






# 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

Patipon Wiangnak<sup>1,\*</sup> , Thin Prabhong<sup>2</sup> , Thiti Phuttaamart<sup>2</sup> ,  
Natthawut Kertkeidkachorn<sup>1</sup> , and Kiyoaki Shirai<sup>1</sup> 

<sup>1</sup>Japan Advanced Institute of Science and Technology, Japan

<sup>2</sup>Chiang Mai University, Thailand

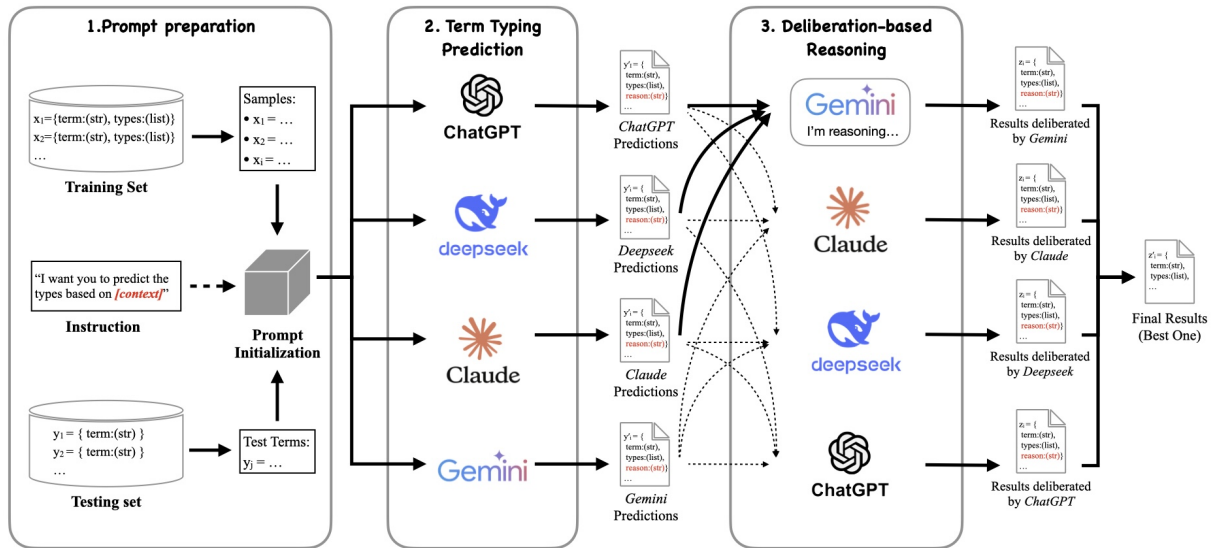
\*Correspondence: Patipon Wiangnak, w.patipon@jaist.ac.jp

**Abstract.** The LLMs4OL Challenge at ISWC 2025 aims to advance the integration of Large Language Models (LLMs) and Ontology Learning (OL) across four key tasks: (1) Text2Onto, (2) Term Typing, (3) Taxonomy Discovery, and (4) Non-Taxonomic Relation Extraction. Our work focuses on the Term Typing Prediction task, where prompting LLMs has shown strong potential. However, in low-resource domains, relying on a single LLM is often insufficient due to domain-specific knowledge gaps and limited exposure to specialized terminology, which can lead to inconsistent and biased predictions. To address this challenge, we propose DREAM-LLMs: a Deliberation-based Reasoning Ensemble Approach with Multiple Large Language Models. Our method begins by crafting few-shot prompts using training examples and querying four advanced LLMs independently: ChatGPT-4o, Claude Sonnet 4, DeepSeek-V3, and Gemini 2.5 Pro. Each model outputs a predicted label along with a brief justification. To reduce model-specific bias, we introduce a deliberation step, in which one LLM reviews the predictions and justifications from the other three to produce a final decision. We evaluate DREAM-LLMs on three low-resource domain datasets: OBI, MatOnto, and SWEET using F1-score as the evaluation metric. The results, 0.908 for OBI, 0.568 for MatOnto, and 0.593 for SWEET, demonstrate that our ensemble strategy significantly improves performance, highlighting the promise of collaborative LLM reasoning in low-resource environments.

