LLMs4OL 2024: The 1st Large Language Models for Ontology Learning Challenge at the 23rd ISWC
LLMs4OL 2024 Task Overview
https://doi.org/10.52825/ocp.v4i.2473
© Authors. This work is licensed under a Creative Commons Attribution 4.0 International License
Published: 02 Oct. 2024

LLMs4OL 2024 Overview: The 1st Large Language Models for Ontology Learning Challenge

Hamed Babaei Giglou, Jennifer D'Souza, and Sören Auer

TIB Leibniz Information Centre for Science and Technology, Hannover, Germany
{hamed.babaei, jennifer.dsouza, auer}@tib.eu

*Correspondence: Hamed Babaei Giglou, hamed.babaei@tib.eu

Abstract: This paper outlines the LLMs4OL 2024, the first edition of the Large Language Models for Ontology Learning Challenge. LLMs4OL is a community development initiative collocated with the 23rd International Semantic Web Conference (ISWC) to explore the potential of Large Language Models (LLMs) in Ontology Learning (OL), a vital process for enhancing the web with structured knowledge to improve interoperability. By leveraging LLMs, the challenge aims to advance understanding and innovation in OL, aligning with the goals of the Semantic Web to create a more intelligent and user-friendly web. In this paper, we give an overview of the 2024 edition of the LLMs4OL challenge[1] and summarize the contributions.

Keywords: LLMs4OL Challenge, Ontology Learning, Large Language Models

1 Introduction

The Semantic Web aims to enrich the current web with structured knowledge and metadata, enabling enhanced interoperability and understanding across diverse systems. At the core of this endeavor is Ontology Learning (OL), a process that automates the extraction of structured knowledge from unstructured data [1], essential for constructing dynamic ontologies that underpin the Semantic Web. The emergence of Large Language Models (LLMs) like GPT-3 [2] and GPT-4 [3] has revolutionized natural language processing (NLP), demonstrating remarkable performance across tasks such as language translation, question answering, and text generation. These models are particularly adept at crystallizing existing textual knowledge from a vast array of sources, making them potentially valuable for OL, where the goal is to extract a shared conceptualization of concepts and relationships from diverse inputs [4]. The introduction of LLMs has thus opened up new avenues of research, including the exploration of their potential in automating the OL process.

In our prior work published in the ISWC 2023 research track proceedings titled "LLMs4OL: Large Language Models for Ontology Learning" [5], marked a notable direction towards employing LLMs in OL, demonstrating their potential in automating knowledge acquisition and representation for the Semantic Web. Based on this research, the **The 1st Large Language Models for Ontology Learning Challenge at**

[1] https://sites.google.com/view/llms4ol

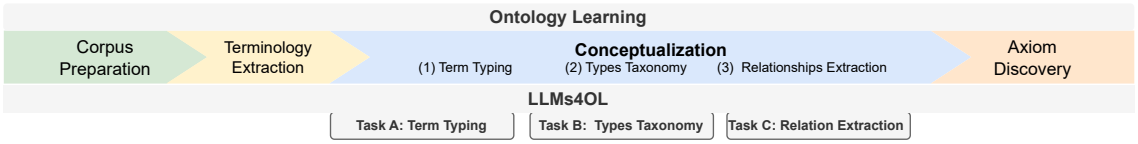

Figure 1. *The LLMs4OL task paradigm is an end-to-end framework for ontology learning. The three OL tasks that empirically validated in the LLMs4OL 2024 challenge, based on our prior research [5], are depicted within the blue arrow, aligned with the greater LLMs4OL paradigm.*

the 23rd ISWC 2024 (1st LLMs4OL Challenge @ ISWC 2024) was introduced as a call for community development. With the LLMs4OL challenge, we aimed to catalyze community-wide engagement in validating and expanding the use of LLMs in OL. This initiative is poised to advance our comprehension of LLMs' roles within the Semantic Web, encouraging innovation and collaboration in developing scalable and accurate ontology learning methods.

LLMs4OL consists of three OL tasks, *Task A – Term Typing*, *Task B – Taxonomy Discovery*, and *Task C – Non-Taxonomic Relation Extraction*. While participation in all three tasks in the LLMs4OL 2024 challenge is stipulated as desirable, but not mandatory. Thus participants choose to enroll only in Task A or B or C, or Task A and B, or Task A and C, or Task B and C. Furthermore, participants are required to implement LLM-based solutions, we did not impose any restrictions on the LLM prompting methods. For instance, they can choose to bring in additional context information from the World Wide Web to enrich the training and test instances. To thoroughly explore the potential of LLMs for OL, we structured the challenge around two distinct evaluation phases: *(1) Few-shot testing phase* and *(2) Zero-shot testing phase*. Through this work, we aim to contribute to the ongoing discourse on the capabilities of LLMs, particularly in the context of OL, and to provide insights into their potential for enhancing the Semantic Web. Thus, in the remainder of this paper, we detail the challenge tasks, what LLMs are being used, participant contributions, and findings.

2 LLMs4OL 2024 Tasks

In the LLMs4OL 2024 challenge, we have organized three main tasks which are centered around the ontology primitives [6] that comprise the following: **1.** a set of strings that describe terminological lexical entries L for conceptual types; **2.** a set of conceptual types T; **3.** a taxonomy of types in a hierarchy H_T; **4.** a set of non-taxonomic relations R described by their domain and range restrictions arranged in a heterarchy of relations H_R; and **5.** a set of axioms A that describe additional constraints on the ontology and make implicit facts explicit.

