LLMs4OL 2025: The 2nd Large Language Models for Ontology Learning Challenge at the 24th ISWC
LLMs4OL 2025 Task Participant Short Papers
https://doi.org/10.52825/ocp.v6i.2888

Published: 01 Oct. 2025

Phoenixes at LLMs4OL 2025 Task A: Ontology Learning With Large Language Models Reasoning

Alireza Esmaeili Fridouni[1] and Mahsa Sanaei[2,*]

[1]Leibniz Universität Hannover, Hannover, Germany

[2]University of Tabriz, Tabriz, Iran

*Correspondence: Mahsa Sanaei, mahsa.san75@gmail.com

Abstract. Recent advances in large language models (LLMs) have demonstrated remarkable capabilities in various natural language understanding tasks, including Ontology Learning (OL), where they automatically or semi-automatically extract knowledge from unstructured data. This work presents our contribution to the LLMs4OL Challenge at the ISWC 2025 conference, focusing on Task A, which comprises two subtasks: term extraction (SubTask A1) and type extraction (SubTask A2). We evaluate three state-of-the-art LLMs — Qwen2.5-72B-Instruct, Mistral-Small-24B-Instruct-2501, and LLaMA-3.3-70B-Instruct — across three domain-specific datasets: Ecology, Scholarly, and Engineering. In this paper, we adopt a Chain-of-Thought (CoT) Few-Shot Prompting strategy to guide the models in identifying relevant domain terms and assigning their appropriate ontology types. CoT prompting enables LLMs to generate intermediate reasoning steps before producing final predictions, which is particularly beneficial for ontology learning tasks that require contextual reasoning beyond surface-level term matching. Model performance is evaluated using the official precision, recall, and F1-score metrics provided by the challenge organizers. The results reveal important insights into the strengths and limitations of LLMs in ontology learning tasks.

Keywords: Large Language Models, Ontology Learning,
Chain-of-Thought Prompting

1. Introduction

The rapid growth of data on the World Wide Web has introduced significant challenges in organizing and interpreting information. The Semantic Web addresses these challenges by structuring content in a form understandable to both humans and machines. Ontologies are the core building blocks of the Semantic Web, representing knowledge through structured, machine-readable models and defining the concepts, relationships, and categories within a specific domain. The manual creation of ontologies is time-consuming and a complex process. To overcome this, Ontology Learning automates or semi-automates ontology construction from unstructured texts [1], playing

[§]Both authors contributed equally to this work. The order of authors is alphabetical.

a vital role in AI and knowledge engineering by transforming textual data into structured knowledge.

The rise of Generative Artificial Intelligence (AI) and Large Language Models has significantly revolutionized Natural Language Processing (NLP), and various other domains, particularly in automating ontology construction processes. The LLMs4OL paradigm [2] and the LLMs4OL 2024 Challenge [3] investigate the potential of these models for ontology learning tasks. In this paper, we present our participation in the LLMs4OL 2025 Challenge [4], which builds upon this framework and the datasets introduced in [3], [5], focusing on evaluating LLM-based methods for automatic ontology construction across multiple domains. Specifically, we contribute to Task A of the challenge, which involves extracting ontological terminologies and types from raw text. Task A consists of two subtasks:

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

To tackle these subtasks, we employ prompt engineering techniques with a focus on Chain-of-Thought prompting [6]. This technique encourages models to break down their reasoning process into interpretable steps, which is especially valuable in ontology learning tasks where understanding domain-specific terms and their semantic categories requires careful contextual interpretation. For this purpose, we utilized state-of-the-art large language models including Qwen2.5-72B-Instruct [7] [1], Mistral-Small-24B-Instruct-2501 [8][2], and LLaMA-3.3-70B-Instruct [9][3]. The rest of this paper is organized as follows: Section 2 reviews some related works. Section 3 describes our proposed methodology. Section 4 presents the experimental results, and Section 5 provides details about the datasets used in our implementation.

