancestors based on strong string similarity. As the number of examples provided increases, GPT-4o is able to predict the correct hierarchy.



(a) Ground Truth Hierarchy

(b) GPT-4o, zero-shot, generic context

(c) GPT-4o, 1-shot, generic context

(d) GPT-4o, 5-shot, generic context

Fig. 4: Ground Truth and predicted hierarchies of 'AccountFacet' taken from individual correctly identified as instance of 'AccountFacet'. Concepts in orange are siblings of 'AccountFacet' (incorrect hierarchy). Concepts in green are correctly identified ancestors. * are incorrect ancestors but belong to the correct hierarchy.

Bibliography

[1] Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G.: Gene ontology: tool for the unification of biology. Nature Genetics **25**(1), 25–29 (May 2000), ISSN 1546-1718, https://doi.org/10.1038/75556, URL http://dx.doi.org/10.1038/75556

[2] Asim, M.N., Wasim, M., Khan, M.U.G., Mahmood, W., Abbasi, H.M.: A survey of ontology learning techniques and applications. Database J. Biol. Databases Curation **2018**, bay101 (2018), https://doi.org/10.1093/DATABASE/BAY101, URL https://doi.org/10.1093/database/bay101

[3] Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.G.: Dbpedia: A nucleus for a web of open data. In: The Semantic Web, 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007, Busan, Korea, November 11-15, 2007, pp. 722–735 (2007), https://doi.org/10.1007/978-3-540-76298-0_52, URL https://doi.org/10.1007/978-3-540-76298-0_52

[4] Baeza-Yates, R.A., Ribeiro-Neto, B.A.: Modern Information Retrieval. ACM Press / Addison-Wesley (1999), ISBN 0-201-39829-X, URL http://www.dcc.ufmg.br/irbook/

[5] Bhattacharya, U., de Boer, M.H., Sosnovsky, S.A.: Automatic Ontology Term Typing by LLMs: The Impact of Prompt and Ontology Variation. CEUR Workshop Proceedings **3967** (2024), URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-105006903401&partnerID=40&md5=ed7b3771dc19477cc1c99014f80d8788, section: 0

[6] Bombieri, M., Fiorini, P., Ponzetto, S.P., Rospocher, M.: Do LLMs Dream of Ontologies? ACM Trans. Intell. Syst. Technol. (Mar 2025), ISSN 2157-6904, https://doi.org/10.1145/3725852, URL https://doi.org/10.1145/3725852, section: 0

[7] Brank, J., Grobelnik, M., Mladenic, D.: A survey of ontology evaluation techniques. In: Proceedings of the conference on data mining and data warehouses (SiKDD 2005), pp. 166–170 (2005)

[8] Buitelaar, P., Cimiano, P., Magnini, B.: Ontology learning from text: An overview. Ontology learning from text: Methods, evaluation and applications **123**, 3–12 (2005)

[9] Casey, E., Barnum, S., Griffith, R., Snyder, J., van Beek, H., Nelson, A.: The Evolution of Expressing and Exchanging Cyber-Investigation Information in a Standardized Form, p. 43–58. Springer International Publishing (2018), ISBN 9783319748726, https://doi.org/10.1007/978-3-319-74872-6_4, URL http://dx.doi.org/10.1007/978-3-319-74872-6_4

[10] Casey, E., Barnum, S., Griffith, R., Snyder, J., van Beek, H.M.A., Nelson, A.: Advancing coordinated cyber-investigations and tool interoperability

using a community developed specification language. Digit. Investig. **22**, 14–45 (2017), https://doi.org/10.1016/J.DIIN.2017.08.002, URL https://doi.org/10.1016/j.diin.2017.08.002

[11] Caufield, J.H., Hegde, H., Emonet, V., Harris, N.L., Joachimiak, M.P., Matentzoglu, N., Kim, H., Moxon, S.A.T., Reese, J.T., Haendel, M.A., Robinson, P.N., Mungall, C.J.: Structured prompt interrogation and recursive extraction of semantics (SPIRES): a method for populating knowledge bases using zero-shot learning. Bioinform. **40**(3) (2024), https://doi.org/10.1093/BIOINFORMATICS/BTAE104, URL https://doi.org/10.1093/bioinformatics/btae104

