

Ontology-Guided On-Device Conversational Knowledge Capture with Large Language Models

Tolga Çöplü, Arto Bendiken, Andrii Skomorokhov, Eduard Bateiko and Stephen Cobb

Haltia, Inc.

Abstract

Generative AI applications must integrate users' personal information into the response generation process to offer an advanced user experience. One of the most effective methods for obtaining accurate and current user information is by capturing this data from AI interactions. This paper examines conversational knowledge capture using ontology and knowledge-graph approaches. We propose enhancing the large language model's (LLM) ability to capture precise and relevant information by training it with a subset of the KNOW ontology, which models personal knowledge. Our paper details the ontology-guided training process and evaluates the success of knowledge capture using a specially constructed dataset. Additionally, we emphasize the importance of privacy in handling personal information and investigate the implementation of knowledge capture with on-device language models. Our findings highlight the potential of on-device solutions to effectively capture personal knowledge while preserving user privacy.

1. Introduction

Expectations for the quality and sophistication of human-AI interactions are steadily increasing. Generative AI applications are now expected to recognize users, understand their characteristics and preferences, and augment this information to enhance interactions. A fundamental challenge in providing this level of user experience is capturing up-to-date knowledge about the user through conversations. This process of identifying and recording personal knowledge and preferences from user interactions is defined as **conversational knowledge capture** (CKC).

CKC presents several critical challenges. Key issues include determining which knowledge from conversations should be captured, how the captured knowledge should be represented, whether the captured knowledge requires updating previous records, and whether the knowledge is duplicate. Fortunately, the emergence of neurosymbolic approaches, which combine large language models (LLMs) and symbolic AI, has provided researchers with new perspectives to address these challenges [1, 2, 3, 4]. LLMs' capabilities in natural language processing can be integrated with the knowledge representation and factual reasoning abilities of knowledge graphs, enhanced by the structure, rules, and inference mechanisms offered by an ontology.


Another significant challenge related to CKC is ensuring the privacy of captured sensitive knowledge. Personal data, which is entirely owned by the user, should be considered vulnerable

KBC-LM'24: Knowledge Base Construction from Pre-trained Language Models workshop at ISWC 2024

✉ tolga@haltia.ai (T. Çöplü); arto@haltia.ai (A. Bendiken); andriy@haltia.ai (A. Skomorokhov); eduard@haltia.ai (E. Bateiko); steve@haltia.ai (S. Cobb)



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

if it is sent to the cloud. On-device AI solutions, which do not require any data to leave the user’s device, provide the most appropriate response to privacy needs [5]. Capturing knowledge with the support of a local language model running on the device, securely storing this information in a local knowledge base, and utilizing it on the device when needed, provides a suitable environment for maintaining the privacy of personal knowledge. However, on-device language models come with their own limitations. The size, capabilities, and power consumption of language models running on personal devices such as smartphones, tablets, and computers must be carefully considered [6]. Fortunately, remarkable developments are emerging every day. Thanks to R&D efforts, LLMs with fewer parameters are now offering faster responses and improved performance compared with older large models.

In this paper, we explore the feasibility of generating personal knowledge graphs on-device through conversational interaction. Our approach focuses on ontology-guided knowledge extraction from prompts in the form of subject-predicate-object triples¹. We have investigated various methods to enable the underlying language model to comprehend a predefined ontology, ensuring effective personal knowledge-graph generation. Subsequently, we selected the most suitable method based on the requirements of on-device execution. Utilizing a specially designed dataset, we evaluate the effectiveness of this method, emphasizing its strengths and identifying potential areas for improvement.

The structure of this paper is as follows: Section 2 discusses various approaches, including in-context learning and fine-tuning for ontology-guided knowledge capture, and focuses on the fine-tuning approach due to its suitability for on-device execution. Section 3 describes the experimental setup, presenting the development framework, language model selection, and the ontology and dataset creation process. Section 4 outlines our performance evaluation framework and the test results. Finally, Section 5 concludes the paper and suggests future directions.

2. Ontology-Guided Symbolic Knowledge Capture

In the literature, language models have demonstrated their capability to transform unstructured text into knowledge graphs [7, 8, 9, 10, 11]. However, the process of populating a knowledge graph from user prompts in alignment with a predefined ontology has been explored only marginally [12, 13, 14, 15, 16, 17]. Except for [17], these studies have enjoyed unconstrained processing and memory capacity. Large models with large context windows have enabled in-context learning methods relying on prompt engineering. However, on-device conversational knowledge capture is not similarly unconstrained. Given current context-window capacities, embedding an entire personal ontology into the system prompt would be unrealistic. Additionally, considering the inference speed of language models running on personal devices, the high token overhead introduced by this would present a barrier to efficient system operation.