**Keywords:** Large Language Models, Ontology Learning, Term Typing Prediction, Deliberation-Based Reasoning, Low-Resource Domains

## 1. Introduction

Large Language Models (LLMs) have advanced many NLP tasks, yet Ontology Learning (OL) remains challenging. Traditional OL methods are based on human-crafted rules, domain expertise, or large labeled datasets, making them slow, costly, and hard to scale. Although LLMs can automate ontology creation [1], they are prone to instability



**Figure 1.** DREAM-LLMs: A Deliberation-based Reasoning Ensemble Approach with Multiple Large Language Models for Term Typing in Low-Resource Domains.

and hallucination in complex tasks. The LLMs4OL 2025 Challenge [2] addresses four tasks: (1) **Text2Onto**, which extracts ontological types and terms from unstructured text; (2) **Term Typing**, which assigns generalized types to terms; (3) **Taxonomy Discovery**, which identifies hierarchical type relations; and (4) **Non-Taxonomic Relation Extraction**, which extracts non-hierarchical relations. This work focuses on Term Typing, which organizes terms into coherent ontological structures. In low-resource domains, a single LLM often fails due to knowledge gaps, leading to inconsistent predictions. To overcome this, we propose DREAM-LLMs, a Deliberation-based Reasoning Ensemble using ChatGPT-4o [3], Claude Sonnet 4 [4], DeepSeek-V3 [5], and Gemini 2.5 Pro [6]. Few-shot prompts query each model independently, after which one model acts as a judge to review peer outputs and decide the final label. This cross-model reasoning mitigates bias and improves accuracy. Resources are available at: <https://github.com/wpatipon-jaist/LLMs4OL2025-Task-B-DREAM-LLMs>.

## 2. Related Work

Large Language Models for Ontology Learning (LLMs4OL) [7] is an end-to-end framework for ontology learning that aims to explore the potential of using LLM to enhance understanding and innovation in OL tasks, aligning with the goals of the Semantic Web to build more intelligent systems. Hybrid methods [8] have often outperformed standalone LLMs, as incorporating external knowledge can significantly improve performance. However, LLMs still face challenges in fully capturing complex domain-specific knowledge. To address this, we utilize a prompt-based approach and introduce a novel deliberation step. This results in an ensemble framework that leverages the strengths of multiple LLMs through collaborative reasoning, ultimately improving term typing performance in low-resource domains.

## 3. Approach

DREAM-LLM designed the basis for the fundamental technique of prompting. There are three major steps as displayed in Figure 1: (1) Prompt Preparation, (2) Term Typing Prediction, and (3) Deliberation-based Reasoning.