[12] Chen, J., He, Y., Geng, Y., Jiménez-Ruiz, E., Dong, H., Horrocks, I.: Contextual semantic embeddings for ontology subsumption prediction. World Wide Web (WWW) **26**(5), 2569–2591 (2023), https://doi.org/10.1007/S11280-023-01169-9, URL https://doi.org/10.1007/s11280-023-01169-9

[13] Corey, D.M., Dunlap, W.P., Burke, M.J.: Averaging correlations: Expected values and bias in combined pearson rs and fisher's z transformations. The Journal of General Psychology **125**(3), 245–261 (1998), https://doi.org/10.1080/00221309809595548, URL https://doi.org/10.1080/00221309809595548

[14] Davies, D.L., Bouldin, D.W.: A cluster separation measure. IEEE Transactions on Pattern Analysis and Machine Intelligence **PAMI-1**(2), 224–227 (1979), https://doi.org/10.1109/TPAMI.1979.4766909

[15] DeepSeek-AI, Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., Zhang, X., Yu, X., Wu, Y., Wu, Z.F., Gou, Z., Shao, Z., Li, Z., Gao, Z., Liu, A., Xue, B., Wang, B., Wu, B., Feng, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., Dai, D., Chen, D., Ji, D., Li, E., Lin, F., Dai, F., Luo, F., Hao, G., Chen, G., Li, G., Zhang, H., Bao, H., Xu, H., Wang, H., Ding, H., Xin, H., Gao, H., Qu, H., Li, H., Guo, J., Li, J., Wang, J., Chen, J., Yuan, J., Qiu, J., Li, J., Cai, J.L., Ni, J., Liang, J., Chen, J., Dong, K., Hu, K., Gao, K., Guan, K., Huang, K., Yu, K., Wang, L., Zhang, L., Zhao, L., Wang, L., Zhang, L., Xu, L., Xia, L., Zhang, M., Zhang, M., Tang, M., Li, M., Wang, M., Li, M., Tian, N., Huang, P., Zhang, P., Wang, Q., Chen, Q., Du, Q., Ge, R., Zhang, R., Pan, R., Wang, R., Chen, R.J., Jin, R.L., Chen, R., Lu, S., Zhou, S., Chen, S., Ye, S., Wang, S., Yu, S., Zhou, S., Pan, S., Li, S.S.: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. CoRR **abs/2501.12948** (2025), https://doi.org/10.48550/ARXIV.2501.12948, URL https://doi.org/10.48550/arXiv.2501.12948

[16] Dong, H., Chen, J., He, Y., Gao, Y., Horrocks, I.: A Language Model Based Framework for New Concept Placement in Ontologies. Lecture Notes in Computer Science **14664**, 79 – 99 (2024), https://doi.org/10.1007/978-3-031-60626-7_5, URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85194222248&doi=10.1007%2F978-3-031-60626-7_5&partnerID=40&md5=7646da39abb6a08ff3ab8c32975bb31e, section: 0

[17] Du, R., An, H., Wang, K., Liu, W.: A short review for ontology learning from text: Stride from shallow learning, deep learning to large language models trend. CoRR **abs/2404.14991** (2024), https://doi.org/10.48550/ARXIV.2404.14991, URL https://doi.org/10.48550/arXiv.2404.14991

