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.2898

Published: 01 Oct. 2025

T-GreC at LLMs4OL 2025 Task B: A Report on Term-Typing Task of OBI Dataset Using LLM With k-Nearest Neighbors

Chavakan Yimmark[1,*] and Teeradaj Racharak[2]

[1]School of Information Science, Japan Advanced Institute of Science and Technology, Japan

[2]Advanced Institute of So-Go-Chi (Convergence Knowledge) Informatics, Tohoku University, Japan

*Correspondence: Chavakan Yimmark, chavakan.yim@jaist.ac.jp

Abstract. This report presents an approach that combines large language models' (LLMs) embedding with k-nearest neighbors (k-NN) for the term-typing task on the OBI (Ontology for Biomedical Investigations) dataset. We investigate the effectiveness of transformer models namely PubMedBERT, BioBERT, DeBERTa-v3, and RoBERTa with k-NN classification using the embedding of each respective model. Our experimental results demonstrate that fine-tuned LLMs not only have the capability to do term typing on their own but also can learn effective embeddings that are exploitable by k-NN for solving the task, with RoBERTa achieving the highest F1 score of 0.827 and k-NN using embedding from the model with score of 0.862. The study reveals that embeddings from transformer models, when used as semantic representations for similarity-based method, improve classification accuracy in this specific case.

Keywords: Ontology Learning, Term-Typing, Large Language Models, K-Nearest Neighbors

1. Introduction

Large language models (LLMs) have demonstrated remarkable capabilities in natural language understanding and classification tasks. In the biomedical domain, transformer-based models such as PubMedBERT [1] and BioBERT [2] have shown effectiveness in processing specialized biomedical texts. However, beyond their direct classification capabilities, these models also generate contextual embeddings that encode semantic information from the text. This raises an important question: *can these embeddings be effectively utilized by traditional machine learning approaches to achieve high classification performance?*

The integration of neural representations with classical machine learning algorithms presents a work around for improving classification performance. Among these algorithms, k-nearest neighbors (k-NN) stands out for its simplicity and effectiveness, particularly when paired with high-quality feature representations [3]. In this work, we leverage k-NN to classify terms by measuring similarity in the embedding space generated by transformer-based language models. The semantically rich embeddings produced by these models enable k-NN to identify nearest neighbors that share meaningful contextual similarity, thereby increasing the likelihood of assigning the correct term to a new instance based on its proximity to known labeled examples.

This report investigates the capability of transformer-based language models for term-typing task stated in [4]: their direct classification performance through fine-tuning and their ability to generate effective embeddings for k-NN classification. We evaluate four transformer models including PubMedBERT, BioBERT, DeBERTa-v3 [5], and RoBERTa [6] on the OBI dataset, comparing their performance as standalone classifiers against their effectiveness as embedding generators for k-NN classification.

2. Related Works

From the LLMs4OL 2024 challenge [7], many teams participated in the Term Typing task using a variety of techniques, including fine-tuning, prompting, retrieval-augmented generation (RAG), prompt-tuning, machine learning, and rule-based strategies. These methods leveraged large language models (LLMs) in different ways to handle the challenges last year. Notably, **silp_nlp** [8] used several machine learning techniques like Random Forest, Logistic Regression, and XGBoost for Term Typing task and got high score on several subtasks.

In many classification tasks, traditional methods like k-Nearest Neighbors (k-NN) have shown competitive performance, particularly when combined with strong feature representations. By utilizing embeddings from pretrained language models, k-NN can potentially serve as a robust non-parametric classifier for text that requires no additional training while still capturing semantic similarities in the embedding space [9].

