





LLMs4OL 2025 Overview: The 2nd Large Language Models for Ontology Learning Challenge

Hamed Babaei Giglou^{1,*} , Jennifer D'Souza^{1,*} , Nandana Mihindukulasooriya³ , and
Sören Auer^{1,2} 

¹TIB Leibniz Information Centre for Science and Technology, Hannover, Germany

²L3S Research Center, Leibniz University of Hannover, Germany

³IBM Research, New York, USA

*Correspondence: {hamed.babaei, jennifer.dsouza}@tib.eu

Abstract. We present the results of the 2nd LLMs4OL 2025 Challenge, a shared task designed to evaluate the effectiveness of large language models (LLMs) for ontology learning. The challenge attracted a diverse set of participants who leveraged a broad spectrum of models, including general-purpose LLMs, domain-specific models, and embedding-based systems. Submissions covered multiple subtasks such as Text2Onto, term typing, taxonomy discovery, and non-taxonomic relationship extractions. The results highlight that hybrid pipelines integrating commercial LLMs with domain-tuned embeddings and fine-tuning approaches achieved the strongest overall performance, while specialized domain models improved results in biomedical and technical datasets. Key insights include the importance of prompt engineering, retrieval-augmented generation (RAG), and ensemble learning. This paper presents the second benchmark of LLM-driven ontology learning, serving as an overview of the participants' contributions to the challenge. Building on this, this overview presents findings, highlights emerging strategies, and offers practical insights for researchers and practitioners seeking to align unstructured language with structured knowledge.

Keywords: Ontology Learning, LLMs4OL Approach, Text2Onto, Generative AI, Large Language Models

1. Introduction

For decades, researchers have studied the complexities of human language—across voice, video, and text—driving advances in natural language processing (NLP) and beyond. These efforts have increasingly expanded into semantic web technologies, where structuring knowledge into formal, machine-readable representations is essential. Beneath unstructured language lies a wealth of latent, meaningful knowledge—often inaccessible without formal structure. Ontologies emerged as a framework to capture this knowledge, providing digital anchors for organizing information. Yet, manual creation and curation proved slow, costly, and ill-suited to the vastness, variability, and subtlety of human language. This bottleneck in traditional ontology engineering persisted despite

the semantic web community's foundational contributions, which—through rigorous methods and reuse—produced a rich ecosystem of resources. Recent advances in large language models (LLMs) bring new momentum to the vision of machine-readable knowledge, leveraging their unprecedented ability to process and generate language.

Language, once static in machines, is now dynamic—capable of understanding, generating, summarizing, and reasoning. The long-standing vision of machines that could truly comprehend and organize knowledge has become a reality. Yet, this new found power raises a critical challenge: *how do we harness the generative capabilities of LLMs without sacrificing the precision, consistency, and logical rigor that ontologies demand?* This question drives the LLMs4OL (Large Language Models for Ontology Learning) challenge series [1], [2], [3], which asks whether LLMs can produce accurate, structured, and reusable knowledge, contribute meaningfully to ontology engineering without compromising semantic integrity, and ultimately reshape how we bridge unstructured text with structured meaning in the era of foundation models.

The LLMs4OL challenge series is based on the following five ontology primitives: 1) Lexical entries L , 2) Conceptual types T , 3) A hierarchical taxonomy H_T , 4) Non-taxonomic relations R within a heterarchy H_R , 5) Axioms A for constraints and rules. This leads to key ontology learning (OL) activities, including corpus preparations, *terminology extraction*, *term typing*, *taxonomy construction*, *relationship extraction*, and axiom discovery. Together, these six tasks constitute the LLMs4OL task framework, aligning with the previously outlined LLMs4OL conceptual model [2].

The 1st LLMs4OL challenge [2] advanced the use of LLMs in OL, showcasing their potential for automated knowledge acquisition. It featured two evaluation phases: a few-shot phase, training on subsets of ontologies before testing on related unseen data, and a zero-shot phase, introducing entirely new ontologies to assess generalizability. Three tasks were addressed: Term Typing, Taxonomy Discovery, and Non-Taxonomic Relation Extraction. Eight teams participated, applying strategies such as prompt engineering, fine-tuning, and hybrid LLM–rule-based or retrieval-augmented models. LLMs performed strongly in Term Typing and Taxonomy Discovery, leveraging their ability to infer hierarchies and generalize across domains, but hybrid approaches often surpassed pure LLMs in simpler tasks by integrating external knowledge and structured reasoning. Results were sensitive to dataset diversity, highlighting the need for robust, well-curated benchmarks. Non-Taxonomic Relation Extraction remained the most challenging, as LLMs struggled with complex, domain-specific relations requiring deep semantic understanding beyond surface cues. This indicates that while LLMs are powerful generalizers, achieving full relational comprehension in OL may demand structured learning, specialized fine-tuning, or enhanced knowledge retrieval.

To further advance strategies in OL, the **2nd LLMs4OL Challenge @ ISWC 2025** takes a deliberate step back—refocusing attention also on a foundational phase of the OL pipeline, while maintaining its core paradigm: the *Text2Onto* task. *Text2Onto* serves as a critical building block in the OL process. It involves the extraction of key terms and candidate types from unstructured text, forming the basis upon which more structured, formal ontologies can later be constructed in subsequent LLMs4OL tasks. By emphasizing this early stage, the challenge aims to strengthen the pipeline from raw language to reusable knowledge. As described in Figure 1, the 2nd LLMs4OL Challenge at ISWC 2025 comprises four core tasks: *Text2Onto*, *Term Typing*, *Taxonomy Discovery*, and *Non-Taxonomic Relation Extraction*. Collectively, these tasks cover key stages in the OL pipeline—from identifying relevant terms in text to constructing structured, semantically rich knowledge representations. This year's challenge attracted

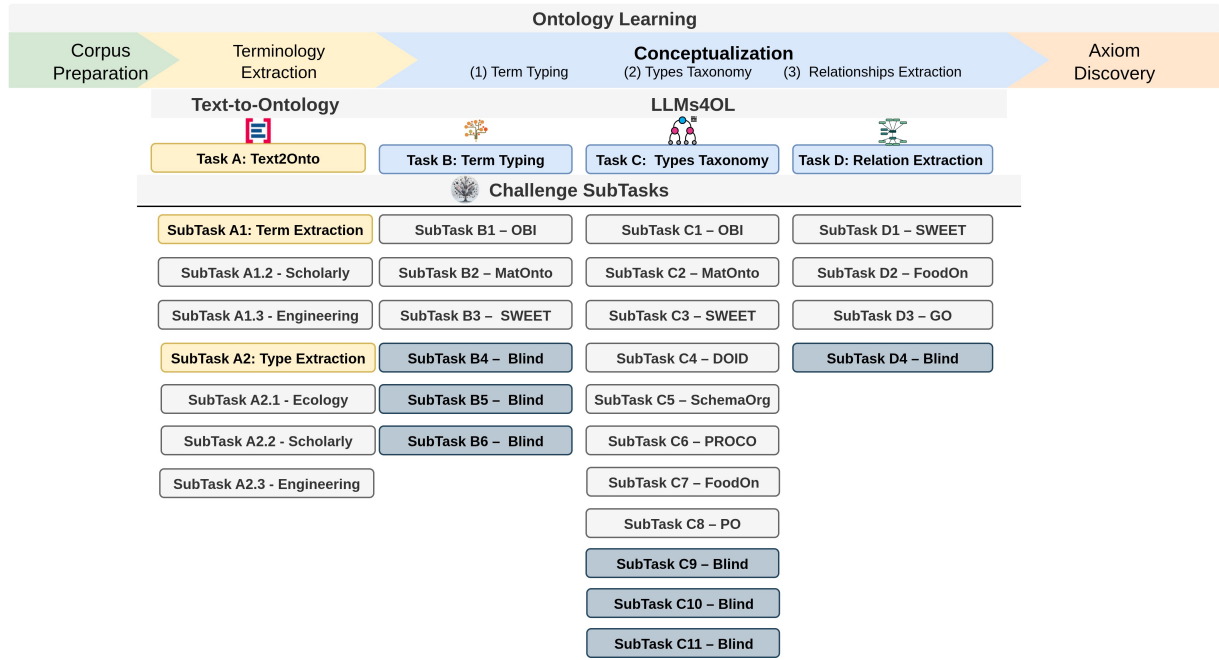


Figure 1. Overview of the LLMs4OL Challenge tasks and subtasks aligned with the Ontology Learning workflow. The process progresses from Corpus Preparation and Terminology Extraction to Conceptualization (including Term Typing, Types Taxonomy, and Relationship Extraction) and ends with Axiom Discovery. Tasks A to D correspond to major challenge categories—Text2Onto, Term Typing, Types Taxonomy, and Relation Extraction—each comprising multiple domain-specific and blind subtasks.