[18] Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., Rao, A., Zhang, A., Rodriguez, A., Gregerson, A., Spataru, A., Rozière, B., Biron, B., Tang, B., Chern, B., Caucheteux, C., Nayak, C., Bi, C., Marra, C., McConnell, C., Keller, C., Touret, C., Wu, C., Wong, C., Ferrer, C.C., Nikolaidis, C., Allonsius, D., Song, D., Pintz, D., Livshits, D., Esiobu, D., Choudhary, D., Mahajan, D., Garcia-Olano, D., Perino, D., Hupkes, D., Lakomkin, E., AlBadawy, E., Lobanova, E., Dinan, E., Smith, E.M., Radenovic, F., Zhang, F., Synnaeve, G., Lee, G., Anderson, G.L., Nail, G., Mialon, G., Pang, G., Cucurell, G., Nguyen, H., Korevaar, H., Xu, H., Touvron, H., Zarov, I., Ibarra, I.A., Kloumann, I.M., Misra, I., Evtimov, I., Copet, J., Lee, J., Geffert, J., Vranes, J., Park, J., Mahadeokar, J., Shah, J., van der Linde, J., Billock, J., Hong, J., Lee, J., Fu, J., Chi, J., Huang, J., Liu, J., Wang, J., Yu, J., Bitton, J., Spisak, J., Park, J., Rocca, J., Johnstun, J., Saxe, J., Jia, J., Alwala, K.V., Upasani, K., Plawiak, K., Li, K., Heafield, K., Stone, K., et al.: The llama 3 herd of models. CoRR **abs/2407.21783** (2024), https://doi.org/10.48550/ARXIV.2407.21783, URL https://doi.org/10.48550/arXiv.2407.21783

[19] Funk, M., Hosemann, S., Jung, J.C., Lutz, C.: Towards Ontology Construction with Language Models. CEUR Workshop Proceedings **3577** (2023), URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85179559244&partnerID=40&md5=33969446bee8becffb8a2e0211cb3651, section: 0

[20] Gangemi, A., Catenacci, C., Ciaramita, M., Lehmann, J.: Ontology evaluation and validation an integrated formal model for the quality diagnostic task (2005), URL https://api.semanticscholar.org/CorpusID:3087032

[21] Giglou, H.B., D'Souza, J., Auer, S.: Llms4ol: Large language models for ontology learning. In: The Semantic Web - ISWC 2023 - 22nd International Semantic Web Conference, Athens, Greece, November 6-10, 2023, Proceedings, Part I, pp. 408–427 (2023), https://doi.org/10.1007/978-3-031-47240-4_22, URL https://doi.org/10.1007/978-3-031-47240-4_22

[22] Giglou, H.B., D'Souza, J., Auer, S.: Llms4ol 2024 overview: The 1st large language models for ontology learning challenge. In: LLMs4OL 2024: The 1st Large Language Models for Ontology Learning Challenge at the 23rd ISWC, Co-located with the 23rd International Semantic Web Conference (ISWC 2024), Baltimore, Maryland, USA, November 11-15, 2024, pp. 3–16 (2024), https://doi.org/10.52825/OCP.V4I.2473, URL https://doi.org/10.52825/ocp.v4i.2473

[23] Giglou, H.B., D'Souza, J., Engel, F.C., Auer, S.: LLMs4OM: Matching Ontologies with Large Language Models. Lecture Notes in Computer Science **15344**, 25 – 35 (2025), https://doi.org/10.1007/978-3-031-78952-6_3, URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85218455564&doi=10.1007%2F978-3-031-78952-6_3&partnerID=40&md5=c3c4db12909f8024518974717c695310, section: 0

[24] He, Y., Chen, J., Antonyrajah, D., Horrocks, I.: Bertmap: A bert-based ontology alignment system. In: Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022, pp. 5684–5691 (2022), https://doi.org/10.1609/AAAI.V36I5.20510, URL https://doi.org/10.1609/aaai.v36i5.20510

[25] He, Y., Chen, J., Dong, H., Horrocks, I.: Exploring Large Language Models for Ontology Alignment. CEUR Workshop Proceedings **3632** (2023), URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85184375684&partnerID=40&md5=43d321af25029eca4ea303b161851303, section: 0

[26] He, Y., Chen, J., Jimenez-Ruiz, E., Dong, H., Horrocks, I.: Language Model Analysis for Ontology Subsumption Inference. Findings of the Association for Computational Linguistics: ACL 2023 pp. 3439–3453 (Jul 2023), https://doi.org/10.18653/v1/2023.findings-acl.213, URL https://aclanthology.org/2023.findings-acl.213/, place: Toronto, Canada Publisher: Association for Computational Linguistics Section: 0