To address these primitives, the tasks for OL [7] are: 1) Corpus preparation – collecting source texts for building ontology. 2) Terminology extraction – extracting relevant terms from the texts. 3) Term typing – grouping similar terms into conceptual types. 4) Taxonomy construction – establishing "is-a" hierarchies between types. 5) Relationship extraction – extracting semantic relationships beyond "is-a" between types. 6) Axiom discovery – finding constraints rules for the ontology. These tasks constitute the LLMs4OL task paradigm as depicted in Figure 1. Assuming the corpus preparation step is done by reusing ontologies publicly released in the community, we introduced the following three main tasks for the first edition of the LLMs4OL challenge.

Table 1. *LLMs4OL 2024 challenge, subtasks, domains, number of participants per subtasks, and evaluation phases.*

Task	***SubTask***	***Domain***	***Participants***	***Phase***
A	A.1 - WordNet	lexicosemantics	7	Few-shot
	A.2 - GeoNames	geographical locations	5	
	A.3 - UMLS - NCI	biomedical	5	
	A.3 - UMLS - MEDCIN		4	
	A.3 - UMLS - SNOMEDCT_US		4	
	A.4 - GO - Biological Process	biological	5	
	A.4 - GO - Cellular Component		5	
	A.4 - GO - Molecular Function		5	
	A.5 - DBO	general knowledge	2	Zero-shot
	A.6 - FoodOn	food	2	
B	B.1 - GeoNames	geographical locations	5	Few-shot
	B.2 - Schema.org	web content types	3	
	B.3 - UMLS	biomedical	3	
	B.4 - GO	biological	1	
	B.5 - DBO	general knowledge	2	Zero-shot
	B.6 - FoodOn	food	1	
C	C.1 - UMLS	biomedical	2	Few-shot
	C.2 - GO	biological	0	
	C.3 - FoodOn	food	0	Zero-shot

2.1 Task A – *Term Typing*

The Table 1 shows 10 subtasks for *Task A* across 6 distinct domains such as lexicosemantics, geographical locations, biomedical, biological, general knowledge, and food domains. This task is defined as "*discover the generalized type for a given lexical term*". For this task, for each ontology, participants are given training instances defined as following formalism.

$$f_{prompt}^{TaskA}(L) := [S?].\ ([L],\ [T])$$

Where S is an optional context sentence (if available in the source ontology), L is the lexical term prompted for, and T is the conceptual term type. In the test phase, types are hidden, and participants predict them for given terms using their trained models.

2.2 Task B – *Taxonomy Discovery*

After grouping terms under a conceptual type, in Task B, the goal is for given types "*discover the taxonomic hierarchy between types*", where the hierarchy between types is defined with an "is-a" relationship. Participants receive training instances for 6 distinct subtasks (described in Table 1) as :

$$f_{prompt}^{TaskB}(a, b) := (T_a,\ T_b)$$

Where T_a is the parent (superclass) of T_b, and T_b is the child (subclass) of T_a. The goal is to train a system to correctly identify the taxonomy between type. The training dataset will include term types and taxonomically related type pairs. In the test phase, participants work with just term types and must use their trained models to identify correct taxonomic relationships (type pairs). The types for the training and test phases are mutually exclusive. Furthermore, for the testing phase participants are required to post-process their outputs to return type pairs that follow the order of superclass-subclass related types.

2.3 Task C – ***Non-Taxonomic Relation Extraction***

Nonetheless, the "is-a" relations are not the only relations in ontologies. So, Task C aims to "*identify non-taxonomic, semantic relations between types*". Training instances are given for three subtasks *C.1 - UMLS*, *C.2 - GO*, and *C.3 - FoodOn* as:

$$f_{prompt}^{TaskC}(h, r, t) := (T_h, r, T_t)$$

Where, T_h and T_t are head and tail taxonomic types, respectively, and r is the non-taxonomic semantic relation between them, chosen from a predefined set R of semantic relations. Participants aimed to train a system to identify pairs of types, and then classify pairs of types into semantic relations. The training phase involves types, relations, and triples of semantic relations; the test phase requires applying the trained system to predict semantically related triples from given types and the set of relations.

The caveat here is that we do not expect participant systems to infer a semantic relation but rather establish semantically related types and classify their relation from a known set of predetermined relations. This implies that any manual ontology specification task predetermines which semantic relations hold for the given ontology. In an alternative scenario, where participants might have had to infer the semantic relation, we realize that the possibilities of semantic relations might have been rather vast. Hence we posit a more realistic task design by predetermining the possible set of semantic relations.

3 Evaluation

There are two main evaluation phases for the challenge, which are the following:

- **Few-shot testing phase.** Each ontology selected for system training will be divided into two parts: one part will be released for the training of the systems and another part will be reserved for the testing of systems in this phase.
- **Zero-shot testing phase.** New ontologies that are unseen during training will be introduced. The objective is to evaluate the generalizability and transferability of the LLMs developed in this challenge.