11 participating teams (each submitting accompanying papers) from across the globe. In total, the challenge received approximately **1,000** submissions spanning **26** individual subtasks, reflecting a strong and growing interest in the intersection of LLMs and ontology engineering.

Through this work, we aim to contribute to the ongoing discourse on the capabilities of LLMs in the context of OL, thus in the remainder of this paper, we detail the challenge tasks, what LLMs are being used, participant contributions, and findings.

2. Challenge Evaluation Overview

2.1 Tasks

The Figure 1 illustrates the tasks and subtasks targeted by the second LLMs4OL challenge at each stage of the OL process.

Task A – Text2Onto¹. Extract ontological terminologies and types from a raw text. This task focuses on extracting ontological types and terms from unstructured text. Given an unstructured text corpus/documents, the goal is to identify foundational elements for ontology construction by recognizing domain-relevant vocabulary and categorizing it appropriately. We aim to tackle two subtasks:

- *SubTask A1 – Term Extraction:* Given a set of documents from one domain, extract all relevant lexical terms L that could form the basis of an ontology.
- *SubTask A2 – Type Extraction:* Using the same set of documents, identify the conceptual types T that would serve as ontology classes.

¹<https://sites.google.com/view/llms4ol2025/task-a-text2onto>

By identifying and extracting these elements (terms and types), the task helps bridge the gap between unstructured natural language and structured ontological knowledge steps.

Task B – Term Typing². *Discover the generalized type for a lexical term.* Once domain-relevant terms and types are extracted (as we explored in Task A - Text2Onto), the next step is to assign a generalized type T to each lexical term L . The term typing task is defined as "given a lexical term L , identify the lexical term types T ". This process involves mapping lexical items to their most appropriate semantic categories or ontological classes. For example, in the biomedical domain, the term "aspirin" should be classified under "Pharmaceutical Drug". This task is crucial for organizing extracted terms into structured ontologies and improving knowledge reuse.

Task C - Taxonomy Discovery³. *Discover the taxonomic hierarchy between type pairs.* Taxonomy discovery focuses on identifying hierarchical relationships between types, enabling the construction of taxonomic structures (i.e., is-a relationships). The task is defined as "given a list of types T , the task is to extract the hierarchical taxonomy H_T that forms an is-a relationship". For example, discovering that "Sedan" is a subclass of "Car" contributes to structuring domain knowledge in a way that supports reasoning and inferencing in ontology-driven applications.

Task D - Non-Taxonomic Relation Extraction⁴. *Identify non-taxonomic, semantic relations between types.* This task aims to extract non-hierarchical (non-taxonomic) semantic relations between concepts in an ontology. The non-taxonomic relation extraction (or Non-Taxonomic RE) task is defined as a "given a set of ontological types T and relationships R , identify (head-type, relation, tail-type) triplets that form a non-taxonomic (none other than is-a) relationship H_R ". Unlike taxonomy discovery, which deals with is-a relationships, this task focuses on other meaningful associations such as part-whole (part-of), causal (causes), functional (used-for), and associative (related-to) relationships. For example, in a medical ontology, discovering that "Aspirin" treats "Headache" adds valuable relational knowledge that enhances the utility of an ontology.

2.2 Datasets

The LLMs4OL 2025 challenge introduces a comprehensive benchmark for OL across diverse domains. The benchmark spans four main tasks—Text2Onto, Term Typing, Taxonomy Discovery, and Non-Taxonomic RE—each containing multiple subtasks constructed from real-world ontologies.

Ontologies within LLMs4OL 2025. The ontologies used in LLMs4OL 2025 span various domains, including biomedicine (OBI, DOID, GO), materials science (MatOnto), environmental science (SWEET), chemistry (PROCO), agriculture (FoodON, PO), general web knowledge (Schema.org), and scholarly data (LexInfo, ENVO, OM). Each ontology provides domain-specific vocabulary and axioms for training and evaluating the models across tasks such as term identification, typing, taxonomy building, and relation extraction. Table 1 provides a detailed list of the ontologies, their respective domains, and descriptions.

Evaluation Phases. All subtasks datasets are divided into two phases: *Seen-Eval*, where participants are given training data along with test sets from the same ontology.

²<https://sites.google.com/view/llms4ol2025/task-b-term-typing>

³<https://sites.google.com/view/llms4ol2025/task-c-taxonomy-discovery>

⁴<https://sites.google.com/view/llms4ol2025/task-d-non-taxonomic-re>

Table 1. List of ontologies in LLMs4OL 2025 challenge with their respective domains and descriptions.

| Ontology | Domain | Description |
|--|----------------------------------|---|
| Ontology for Biomedical Investigations (OBI) [4] | Medicine | The OBI is a comprehensive, community-driven ontology that provides a structured framework for representing all aspects of biomedical and clinical investigations. It facilitates consistent annotation and integration of experimental data across diverse biomedical disciplines. |
| Material Ontology (MatOnto) [5] | Material Science and Engineering | The MatOnto is a domain-specific ontology designed to represent knowledge about materials, their properties, structures, and processing methods, primarily for use in materials science and engineering applications. |
| Semantic Web for Earth and Environment Technology Ontology (SWEET) [6] | Environment | The SWEET is an investigation in improving the discovery and use of Earth science data, through software understanding of the semantics of web resources. SWEET is a collection of ontologies conceptualizing a knowledge space for Earth system science and includes both orthogonal concepts (space, time, Earth realms, physical quantities, etc.) and integrative science knowledge concepts (phenomena, events, etc.). |
| Human Disease Ontology (DOID) [7] | Medicine | The Disease Ontology has been developed as a standardized ontology for human disease with the purpose of providing the biomedical community with consistent, reusable, and sustainable descriptions of human disease terms, phenotype characteristics, and related medical vocabulary disease concepts. |
| Schema.org Ontology (SchemaOrg) [8], [9] | General Knowledge | Schema.org is a collaborative, community activity with a mission to create, maintain, and promote schemas for structured data on the Internet, on web pages, in email messages, and beyond. |
| PROcess Chemistry Ontology (PROCO) [10] | Chemistry | PROCO is a formal ontology that aims to standardly represent entities and relations among entities in the domain of process chemistry. |
| Food Ontology (FoodON) [11] | Agricultural | FoodOn, the food ontology, contains vocabulary for naming food materials and their anatomical and taxonomic origins, from raw harvested food to processed food products, for humans and domesticated animals. It provides a neutral and ontology-driven standard for government agencies, industry, nonprofits, and consumers to name and reference food products and their components throughout the food supply chain. |
| Plant Ontology (PO) [12] | Agricultural | The Plant Ontology (PO) is a structured vocabulary and database resource that links plant anatomy, morphology, and growth and development to plant genomics data. |
| Gene Ontology (GO) [13] | Biology and Life Sciences | The Gene Ontology (GO) provides structured controlled vocabularies for the annotation of gene products with respect to their molecular function, cellular component, and biological role. |

Blind-Eval, where systems are being evaluated on a hidden test set from an unseen ontology with no training data provided. This tests generalization across ontological domains and structures.

Text2Onto. First, we retrieve ontology elements such as terms, types, and their associated taxonomic and non-taxonomic relations. These elements are then partitioned into subsets using the Capacitated Minimum Spanning Tree Problem algorithm, ensuring that nodes within each subset remain connected through one-hop taxonomic relations, resulting in more semantically homogeneous text. Once partitioned, synthetic text is produced through a two-step process: ontology axioms are first verbalized using templates aligned with axiom structures, and then paraphrased by LLM into natural text while preserving semantic accuracy. The resulting text documents are then used for term and type extraction subtasks. The statistics are presented in Table 2.