2. Related Works

Recent advancements in LLMs have opened new avenues for automating ontology engineering tasks such as modeling, alignment, and population. Shimizu et al. [10] advocate for modular ontology design to better align with LLM capabilities, showing that breaking down ontologies into smaller modules significantly improves LLM performance in tasks like alignment and entity matching. Similarly, Lippolis et al. [11] explore prompt-based methods—Memoryless CQbyCQ and Ontogenia—for generating ontologies from competency questions. Their results show that LLMs, particularly commercial models, can outperform novice human modellers, though outputs often require manual refinement due to structural inconsistencies and redundant elements. While prior work often addresses isolated subtasks, Luo et al. [12] propose OLLM, a scalable method for building taxonomic backbones by fine-tuning LLMs with custom regularization. Their approach outperforms subtask-based methods both semantically and structurally. Complementary to this, Bakker et al. [13] evaluate LLMs in real-world ontology construction from news articles, noting strengths in class and individual identification but weaknesses in relation modeling and consistency. In the educational domain, Li et al. [14] use LLMs to extract domain concepts from lecture materials and generate ontologies that improve student performance prediction when integrated with

[1] https://huggingface.co/Qwen/Qwen2.5-72B-Instruct
[2] https://huggingface.co/mistralai/Mistral-Small-24B-Instruct-2501
[3] https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct

graph neural networks. Collectively, these studies underscore the promise of LLMs in semi-automating ontology learning while highlighting ongoing challenges in structural accuracy, evaluation, and prompt sensitivity.

3. Methodology

As mentioned earlier, in this work we participate in task A which includes 2 subtasks: term extraction (SubTask A1) and type extraction (SubTask A2). The core of our method focuses on leveraging the power of LLMs for Ontology Learning tasks. To address these subtasks, we employ prompt engineering techniques. LLMs often struggle with complex reasoning, which can limit their effectiveness in specialized tasks. To overcome this, we incorporate Chain-of-Thought prompting, a technique that encourages LLMs to perform step-by-step reasoning before generating final answers. This allows the models to produce more structured, thoughtful, and interpretable responses compared to traditional direct prompting methods.

Considering the complexity of identifying domain-specific terms, accurately classifying them into appropriate ontological categories, and extracting their types, CoT prompting was particularly well-suited to these subtasks. It enables models to reason contextually about the meaning of terms and their relationships within a given text. Furthermore, we enhance our prompts by combining CoT prompting with few-shot examples, providing the models with task-specific demonstrations to help improve accuracy and consistency.

3.1 Prompt Design

For each subtask, we designed prompts related to its specific objectives. The prompts were carefully structured to satisfy with the requirements of ontology learning while leveraging Chain-of-Thought prompting and few-shot examples to improve reasoning and output quality.

3.1.1 SubTask A1: Term Extraction

In this subtask, the prompt instructs the model to carefully read a given text and extract relevant domain-specific terms suitable for ontology construction.. The prompt includes reasoning steps and examples to guide the extraction process. The prompt in Figure 1 used in this subtask.

3.1.2 SubTask A2: Type Extraction

In this subtask, the prompt guides the model to assign appropriate types or categories to a list of terms extracted from a text. Similar to SubTask A1, the prompt encourages step-by-step reasoning, asking the model to infer the semantic role of each term within the given context.

3.2 Large Language Models

In this study, we utilize three state-of-the-art LLMs to evaluate the effectiveness of our approach – Qwen2.5-72B-Instruct [6], Mistral-Small-24B-Instruct-2501, and LLaMA-3.3-70B-Instruct [8]. These models were selected based on their strong performance in natural language understanding and generation tasks, making them well-suited for prompt-based ontology learning applications. We run each subtask using all three LLMs

Your task is to extract key domain-specific terms from the following text. These terms will be used as candidates for building an ontology. To ensure high quality and relevance, follow this step-by-step reasoning process:
Step-by-Step Instructions:
1. Understand the Topic Carefully read the title and text. Identify the main subject area or domain it belongs to (e.g., machine learning, astrophysics, medicine). 2. Identify Technical and Domain-Specific Terms Look through the text for words and phrases that are specific to the domain. Focus on technical nouns, noun phrases, and expressions that carry specialized meaning within the context. These can be: - Single-word terms (e.g., "classifier") - Multi-word compound terms (e.g., "support vector machine") 3. Filter Out Generic and Non-Informative Terms Exclude: - Common words (e.g., "thing", "many") - Function words (e.g., "the", "of", "in") - Verbs, adjectives, and irrelevant modifiers unless part of a technical phrase 4. Rank Terms by Relevance From the filtered list: - Sort the terms by relevance: most important first - Select 5 terms that are most important and central to the topic. 5. Present Only the Terms Do not include explanations, definitions, or duplicates. Only list the final ranked terms.
Example: Title: Supervised Machine Learning Algorithms for High-Dimensional Data Classification Text: The system uses supervised machine learning algorithms such as decision trees, random forests, and support vector machines to classify high-dimensional data.
Chain-of-Thought: - Topic: Machine learning - Identified technical terms: supervised machine learning, decision trees, random forests, support vector machines, high-dimensional data - Filtered out generic words like "system", "uses", "such as", "to classify" - Ranked terms by relevance
Output: 1. Supervised machine learning 2. Decision trees 3. Random forests 4. Support vector machines 5. High-dimensional data
Now follow the same reasoning process to extract terms from the text below.
Title: "'title"'
Text: "'text"'
The output format should be as a list: ["term-1", ...]
Output:

Figure 1. *Prompt Template for Subtask A1 - Term Extraction*

to assess their comparative performance, reasoning capabilities, and generalization across different knowledge domains.