[27] Hlomani, H., Stacey, D.A.: Approaches , methods , metrics , measures , and subjectivity in ontology evaluation : A survey (2014), URL https://api.semanticscholar.org/CorpusID:51371006

[28] Khadir, A.C., Aliane, H., Guessoum, A.: Ontology learning: Grand tour and challenges. Comput. Sci. Rev. **39**, 100339 (2021), https://doi.org/10.1016/J.COSREV.2020.100339, URL https://doi.org/10.1016/j.cosrev.2020.100339

[29] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., Kiela, D.: Retrieval-augmented generation for knowledge-intensive NLP tasks. In: Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual (2020), URL https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html

[30] Liu, H., Perl, Y., Geller, J.: Concept placement using BERT trained by transforming and summarizing biomedical ontology structure. J. Biomed. Informatics **112**, 103607 (2020), https://doi.org/10.1016/J.JBI.2020.103607, URL https://doi.org/10.1016/j.jbi.2020.103607

[31] Liu, H., Tam, D., Muqeeth, M., Mohta, J., Huang, T., Bansal, M., Raffel, C.: Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. In: Advances in Neural Information Processing Sys-

tems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022 (2022), URL http://papers.nips.cc/paper_files/paper/2022/hash/0cde695b83bd186c1fd456302888454c-Abstract-Conference.html

[32] Liu, Y., Deng, G., Xu, Z., Li, Y., Zheng, Y., Zhang, Y., Zhao, L., Zhang, T., Liu, Y.: Jailbreaking chatgpt via prompt engineering: An empirical study. CoRR **abs/2305.13860** (2023), https://doi.org/10.48550/ARXIV.2305.13860, URL https://doi.org/10.48550/arXiv.2305.13860

[33] Lubani, M., Noah, S.A.M., Mahmud, R.: Ontology population: Approaches and design aspects. J. Inf. Sci. **45**(4) (2019), https://doi.org/10.1177/0165551518801819, URL https://doi.org/10.1177/0165551518801819

[34] Mai, H.T., Chu, C.X., Paulheim, H.: Do llms really adapt to domains? an ontology learning perspective. In: The Semantic Web - ISWC 2024 - 23rd International Semantic Web Conference, Baltimore, MD, USA, November 11-15, 2024, Proceedings, Part I, pp. 126–143 (2024), https://doi.org/10.1007/978-3-031-77844-5_7, URL https://doi.org/10.1007/978-3-031-77844-5_7

[35] Mascardi, V., Cordì, V., Rosso, P.: A comparison of upper ontologies. In: WOA 2007: Dagli Oggetti agli Agenti. 8th AI*IA/TABOO Joint Workshop "From Objects to Agents": Agents and Industry: Technological Applications of Software Agents, 24-25 September 2007, Genova, Italy, pp. 55–64 (2007), URL http://woa07.disi.unige.it/papers/mascardi.pdf

[36] Mateiu, P., Groza, A.: Ontology engineering with Large Language Models. 2023 25th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC) pp. 226–229 (Sep 2023), https://doi.org/10.1109/SYNASC61333.2023.00038, URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85193849216&doi=10.1109%2FSYNASC61333.2023.00038&partnerID=40&md5=82842416ae615f60e5776b5c7fd1ef98, section: 0

[37] Mosbach, M., Pimentel, T., Ravfogel, S., Klakow, D., Elazar, Y.: Few-shot fine-tuning vs. in-context learning: A fair comparison and evaluation. In: Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023, pp. 12284–12314 (2023), https://doi.org/10.18653/V1/2023.FINDINGS-ACL.779, URL https://doi.org/10.18653/v1/2023.findings-acl.779

[38] Norouzi, S.S., Barua, A., Christou, A., Gautam, N., Eells, A., Hitzler, P., Shimizu, C.: Ontology population using llms. CoRR **abs/2411.01612** (2024), https://doi.org/10.48550/ARXIV.2411.01612, URL https://doi.org/10.48550/arXiv.2411.01612