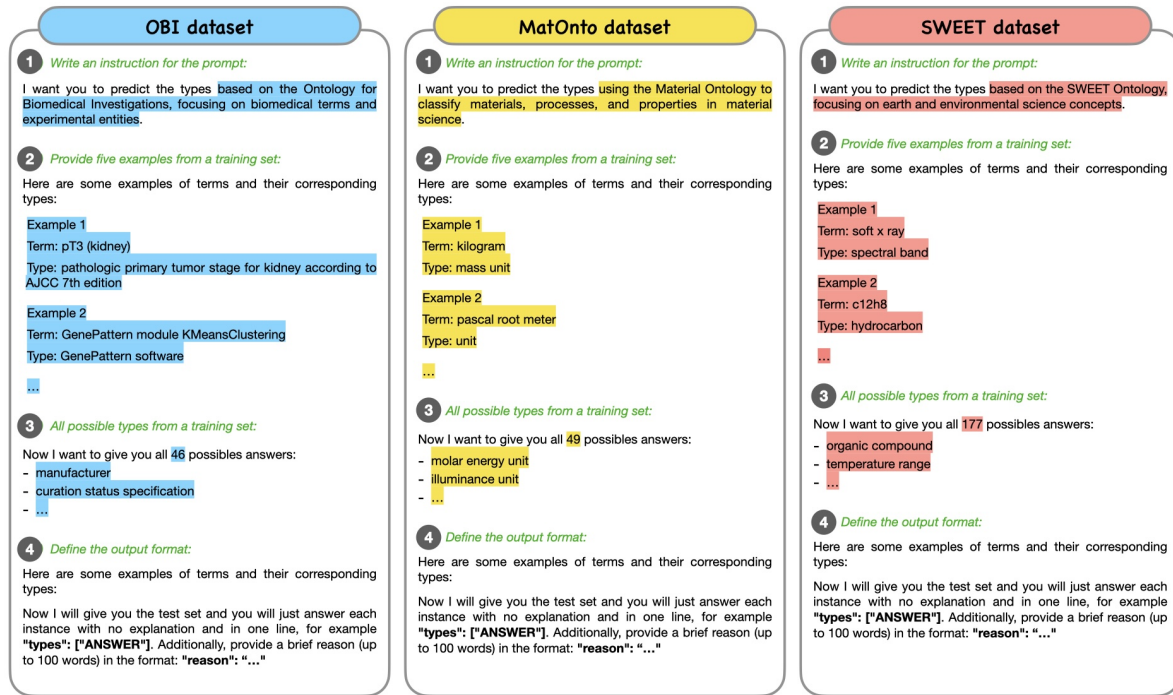
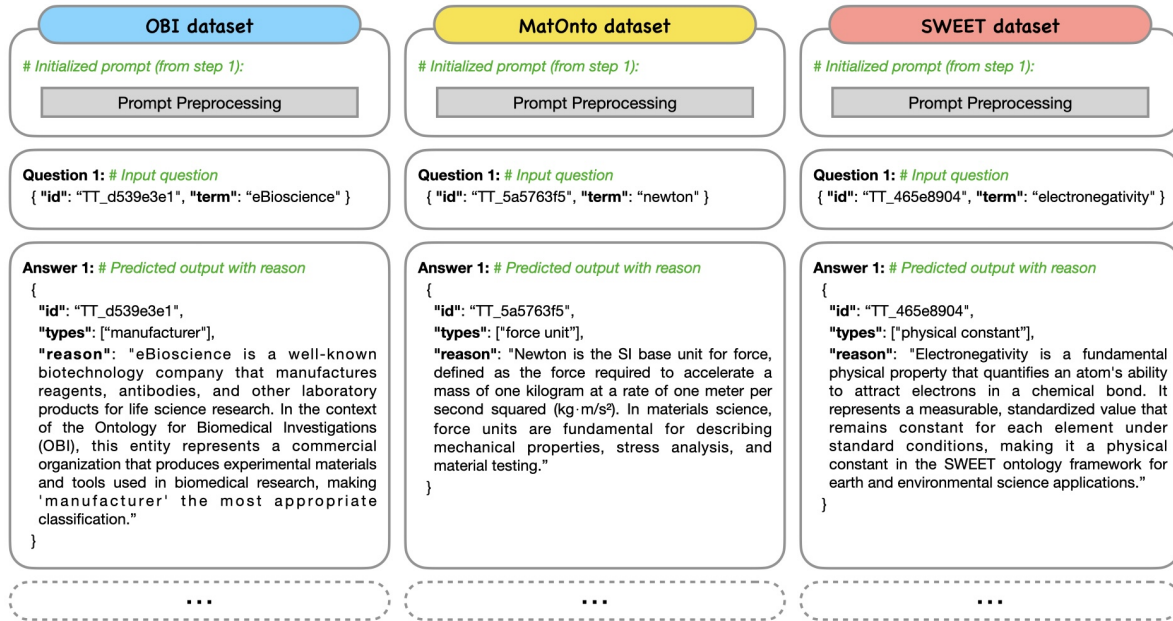


Figure 2. Example few-shot prompts for the LLMs4OL challenge datasets OBI, MatOnto, and SWEET.