For evaluation, we used the challenge datasets [8] – available at challenge GitHub[2] repository – with *standard evaluation metrics* used for all tasks. Given $\mathcal{G}(i)$ as a set of ground truth labels for sample i, and $\mathcal{P}(i)$ as a set of predicted labels for sample i, the precision P, recall R, and F1-score $F1$ are being calculated as follows:

$$P = \frac{\sum_i |\mathcal{G}(i) \cap \mathcal{P}(i)|}{\sum_i |\mathcal{P}(i)|}, \quad R = \frac{\sum_i |\mathcal{G}(i) \cap \mathcal{P}(i)|}{\sum_i |\mathcal{G}(i)|}, \quad F1 = \frac{2 \times P \times R}{P + R}$$

With precision, we assessed the percentage of the returned related pairs, while recall was used to measure the proportion of correct pairs that were accurately retrieved. In the end, the F1-score was calculated as the harmonic mean of precision and recall, serving as a comparison metric for the participants' submissions. We used Codalab[3] [9] submission platform to organize participants submissions and scoring.

4 Participant Systems and Results

The LLMs4OL 2024 challenge has inspired diverse solutions, showcasing the growing potential of LLMs for OL tasks. Using the Codalab submissions platform, for this

[2] https://github.com/HamedBabaei/LLMs4OL-Challenge-ISWC2024
[3] https://codalab.lisn.upsaclay.fr/competitions/19547

Table 2. *LLMs4OL 2024 challenge participants methods. * refers to the subtask that did not make the submission to the leaderboard but was reported in the paper. `MF` refers to "Molecular Function", `CC` refers to "Cellular Component", and `BF` refers to "Biological Process". `NCI`, `SNOMEDCT_US`, and `MEDCIN` are from "UMLS".*

Team Name	LLM of Use	Approach	Code	A.1 - WordNet	A.2 - GeoNames	A.3 - NCI	A.3 - MEDCIN	A.3 - SNOMEDCT_US	A.4 - GO - BP	A.4 - GO - CC	A.4 - GO - MF	A.5 - DBO	A.6 - FoodOn	B.1 - GeoNames	B.2 - Schema.org	B.3 - UMLS	B.4 - GO	B.5 - DBO	B.6 - FoodOn	C.1 - UMLS	C.2 - GO	C.3 - FoodOn
DSTI [10]	Flan-T5 GTE-Large	Fine-tuning RAG		✓	*																	
DaSeLab [11]	GPT-3.5-Turbo	Fine-tuning		✓	✓	✓	✓	✓														
RWTH-DBIS [12]	GPT-3.5-Turbo LLaMA-3-8B	Prompting Fine-Tuning		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓							
SKH-NLP [13]	LLaMA-3-70B Sentence-BERT	Prompting Fine-Tuning												✓								
TheGhost [14]	BLOOM-1B7 BLOOM-3B BLOOM- 7B1 LLaMA-7B LLaMA-2-7B LLaMA-3-8B BioMistral-7B OpenBioLLM-8B	Prompt-Tuning		✓	✓	✓	✓	✓	✓	✓	✓											
silp_nlp [15]	GPT-4o Mixtral-8x7B LLaMA-3-8B BERT Sentence-BERT	Prompting Fine-Tuning ML		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓		✓		
Phoenixes [16]	Mistral-7B Sentence-BERT	RAG		✓		✓			✓	✓	✓			✓	✓	✓	✓	✓	✓	✓		
TSOTSALearning [17]	GPT-4 BERT	RAG Rules		✓	✓				✓	✓	✓			✓								

challenge we set a limit of 10 submissions per day and a total of 30 submissions per subtask. We received 272 total submissions from 14 participants. In final, this challenge attracted the interest of the final eight research teams, as demonstrated by the various approaches they submitted for the subtasks. Each subtask of the competition depicted a rigorous field inherent to OL, which helped facilitate breakthroughs in finding generalized types (Task A), identifying taxonomic hierarchies (Task B), and extracting non-taxonomic relations (Task C), further scaffolding future AI advancements. Notably, teams employed varied strategies to tackle subtasks, such as fine-tuning, prompt-tuning, and retrieval-augmented generation (RAG). These approaches were used to analyze OL tasks across domains like lexicosemantics, geographical locations, biomedical concepts, and more (see Table 1 for subtasks and domains involved in this challenge). The summary of explored LLMs and subtasks are presented in Table 2 and in the following we will detail contributions and findings.

4.1 Participants Contributions

The results for Task A are presented in Table 3, for Task B in Table 5, and for Task C in Table 4.

DSTI [10]. DSTI fine-tuned Flan-T5-Small [18] model for *SubTasks A.1 - WordNet* and *A.2 - GeoNames*. Obtained F1-score of 0.9716 for *SubTask A.1* and ranked as a second team. But for GeoNames they did not submit the model to the leaderboard due to the larger nature of GeoNames dataset that required more computational resources. They introduced two approaches for OL. The first approach is fine-tuning LLMs using the zero-shot prompting method, the second approach is using a RAG pipeline using the General Text Embeddings (GTE)-Large [19] model as a retriever and fine-tuned LLM as a retriever. Due to the computational resources they preferred to use the Flan-T5-small model, and the results showed the effectiveness of their approach.