An alternative to in-context learning involves training a language model with a predefined ontology so that the model internalizes it. There are two strategies to consider: pretraining the LLM on the ontology or fine-tuning it. This paper does not explore pretraining due to its extensive data, computational resource, energy, and time requirements. Additionally, pretraining

¹<https://www.w3.org/TR/rdf12-concepts/>

does not offer a flexible response to ongoing changes or expansions in the ontology. Therefore, this paper focuses on fine-tuning as a method to train language models on personal ontologies, highlighting its advantages in feasibility and maintainability.

Fine-tuning is a process whereby a pretrained language model is further trained on a specific dataset to tailor its capabilities to a particular task. In our study, the language model is expected to understand the ontology classes and their properties, and use them to populate a knowledge graph from user prompts. The first step involves preparing a fine-tuning dataset, which includes user prompts, system prompts, and expected model responses for each concept in the ontology. This dataset is used to fine-tune the language model, which is then evaluated by testing it with new prompts to assess the effectiveness of the CKC process.

The following points highlight the key aspects of ontology fine-tuning:

- The training dataset’s **coverage and diversity** are vital for successful fine-tuning. These characteristics greatly influence the LLM’s ability to capture knowledge effectively. Details about the dataset and how it is constructed are discussed in Section 3.4.
- The training dataset must include a **variety of examples** for the predefined ontology. Research related to the structure of the examples prepared for ontology concepts is detailed in Section 4.
- If the language model encounters a user prompt that is not relevant to the predefined ontology concepts, it should not attempt to capture knowledge. Therefore, the dataset should also contain sufficient **out-of-context samples** to enable the language model to distinguish between relevant and irrelevant information for capture.

3. Experimental Setup

3.1. Development Framework

The methods suggested in this paper have been implemented using the Apple MLX framework [18]. MLX is a specialized array framework designed for machine learning applications, akin to NumPy, PyTorch, or JAX, with the distinction of being exclusive to Apple silicon.

Ontology fine-tuning has been conducted using the parameter-efficient QLoRA adapters [19] on our custom dataset, comprising randomly selected, non-overlapping sets of training, validation, and test samples.

3.2. Language Model

Due to the constraint of on-device execution, our study does not use state-of-the-art large-parameter cloud-based language models. Instead, we opted for a relatively low-parameter model with proven effectiveness across diverse domains. Based on its performance in the Hugging Face Open LLM Leaderboard [20] and its robust ecosystem, we selected Mistral-7B-Instruct-v0.2 [21], which is based on the Llama 2 [22] architecture. The MLX 4-bit quantized version, with a disk size of 4.26 GB, stands out as a suitable model for many personal computers, tablets, and even new-generation smartphones.

3.3. Applied Ontology

Our study is inspired by KNOW[23]—the Knowledge Navigator Ontology for the World—and utilizes it for representing personal information. KNOW was introduced as a pioneering framework designed to capture everyday knowledge to enhance language models in real-world generative AI applications such as personal AI assistants. The ontology focuses on human life, encompassing everyday concerns and significant milestones, and limits its initial scope to established human universals, including spacetime (places, events) and social dimensions (people, groups, organizations). This pragmatic approach emphasizes universality and utility, contrasting with previous works like Schema.org[24] and Cyc[25] by building on language models' inherent encoding of salient commonsense knowledge.

Because of the requirement that each element in the ontology be associated with a diverse set of prompt and response samples within the training dataset, our research focuses on a specific subset of the KNOW ontology. This subset concentrates on core family relationships with four ontology classes, eleven object properties, and one data property. A visual depiction of this subset is presented in Figure 1.