### 3.1 Prompt Preparation

In the first step, we need to prepare a few-shot prompt that is general and suitable for low-resource settings, such as the challenge datasets OBI, MatOnto, and SWEET. As in the example prompts for each dataset in Figure 2, there are four parts involved in designing the few-shot prompt. Using these four components of prompt preparation, we construct a dataset-specific prompt and provide it as input to the LLMs:

- Prompt Instruction:** We provide a contextual instruction prompt to LLMs for each dataset, guiding them by explicitly stating the task and setting the context. Without this instruction, the model may interpret the task ambiguously, leading to inconsistent or off-topic outputs. The details are as follows:
  - OBI:** "I want you to predict the types based on the Ontology for Biomedical Investigations, focusing on biomedical terms and experimental entities."
  - MatOnto:** "I want you to predict the types using the Material Ontology to classify materials, processes, and properties in material science."
  - SWEET:** "I want you to predict the types based on the SWEET Ontology, focusing on earth and environmental science concepts."
- Provide the Examples:** We include a few-shot example, each randomly sampled from the training set of the corresponding dataset, ensuring that each example represents a unique term type. These examples help the LLM infer the expected input–output format and align its predictions with the intended output. Each example consists of an example number, a term, and its corresponding type. In this study, we used five-shot prompting.
- Provide all possible answers:** We then provide the LLM with the number of unique type labels, along with a complete list of these labels found in the training set for each dataset. This helps the LLM scope its outputs appropriately.



**Figure 3.** Input-output examples of term type prediction using dataset-specific prompts

- 4. Define the output format:** To ensure consistency, we specify a standardized output format for the LLM to follow during prediction. This minimizes ambiguity in the model responses and ensures that the outputs align with the expected structure required for the downstream tasks or scoring scripts.

### 3.2 Term Typing Prediction

In the second step, we use the dataset-specific prompt constructed in the previous step and provide it, along with a test term, as input to the LLM. The model then generates a prediction for each term individually. As shown in Figure 3, we illustrate examples of input-output pairs for a few-shot prompt.

### 3.3 Deliberation-Based Reasoning

In this step, we introduce a deliberation process in which one LLM reviews the predictions and explanations provided by the other LLMs and makes the final decision. This approach helps reduce model-specific biases and improves the robustness of predictions by using multiple perspectives. By having each LLM evaluate the outputs of the others and justify its selection, the system promotes consensus and critical assessment, leading to more reliable and consistent final decisions, especially important feature in low-resource or ambiguous scenarios. As illustrated by the example prompts for each dataset in Figure 4, the deliberation prompt consists of three key components:

- 1. Prompt Instruction:** We provide a contextual instruction prompt to LLMs, directing them to act as a judge by evaluating the predictions generated by other LLMs for a given term across different datasets. Subsequently, we supply the LLM with all possible type labels found in the training set for the corresponding datasets:
  - a) OBI:** *"I want you to judge predictions from other 3 LLMs from terms based on the Ontology for Biomedical Investigations, focusing on biomedical terms and experimental entities."*