Table 3. *Task A - Term Typing Results for SubTasks*

SubTask	Team Name	F1-Score	Precision	Recall
A.1 (FS) - WordNet	TSOTSALearning	**0.9938**	**0.9938**	**0.9938**
	DSTI	0.9716	0.9716	0.9716
	DaseLab	0.9697	0.9689	0.9704
	RWTH-DBIS	0.9446	0.9446	0.9446
	TheGhost	0.9392	0.9389	0.9395
	Silp_nlp	0.9037	0.9037	0.9037
	Phoenixes	0.8158	0.7689	0.8687
A.2 (FS) - GeoNames	DaseLab	**0.5906**	0.5906	**0.5906**
	Silp_nlp	0.4433	**0.7503**	0.3146
	RWTH-DBIS	0.4355	0.4355	0.4355
	TSOTSALearning	0.2937	0.2937	0.2937
	TheGhost	0.1489	0.1461	0.1519
A.3 (FS) - UMLS - NCI	DaseLab	**0.8249**	0.8161	**0.8340**
	Silp_nlp	0.6974	**0.8792**	0.5779
	TheGhost	0.5370	0.4450	0.6769
	RWTH-DBIS	0.1691	0.1821	0.1579
	Phoenixes	0.0737	0.0562	0.1070
A.3 (FS) - UMLS - MEDCIN	Silp_nlp	**0.9382**	**0.9591**	0.9181
	DaseLab	0.9373	0.9379	**0.9366**
	TheGhost	0.5328	0.4183	0.7336
	RWTH-DBIS	0.4566	0.4607	0.4526
A.3 (FS) - UMLS - SNOMEDCT_US	DaseLab	**0.8829**	**0.8810**	**0.8848**
	Silp_nlp	0.7552	0.8583	0.6742
	TheGhost	0.5275	0.4266	0.6910
	RWTH-DBIS	0.4747	0.4888	0.4613
A.4 (FS) - GO - Cellular Component	Silp_nlp	**0.2726**	**0.4279**	0.2000
	RWTH-DBIS	0.2178	0.1846	**0.2656**
	TheGhost	0.1877	0.1653	0.2171
	TSOTSALearning	0.0638	0.0767	0.0545
	Phoenixes	0.0158	0.0124	0.0217
A.4 (FS) - GO - Biological Process	Silp_nlp	**0.2691**	**0.4006**	**0.2026**
	TheGhost	0.1025	0.0964	0.1095
	RWTH-DBIS	0.0881	0.0693	0.1207
	TSOTSALearning	0.0648	0.0806	0.0542
	Phoenixes	0.0319	0.0214	0.0622
A.4 (FS) - GO - Molecular Function	Silp_nlp	**0.2970**	**0.4185**	**0.2302**
	RWTH-DBIS	0.1418	0.1670	0.1231
	TheGhost	0.1270	0.1278	0.1261
	TSOTSALearning	0.0910	0.1072	0.0790
	Phoenixes	0.0700	0.0485	0.1256
A.5 (ZS) - DBO	RWTH-DBIS	**0.4270**	**0.4270**	**0.4270**
	Silp_nlp	0.3009	0.3009	0.3009
A.6 (ZS) - FoodOn	RWTH-DBIS	**0.8068**	**0.8068**	**0.8068**
	Silp_nlp	0.7278	0.7278	0.7278

RWTH-DBIS [12]. This team participated in tasks A and B (12 subtasks in total). For both tasks, they proposed a domain-specific continual training, fine-tuning, and knowledge-enhanced prompt-tuning approach. The models are firstly enriched with conceptual information related to terms and types. This is followed by CausalLM man-

ner and task-specific fine-tuning using LLaMA-3-8B [20]. The proposed approach performs well on several subtasks, showcasing that incorporating domain-specific information and providing a list of classification types enhances inference performance. They concluded that in Task A, GPT-3.5-Turbo [21] outperformed fine-tuned open-source LLM, and incorporating domain-specific information and providing a list of types at prompt significantly enhances the performance.

DaSeLab [11]. The DaSeLab team participated in *UMLS, GeoNames, and WordNet* subtasks. This team approach is based on fine-tuning a GPT-3.5-Turbo model. The result of fine-tuning on *UMLS* and *GeoNames* domains showed that fine-tuning of such model can achieve superior performance. The DaSeLab ranked first place in *NCI* (0.8249), *GeoNames* (0.5906), and *SNOMEDCT_US* (0.8829) subtasks (scores inside practices are F1-scores).

TheGhost [14]. The TheGhost team investigated a variety of LLMs with a prompt-tuning approach. They are the first team in the challenge that explored 11 LLMs (the LLM list depicted in Table 2) for 8 subtasks of term typing tasks within a few-shot testing evaluation scenario. They showed the viability of soft prompt tuning for OL and the challenge of imbalanced class prompt tuning. Their finding supports the complexity of geographical and biological domains at the term typing task of OL.

silp_nlp [15]. The silp_nlp team participated in all three tasks with a total of 15 subtasks. They ranked in first place in several subtasks including *A.3 (FS) - UMLS - MEDCIN* (0.9382), *A.4 (FS) - GO - Cellular Component* (0.2726), *A.4 (FS) - GO - Biological Process* (0.2691), *A.4 (FS) - GO - Molecular Function* (0.2970), *B.2 (FS) - Schema.org* (0.6157), *B.3 (FS) - UMLS*, *B.5 (FS) - DBO* (0.2109), and *C.1 (FS) - UMLS* (0.0783). They employed several machine learning techniques, such as Random Forest, Logistic Regression, and XGBoost, alongside advanced generative models like LLaMA-3-8B, Mixtral [22], and GPT-4o [3]. The results revealed that prompt-based methods were effective in some domains but not universally applicable. Notably, Random Forest models excelled in subtasks A.1 through A.4, while GPT-4o dominated the zero-shot tasks A.5 and A.6, as well as relation extraction tasks B and C. This team obtained in first-place in six subtasks and second place in five subtasks.