LLMs4OL Tasks Paradigm. The [3] describes the dataset construction procedure for LLMs4OL tasks paradigm [1]. The constructed dataset statistics for tasks B, C, and D are represented in Table 2. The Task B spans six subtasks—three in the Seen-Eval phase (B1 to B3) and three in the Blind-Eval phase (B4 to B6). Similarly, Task C includes 11 subtasks—eight seen and three blind. Finally, Task D includes 4 subtasks—three seen and one blind.

Table 2. LLMs4OL 2025 challenge, subtasks, domains, participants, and evaluation phases stats. The "PT" refers to the number of participants per subtask. In the dataset columns, for Task A, L refers to the number of lexical terms, T refers to the number of types (similarly for Task B and C), and for Task D, R refers to the number of non-taxonomic relations.

| Task | SubTask | Domain | Train | Test | PT | Phase |
|------------------------|------------------------|---------------------------------|---------------------------|--------------------------|----|-------|
| (A) Text2Onto | A1.2 - Term Extraction | Scholarly | 40 ($L=246$) | 10 ($L=32$) | 9 | |
| | A1.3 - Term Extraction | Engineering | 83 ($L=1,143$) | 21 ($L=547$) | 9 | |
| | A2.1 - Type Extraction | Ecology | 2,000 ($T=5,734$) | 482 ($T=994$) | 7 | Seen |
| | A2.2 - Type Extraction | Scholarly | 40 ($T=260$) | 10 ($T=30$) | 8 | |
| | A2.3 - Type Extraction | Engineering | 83 ($T=772$) | 21 ($T=36$) | 8 | |
| (B) Term Typing | B1 - OBI | Medicine | 201 ($T=46$) | 87 ($T=36$) | 6 | |
| | B2 - MatOnto | Materials Science & Engineering | 85 ($T=49$) | 36 ($T=30$) | 8 | Seen |
| | B3 - SWEET | Ecology and Environment | 1,707 ($T=177$) | 732 ($T=135$) | 6 | |
| | B4 - Blind | Ecology | - | 46 | 2 | |
| | B5 - Blind | Scholarly | - | 288 | 2 | Blind |
| | B6 - Blind | Engineering | - | 1,953 | 1 | |
| (C) Taxonomy Discovery | C1 - OBI | Medicine | 8,249 ($T=4,237$) | 3,536 ($T=2,821$) | 5 | |
| | C2 - MatOnto | Materials Science & Engineering | 840 ($T=653$) | 361 ($T=370$) | 7 | |
| | C3 - SWEET | Ecology and Environment | 11,137 ($T=7,542$) | 4,774 ($T=4,118$) | 3 | |
| | C4 - DOID | Medicine | 28,924 ($T=10,254$) | 12,396 ($T=7,411$) | 2 | Seen |
| | C5 - SchemaOrg | General Knowledge | 723 ($T=692$) | 311 ($T=359$) | 5 | |
| | C6 - PROCO | Chemistry | 1,313 ($T=790$) | 563 ($T=530$) | 3 | |
| | C7 - FoodOn | Agriculture | 53,020 ($T=31,076$) | 22,723 ($T=20,148$) | 3 | |
| | C8 - PO | Agriculture | 2,005 ($T=1,444$) | 860 ($T=916$) | 4 | |
| | C9 - Blind | Ecology | - | 16,273 | 1 | |
| | C10 - Blind | Scholarly | - | 276 | 1 | Blind |
| | C11 - Blind | Engineering | - | 1,124 | 3 | |
| (D) Non-Taxonomic RE | D1 - SWEET | Ecology and Environment | 360 ($T=662, R=3$) | 155 ($T=289, R=2$) | 3 | |
| | D2 - FoodOn | Agriculture | 1,450 ($T=2,838, R=6$) | 622 ($T=1,233, R=4$) | 2 | Seen |
| | D3 - GO | Biology and Life Sciences | 11,606 ($T=9,622, R=6$) | 4,975 ($T=5,189, R=6$) | 0 | |
| | D4 - Blind | Ecology | - | 147 | 1 | Blind |

2.3 Evaluation Metrics

Each task in the challenge is evaluated using precision (P), recall (R), and F1-score, defined as follows:

$$P = \frac{|\text{Correct}|}{|\text{Predicted}|}, \quad R = \frac{|\text{Correct}|}{|\text{Ground Truth}|}, \quad F1 = \frac{2 \times P \times R}{P + R}$$

The **Task A** evaluates string label prediction using Jaccard similarity with a threshold of 0.8. A prediction is considered correct if it matches a ground truth label above the threshold, with each label matched at most once. The **Task B** involves multi-label classification where each instance can have multiple types. Predicted and true types are treated as sets, and correctness is determined by the intersection across all instances. Moreover, **Task C** focuses on hierarchical relation extraction, where predictions are parent-child pairs. A match is correct if the predicted pair appears exactly in the ground truth set. Finally, **Task D** evaluates triple-based relation extraction, including symmetric relations (e.g., `equivalentClass`, `sameAs`, `disjointWith`). Triples are normalized, and symmetric counterparts are added to both prediction and ground truth before computing the scores.

3. Participant Systems and Results

For the challenge evaluation platform, we used the CodaLab submission platform [14] (the challenge can be accessed via <https://codalab.lisn.upsaclay.fr/competitions/23065>), which provides a standardized environment for managing challenge submissions and leaderboard tracking. Throughout the duration of the competition, a total of **1,038** submissions were received from **35** participants, who engaged across 26 diverse subtasks. The Figure 2 illustrates the high level of engagement throughout the competition, emphasizing the strong interest and widespread participation it attracted. In

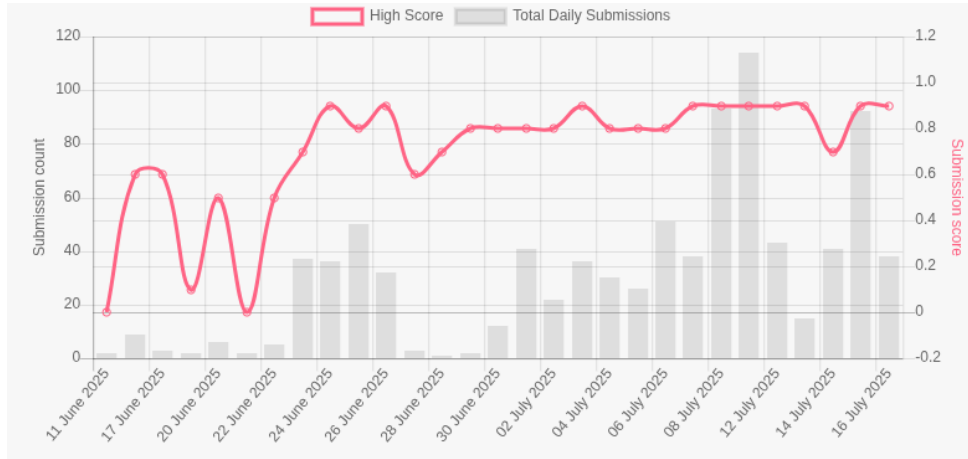


Figure 2. Submissions Statistics

Table 3. Teams performance across various subtasks using the F1-Score metric.