3. Methodology

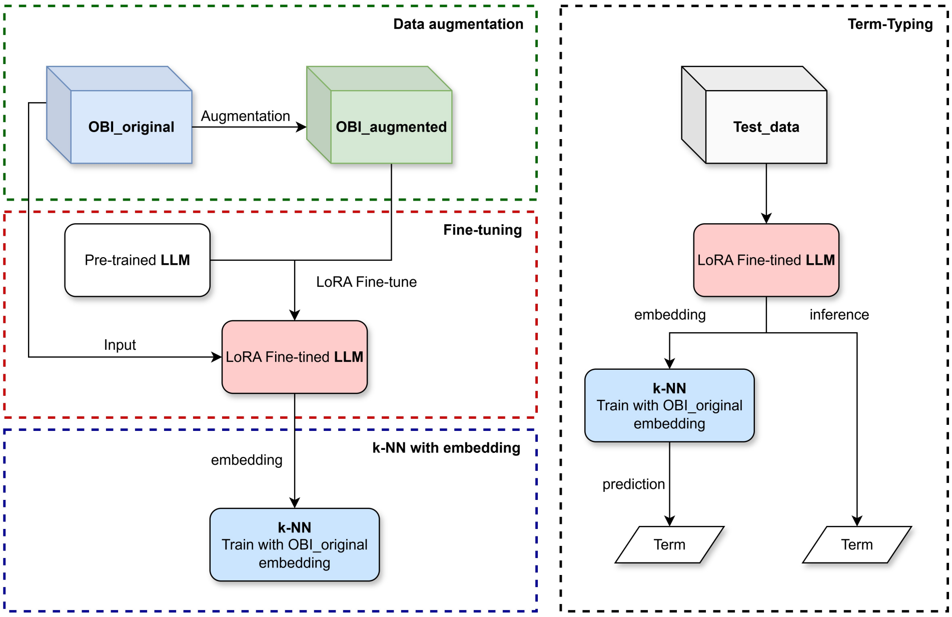

Figure 1. *Overview of our methodology. (Left) The training pipeline includes data augmentation, fine-tuning, and k-NN with embeddings. (Right) Term-typing process using the trained model.*

The methodology consists of three main components (shown in Figure 1): data augmentation, pre-trained language models and fine-tuning, and using k-NN with embedding generated from LoRA-adapted model.

Data augmentation: The used OBI dataset is given by the 2nd LLMs4OL Challenge [10]. The original dataset is then used for making augmented dataset. Because of the simplicity of terms in the OBI dataset, we chose to perform augmentation by randomly inserting, deleting, swapping, or substituting a character in each term. In our experiment, we randomly applied augmentation three times for each term in the original dataset. We assume that these augmentations could help the model learn the representation of each term more effectively.

Pre-trained language models and fine-tuning: We use several pre-trained encoder-only transformer models which build context-aware representations of the input, making them ideal for our text classification task. The models include:

- **PubMedBERT** (`microsoft/BiomedNLP-PubMedBERT-base-uncased-abstract`): Pretrained from scratch on PubMed abstracts using an uncased vocabulary, optimized for biomedical NLP tasks.
- **BioBERT** (`dmis-lab/biobert-base-cased-v1.1`): Initialized from BERT-base and further pre-trained on PubMed and PMC corpora, retains cased vocabulary for biomedical entity preservation.
- **DeBERTa-v3-base** (`microsoft/deberta-v3-base`): Enhances DeBERTa with disentangled attention and ELECTRA-style pretraining; provides improved contextual representation across multiple tasks.
- **RoBERTa-base** (`roberta-base`): Robustly optimized variant of BERT using dynamic masking, larger batch sizes, and more training data, serves as a strong general-purpose encoder baseline.

For each model, we fine-tune all the models using Low-Rank Adaptation (LoRA) technique [11] with the augmented dataset. We also fine-tune PubMedBERT with the original dataset to see the effect of the augmentation.

Using k-NN with embedding generated from LoRA-adapted model: We extract embedding from the last hidden layer of each LoRA adapted model to use with k-NN model. We only use original dataset embedding to ensure the quality of the vectors being used for k-NN training. The k-NN uses cosine similarity as the distance metric, with k=1 due to the small dataset with lots of similarity between terms. With good quality embedding, k-NN is assumed to predict the closest type to the embedding of the input term.

4. Experiments and Results

We evaluate the performance of transformer-based language models on the OBI term-typing task using two approaches: direct fine-tuning with LoRA adaptation and k-NN classification using embeddings from LoRA-adapted models. We only do single label classification to make it simple for the model without considering the confidence score threshold. Table 1 presents the F1 scores for both approaches across different transformer models.

Impact of Data Augmentation: The comparison between PubMedBERT performance on original and augmented datasets demonstrates the effectiveness of our character-level augmentation strategy. PubMedBERT shows consistent improvement with data augmentation, achieving F1 scores of 0.816 (vs. 0.770 on original) for direct

Table 1. *F1 scores of each model using as a classification model (LLM) and using as encoder for k-NN model (k-NN with embedding)*

Model	**LLM**	**k-NN with embedding**
PubMedBERT (OBI_original)	0.770	0.747
PubMedBERT (OBI_augmented)	0.816	0.804
biobert-base-cased-v1.1	0.793	0.816
deberta-v3-base	0.022	0.850
roberta-base	**0.827**	**0.862**

fine-tuning and 0.804 (vs. 0.747 on original) for k-NN classification. This indicates that the simple character-level modifications (insertion, deletion, swap, substitution) effectively help the model learn more robust representations of biomedical terms.