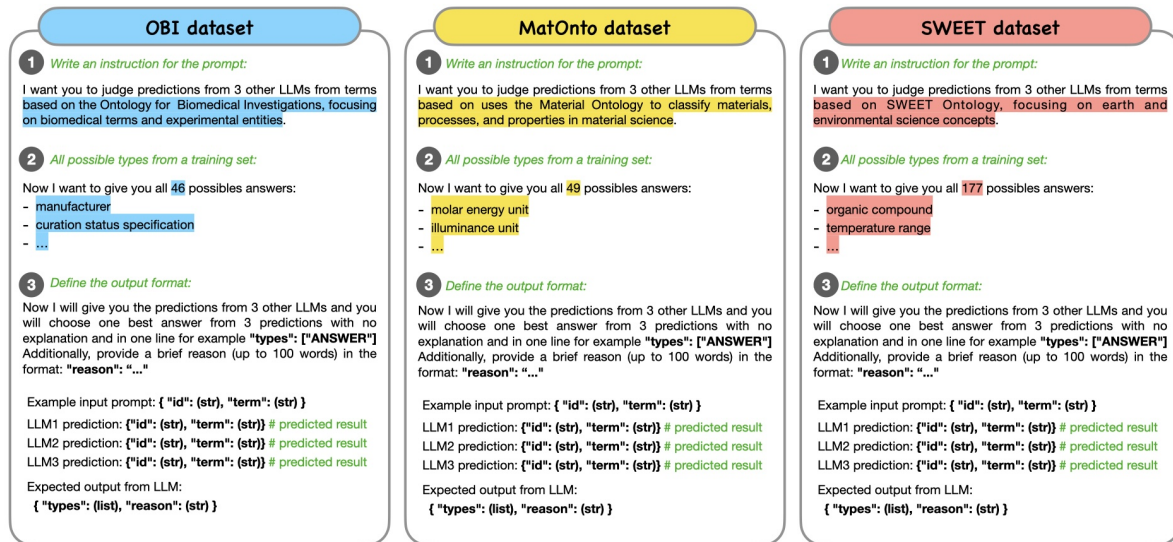


Figure 4. Example templates of the deliberation prompt for each dataset.

- MatOnto:** "I want you to judge predictions from other 3 LLMs from terms based on uses the Material Ontology to classify materials, processes, and properties in material science."
  - SWEET:** "I want you to judge predictions from 3 other LLMs from terms based on SWEET Ontology, focusing on earth and environmental science concepts."
- Provide all possible answers:** We then provide the LLM with the number of unique type labels, along with a complete list of these labels found in the training set for each dataset.
  - Define the Output Format:** In this study, we used four LLMs. We define a required output format in which an LLM selects the best prediction from the options provided by the other LLMs and includes a brief explanation for its choice. This process is repeated for each LLM, enabling a cross-check mechanism between models.

## 4. Experiment

### 4.1 Datasets

In the LLMs4OL 2025 Challenge, three domain-specific low-resource datasets are used across the tasks: (1) **OBI**, which is a dataset focused on biomedical investigations and experimental entities; (2) **MatOnto**, which covers the classification of materials, processes, and properties in materials science; and (3) **SWEET**, which encompasses concepts in earth and environmental sciences. The statistical details of each dataset are provided in Table 1.

Table 1. Statistics of the datasets used in the LLMs4OL 2025 Challenge.

Dataset	#Train	#Test	#Unique Labels
OBI	201	87	46
MatOnto	85	37	49
SWEET	1,558	626	177

**Table 2.** The results of the term typing prediction step.

Dataset	F1-score			
	ChatGPT-4o	Deepseek-V3	Claude Sonnet 4	Gemini 2.5 Pro
OBI	0.77	0.793	0.816	<b>0.851</b>
MatOnto	0.46	0.474	<b>0.568</b>	<b>0.568</b>
SWEET	0.375	0.436	0.484	<b>0.548</b>

## 4.2 Experimental Setup

We employ four advanced LLMs in our ensemble-based approach for term typing: ChatGPT-4o by OpenAI, known for its strong general-purpose reasoning capabilities; Claude Sonnet 4 by Anthropic, recognized for its robust language understanding and safety alignment; DeepSeek-V3, an open-source model optimized for retrieval-augmented tasks; and Gemini 2.5 Pro by Google DeepMind, which offers multimodal support and competitive performance across various tasks. Each LLM is queried independently using a few-shot prompting strategy, where the prompt consists of four components, as detailed in Sections 3.1 and 3.2. In the deliberation step, one LLM is designated as the judge and receives the predictions and justifications generated by the other three models as the explainers, selecting the most appropriate label based on their reasoning. The structure of the deliberation prompt is described in Section 3.3. The prediction of each model is either directly evaluated in the few-shot setting or further refined through the deliberation presented in subsequent sections.