TSOTSALearning [17]. The TSOTSALearning team focused on LLMs such as BERT [23] and GPT-4. Through experimentation on *SubTask A.1 - WordNet* dataset, they achieved an F1-score of 0.9264 with GPT-4, but significantly improved results when they combined BERT with rule-based strategies, leading to an F1-score of 0.9938 and ranked first place in *WordNet* dataset. Their findings showed the importance of incorporating rules into LLMs for enhanced accuracy in OL. However, they highlight the challenge of identifying appropriate rules, suggesting that future work should focus on automating rule detection and integrating them seamlessly into LLMs. The *WordNet* dataset is being considered as a low number of types and having a higher number of types makes it challenging to obtain highly accurate rules.

SKH-NLP [13]. Team SKH-NLP participated in *SubTask B.1 - GeoNames*, where they developed a fine-tuning approach using the LLaMA-3-70B and BERT-Large [24]. This team obtained the first place in *SubTask B.1 - GeoNames* with an F1-score of 0.6557. Their comprehensive analysis demonstrates that BERT-Large, when fine-tuned, achieves performance close to the larger LLaMA-3-70B model.

Phoenixes [16]. The Phoenixes team explored the application of a Retrieval Augmented Generation (RAG) approach within the 12 subtaks of the challenge. They introduced a promising RAG-specific formulation over all three tasks of OL, where a

Table 4. *Task B - Taxonomy Discovery Results for SubTasks*

SubTask	Team Name	F1-Score	Precision	Recall
B.1 (FS) - GeoNames	SKH-NLP	**0.6557**	**0.6318**	**0.6814**
	RWTH-DBIS	0.3409	0.2400	0.5882
	Silp_nlp	0.0830	0.0446	0.5931
	TSOTSALearning	0.0104	0.0052	0.5294
	Phoenixes	0.0036	0.0019	0.0294
B.2 (FS) - Schema.org	Silp_nlp	**0.6157**	0.4578	**0.9396**
	RWTH-DBIS	0.5733	**0.5475**	0.6016
	Phoenixes	0.0155	0.0079	0.3901
B.3 (FS) - UMLS	Silp_nlp	**0.3544**	**0.4118**	0.3111
	Phoenixes	0.0960	0.0550	0.3778
	RWTH-DBIS	0.0491	0.0257	**0.5556**
B.4 (FS) - Gene Ontology (GO)	Phoenixes	0.0164	0.0180	0.0149
B.5 (FS) - DBpedia Ontology (DPO)	Silp_nlp	**0.2109**	0.1412	0.4164
	Phoenixes	0.0164	0.0180	0.0149
B.6 (ZS) - Food Ontology (FoodOn)	Phoenixes	0.0308	0.0243	0.0420

Table 5. *Task C - Non-Taxonomic Relation Extraction Results for SubTasks*

SubTask	Team Name	F1-Score	Precision	Recall
C.1 (FS) - UMLS	Silp_nlp	0.0783	**0.0494**	**0.1888**
	Phoenixes	0.0273	0.0433	0.0199

RAG system with minor changes was developed for both tasks A and B, later can be used as a two-step approach for task C. Task C consists of the following steps: Step 1 – runs the Task B approach for finding child-parent pairs and step 2 – applying the Task A approach for assigning the relations to the pairs. They incorporated Mistral-7B [25] as LLM and Dense Passage Retrieval (DPR) [26] model as the retriever model in the RAG framework. However, their results in both zero-shot and few-shot fall shorter than the fine-tuned models and this suggests that still fine-tuning is the key to obtain a high performance within OL.

4.2 Large Language Models

The participants in the challenge utilized a diverse array of LLMs, each bringing distinct strengths to the tasks. We detailed a breakdown of the key strengths of the prominent LLMs used.

GPT FAMILY – GPT-3.5-Turbo, GPT-4, and GPT-4o: GPT based LLMs, developed by OpenAI, are renowned for their advanced natural language understanding and generation capabilities. These models excel in context comprehension and can handle a variety of queries effectively, making them particularly suitable for tasks that require deep semantic understanding and detailed generation. Their ability to generalize from a wide range of training data allows them to perform well across various knowledge domains relevant ontologies [5], [27]. GPT-3.5-Turbo was a popular choice among participants, with teams such as DaSeLab, RWTH-DBIS, and silp_nlp using the model and demonstrating its high adaptability and effectiveness across the various challenge subtasks. Furthermore, GPT-4 and GPT-4o as more advanced models over GPT-3, were explored by the teams: TSOTSA Learning and *silp_nlp*.