Direct Fine-tuning Performance: Among the transformer models evaluated, RoBERTa-base achieved the highest F1 score of 0.827 when used as a direct classifier through LoRA fine-tuning. BioBERT and PubMedBERT (augmented) also showed strong performance with F1 scores of 0.793 and 0.816, respectively. Notably, DeBERTa-v3-base performed extremely poorly in direct fine-tuning (F1 = 0.022), suggesting that the LoRA adaptation may not be suitable for this particular model architecture on the OBI dataset, possibly due to overfitting or incompatibility with the fine-tuning approach.

k-NN with Embedding Performance: The k-NN approach using embeddings from LoRA-adapted models revealed interesting patterns. RoBERTa-base achieved the highest performance with an F1 score of 0.862, which notably exceeds its direct fine-tuning performance (0.827). This suggests that RoBERTa's learned representations are highly effective for similarity-based classification and that the k-NN approach with k=1 and cosine similarity is well-suited for capturing term relationships in the embedding space.

The most dramatic improvement was observed with DeBERTa-v3-base, which achieved an F1 score of 0.850 with k-NN despite its poor direct fine-tuning performance. This indicates that while DeBERTa-v3 may struggle with the LoRA fine-tuning process, it generates high-quality embeddings that encode meaningful semantic information for biomedical terms. BioBERT also showed improvement with k-NN (0.816 vs. 0.793), while PubMedBERT models showed slight decreases when using k-NN compared to direct fine-tuning.

Embedding Quality Analysis: The superior performance of k-NN with certain models (RoBERTa, DeBERTa-v3, BioBERT) suggests that these transformer architectures learn embedding representations that effectively capture semantic similarities between biomedical terms. The use of k=1 in our k-NN approach appears justified given the small dataset size and high similarity between terms in the OBI dataset. The cosine similarity metric effectively captures the semantic relationships encoded in the high-dimensional embedding space.

These results demonstrate that transformer embeddings can serve as powerful feature representations for traditional machine learning approaches, sometimes outperforming the models' own fine-tuned classification capabilities in the case of OBI dataset provided.

MatOnto and SWEET: We also do experiments on other datasets provided by the 2nd LLMs4OL challenge, MatOnto and SWEET. We train RoBERTa, which gives us the best result on the OBI dataset, with both the original and augmented version of both

Table 2. *F1 scores of RoBERTa using as a classification model (LLM) and using as encoder for k-NN model (k-NN with embedding) on MatOnto and SWEET dataset*

Dataset	RoBERTa	k-NN with embedding
MatOnto_original	0.108	0.027
MatOnto_augmented	0.189	0.027
SWEET_original	**0.519**	**0.062**
SWEET_augmented	0.504	0.060

the datasets. The experiment setting stays the same as the previous one of OBI, using RoBERTa as the classifier and the embedding with k-NN (k=1).

From the results in Table 2, we can conclude that our augmentation methods do not suit the MatOnto and SWEET datasets as effectively as they did for OBI. In fact, for MatOnto, the RoBERTa model performed very poorly overall, with an F1 score of only 0.108 and 0.189 on the original dataset and augmented dataset accordingly with no improvement when using augmented data or k-NN classification. The embeddings extracted for k-NN classification yielded even worse performance (F1 = 0.027).

For the SWEET dataset, direct fine-tuning of RoBERTa resulted in moderate performance (F1 = 0.519 on the original and 0.504 on the augmented version), while k-NN classification severely underperformed. This gap between direct classification and embedding-based classification suggests that although the model can be adapted to the task to some extent through LoRA fine-tuning, the embeddings extracted may not effectively encode semantic similarity among SWEET terms. The poor performance of k-NN could stem from two main factors: (1) the model may fail to produce meaningful and discriminative embeddings for this domain, and (2) the use of a low k-value (k=1) may not be robust enough given the diversity or imbalance in the label distribution.

Overall, these findings suggest that the effectiveness of transformer-based embeddings for k-NN classification is highly dataset-dependent. While the OBI dataset benefited greatly from both augmentation and embedding-based k-NN, the same strategies fail to generalize to MatOnto and SWEET, underscoring the need for tailored adaptation and deeper understanding of dataset-specific characteristics.

5. Conclusion and Limitations

This paper presents a study on combining large language models (LLMs) with k-NN for term-typing in the OBI dataset. Our experiments demonstrate that embeddings derived from fine-tuned transformer models can be highly effective for k-nearest neighbors (k-NN) classification, sometimes even outperforming the models’ own direct classification capabilities.

Key findings include: (1) Domain-specific models such as PubMedBERT and BioBERT serve as strong baselines for biomedical text classification tasks; (2) Simple character-level data augmentation improves model robustness and classification performance in OBI dataset, and (3) Combining LLM embeddings with k-NN leads to notable gains in classification accuracy, particularly in models like DeBERTa-v3, which struggled in direct fine-tuning with OBI dataset.