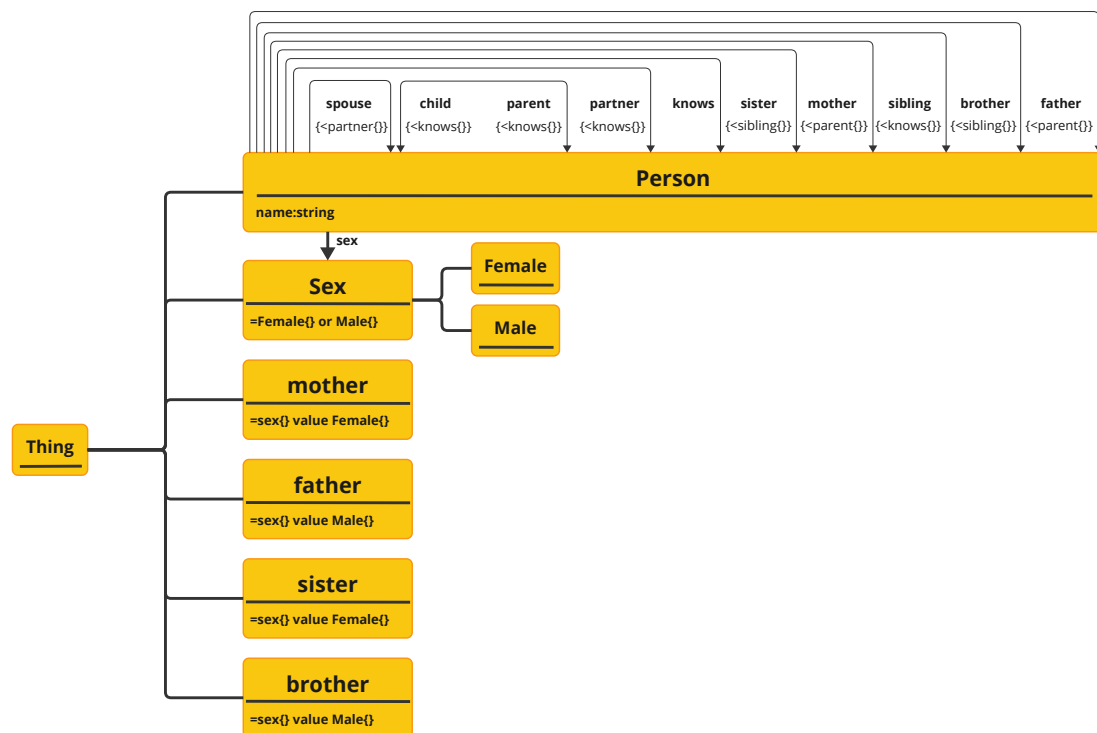


Figure 1: A visual representation of the ontology design used in this paper

3.4. Dataset

For a language model to effectively learn a predefined ontology and use it to perform knowledge extraction and capture, a robust and diverse training dataset is essential. Our paper focuses on a subset of the KNOW ontology that includes the concepts of ‘person’, ‘name’, ‘sex’, ‘child’, ‘father’, ‘mother’, ‘sibling’, ‘sister’, ‘brother’, ‘spouse’, ‘partner’, and ‘knows’. We created 145 manually crafted user prompts along with their respective ontology responses for training. Additionally, to manage inputs that fall outside these ontology concepts, we included 32 generic user prompts in the dataset. The composition of this training dataset, which consists of 177 user prompts, is illustrated in Figure 2. Concepts not associated with the ontology are labeled as the ‘none’ in the figure. Since each sample prompt usually includes multiple concepts, the chart shows more concept occurrences than prompts.

For the test set to be used in the evaluation, we generated 100 manual test prompts and their expected responses about family relations based on the television series ‘The Waltons’. The Waltons is a classic American television series that aired from 1972 to 1981, depicting the life and challenges of a close-knit family in rural Virginia during the Great Depression and World War II. The show focuses on the daily experiences, values, and growth of the Walton family, emphasizing themes of love, perseverance, and community. The composition of this test dataset, which consists of 100 user prompts, is illustrated in Figure 2.