| | SBU-NLP | Alexbek | Siip.nlp | LABKAG | IRIS | ELLMO | DREAM-LLMs | Phoenixes | T-GreC | DaseLab | CUET Zenith | SEMA |
|---------|---------------|---------------|---------------|---------------|---------------|---------------|------------|-----------|--------|---------|-------------|--------|
| A1.2 | 0.5870 | 0.6471 | 0.4578 | 0.7000 | - | 0.3652 | - | 0.3951 | - | - | - | - |
| A1.3 | 0.6196 | 0.4418 | 0.4302 | 0.6661 | - | 0.6073 | - | 0.2556 | - | - | - | - |
| A2.1 | 0.6602 | 0.5895 | 0.5535 | 0.5595 | - | - | - | 0.4309 | - | - | - | - |
| A2.2 | 0.6585 | 0.7586 | 0.2500 | 0.8308 | - | 0.5524 | - | 0.3913 | - | - | - | - |
| A2.3 | 0.6585 | 0.4688 | 0.2545 | 0.4694 | - | 0.6750 | - | 0.1846 | - | - | - | - |
| B1 | 0.9425 | 0.7709 | 0.8021 | - | 0.8387 | - | 0.9080 | - | 0.8621 | - | - | - |
| B2 | 0.5676 | 0.6053 | 0.4872 | - | 0.6667 | - | 0.5676 | - | 0.1892 | 0.3243 | - | - |
| B3 | 0.6935 | 0.6557 | 0.3297 | - | 0.6529 | - | 0.5927 | - | 0.5192 | - | - | - |
| B4 | 0.7563 | 0.6560 | - | - | - | - | - | - | - | - | - | - |
| B5 | 0.9271 | 0.4722 | - | - | - | - | - | - | - | - | - | - |
| B6 | - | 0.1715 | - | - | - | - | - | - | - | - | - | - |
| C1 | 0.3534 | 0.2943 | 0.2273 | - | 0.3972 | - | - | - | - | - | 0.1142 | - |
| C2 | 0.6621 | 0.5590 | 0.4473 | 0.4836 | 0.4472 | - | - | - | - | - | - | 0.1441 |
| C3 | 0.4997 | 0.2549 | - | - | 0.2520 | - | - | - | - | - | - | - |
| C4 | - | 0.1806 | - | 0.0016 | - | - | - | - | - | - | - | - |
| C5 | 0.6567 | 0.3296 | 0.2609 | 0.6501 | - | - | - | - | - | - | 0.0866 | - |
| C6 | 0.2146 | 0.3865 | 0.2601 | - | - | - | - | - | - | - | - | - |
| C7 | - | 0.1171 | - | 0.0215 | - | - | - | - | - | - | - | - |
| C8 | 0.2702 | 0.4817 | 0.2106 | 0.0357 | - | - | - | - | - | - | - | - |
| C9 | - | - | - | 0.0485 | - | - | - | - | - | - | - | - |
| C10 | - | - | 0.5735 | - | - | - | - | - | - | - | - | - |
| C11 | - | - | 0.4684 | - | - | - | - | - | - | - | - | - |
| D1 | - | - | 0.6263 | - | 0.5323 | 0.1448 | - | - | - | - | - | - |
| D2 | - | - | 0.0084 | - | - | 0.0007 | - | - | - | - | - | - |
| D3 | - | - | - | - | - | - | - | - | - | - | - | - |
| D4 | - | - | 0.5051 | - | - | - | - | - | - | - | - | - |
| Mean F1 | 0.3741 | 0.3400 | 0.2751 | 0.1718 | 0.1457 | 0.0902 | 0.0796 | 0.0638 | 0.0604 | 0.0125 | 0.0077 | 0.0055 |

total, we received 13 paper submissions describing participant systems. One submission was desk-rejected due to the absence of both the system paper and evaluation results. Of the remaining 12 teams, 11 team papers are accepted for the 2nd LLMs4OL challenge proceedings, with their evaluations concluded and included in the final rankings. Additionally, the DaseLab team [15] submitted their system for evaluation on SubTask B2 – MatOnto and achieved a ranked position within that challenge. However, since no accompanying system description paper was submitted, their contribution is reflected solely in the leaderboard of SubTask B2 and not considered in the overall analysis of the LLMs4OL Challenge. This underscores the importance of both technical contributions and accompanying documentation for full inclusion in challenge outcomes.

3.1 Leaderboard

The Table 3 represents the finalized leaderboard for participants based on F1 scores. It highlights that the **SBU-NLP** team achieved the highest mean F1 score (0.3741), followed by **Alexbek** (0.3400), while the rest of the teams scored lower. Some teams attempted a wide range of subtasks, while others only submitted to a limited set, which explains the

skew in the mean F1 scores. Moreover, **LABKAG** stood out several A-series subtasks (e.g., A1.2 with 0.7000, A2.1 with 0.6661, and A2.2 with 0.8308, the highest in the Text2Onto). **SBU-NLP** performed strongly in B-series subtasks, with near state-of-the-art F1 in B1 (0.9425) and B5 (0.9271). **IRIS** had competitive scores in the B-series (e.g., B1 at 0.8387) and even stood out at B2 (0.6667). **Alexbek** has participated well in Task C with nearly 4 top rankings in Task C. The D-series tasks attracted fewer submissions. While some subtasks had F1 scores above 0.9 (e.g., B1, B5), others had scores close to zero (e.g., C4, C9, D2), showing the varying difficulty levels across the benchmark.

3.2 Contributions

SBU-NLP. The SBU-NLP [16] team participated in Tasks A, B, and C, employing prompt engineering. This team study demonstrated that prompt-based strategies, utilizing LLMs to enable effective, scalable, and domain-independent automated ontology construction, successfully overcome LLM context window limitations through careful prompt engineering and sampling techniques like stratified random sampling, simple random sampling, and chunking. A significant finding was that batch-prompted LLMs frequently matched or outperformed non-batch models across various subtasks, with Claude Sonnet 4 (Batch) [17] consistently achieving top F1 scores in Task B across all domains and blind test sets, and showing considerable gains in Task C subtasks like MatOnto and Schema.org. Conversely, for the OBI subtask, the research revealed that pretrained sentence embedding models [18] (e.g., BGE-M3⁵ [19], all-mpnet-base-v2⁶, all-MiniLM-L6-v2⁷, and Stella⁸ [20]) performed comparably to a simpler token overlap baseline, suggesting that embedding-based methods may not always offer substantial advantages in highly lexically overlapping structured ontology tasks. These results highlight promising directions for optimizing resource usage in knowledge representation tasks by leveraging LLMs without computationally expensive training pipelines. The experiments utilized several state-of-the-art LLMs, including Gemini-2.5-Flash [21], Grok-3 [22], Grok-3-mini [22], DeepSeek-V3 [23], GPT-4o-mini [24], and Claude Sonnet 4 [17].

Alexbek. The Alexbek team [25] presents a unified, modular, and lightweight LLM-based framework for OL, demonstrating its success in Tasks A, B, and C without requiring large-scale finetuning. A core finding is the high effectiveness of few-shot prompting combined with retrieval-augmented generation (RAG) [26], which consistently boosted performance across tasks, particularly in Task A for joint term and type extraction, where chained term and type exemplars proved superior. For Task B, a dual strategy involving RAG for known domains and a zero-shot classifier for unseen domains was highly effective, with an ensemble of embedding models outperforming baselines. In Task C, the study successfully modeled hierarchical relationships through a simple yet effective dedicated cross-attention layer applied to type embeddings, which was trained on frozen embeddings or with lightweight LoRA finetuning [27]. The system consistently achieved top-ranking results across various subtasks and domains, highlighting its adaptability, generalizability, and robustness even in blind settings. However, limitations were noted in domains with sparse lexical cues or fine-grained semantic distinctions, leading to reduced precision. The LLMs central to this study included Qwen3-Embedding-4B [28] for generating embeddings across Tasks A, B, and C, all-mpnet-base-v2, and BGE-

⁵<https://huggingface.co/BAAI/bge-m3>

⁶<https://huggingface.co/sentence-transformers/all-mpnet-base-v2>

⁷<https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2>

⁸https://huggingface.co/NovaSearch/stella_en.1.5B_v5

Large⁹ [29] as components of the ensemble classifier for zero-shot term typing, and a lighter Qwen3-0.6B [30] encoder utilized with LoRA adapters for taxonomy discovery

silp_nlp. The *silp_nlp* team [31] presented a clustering-enhanced methodology grounded in LLMs for OL across all four tasks. A key finding is the efficacy of combining lexical and semantic clustering with adaptive prompting and domain-adapted transformer models to achieve robust and scalable OL. For Task A, both heuristic-guided direct extraction and RAG (specifically called as a RAE –Retrieval-Augmented Extraction) were employed, with RAG dynamically enriching prompts with domain-specific exemplars to enhance term-type extraction. Task B utilized a multi-stage hybrid approach combining deterministic lexical clustering with LLM-based semantic disambiguation, including an expert persona and few-shot examples in prompts. In Task C, semantic-based clustering was preferred for its ability to capture conceptual similarity, followed by LLM-based relation extraction within clusters. For Task D, a fully LLM-centric knowledge discovery method and a hybrid method combining semantic embeddings and algorithmic clustering were developed for non-taxonomic relation extraction. Despite overall success, lower F1-scores in subtasks with domain-specific jargon (e.g., Scholarly and SWEET ontologies) highlighted limitations in contextual understanding and the need for deeper domain alignment. The LLMs used in the study included proprietary LLMs from Grok, Gemini-2.5-flash (for Tasks A, C, and D’s domain inference), and Gemini-2.5-pro (for Task B and Task D’s high-precision relation extraction). Additionally, domain-specialized transformer models such as MaterialsBERT [32] and BioBERT [33] were employed for generating semantic embeddings.