## 5. Result and Discussion

In this study, we evaluated each step using precision, recall, and F1-score. Since all three metrics exhibited identical trends across evaluations, we report only the F1-score in the tables for clarity and conciseness. Table 2 presents the results of the term typing prediction step, showing the F1-scores of four models: ChatGPT-4o, DeepSeek-V3, Claude Sonnet 4, and Gemini 2.5 Pro on three datasets: OBI, MatOnto, and SWEET, based on the prompting process. Overall, Gemini 2.5 Pro achieved the highest F1-scores across all datasets, performing particularly well on OBI with 0.851, followed by Claude Sonnet 4 at 0.816. DeepSeek-V3 and ChatGPT-4o scored slightly lower. For MatOnto, Gemini 2.5 Pro and Claude Sonnet 4 tied for the top score of 0.568, while ChatGPT-4o and DeepSeek-V3 trailed. SWEET proved to be the most challenging, with all models scoring lower; however, Gemini 2.5 Pro again led with 0.548, while ChatGPT-4o had the lowest score of 0.375.

Table 3 summarizes the deliberation step, where multiple models act as explainers and a separate model serves as judge. For OBI, the highest F1-score of 0.908 was achieved when ChatGPT-4o, DeepSeek-V3, and Claude Sonnet 4 acted as explainers and Gemini 2.5 Pro served as judge, surpassing the best single-model score in Table 2. This demonstrates that combining diverse reasoning perspectives and centralizing decision-making in a strong judge can enhance predictive performance. For SWEET, the highest score was 0.593 with Claude Sonnet 4 as judge, slightly improving upon the top single-model result. In contrast, MatOnto's best deliberation score of 0.568 matched the single-model peak, suggesting that the leading model was already highly confident and accurate, leaving little room for improvement, or that weaker peer output introduced noise the judge could not overcome.

**Table 3.** The results of the deliberation-based reasoning step. Model abbreviations:  $M_1$  = ChatGPT-4o,  $M_2$  = Claude Sonnet 4,  $M_3$  = DeepSeek-V3,  $M_4$  = Gemini 2.5 Pro.

Dataset	Configuration (Explainers → Judge)	F1-score
OBI	$M_3 + M_2 + M_4 \rightarrow M_1$	0.874
	$M_1 + M_3 + M_4 \rightarrow M_2$	0.862
	$M_1 + M_2 + M_4 \rightarrow M_3$	0.874
	$M_1 + M_3 + M_2 \rightarrow M_4$	<b>0.908</b>
MatOnto	$M_3 + M_2 + M_4 \rightarrow M_1$	<b>0.568</b>
	$M_1 + M_3 + M_4 \rightarrow M_2$	0.460
	$M_1 + M_2 + M_4 \rightarrow M_3$	<b>0.568</b>
	$M_1 + M_3 + M_2 \rightarrow M_4$	0.514
SWEET	$M_3 + M_2 + M_4 \rightarrow M_1$	0.534
	$M_1 + M_3 + M_4 \rightarrow M_2$	<b>0.593</b>
	$M_1 + M_2 + M_4 \rightarrow M_3$	0.511
	$M_1 + M_3 + M_2 \rightarrow M_4$	0.514

**Table 4.** Token usage and cost for different LLMs.

Model	Token Input	Token Output	Cost (USD)
DeepSeek-V3	1.6M	107K	0.29
ChatGPT-4o	1.5M	110K	4.25
Claude Sonnet 4	1.8M	129K	7.61
Gemini 2.5 Pro	1.2M	145K	5.32

As shown in Figure 5, deliberation can even reverse an individual model’s incorrect prediction by leveraging peer reasoning. In this OBI example, Claude Sonnet 4, ChatGPT-4o, and DeepSeek-V3 all underperformed relative to Gemini 2.5 Pro in single-model evaluation. Notably, Gemini 2.5 Pro misclassified the term “*term imported*” when acting alone, selecting “*curation status specification*” instead of the correct “*obsolescence reason specification*”. However, when acting as judge and receiving reasoning from all three peers, including DeepSeek-V3’s correct explanation, Gemini 2.5 Pro produced the correct prediction. This occurs because the judge is no longer constrained to its own initial reasoning path. By reviewing multiple chains of thought, it can identify stronger evidence, reconcile contradictions, and avoid heuristic biases that misled it in isolation. In this case, DeepSeek-V3’s rationale provided the critical signal that enabled Gemini 2.5 Pro to override its earlier misconception. More broadly, requiring models to articulate reasoning and then combining these rationales in a context-driven ensemble allows the final model to integrate both correct and incorrect logic paths into a more reliable decision, often surpassing the performance of any single model.