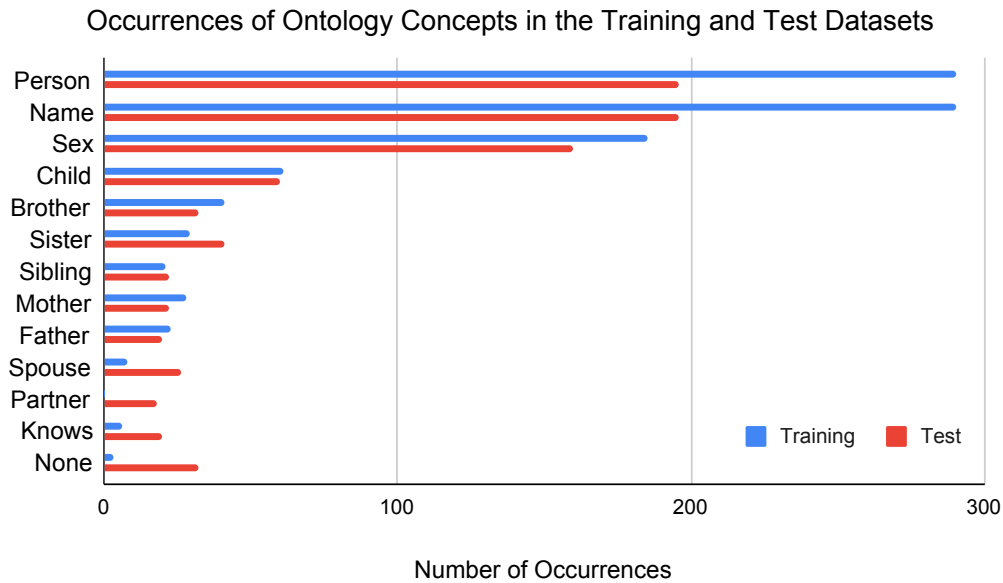


Figure 2: Occurrences of ontology concepts in the training and test datasets

The Turtle format was chosen for serializing the ontology population in our research because of its straightforward structure, readability, and prevalent use in existing pretraining datasets for LLMs.

4. Performance Evaluation

Our research focuses on fine-tuning an on-device language model with predefined ontology concepts and capturing knowledge from user prompts that fit the ontology. This section will detail the fine-tuning approaches and the relevant performance evaluations.

One of our main objectives is to comprehend the training dynamics of ontology-guided knowledge capture. Consequently, we will not emphasize the optimization of fine-tuning hyperparameters. For a fair performance evaluation, all tests in this section were conducted using the default QLoRA hyperparameters specified in the MLX framework. To ensure consistency in the test results, each training session was configured to run for 18 epochs.

Firstly, we investigated whether training data associated with each ontology concept is necessary for the successful fine-tuning process. For instance, we evaluated the ability of a language model, trained with only one of two semantically related concepts (e.g., 'brother'), to capture knowledge related to the concept that was not included in the training data (e.g., 'sister'). During the evaluations, the generated prompt responses were processed triple by triple and compared against the ground truth established for the test set. The findings are presented in Figure 3.

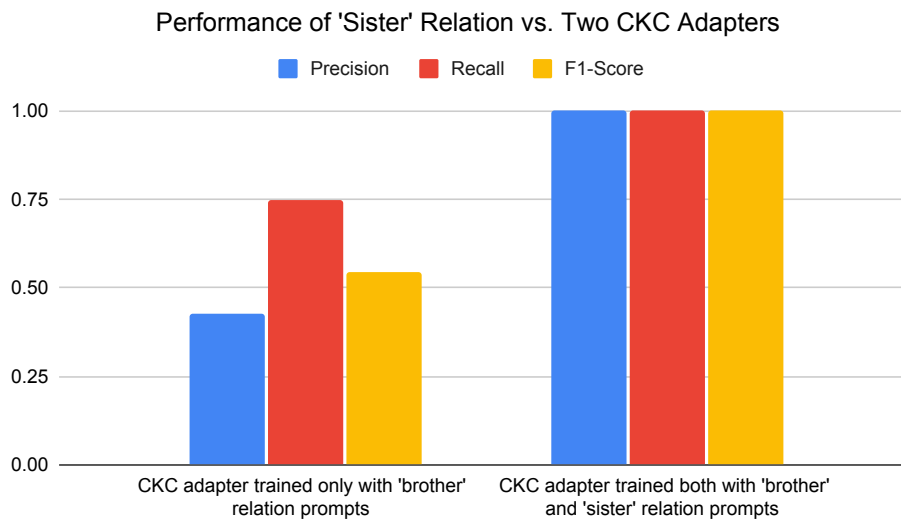


Figure 3: CKC performance (precision, recall, and f1-score) of 'sister' relation prompts for two differently fine-tuned QLoRA adapters

As illustrated in Figure 3, a CKC QLoRA adapter trained exclusively on the 'brother' concept shows a significant decline in performance when tested with the 'sister' relation prompts. In contrast, an adapter trained on both the 'brother' and 'sister' concepts demonstrates excellent performance on the same test set. Although the language model's pretraining is sufficient to distinguish between the 'brother' and 'sister' concepts, our tests reveal that this is inadequate for effective knowledge capture.