LABKAG. The LABKAG [34] studied the effectiveness of prompt design as a primary strategy for structured knowledge acquisition, without relying on fine-tuning or external knowledge. The study consistently demonstrated that incorporating in-domain examples and providing richer context within prompts significantly enhances performance for challenge tasks. For Task A, in-domain few-shot prompts consistently outperformed generic one-shot prompts for entity extraction, notably improving recall, and providing full document context during entity classification led to substantial F1 score increases. An additional term expansion step was crucial for boosting recall in the Engineering subset of Task A, although it reduced precision. Conversely, the inclusion of noise in examples or input consistently degraded performance, underscoring the importance of careful prompt selection based on data characteristics. For Task C, which focused on identifying taxonomic hierarchies, the study found that a greater number of in-context examples (few-shot) generally improved performance. To manage challenges with long input lists in Task C, pre-processing strategies like length-based chunking were adopted to address token limits, and category regrouping (classifying terms into semantically coherent groups) consistently improved output quality by filtering irrelevant terms and enhancing intra-group coherence. The LLMs used in the study were Qwen3-8B for Task A, which was deployed and run locally without additional training or parameter updates, and GPT-4o-mini, along with Gemini 2.5 Pro for evaluating performance in Task C.

IRIS. The IRIS team’s [35] demonstrates that model-agnostic data manipulations significantly enhance the performance of LLMs in OL, specifically Task B, C, and D. After careful curation of inputs through input-enrichment techniques and a pruning technique yields substantial performance improvements are yielded. These techniques show synergistic benefits when applied together. For Task B, data augmentation (using rule-based generators, Wiktionary, and GPT-4o mini synonyms) combined with automatic

⁹<https://huggingface.co/BAAI/bge-large-en-v1.5>

definition mining (from Wikipedia, Wiktionary¹⁰, domain APIs, or GPT-4o fallback) was found to substantially boost performance, particularly in addressing rare-class sparsity and limited lexical variety. For Tasks C and D, similarity-based candidate filtering (using all-MiniLM-L6-v2 Sentence-BERT embeddings) was deemed indispensable, drastically improving F1-scores by pruning the search space and reducing noise. While type definitions alone offered little benefit in these latter tasks, they provided further gains when combined with filtering, acting as a fine-grained booster once the search space was denoised. These three data-layer heuristics collectively address issues of rare-class sparsity, context deficit, and quadratic explosion without altering the model architecture or hyperparameters. The primary LLM used for the core classification tasks (B, C, and D) was a fine-tuned DeBERTa-v3-large [36] encoder. Additionally, GPT-4o was employed as a fallback for automatic definition mining when other web sources failed, and the all-MiniLM-L6-v2 Sentence-BERT model was used for generating embeddings in the similarity-based candidate filtering for Subtasks C and D.

ELLMO. The ELLMO team [37] for Task A found that simpler prompts defining pattern-like rules performed better than elaborate strategies. The optimal approach (introduced LLM-centric vs. classification) was problem-specific, with the Engineering dataset benefiting from the classification approach, while the Scholarly dataset generally benefited from the LLM-centric approach. For the classification approach, including a "neither" class significantly improved F1 scores by boosting precision, despite a general decrease in recall. The choice of LLM was also critical, as performance differences were observed across datasets when comparing smaller models to large-scale LLMs, where Claude Sonnet 4 outperformed Mistral-Small-3.1¹¹ on the Engineering dataset. However, for the Scholarly dataset, this was not true as it showed that Mistral-Small-3.1 outperformed w.r.t. Claude Sonnet 4. For Task D, the primary approach involved reducing the number of potential edges using either clustering or vector databases, followed by querying an LLM for edge probabilities. It was found that clustering performed well for the SWEET dataset but yielded low recall for the GO dataset. Vector database methods resulted in lower recall and were not scalable to larger datasets compared to clustering. A crucial insight was that asking the LLM to output probabilities for edge existence significantly improved recall and F1 scores. However, providing examples in the prompt influenced the distribution of LLM-generated probabilities, sometimes leading to lower performance if examples caused the LLM to assign too many high probabilities. The LLMs utilized in this work included Mistral-Small-3.1 for Task A, Claude Sonnet 4 for comparative testing in Task A, and Llama-3.2-90B-Vision-Instruct¹² for clustering in Task D. The authors also noted a substantial performance decrease between training and testing datasets, possibly due to overfitting or inconsistencies in the test data.

DREAM-LLMs. The DREAM-LLMs team [38] introduces a deliberation-based reasoning ensemble approach with multiple LLMs for Task B in low-resource domains within subtasks B1, B2, and B3. They found that relying on a single LLM in low-resource environments is often insufficient due to domain-specific knowledge gaps and limited exposure to specialized terminology, resulting in inconsistent and biased predictions. To overcome this, DREAM-LLMs involves crafting few-shot prompts and independently querying several LLMs, with each model providing a predicted label and a brief justification. A promising contribution is the deliberation step, where one LLM reviews the predictions and explanations from the others to make a final decision, effectively mitigating individual model biases and outperforming standard prompting approaches.

¹⁰<https://www.wiktionary.org/>

¹¹<https://huggingface.co/mistralai/Mistral-Small-3.1-24B-Instruct-2503>

¹²<https://huggingface.co/meta-llama/Llama-3.2-90B-Vision-Instruct>

This collaborative reasoning was found to not only enhance predictive accuracy but also encourage weaker models to align their outputs with stronger counterparts, thereby improving decision consistency in low-resource settings. The specific LLMs utilized in this work were ChatGPT-4o, Claude Sonnet 4, DeepSeek-V3, and Gemini-2.5-Pro.

Phoenixes. The Phoenixes team [39] methodology centers on the effectiveness of Chain-of-Thought (CoT) Few-Shot Prompting strategies in Text2Onto subtasks. The study demonstrates that combining reasoning-based prompting with instruction-tuned models can effectively support Text2Onto tasks across diverse domains without task-specific fine-tuning, enabling models to perform step-by-step reasoning and contextual interpretation. Performance varied by domain, with the approach showing particular strength in domains with clearer conceptual structures like Ecology and Scholarly communication, compared to more complex fields like Engineering, where models generally struggled. Qwen2.5-72B-Instruct¹³ [40] often emerged as a strong performer, achieving the highest F1 score for term extraction in the Scholarly domain (0.3950) and outperforming other models in type extraction for the Ecology dataset (F1 score of 0.4309). LLaMA-3.3-70B-Instruct¹⁴ [41] demonstrated high recall across various subtasks and domains, indicating its capability in identifying a broader set of relevant terms and types. The LLMs evaluated were Qwen2.5-72B-Instruct, Mistral-Small-24B-Instruct-2501¹⁵ [42], and LLaMA-3.3-70B-Instruct across Ecology, Scholarly, and Engineering datasets.

T-GreC. The T-GreC team [43] investigates the effectiveness of combining embeddings with k-nearest neighbors (k-NN) for Task B. They found that embeddings derived from fine-tuned transformer models can be highly effective for k-NN classification. The RoBERTa-base [44] model achieved the highest F1 score of 0.862 using k-NN with embeddings, notably exceeding its direct fine-tuning performance of 0.827. A dramatic improvement was observed with DeBERTa-v3-base, which performed extremely poorly in direct fine-tuning (F1 score of 0.022) but achieved a high F1 score of 0.850 with k-NN using its embeddings, suggesting it generates high-quality semantic embeddings despite issues with LoRA fine-tuning. Furthermore, simple character-level data augmentation (insertion, deletion, swapping, or substitution) significantly improved performance for PubMedBERT [45] on the OBI dataset. However, a crucial limitation discovered was that these strategies, including augmentation and embedding-based k-NN, failed to generalize effectively to the MatOnto and SWEET datasets, underscoring their dataset-dependency.