[39] Noy, N.F., McGuinness, D.L., et al.: Ontology development 101: A guide to creating your first ontology (2001)

[40] OpenAI: Hello GPT-4o (2024), URL https://openai.com/index/hello-gpt-4o/

[41] OpenAI: Introducing OpenAI O1 (2024), URL https://openai.com/o1/

[42] OpenAI: New embedding models and API updates — openai.com. https://openai.com/index/new-embedding-models-and-api-updates/ (25-01-2024), [Accessed 06-05-2025]

[43] Petasis, G., Karkaletsis, V., Paliouras, G., Krithara, A., Zavitsanos, E.: Ontology population and enrichment: State of the art. In: Knowledge-Driven Multimedia Information Extraction and Ontology Evolution - Bridging the Semantic Gap, pp. 134–166 (2011), https://doi.org/10.1007/978-3-642-20795-2_6, URL https://doi.org/10.1007/978-3-642-20795-2_6

[44] Qiao, S., Ou, Y., Zhang, N., Chen, X., Yao, Y., Deng, S., Tan, C., Huang, F., Chen, H.: Reasoning with language model prompting: A survey. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023, pp. 5368–5393 (2023), https://doi.org/10.18653/V1/2023.ACL-LONG.294, URL https://doi.org/10.18653/v1/2023.acl-long.294

[45] Sahbi, A., Alec, C., Beust, P.: Semantic vs. LLM-based approach: A case study of KOnPoTe vs. Claude for ontology population from French advertisements. Data and Knowledge Engineering **156** (2025), https://doi.org/10.1016/j.datak.2024.102392, URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85211195979&doi=10.1016%2Fj.datak.2024.102392&partnerID=40&md5=7adba98e00024b722539b1e6c7c8ebdc, section: 0

[46] Schulhoff, S., Ilie, M., Balepur, N., Kahadze, K., Liu, A., Si, C., Li, Y., Gupta, A., Han, H., Schulhoff, S., et al.: The prompt report: a systematic survey of prompt engineering techniques. arXiv preprint arXiv:2406.06608 (2024)

[47] Shaya, E.: Astronomy Ontology (2012), URL https://www.astro.umd.edu/~eshaya/astro-onto/ontologies/astronomy.html

[48] Wang, G., Sun, Z., Gong, Z., Ye, S., Chen, Y., Zhao, Y., Liang, Q., Hao, D.: Do advanced language models eliminate the need for prompt engineering in software engineering? arXiv preprint arXiv:2411.02093 (2024)

[49] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E.H., Le, Q.V., Zhou, D.: Chain-of-thought prompting elicits reasoning in large language models. In: Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022 (2022), URL http://papers.nips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html

[50] White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., Schmidt, D.C.: A prompt pattern catalog to enhance prompt engineering with chatgpt. CoRR **abs/2302.11382** (2023), https://doi.org/10.48550/ARXIV.2302.11382, URL https://doi.org/10.48550/arXiv.2302.11382

[51] Wilson, R.S.I., Goonetillake, J.S., Ginige, A., Walisadeera, A.I.: Ontology quality evaluation methodology. In: Computational Science and Its Appli-

cations - ICCSA 2022 - 22nd International Conference, Malaga, Spain, July 4-7, 2022, Proceedings, Part I, pp. 509–528 (2022), https://doi.org/10.1007/978-3-031-10522-7_35, URL https://doi.org/10.1007/978-3-031-10522-7_35

[52] Wong, W., Liu, W., Bennamoun, M.: Ontology learning from text: A look back and into the future. ACM Comput. Surv. **44**(4), 20:1–20:36 (2012), https://doi.org/10.1145/2333112.2333115, URL https://doi.org/10.1145/2333112.2333115

[53] Zhou, L.: Ontology learning: state of the art and open issues. Inf. Technol. Manag. **8**(3), 241–252 (2007), https://doi.org/10.1007/S10799-007-0019-5, URL https://doi.org/10.1007/s10799-007-0019-5