The second issue we examined was whether it is necessary for each concept to be used in conjunction with other concepts in the fine-tuning dataset. To provide a concrete example, we assessed the performance of fine-tuning using samples that included only the ‘brother’ or only the ‘sister’ relationship when prompted with contexts where both the ‘brother’ and ‘sister’ concepts co-occur. We compared the performance of a knowledge capture adapter fine-tuned with samples containing only the ‘brother’ and ‘sister’ concepts against another adapter trained with the same data set, with extra samples where both ‘brother’ and ‘sister’ concepts appear together. The results of this comparison are presented below in Figure 4.

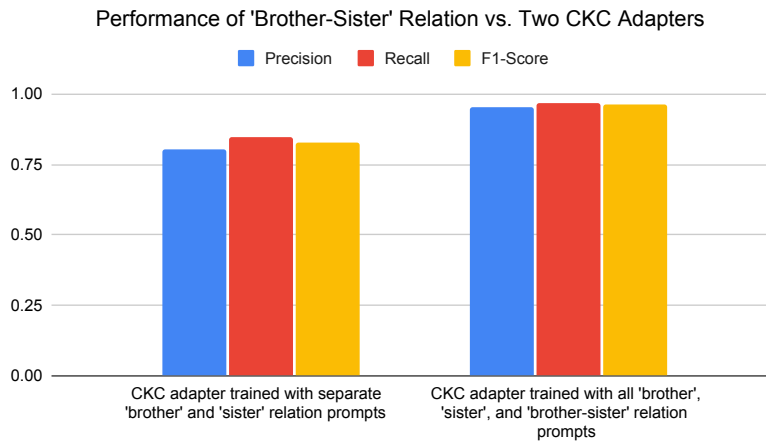


Figure 4: CKC performance (precision, recall, and f1-score) of ‘brother-sister’ relation prompts for two differently fine-tuned QLoRA adapters

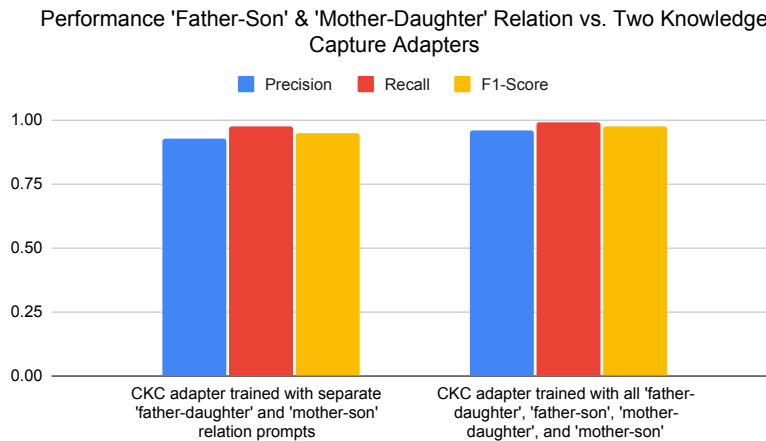


Figure 5: CKC performance (precision, recall, and f1-score) of ‘father-son’ and ‘mother-daughter’ relation prompts for two differently fine-tuned QLoRA adapters

Similarly, in Figure 5, we presented the knowledge capture performance of two different adapters, one trained with only ‘father-daughter’ and ‘mother-son’ prompts and the other trained with all ‘father-daughter’, ‘father-son’, ‘mother-daughter’, and ‘mother-son’ prompts. Both Figure 4 and Figure 5 demonstrate that having cross relations between ontology concepts in the training set increases knowledge capture performance. In the future, we aim to repeat these tests with various ontologies and language models to obtain more general results regarding this approach.