CUET Zenith. The CUET Zenith team [46] proposed a hybrid methodology for taxonomy discovery, focusing on biomedical (OBI) and general-purpose (SchemaOrg) knowledge domains. A key finding is that the judicious integration of classical machine learning with LLMs yields efficient and scalable solutions for ontology structure induction. For Subtask C1, a hybrid approach combining semantic clustering of Sentence-BERT embeddings with few-shot prompting using Qwen3-14B was introduced. For Subtask C5, the team introduced a cascaded validation framework that harmonizes deep semantic representations from sentence transformer all-mpnet-base-v2 embeddings, ensemble classification via XGBoost, and a hierarchical LLM-based reasoning pipeline. This framework involved a two-tier LLM validation system for medium-confidence predictions, employing TinyLlama-1.1B for binary validation and GPT-4o for probabilistic verification, resulting in the highest F1-score of 0.0866 for this subtask. Notably, smaller LLMs like TinyLlama-1.1B, when optimally coupled with XGBoost, were found to outperform

¹³<https://huggingface.co/Qwen/Qwen2.5-72B-Instruct>

¹⁴<https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct>

¹⁵<https://huggingface.co/mistralai/Mistral-Small-24B-Instruct-2501>

Table 4. LLMs4OL 2025 challenge participants' methods.

| Rank | Team Name | LLM of Use | Approach | Code | A1.2 - Scholarly | A1.3 - Engineering | A2.1 - Ecology | A2.2 - Scholarly | A2.3 - Engineering | B1 - OBI | B2 - MaOnto | B3 - SWEET | B4 - Ecology** | B5 - Scholarly** | B6 - Engineering** | C1 - OBI | C2 - MaOnto | C3 - SWEET | C4 - DOID | C5 - SchemaORG | C6 - PROCO | C7 - FoodOn | C8 - PO | C9 - Ecology** | C10 - Scholarly** | C11 - Engineering** | D1 - SWEET | D2 - FoodOn | D3 - GO | D4 - Ecology** | |
|------|------------------|--|---|---|------------------|--------------------|----------------|------------------|--------------------|----------|-------------|------------|----------------|------------------|--------------------|----------|-------------|------------|-----------|----------------|------------|-------------|---------|----------------|-------------------|---------------------|------------|-------------|---------|----------------|--|
| 1 | SBU-NLP [16] | Gemini-2.5-Flash DeepSeek-V3-0324 Claude Sonnet 4 Grok-3-mini GPT4o-mini Grok-3 BGE-M3 all-mpnet-base-v2 all-MiniLM-L6-v2 stella_en.1.5B_v5 | Prompt Engineering Embeddings |  | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | ✓ | ✓ | | ✓ | | | | | | | | |
| 2 | Alexbek [25] | Qwen3-Embedding-4B all-mpnet-base-v2 bge-large-en-v1.5 Qwen3-0.6B | RAG Prompt Engineering Few-Shot Prompting Finetuning |  | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | | | | | | | |
| 3 | silp.nlp [31] | Grok Gemini-2.5-Pro MaterialsBERT BioBERT | Prompt Engineering RAG Clustering Emeddings |  | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | | | ✓ | ✓ | | | ✓ | ✓ | ✓ | ✓ | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | |
| 4 | LABKAG [34] | GPT-4o-mini Gemini-2.5-Pro Qwen3-8B | Prompt Engineering Few-Shot Prompting Embeddings |  | ✓ | ✓ | ✓ | ✓ | ✓ | | | | | | | | ✓ | | ✓ | ✓ | | ✓ | ✓ | ✓ | | | | | | | |
| 5 | IRIS [35] | DeBERTa-v3-large GPT-4o all-MiniLM-L6-v2 | Data Augmentation Finetuning Embeddings |  | | | | | | ✓ | ✓ | ✓ | | | | ✓ | ✓ | ✓ | | | | | | | | | ✓ | | | | |
| 6 | ELLMO [37] | Mistral-Small-3.1 BERT LLaMA-3.2-90B-Vision-Instruct Claude Sonnet 4 | Finetuning Prompt Engineering Embeddings Clustering |  | ✓ | ✓ | | ✓ | ✓ | | | | | | | | | | | | | | | | | | ✓ | ✓ | | | |
| 7 | DREAM-LLMs [38] | GPT-4o Claude Sonnet 4 DeepSeek-V3 Gemini 2.5 Pro | Prompt Engineering Ensemble Learning |  | | | | | | ✓ | ✓ | ✓ | | | | | | | | | | | | | | | | | | | |
| 8 | Phoenixes [39] | Qwen2.5-72B-Instruct Mistral-Small-24B-Instruct-2501 LLaMA-3.3-70B-Instruct | Prompt Engineering |  | ✓ | ✓ | ✓ | ✓ | ✓ | | | | | | | | | | | | | | | | | | | | | | |
| 9 | T-GreC [43] | PubMedBERT BioBERT DeBERTa-v3-base RoBERTa-base | Data Augmentation Embeddings Finetuning |  | | | | | | ✓ | ✓ | ✓ | | | | | | | | | | | | | | | | | | | |
| 10 | DaseLab [15] | GPT-3.5-Turbo all-mpnet-base-v2 TinyLlama-1.1B | Finetuning |  | | | | | | ✓ | | | | | | | | | | | | | | | | | | | | | |
| 11 | CUET_Zenith [46] | GPT-4o Qwen3-14B Mistral-7B BioBERT | RAG Ensemble Learning Prompt Engineering |  | | | | | | | | | | | | ✓ | | | | ✓ | | | | | | | | | | | |
| 12 | SEMA [47] | LLaMA 3.1-8B | Data Augmentation Finetuning |  | | | | | | | | | | | | | ✓ | | | | | | | | | | | | | | |

larger counterparts in some configurations, and strategic threshold-gated LLM validation significantly reduced inference costs while maintaining precision. The LLMs utilized in this work included Qwen3-14B, TinyLlama-1.1B, GPT-4o, Mistral-7B, and BioBERT.

SEMA. The SEMA team [47] participated in task C of the challenge, specifically focusing on the MatOnto subtask. They introduced prompt-decoupled based finetuning, which does the finetuning with one prompt format and testing it with a semantically equivalent but structurally different one. They found that prompt decoupling can improve generalization and mitigate overfitting to specific phrasings. The study also found that a 6:1 false-to-true ratio for contrastive negative sampling was optimal, providing the best balance for training signal and label imbalance, while higher ratios led to reduced performance. Furthermore, domain-specific system prompts (e.g., "expert in material science") consistently outperformed generic ones, suggesting that expert framing can guide the model towards more domain-sensitive interpretations. The LLMs utilized in this work included LLaMA-3.1-8B, which was fine-tuned using LoRA adapters.

3.3 LLM Usage

Participants leveraged a mix of large and smaller generative models, embedding models, and domain-specific models:

Larger LLMs. The challenge experienced an extensive use of general-purpose instruction-tuned and chat models, including GPT variants (GPT-3.5-Turbo, GPT-4o, GPT-4o-mini), Claude Sonnet 4, Gemini models (Gemini-2.5-Pro, Gemini-2.5-Flash), Grok models (Grok-3, Grok-3-mini), LLaMA variants (LLaMA-3.2-90B-Vision-Instruct,

LLaMA-3.3-70B-Instruct), Mistral (Mistral-Small-3.1, Mistral-Small-24B-Instruct-2501), Qwen3 variants (14B), and DeepSeek-V3-0324. These models demonstrated the highest engagement among participants for ontology learning tasks. Overall, teams that extensively used large instruction-tuned LLMs consistently achieved higher F1 scores across multiple subtasks.

Smaller LLMs. The smaller LLMs, including LLaMA variants (TinyLlama-1.1B, LLaMA 3.1–8B) and Qwen3 variants (0.6B, 8B), are used due to their resource-efficiency, which makes them suitable for low-resource finetuning, but they require careful adaptation to match the performance of larger models.