4. Results

This section presents the performance of our proposed method across the three datasets provided in the LLMs4OL 2025 Challenge. We evaluate the system's effectiveness on both subtasks using the official metrics: Precision, Recall, and Macro-F1.

4.1 Datasets

As described in the task definition, three datasets are defined for Task A, that each corresponds to a specific domain, as outlined below:

- Ecology: A dataset that considers the construction of an ontology based on concepts and terminology in the ecology domain.

Table 1. *Phoenixes at LLMs4OL Challenge Results Across LLMs4OL Task A.*

Dataset	Model	F1-score	Precision	Recall
SubTask A1 - Term Extraction				
A1.2 - Scholarly	LLaMa-3.3-70B-Instruct	0.3617	0.2741	**0.5312**
	Mistral-Small-24B-Instruct-2501	0.3750	0.3125	0.4687
	Qwen2.5-72B-Instruct	**0.3950**	**0.3265**	0.5000
A1.3 - Engineering	LLaMa-3.3-70B-Instruct	**0.2556**	0.4063	**0.1864**
	Mistral-Small-24B-Instruct-2501	0.1031	0.3548	0.0603
	Qwen2.5-72B-Instruct	0.1435	**0.4893**	0.0840
SubTask A2 - Type Extraction				
A2.1 - Ecology	LLaMa-3.3-70B-Instruct	0.4031	**0.3767**	0.4336
	Mistral-Small-24B-Instruct-2501	0.2820	0.2291	0.3666
	Qwen2.5-72B-Instruct	**0.4309**	0.3566	**0.5442**
A2.2 - Scholarly	LLaMa-3.3-70B-Instruct	0.3913	0.2903	**0.6000**
	Mistral-Small-24B-Instruct-2501	**0.4272**	**0.3634**	0.5181
	Qwen2.5-72B-Instruct	0.3037	0.2448	0.4000
A2.3 - Engineering	LLaMa-3.3-70B-Instruct	0.1525	0.0900	**0.5000**
	Mistral-Small-24B-Instruct-2501	0.1655	0.1100	0.3333
	Qwen2.5-72B-Instruct	**0.1846**	**0.1276**	0.3333

- Scholarly: A dataset that considers the construction of an ontology grounded in scholarly communication and the academic publishing domain.
- Engineering: A dataset that considers the construction of an ontology derived from the engineering domain, including relevant structures, processes, and terminologies.

Each dataset contains domain-specific texts for which ontology terms and their types need to be extracted.

4.2 Results Analysis

We report the performance of the three used LLMs on both subtasks across the three datasets. Table 1 summarizes the results for SubTask A1 and SubTask A2.

4.2.1 SubTask A1 — Term Extraction

In the Scholarly domain, Qwen2.5-72B-Instruct obtained the highest Macro-F1 score 0.3950. Mistral-Small-24B followed closely, with stronger precision 0.3125. LLaMA-3.3-70B demonstrated the highest recall 0.5312 but low recall 0.0840, effectively identifying a broader set of candidate terms. In the Engineering dataset, results varied across models, reflecting domain-specific characteristics. LLaMA-3.3-70B achieved the highest F1-score (0.2556) and recall (0.1864), while Qwen2.5-72B produced the highest precision (0.4893), indicating selective term identification. These observations suggest that combining Chain-of-Thought prompting with few-shot examples supports term extraction, particularly in domains like Scholarly communication, where terminological boundaries are relatively well-defined, compared to highly technical fields like Engineering.

4.2.2 SubTask A2 — Type Extraction

In the Ecology dataset, Qwen2.5-72B-Instruct outperformed the other models, achieved the highest Macro-F1 score 0.4309 and recall 0.5442, followed by LLaMA-3.3-70B

with F1-score 0.4031, showing balanced performance. For Scholarly, Mistral-Small-24B obtained the highest Macro-F1 0.4272 and precision 0.3634, while LLaMA-3.3-70B achieved the highest recall 0.6000, indicating strong retrieval capability. In Engineering, models struggled and provided competitive results, reflecting the complexity and domain-specific nature of type assignment in this field. Qwen2.5-72B-Instruct achieved the highest F1-score 0.1846 and precision 0.1276, and LLaMA-3.3-70B delivered the highest recall 0.5.