LLAMA FAMILY – LLaMA-7B, LLaMA-2-7B, LLaMA-3-8B, and LLaMA-3-70B: The LLaMA models were another prominent choice among participants. With models like LLaMA-2 and LLaMA-3 featured by TheGhost, RWTH-DBIS, SKH-NLP, and silp_nlp, their popularity stems from their open-source, efficiency, and scalability. These models' strengths in handling large-scale data and intricate details made them well-suited for comprehensive multi-dimensional data interpretation.

BLOOM FAMILY – BLOOM-1B7, BLOOM-3B, and BLOOM-7B1: BLOOM [28] models, featured in our original research work [5], gained traction due to their open-access nature and collaborative development. TheGhost, in particular, utilized a range of BLOOM models for their flexibility and multilingual capabilities.

BIOMEDICAL FAMILY – BioMistral-7B and OpenBioLLM-8B: BioMistral-7B [29], as a domain-specific fine-tuned variant of Mistral-7B, and OpenBioLLM-8B [30], as a domain-specific fine-tuned variant of LLaMA-3-8B, were utilized for their domain-specific strengths in biomedical contexts. TheGhost's use of these models highlights their importance in tasks requiring detailed biomedical terminology and concepts, emphasizing their significance in the specialized subfields of the challenge.

MISTRAL FAMILY – Mistral-7B and Mixtral-8x7B: Mistral-7B, part of the Mistral family of models, was noted for its performance in the challenge by teams like Phoenixes and TheGhost. Moreover, Mixtral-8x7B was utilized by the team silp_nlp.

OTHERS – Flan-T5, GTE-Large, Sentence-BERT, and DPR: Flan-T5 and GTE-Large were chosen for their adaptability and fine-tuning capabilities. DSTI recognized their potential in fine-tuning and handling diverse NLP tasks efficiently when there are limited computational resources. Sentence-BERT was prominently used for tasks involving semantic similarity and sentence-level embeddings. Its popularity among participants like SKH-NLP and Phoenixes. Phoenixes used DPR for the retrieval model of the RAG approach.

4.3 Trade-offs Between Precision and Recall

Across the tasks, a clear trend emerges among the participating teams. Teams like silp_nlp often exhibit high precision but lower recall, particularly in subtasks related to GO and UMLS ontologies. This suggests that while silp_nlp is adept at avoiding false positives and making accurate predictions, it frequently misses relevant instances, indicating a more conservative approach. However, teams such as RWTH-DBIS and Phoenixes display a different trend, where recall is relatively higher than precision. These teams retrieve a larger number of relevant results but at the cost of precision, indicating that they tend to capture a broad set of possible answers, including many false positives. Their approach may be useful in tasks where coverage is prioritized over accuracy, but it also introduces challenges in filtering out noise.

Teams that manage to balance both precision and recall, such as DaSeLab and SKH-NLP, stand out for their well-rounded performance. These teams perform consistently across different subtasks by finding a middle ground between retrieving enough relevant results and minimizing false positives. DaSeLab, for example, shows balanced performance across multiple subtasks, especially in UMLS-related tasks, suggesting a more effective strategy. Meanwhile, SKH-NLP stands out in the GeoNames taxonomy discovery task, where it achieves high precision and recall, demonstrating its capability to capture relevant information without sacrificing accuracy.

In more challenging tasks, such as non-taxonomic relation extraction, the disparity between precision and recall becomes particularly pronounced. For example, both silp_nlp and Phoenixes struggle, with silp_nlp showing low precision but managing to retrieve more relevant results than Phoenixes, which has very low recall. This suggests that these tasks may require more sophisticated models or techniques to achieve higher performance. Overall, the results reflect that teams vary significantly in how they prioritize precision and recall, depending on the specific subtask, with some teams excelling in precision-oriented tasks while others show better results in recall-sensitive subtasks.

5 Discussion

Performance Analysis. As the participating teams navigated through the zero-shot and few-shot testing phases of the LLMs4OL 2024 challenge, notable variations in performance underscored the importance of model adaptability and data-specific adjustments. Few-shot tasks, particularly those involving geographical, biological, and biomedical domains, highlighted the critical need for specialized model tuning and the strategic use of training data to achieve high precision and recall rates. This indicates that achieving optimal performance in real-world ontology challenges requires not only selecting the right LLMs but also fine-tuning them to align with the specific characteristics of the domains and tasks at hand. Additionally, studies show that for Task A, even smaller models like Flan-T5-Small with 80M parameters can perform well when there are fewer types. However, as the number of types increases, larger models, such as those with 7B parameters, tend to perform better. One reason for the popularity of 7B models is that Parameter-Efficient Fine-Tuning (PEFT) [31] fine-tuning requires less memory compared to traditional fine-tuning methods. Many participants also incorporated external knowledge, such as type definitions, synthesis data using LLMs, or general knowledge graphs (KGs) to build answer sets. These strategies have demonstrated a positive impact on fine-tuning performance.

Complexity Across Domains and Tasks. The results indicated that certain domains and tasks, such as biomedical term typing and non-taxonomic relation extraction, were more challenging than others. The variation in performance across tasks, particularly in relation to term complexity (e.g., Gene Ontology), highlights the complexity of certain knowledge domains. This still requires specialized approaches. The Phoenixes (on all three tasks) and DSTI (on task A only) teams introduced a formulation based on Retrieval-Augmented Generation (RAG) approaches with success, indicating that combining LLM generation capabilities with retrieval mechanisms can enhance accuracy in OL tasks. This approach is particularly suitable due to the hybrid framework with high adaptability to be extended with different components.