Table 4 compares the model cost profiles. DeepSeek-V3 stands out for its low cost despite processing a large number of tokens. Claude Sonnet 4 achieves strong performance in some tasks, but is the most expensive. Gemini 2.5 Pro produces the most output tokens, resulting in a mid-range cost, while ChatGPT-4o remains balanced in both usage and price. These differences highlight the importance of weighing performance gains against operational costs when selecting an LLM for the judge role.

## 6. Conclusion

In this paper, we explore the effectiveness of LLMs for ontology-based term typing in low-resource domains. Although individual LLMs perform well with prompting, their

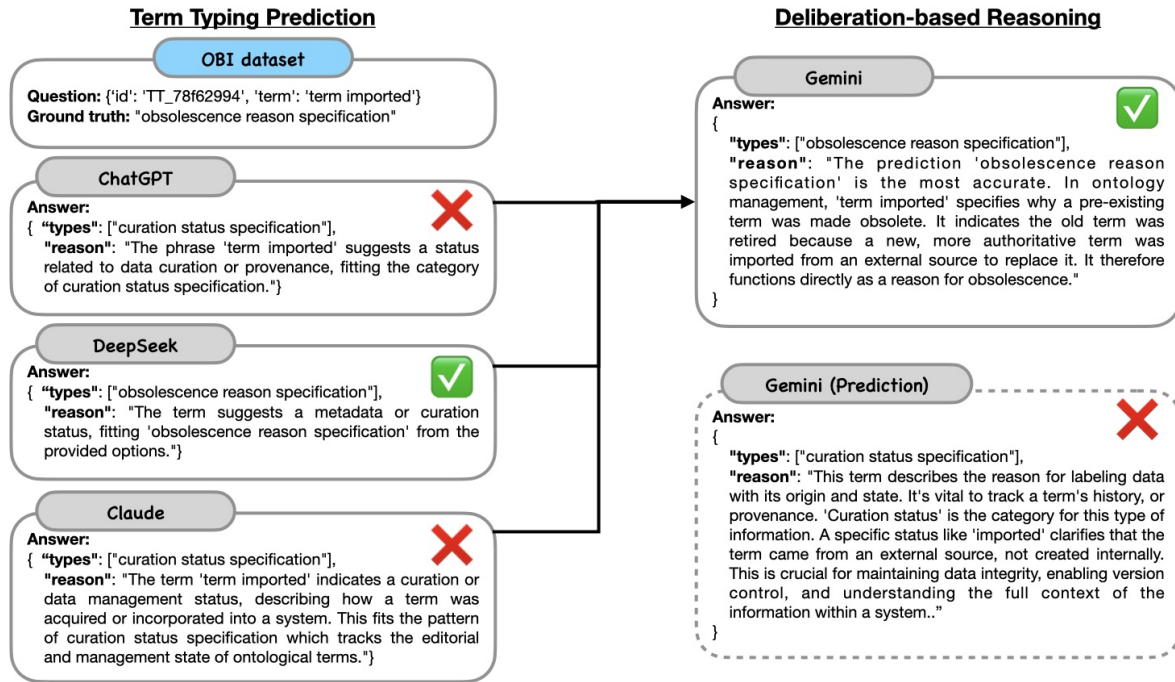


Figure 5. Deliberation-based reasoning example for the OBI dataset term.

output can be inconsistent due to limited domain-specific knowledge. To address this, we proposed DREAM-LLMs, a deliberation-based approach in which multiple LLMs act as explainers and a separate model serves as a judge to improve reasoning and prediction quality. DREAM-LLMs achieve improved overall performance. In future work, we aim to extend our approach to multi-label term typing, which will better capture the nuanced semantics of terms in ontology-based tasks and enhance the practical utility of the models in real-world applications. Furthermore, we will consider a diversity-based selection strategy [9] to improve performance by providing more representative and informative examples in a few-shot prompt, thus reducing redundancy compared to the random method.