Embedding Models. Several participants leveraged embedding and vector representation models, including all-mpnet-base-v2, all-MiniLM-L6-v2, bge-large-en-v1.5, and Qwen3-Embedding-4B. These models were primarily used for semantic similarity computations, RAG, and clustering of ontology concepts, providing structured representations to support LLM-based reasoning.

Domain-specific Models. Domain-specific models such as PubMedBERT, BioBERT, MaterialsBERT, DeBERTa-v3-base, DeBERTa-v3-large, and RoBERTa-base were employed to capture specialized knowledge in biomedical, materials, and engineering domains. Their use highlights the importance of leveraging pre-trained knowledge from relevant fields to improve ontology learning performance in highly technical datasets.

4. Observations and Lessons Learned

Table 4 provides an overview of the methods employed by the LLMs4OL 2025 challenge participants. Across all submissions, LLMs were combined with a diverse set of strategies such as prompt engineering, RAG, embeddings, clustering, ensemble learning, data augmentation, and fine-tuning. We observe that:

- *Hybrid Solutions.* The top-ranked teams (*SBU-NLP*, *Alexbek*, and *silp_nlp*) relied on *multi-model pipelines* that integrated state-of-the-art commercial LLMs (e.g., Gemini-2.5, Claude Sonnet 4, GPT-4o) with open-source models (e.g., Qwen3, BioBERT, MaterialsBERT, all-mpnet-base-v2). These systems demonstrated the importance of *hybrid solutions*, where instruction-based prompting and embeddings were combined with retrieval or clustering to handle the variety of ontology tasks.
- *Domain Adaptation through Finetuning and Augmentation.* Teams in the mid-ranking group (e.g., *LABKAG*, *IRIS*, *ELLMO*, *DREAM-LLMs*) favored *finetuning and embeddings*, often targeting domain-specific subtasks. For instance, *IRIS* used DeBERTa-v3-large and MiniLM embeddings with data augmentation, while *ELLMO* combined finetuning with clustering and multimodal LLMs (e.g., LLaMA-3.2-90B-Vision-Instruct). These approaches show a tendency to balance general-purpose prompting with domain adaptation using finetuning or data augmentation.
- *Resource-efficient Models with Specialized Task Design.* Lower-ranked submissions, such as *CUET_Zenith* and *SEMA*, often emphasized *specialization*, focusing on a narrow set of tasks or using lightweight models (e.g., TinyLlama, PubMedBERT) enhanced with RAG. While less competitive overall, these efforts highlighted the potential of resource-efficient models when combined with targeted task designs.

Overall, there is a clear methodological spectrum: 1) High-performing systems integrated multiple LLMs with prompt engineering, RAG, and embeddings. 2) Middle-performing systems leaned on finetuning, ensemble strategies, and data augmentation.

3) Specialized teams showcased efficient or domain-specific models with focused methods. This diversity underlines not only the flexibility of LLMs in ontology learning tasks but also the importance of strategic method integration over reliance on a single modeling paradigm.

The contributions reveal several important lessons, summarized below:

- *Prompt engineering remains a cornerstone.* Carefully designed prompts, few-shot prompting, step-by-step reasoning via CoT prompting, and domain framing consistently boost performance (SBU-NLP, LABKAG, Alexbek, Phoenixes).
- *LLM ensembles improve low-resource domain performance.* Combining multiple models with deliberation mitigates individual biases (DREAM-LLMs).
- *RAG enhances adaptability.* Using RAG helps LLMs generalize to unseen domains and sparse lexical settings (Alexbek, silp_nlp, CUET_Zenith).
- *Domain-specific embeddings complement LLMs.* Embeddings like Sentence-BERT, BGE-M3, and BioBERT help in semantic clustering and k-NN classification (T-GreC, silp_nlp, SBU-NLP, LABKAG, IRIS, T-GreC).
- *Data augmentation and enrichment improve results.* Synonym expansion, definition mining, and input pruning help tackle rare-class sparsity (IRIS, T-GreC, SEMA).
- *Smaller LLMs can compete effectively.* With LoRA finetuning, smaller models can rival larger ones in performance and cost-efficiency (CUET_Zenith, Alexbek).
- *Clustering can aid non-taxonomic relation extraction.* Lexical or semantic clustering reduces search space and improves scalability (ELLMO, silp_nlp).
- *Context window management matters.* Long documents require chunking, stratified sampling, or batching to maintain LLM performance (SBU-NLP, LABKAG).
- *Batch prompting is efficient.* Batch-prompted LLMs can match or outperform non-batched approaches while reducing computation (SBU-NLP).
- *Ensembling reduces errors.* Cascaded LLM validation and voting schemes mitigate mistakes from individual models, especially in hierarchical or low-resource tasks (DREAM-LLMs, CUET_Zenith).

5. Conclusion

The LLMs4OL 2025 Challenge demonstrates that LLMs are already capable of contributing meaningfully to ontology learning, but no single approach suffices across all tasks. General-purpose LLMs proved valuable for broad coverage, while domain-specific models captured specialized knowledge, and embeddings supported semantic similarity computations. The most successful systems combined these components in hybrid pipelines, balancing instruction-following abilities with domain adaptation and retrieval-based strategies. Despite these advances, results also reveal limitations: performance varied across subtasks, robustness issues emerged, and resource-efficient finetuning remained a challenge. Overall, the challenge highlights the potential of LLMs to accelerate ontology engineering while underscoring the need for continued research into scalable, interpretable, and domain-sensitive approaches. Future work will extend evaluation datasets, improve reproducibility, and foster collaborations between the semantic web and NLP communities to realize the vision of autonomous ontology learning.

Data availability statement

The datasets supporting the findings of this article are publicly available and can be accessed via GitHub repository at <https://github.com/sciknoworg/LLMs4OL-Challenge/tree/main/2025>.

Author contributions

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

Jennifer D'Souza: Conceptualization, Investigation, Resources, Supervision, Project administration, Funding acquisition, Review & Editing.

Nandana Mihindukulasooriya: Conceptualization, Investigation, Resources, Review & Editing.

Soren Auer: Conceptualization, Review & Editing, Supervision, Project administration, Funding acquisition.

Competing interests

The authors declare that they have no competing interests

Funding

The 2nd LLMs4OL Challenge @ ISWC 2025 was jointly supported by the [SCINEXT project](#) (BMFTR, German Federal Ministry of Research, Technology and Space, Grant ID: 01IS22070) and the [NFDI4DataScience initiative](#) (DFG, German Research Foundation, Grant ID: 460234259).

Acknowledgements

We would like to thank Andrei Aione, the TIB Leibniz Information Centre for Science and Technology, and the ISWC community for their support and contributions to this challenge.

References

- [1] H. Babaei Giglou, J. D'Souza, and S. Auer, "Llms4ol: Large language models for ontology learning", in *International Semantic Web Conference*, Springer, 2023, pp. 408–427.
- [2] H. Babaei Giglou, J. D'Souza, and S. Auer, "Llms4ol 2024 overview: The 1st large language models for ontology learning challenge", *Open Conference Proceedings*, vol. 4, pp. 3–16, Oct. 2024. DOI: [10.52825/ocp.v4i.2473](https://doi.org/10.52825/ocp.v4i.2473). [Online]. Available: <https://www.tib-op.org/ojs/index.php/ocp/article/view/2473>.
- [3] H. B. Giglou, J. D'Souza, S. Sadruddin, and S. Auer, "Llms4ol 2024 datasets: Toward ontology learning with large language models", in *Open Conference Proceedings*, vol. 4, 2024, pp. 17–30.
- [4] A. Bandrowski et al., "The ontology for biomedical investigations", *PloS one*, vol. 11, no. 4, e0154556, 2016.
- [5] R. G. B. Miller and B. Heussler, *Matonto*, <https://github.com/inovexcorp/MatOnto-Ontologies>, Ontology repository for material science, 2014. [Online]. Available: <https://github.com/inovexcorp/MatOnto-Ontologies>.