Overall, the findings highlight how domain characteristics and conceptual complexity influence model behavior, and highlight the effectiveness of combining Chain-of-Thought reasoning and few-shot prompting in improving type extraction accuracy, particularly in domains with well-defined conceptual structures like Ecology and Scholarly communication.

5. Conclusion

In this work, we explored the use of large language models with Chain-of-Thought reasoning and few-shot prompting for ontology learning tasks in the LLMs4OL 2025 Challenge. The results demonstrate that combining reasoning-based prompting with instruction-tuned models can effectively support term and type extraction across diverse domains. While performance varies by domain, the approach shows particular strength and encouraging performance in areas with clearer conceptual structures. These findings highlight the potential of prompt engineering strategies for ontology learning without task-specific fine-tuning.

Underlying and related material

The implementation of this work is published in the GitHub repository for the research community at https://github.com/MahsaSanaei/Phoenixes-LLMs4OL2025.

Author contributions

Alireza Esmaeili Fridouni. Conceptualization, Methodology, Software, Writing – original draft.

Mahsa Sanaei. Conceptualization, Methodology, Software, Formal analysis, Visualization, Writing – original draft, Writing – review & editing.

Competing interests

The authors declare that they have no competing interests.

References

[1] W. Wong, W. Liu, and M. Bennamoun, "Ontology learning from text: A look back and into the future", *ACM computing surveys (CSUR)*, vol. 44, no. 4, pp. 1–36, 2012.

[2] J. D. H. Babaei Giglou and S. Auer, "Llms4ol: Large language models for ontology learning," *in The Semantic Web – ISWC 2023*, vol. 4, no. Y, pp–pp, Oct. 2024.

[3] 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. [Online]. Available: https://www.tib-op.org/ojs/index.php/ocp/article/view/2473.

[4] H. Babaei Giglou, J. D'Souza, N. Mihindukulasooriya, and S. Auer, "Llms4ol 2025 overview: The 2nd large language models for ontology learning challenge", *Open Conference Proceedings*, 2025.

[5] 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.

[6] J. Wei et al., "Chain-of-thought prompting elicits reasoning in large language models", *Advances in neural information processing systems*, vol. 35, pp. 24 824–24 837, 2022.

[7] A. Yang et al., "Qwen2. 5-1m technical report", *arXiv preprint arXiv:2501.15383*, 2025.

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

[9] A. Grattafiori et al., *The llama 3 herd of models*, 2024. arXiv: 2407.21783 [cs.AI]. [Online]. Available: https://arxiv.org/abs/2407.21783.

[10] C. Shimizu and P. Hitzler, "Accelerating knowledge graph and ontology engineering with large language models", *Journal of Web Semantics*, vol. 85, p. 100 862, 2025, ISSN: 1570-8268. DOI: https://doi.org/10.1016/j.websem.2025.100862. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1570826825000022.

[11] A. S. Lippolis et al., "Ontology generation using large language models", in *The Semantic Web: 22nd European Semantic Web Conference, ESWC 2025, Portoroz, Slovenia, June 1–5, 2025, Proceedings, Part I*, Portoroz, Slovenia: Springer-Verlag, 2025, pp. 321–341, ISBN: 978-3-031-94574-8. DOI: 10.1007/978-3-031-94575-5_18. [Online]. Available: https://doi.org/10.1007/978-3-031-94575-5_18.

[12] A. Lo, A. Q. Jiang, W. Li, and M. Jamnik, "End-to-end ontology learning with large language models", in *Advances in Neural Information Processing Systems*, A. Globerson et al., Eds., vol. 37, Curran Associates, Inc., 2024, pp. 87 184–87 225. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2024/file/9e89f068a62f6858c661a8abecf5bb0a-Paper-Conference.pdf.

[13] R. M. Bakker, D. L. Di Scala, and M. de Boer, "Ontology learning from text: An analysis on llm performance", in *Proceedings of the 3rd NLP4KGC International Workshop on Natural Language Processing for Knowledge Graph Creation, colocated with Semantics*, 2024, pp. 17–19.

[14] G. Li, C. Tang, L. Chen, D. Deguchi, T. Yamashita, and A. Shimada, "Llm-driven ontology learning to augment student performance analysis in higher education", in *Knowledge Science, Engineering and Management*, C. Cao, H. Chen, L. Zhao, J. Arshad, T. Asyhari, and Y. Wang, Eds., Singapore: Springer Nature Singapore, 2024, pp. 57–68, ISBN: 978-981-97-5498-4.