## Author contributions

**Patipon Wiangnak, Thin Prabhong, and Thiti Phuttaamart:** Conceptualization; Software Development; Writing – Original Draft Preparation.

**Natthawut Kertkeidkachorn:** Conceptualization; Project Administration; Supervision; Writing – Review and Editing.

**Kiyoaki Shirai:** Supervision; Project Administration; Writing – Review and Editing.

## Competing interests

The authors declare that they have no competing interests.

## References

- [1] S. Pan, L. Luo, Y. Wang, C. Chen, J. Wang, and X. Wu, "Unifying Large Language Models and Knowledge Graphs: A Roadmap", *IEEE Transactions on Knowledge and Data Engineering*, vol. 36, no. 7, pp. 3580–3599, Jul. 2024, arXiv:2306.08302 [cs], ISSN: 1041-4347, 1558-



- 2191, 2326-3865. DOI: [10.1109/TKDE.2024.3352100](https://doi.org/10.1109/TKDE.2024.3352100). Accessed: Aug. 6, 2025. [Online]. Available: <http://arxiv.org/abs/2306.08302>.
- [2] H. Babaei Giglou, J. D'Souza, A. C. Aioanei, N. Mihindukulasooriya, and S. Auer, "Llms4ol 2025 overview: The 2nd large language models for ontology learning challenge", *Open Conference Proceedings*, 2025.
- [3] OpenAI et al., *GPT-4o System Card*, arXiv:2410.21276 [cs], Oct. 2024. DOI: [10.48550/arXiv.2410.21276](https://doi.org/10.48550/arXiv.2410.21276). Accessed: Aug. 6, 2025. [Online]. Available: <http://arxiv.org/abs/2410.21276>.
- [4] Anthropic, *System Card: Claude Opus 4 & Claude Sonnet 4*, May 2025. [Online]. Available: <https://www-cdn.anthropic.com/07b2a3f9902ee19fe39a36ca638e5ae987bc64dd.pdf>.
- [5] DeepSeek-AI et al., *DeepSeek-V3 Technical Report*, arXiv:2412.19437 [cs], Feb. 2025. DOI: [10.48550/arXiv.2412.19437](https://doi.org/10.48550/arXiv.2412.19437). Accessed: Aug. 6, 2025. [Online]. Available: <http://arxiv.org/abs/2412.19437>.
- [6] G. Comanici et al., *Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities*, arXiv:2507.06261 [cs], Jul. 2025. DOI: [10.48550/arXiv.2507.06261](https://doi.org/10.48550/arXiv.2507.06261). Accessed: Aug. 6, 2025. [Online]. Available: <http://arxiv.org/abs/2507.06261>.
- [7] H. B. Giglou, J. D'Souza, and S. Auer, *LLMs4OL: Large Language Models for Ontology Learning*, arXiv:2307.16648 [cs], Aug. 2023. DOI: [10.48550/arXiv.2307.16648](https://doi.org/10.48550/arXiv.2307.16648). Accessed: Jul. 8, 2025. [Online]. Available: <http://arxiv.org/abs/2307.16648>.
- [8] H. B. Giglou, J. D'Souza, and S. Auer, *LLMs4OL 2024 Overview: The 1st Large Language Models for Ontology Learning Challenge*, arXiv:2409.10146 [cs], Sep. 2024. DOI: [10.48550/arXiv.2409.10146](https://doi.org/10.48550/arXiv.2409.10146). Accessed: Jul. 8, 2025. [Online]. Available: <http://arxiv.org/abs/2409.10146>.
- [9] A. Alcoforado et al., "From Random to Informed Data Selection: A Diversity-Based Approach to Optimize Human Annotation and Few-Shot Learning", in *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, P. Gamallo et al., Eds., Santiago de Compostela, Galicia/Spain: Association for Computational Linguistics, Mar. 2024, pp. 492–502. Accessed: Aug. 8, 2025. [Online]. Available: <https://aclanthology.org/2024.propor-1.50/>.