However, there are important limitations to this study. Firstly, all evaluations were conducted under the constraints of the LLMs4OL 2025 challenge, which included limited evaluation quota. This restricted the ability to perform broader, more comprehensive testing of multiple fine-tuning strategies and validation runs. Secondly, we only consider single label classification, meaning that our method is not suitable for the scenario where there exist more than one types for a specific term. Finally, our method performs well with only a specific dataset (OBI) and badly on the others.

Future work should expand this approach to other datasets and tasks in ontology learning and consider scaling k-NN for larger datasets and test the effect of k-value systematically [12]. Exploring alternative similarity metrics, integrating ensemble strategies, and improving interpretability of classification results. Other augmentation methods should also be considered to scale-up the dataset in a more meaningful way as the current simple method does not apply well with MatOnto and SWEET dataset. The future work should also try implementing this method with newer models like Qwen3 or Nomic-AI-based embedding models to observe whether newer embedding-based models have the same behavior as those used in this paper or not.

Despite these limitations, this study provides evidence that LLM embeddings can serve as strong semantic representations for hybrid classification approaches and may complement or even outperform direct fine-tuning methods in certain cases.

Data availability statement

The datasets used in this study are publicly available through the 2nd LLMs4OL challenge.

Underlying and related material

Model implementations and experimental configurations are available in our GitHub repository[1].

Author contributions

Chavakan Yimmark: Conceptualization, Methodology, Software Development, Data Analysis, and Writing – Original Draft.
Teeradej Racharak: Supervision, Guidance on Methodology, Reviewing, and Editing – Final Manuscript.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

We thank the organizers of the 2nd LLMs4OL challenge for providing the dataset. We also acknowledge the computational resources provided by Japan Advanced Institute of Science and Technology. We thank Gregory Schwartzman for helpful discussions and advice.

[1] https://github.com/chavakan209/T-GreC-at-LLMs4OL-2025-Task-B

References

[1] 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.

[2] J. Lee, W. Yoon, S. Kim, D. Kim, C. H. So, and J. Kang, “Biobert: A pre-trained biomedical language representation model for biomedical text mining”, *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.

[3] P. Cunningham and S. J. Delany, “K-Nearest Neighbour Classifiers: 2nd Edition (with Python examples)”, en, *ACM Computing Surveys*, vol. 54, no. 6, pp. 1–25, Jul. 2022, arXiv:2004.04523 [cs], ISSN: 0360-0300, 1557-7341. DOI: 10.1145/3459665. Accessed: Aug. 8, 2025. [Online]. Available: http://arxiv.org/abs/2004.04523.

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

[5] P. He, X. Gao, J. Chen, and J. Gao, “Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing”, *arXiv preprint arXiv:2111.09543*, 2021.

[6] Y. Liu et al., “Roberta: A robustly optimized bert pretraining approach”, *arXiv preprint arXiv:1907.11692*, 2019.

[7] 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. Accessed: Aug. 6, 2025. [Online]. Available: http://arxiv.org/abs/2409.10146.

[8] P. Kumar Goyal, S. Singh, and U. Shanker Tiwary, “Silp_nlp at LLMs4OL 2024 Tasks A, B, and C: Ontology Learning through Prompts with LLMs”, en, *Open Conference Proceedings*, vol. 4, pp. 31–38, Oct. 2024, ISSN: 2749-5841. DOI: 10.52825/ocp.v4i.2485. Accessed: Aug. 7, 2025. [Online]. Available: https://www.tib-op.org/ojs/index.php/ocp/article/view/2485.

[9] C. Eang and S. Lee, “Improving the Accuracy and Effectiveness of Text Classification Based on the Integration of the Bert Model and a Recurrent Neural Network (RNN_bert_based)”, en, *Applied Sciences*, vol. 14, no. 18, p. 8388, Sep. 2024, ISSN: 2076-3417. DOI: 10.3390/app14188388. Accessed: Aug. 7, 2025. [Online]. Available: https://www.mdpi.com/2076-3417/14/18/8388.

[10] 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.

[11] E. J. Hu et al., “Lora: Low-rank adaptation of large language models”, *arXiv preprint arXiv:2106.09685*, 2022.

[12] S. Okamoto and K. Satoh, “An average-case analysis of k-nearest neighbor classifier”, en, in *Case-Based Reasoning Research and Development*, J. G. Carbonell et al., Eds., vol. 1010, Series Title: Lecture Notes in Computer Science, Berlin, Heidelberg: Springer Berlin Heidelberg, 1995, pp. 253–264, ISBN: 978-3-540-60598-0 978-3-540-48446-2. DOI: 10.1007/3-540-60598-3_23. Accessed: Aug. 8, 2025. [Online]. Available: http://link.springer.com/10.1007/3-540-60598-3_23.