Few-Shot and Zero-Shot Testing Phases. While many models performed well in the few-shot phase, the zero-shot testing phase exposed limitations in the generalization capabilities of LLMs. Models like GPT-3.5 and GPT-4 demonstrated strong performance, but there were notable drops when transitioning from few-shot to zero-shot testing phases. More research is needed to improve the transferability and robustness of LLMs across unseen domains and ontologies.

Task A vs Task C. From a task perspective, Task C attracted only two teams, indicating it was perceived as highly challenging. Non-taxonomic relation extraction requires identifying complex relationships between terms that go beyond hierarchical (taxonomy-based) relations, which is a significantly more intricate task. Unlike sim-

ple is-a relationships, non-taxonomic relations are more diverse, context-dependent, and require a deeper understanding of the subject matter. Extracting these relations often involves dealing with ambiguous or implicit connections, requiring models to infer meanings that might not be explicit. This complexity might have discouraged more teams from participating, as success in this task requires advanced techniques, often combining deep semantic understanding with domain-specific knowledge. On the other hand, Task A, term typing, had much higher participation compared to Task C. This task involves classifying terms into predefined categories, a more familiar task for many researchers. Term typing is conceptually simpler because it involves assigning a label to a term, which is something that even general-purpose LLMs can do relatively well. There is a clear, finite set of categories or types, and many participants experimented with text classification approaches.

6 Conclusion

The 1st Large Language Models for Ontology Learning Challenge at ISWC 2024 has revealed the emerging potential of LLMs beyond previous studies of OL tasks. The diverse range of participant systems, including fine-tuning, prompt-tuning, and retrieval-augmented generation approaches, demonstrated how adaptable LLMs can be when handling complex ontological data across various domains. The integration of diverse LLMs like GPT-4o, GPT-3.5, LLaMA-3, and Mistral underscored the versatility of LLMs.

Through this challenge, key insights were garnered regarding the strengths and limitations of current LLMs for OL. Notably, while LLMs have shown a remarkable capacity to generalize across unseen tasks (as evidenced by their performance in few-shot and zero-shot scenarios), certain domains such as biomedical and geographical ontologies posed unique challenges, particularly in terms of class imbalance and complex taxonomies. These challenges opened pathways for future research, emphasizing the need for scalable LLM training and the refinement of prompt-based methods to handle highly specialized ontologies.

Moreover, the variety of approaches suggests that hybrid methods combining LLMs with domain-specific knowledge are particularly effective. Moving forward, research should focus on improving the interpretability and scalability of LLM-based OL systems to enable even more accurate and dynamic knowledge extraction. This challenge has laid the groundwork for expanding LLM capabilities in the context of the Semantic Web, fostering innovation and collaboration in building the next generation of intelligent web technologies.

Data Availability Statement

The datasets supporting the findings of this article are publicly available and can be accessed via Zenodo at https://doi.org/10.5281/zenodo.13851373, or through the GitHub repository: https://github.com/HamedBabaei/LLMs4OL-Challenge-ISWC2024.

Authors Contributions

Hamed Babaei Giglou: Conceptualization, Methodology, Software, Validation, Investigation, Resources, Data Curation, Writing - Original Draft, Writing – Review & Editing, Visualization.
Jennifer D'Souza: Conceptualization, Methodology, Investigation, Resources, Super-

vision, Project administration, Funding acquisition, Writing – Review & Editing, Visualization.
Sören Auer: Conceptualization, Methodology, Review & Editing, Supervision, Project administration, Funding acquisition.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

The 1st LLMs4OL Challenge @ ISWC 2024 jointly supported by the NFDI4DataScience initiative (DFG, German Research Foundation, Grant ID: 460234259) and the SCINEXT project (BMBF, German Federal Ministry of Education and Research, Grant ID: 01IS22070).

References

[1] A. Konys, "Knowledge repository of ontology learning tools from text," *Procedia Computer Science*, vol. 159, pp. 1614–1628, 2019.

[2] T. B. Brown, B. Mann, N. Ryder, *et al.*, *Language models are few-shot learners*, 2020. arXiv: 2005.14165 [cs.CL].

[3] OpenAI, J. Achiam, S. Adler, S. Agarwal, and et al., *Gpt-4 technical report*, 2024. arXiv: 2303.08774 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2303.08774.

[4] T. R. Gruber, "Toward principles for the design of ontologies used for knowledge sharing?" *International journal of human-computer studies*, vol. 43, no. 5-6, pp. 907–928, 1995.

[5] H. Babaei Giglou, J. D'Souza, and S. Auer, "Llms4ol: Large language models for ontology learning," in *The Semantic Web – ISWC 2023*, T. R. Payne, V. Presutti, G. Qi, *et al.*, Eds., Cham: Springer Nature Switzerland, 2023, pp. 408–427, ISBN: 978-3-031-47240-4.

[6] A. Maedche and S. Staab, "Ontology learning for the semantic web," *IEEE Intelligent systems*, vol. 16, no. 2, pp. 72–79, 2001.