- [6] R. G. Raskin and M. J. Pan, "Knowledge representation in the semantic web for earth and environmental terminology (sweet)", *Computers & geosciences*, vol. 31, no. 9, pp. 1119–1125, 2005.
- [7] L. M. Schriml et al., "Human disease ontology 2018 update: Classification, content and workflow expansion", *Nucleic acids research*, vol. 47, no. D1, pp. D955–D962, 2019.
- [8] P. Barker and L. M. Campbell, "What is schema. org", *LRMI*. Retrieved April, vol. 21, p. 2015, 2014.
- [9] Schema.org Community, *Schema.org - structured data on the web*, <https://schema.org>, Accessed: 2025, 2011. [Online]. Available: <https://schema.org>.
- [10] W. Schafer, O. He, A. L. Dunn, and Z. E. X. Dance, *Ontology for process chemistry – giving context to instrument data structured by the allotrope data model*, Presented at the Allotrope Connect Virtual Conference, April 19–26, 2021, Virtual meeting, 2021. [Online]. Available: <https://www.youtube.com/watch?v=HVv8TJc7p9c>.
- [11] D. M. Dooley et al., "Foodon: A harmonized food ontology to increase global food traceability, quality control and data integration", *npj Science of Food*, vol. 2, no. 1, p. 23, 2018.
- [12] P. Jaiswal et al., "Plant ontology (po): A controlled vocabulary of plant structures and growth stages", *Comparative and functional genomics*, vol. 6, no. 7-8, pp. 388–397, 2005.
- [13] G. O. Consortium, "The gene ontology (go) database and informatics resource", *Nucleic acids research*, vol. 32, no. suppl_1, pp. D258–D261, 2004.
- [14] A. Pavao 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>.
- [15] A. Barua, S. S. Norouzi, and P. Hitzler, "Daselab at llms4ol 2024 task a: Towards term typing in ontology learning", in *Open Conference Proceedings*, vol. 4, 2024, pp. 77–84.
- [16] R. Rahnamoun and M. Shamsfard, "Sbu-nlp at llms4ol 2025 tasks a, b, and c: Stage-wise ontology construction through llms without any training procedure", *Open Conference Proceedings*, 2025.
- [17] Anthropic. "Introducing claude 4". Accessed: 2025-08-08. [Online]. Available: <https://www.anthropic.com/news/claude-4>.
- [18] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks", in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Nov. 2019. [Online]. Available: <https://arxiv.org/abs/1908.10084>.
- [19] J. Chen, S. Xiao, P. Zhang, K. Luo, D. Lian, and Z. Liu, *Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation*, 2024. arXiv: [2402.03216](https://arxiv.org/abs/2402.03216) [cs.CL].
- [20] D. Zhang, J. Li, Z. Zeng, and F. Wang, *Jasper and stella: Distillation of sota embedding models*, 2025. arXiv: [2412.19048](https://arxiv.org/abs/2412.19048) [cs.IR]. [Online]. Available: <https://arxiv.org/abs/2412.19048>.
- [21] G. Comanici et al., "Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities", *arXiv preprint arXiv:2507.06261*, 2025.
- [22] xAI, *Grok*, n.d. [Online]. Available: <https://grok.com/>.
- [23] DeepSeek-AI, *Deepseek-v3 technical report*, 2024. arXiv: [2412.19437](https://arxiv.org/abs/2412.19437) [cs.CL]. [Online]. Available: <https://arxiv.org/abs/2412.19437>.
- [24] OpenAI. "Gpt-4o mini: Advancing cost-efficient intelligence". Accessed: 2025-08-08. [Online]. Available: <https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/>.
- [25] A. Beliaeva and T. Rahmatullaev, "Alexbek at llms4ol 2025 tasks a, b, and c: Heterogeneous llm methods for ontology learning (few-shot prompting, ensemble typing, and attention-based taxonomies)", *Open Conference Proceedings*, 2025.

- [26] P. Lewis et al., *Retrieval-augmented generation for knowledge-intensive nlp tasks*, 2021. arXiv: [2005.11401](https://arxiv.org/abs/2005.11401) [cs.CL]. [Online]. Available: <https://arxiv.org/abs/2005.11401>.
- [27] Z. Han, C. Gao, J. Liu, J. Zhang, and S. Q. Zhang, "Parameter-efficient fine-tuning for large models: A comprehensive survey", *arXiv preprint arXiv:2403.14608*, 2024.
- [28] Y. Zhang et al., "Qwen3 embedding: Advancing text embedding and reranking through foundation models", *arXiv preprint arXiv:2506.05176*, 2025.
- [29] S. Xiao, Z. Liu, P. Zhang, and N. Muennighoff, *C-pack: Packaged resources to advance general chinese embedding*, 2023. arXiv: [2309.07597](https://arxiv.org/abs/2309.07597) [cs.CL].
- [30] Q. Team, *Qwen3 technical report*, 2025. arXiv: [2505.09388](https://arxiv.org/abs/2505.09388) [cs.CL]. [Online]. Available: <https://arxiv.org/abs/2505.09388>.
- [31] P. Goyal, S. Singh, and U. S. Tiwary, "Silp_nlp at llms4ol 2025 tasks a, b, c, and d: Clustering-based ontology learning using llms", *Open Conference Proceedings*, 2025.
- [32] P. Shetty et al., "A general-purpose material property data extraction pipeline from large polymer corpora using natural language processing", *npj Computational Materials*, vol. 9, no. 1, p. 52, 2023.
- [33] J. Lee et al., "Biobert: A pre-trained biomedical language representation model for biomedical text mining", *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.
- [34] X. Zhao, K. Drake, C. Watanabe, Y. Sasaki, and H. Hando, "Labkag at llms4ol 2025 tasks a and c: Context-rich prompting for ontology construction", *Open Conference Proceedings*, 2025.
- [35] I.-A. Latipov, M. Holenderski, and N. Meratnia, "Iris at llms4ol 2025 tasks b, c, and d: Enhancing ontology learning through data enrichment and type filtering", *Open Conference Proceedings*, 2025.
- [36] P. He, J. Gao, and W. Chen, *Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing*, 2021. arXiv: [2111.09543](https://arxiv.org/abs/2111.09543) [cs.CL].
- [37] R. Roche, K. Gray, J. Murdock, and D. C. Crowder, "Ellmo at llms4ol 2025 tasks a and d: Llm-based term, type, and relationship extraction", *Open Conference Proceedings*, 2025.
- [38] P. Wiangnak, T. Prabhong, T. Phuttaamart, N. Kertkeidkachorn, and K. Shirai, "The dream-llms at llms4ol 2025 task b: A deliberation-based reasoning ensemble approach with multiple large language models for term typing in low-resource domains", *Open Conference Proceedings*, 2025.
- [39] A. E. Fridouni and M. Sanaei, "Phoenixes at llms4ol 2025 task a: Ontology learning with large language models reasoning", *Open Conference Proceedings*, 2025.
- [40] Q. Team, *Qwen2.5: A party of foundation models*, Sep. 2024. [Online]. Available: <https://qwenlm.github.io/blog/qwen2.5/>.
- [41] A. Dubey et al., "The llama 3 herd of models", *arXiv e-prints*, arXiv–2407, 2024.
- [42] A. Q. Jiang et al., *Mistral 7b*, 2023. arXiv: [2310.06825](https://arxiv.org/abs/2310.06825) [cs.CL]. [Online]. Available: <https://arxiv.org/abs/2310.06825>.
- [43] C. Yimmark and T. Racharak, "T-grec at llms4ol 2025 task b: A report on term-typing task of obi dataset using llm with k-nearest neighbors", *Open Conference Proceedings*, 2025.
- [44] Y. Liu et al., "Roberta: A robustly optimized bert pretraining approach", *arXiv preprint arXiv:1907.11692*, 2019.
- [45] Y. Gu et al., "Domain-specific language model pretraining for biomedical natural language processing", *ACM Transactions on Computing for Healthcare (HEALTH)*, vol. 3, no. 1, pp. 1–23, 2021.
- [46] R. Ilman, M. Rahman, and S. Rahman, "Cuet zenith at llms4ol 2025 task c: Hybrid embedding-llm architectures for taxonomy discovery", *Open Conference Proceedings*, 2025.
- [47] M. Canal, J. I. Abreu, and Y. Gutiérrez, "Sema at llms4ol 2025 task c: Prompt-decoupled fine-tuning on matonto with llama", *Open Conference Proceedings*, 2025.