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

[8] H. Babaei Giglou, J. D'Souza, S. Sadruddin, and S. Auer, "Llms4ol 2024 datasets: Toward ontology learning with large language models," *Open Conference Proceedings*, vol. 4, Oct. 2024.

[9] A. Pavao, I. Guyon, A.-C. Letournel, *et al.*, "Codalab competitions: An open source platform to organize scientific challenges," *Journal of Machine Learning Research*, vol. 24, no. 198, pp. 1–6, 2023. [Online]. Available: http://jmlr.org/papers/v24/21-1436.html.

[10] H. Abi Akl, "Dsti at llms4ol 2024 task a: Intrinsic versus extrinsic knowledge for type classification, Applications on wordnet and geonames datasets," *Open Conference Proceedings*, vol. 4, Oct. 2024.

[11] A. Barua, S. Saki Norouzi, and P. Hitzler, "Daselab at llms4ol 2024 task a: Towards term typing in ontology learning," *Open Conference Proceedings*, vol. 4, Oct. 2024.

[12] Y. Peng, Y. Mou, B. Zhu, S. Sowe, and S. Decker, "Rwth-dbis at llms4ol 2024 tasks a and b, Knowledge-enhanced domain-specific continual learning and prompt-tuning of large language models for ontology learning," *Open Conference Proceedings*, vol. 4, Oct. 2024.

[13] S. Hashemi, M. Karimi Manesh, and M. Shamsfard, "Skh-nlp at llms4ol 2024 task b: Taxonomy discovery in ontologies using bert and llama 3," *Open Conference Proceedings*, vol. 4, Oct. 2024.

[14] T. Phuttaamart, N. Kertkeidkachorn, and A. Trongratsameethong, "The ghost at llms4ol 2024 task a: Prompt-tuning-based large language models for term typing," *Open Conference Proceedings*, vol. 4, Oct. 2024.

[15] P. Kumar Goyal, S. Singh, and U. Shanker Tiwari, "Silp_nlp at llms4ol 2024 tasks a, b, and c: Ontology learning through prompts with llms," *Open Conference Proceedings*, vol. 4, Oct. 2024.

[16] M. Sanaei, F. Azizi, and H. Babaei Giglou, "Phoenixes at llms4ol 2024 tasks a, b, and c: Retrieval augmented generation for ontology learning," *Open Conference Proceedings*, vol. 4, Oct. 2024.

[17] C. Ymele and A. Jiomekong, "Combining rules to large language model for ontology learning," *Open Conference Proceedings*, vol. 4, Oct. 2024.

[18] H. W. Chung, L. Hou, S. Longpre, *et al.*, *Scaling instruction-finetuned language models*, 2022. arXiv: 2210.11416 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2210.11416.

[19] Z. Li, X. Zhang, Y. Zhang, D. Long, P. Xie, and M. Zhang, *Towards general text embeddings with multi-stage contrastive learning*, 2023. arXiv: 2308.03281 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2308.03281.

[20] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, and et al., *The llama 3 herd of models*, 2024. arXiv: 2407.21783 [cs.AI]. [Online]. Available: https://arxiv.org/abs/2407.21783.

[21] OpenAI, *Openai gpt-3.5 api [gpt-3.5-turbo]*, https://platform.openai.com/docs/models/gpt-3-5, Available at: https://platform.openai.com/docs/models/gpt-3-5, 2024.

[22] A. Q. Jiang, A. Sablayrolles, A. Roux, *et al.*, *Mixtral of experts*, 2024. arXiv: 2401.04088 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2401.04088.

[23] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds., Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. [Online]. Available: https://aclanthology.org/N19-1423.

[24] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, *Bert: Pre-training of deep bidirectional transformers for language understanding*, 2019. arXiv: 1810.04805 [cs.CL]. [Online]. Available: https://arxiv.org/abs/1810.04805.

[25] A. Q. Jiang, A. Sablayrolles, A. Mensch, *et al.*, *Mistral 7b*, 2023. arXiv: 2310.06825 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2310.06825.

[26] V. Karpukhin, B. Oğuz, S. Min, *et al.*, *Dense passage retrieval for open-domain question answering*, 2020. arXiv: 2004.04906 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2004.04906.

[27] H. B. Giglou, J. D'Souza, F. Engel, and S. Auer, *Llms4om: Matching ontologies with large language models*, 2024. arXiv: 2404.10317 [cs.AI]. [Online]. Available: https://arxiv.org/abs/2404.10317.

[28] B. Workshop, : T. L. Scao, A. Fan, and et al., *Bloom: A 176b-parameter open-access multilingual language model*, 2023. arXiv: 2211.05100 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2211.05100.

[29] Y. Labrak, A. Bazoge, E. Morin, P.-A. Gourraud, M. Rouvier, and R. Dufour, *Biomistral: A collection of open-source pretrained large language models for medical domains*, 2024. arXiv: 2402.10373 [cs.CL].

[30] M. S. Ankit Pal, *Openbiollms: Advancing open-source large language models for healthcare and life sciences*, https://huggingface.co/aaditya/OpenBioLLM-Llama3-70B, 2024.

[31] L. Xu, H. Xie, S.-Z. J. Qin, X. Tao, and F. L. Wang, *Parameter-efficient fine-tuning methods for pretrained language models: A critical review and assessment*, 2023. arXiv: 2312.12148 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2312.12148.