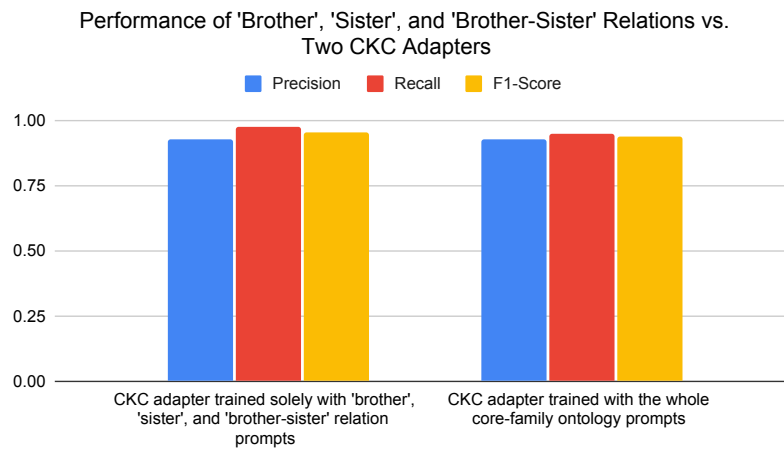


Figure 6: CKC performance (precision, recall, and f1-score) of ‘brother’, ‘sister’, and ‘brother-sister’ relation prompts for two differently fine-tuned QLoRA adapters

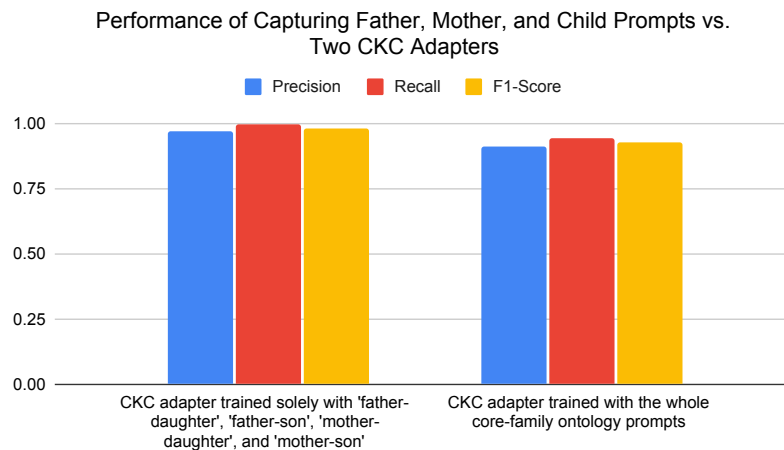


Figure 7: CKC performance (precision, recall, and f1-score) of ‘father-daughter’, ‘father-son’, ‘mother-daughter’, and ‘mother-son’ relation prompts for two differently fine-tuned QLoRA adapters

Another aspect we researched in our study was the impact of the ontology size used in training on knowledge capture performance. In our previous tests, we trained the language model on specific ontology concepts and then examined the knowledge capture performance of the resulting adapter. At this stage, we compared the performance of a knowledge capture adapter trained solely on the ontology concepts present in the test set with an adapter trained on the entire core-family ontology. Our objective was to investigate whether adapters experience any degradation in performance as they are trained on more ontology concepts. In other words, we observed the scalability of ontology learning for the on-device language model we used. The results for the two different groups are presented in Figures 6 and 7. As seen from the figures, as the number of concepts in ontology training increases, there is a slight decrease in the performance of the knowledge capture adapter. However, the observed degradation within the scope of the tests does not appear to lead to a major scalability problem.

In the subsequent phase, we explored the optimal number of training epochs required to achieve maximum performance for the training set. For this analysis, we continued using the default MLX QLoRA hyperparameters, but trained the QLoRA adapter over various epoch lengths. We then conducted evaluations on the test set using each trained adapter, and the findings are presented in Figure 8.



Figure 8: Ontology population performance (precision, recall, and f1-score) of the core-family ontology for various epochs

As depicted in Figure 8, the success rate of the ontology population increases with longer training. However, considering the resource usage and energy consumption, we observe that 18 epochs is sufficient for fine-tuning.

5. Conclusion

In this paper, we first explored on-device ontology-guided conversational knowledge capture and its importance in the generative AI domain. We then discussed the ontology approach and how to train an on-device LLM with ontology concepts. The language model was fine-tuned using a custom dataset focused on core family relationships, and we evaluated the model's ability to learn personal ontology concepts.

Our findings indicate that fine-tuning is particularly effective for training an on-device language model with ontology concepts for conversational knowledge capture. In our future work, we aim to integrate the generated knowledge graph with the language model for knowledge utilization, combining the strengths of the neural and symbolic AI approaches.

References

- [1] A. Sheth, K. Roy, M. Gaur, Neurosymbolic AI – Why, What, and How, 2023. URL: <http://arxiv.org/abs/2305.00813>, arXiv:2305.00813 [cs].
- [2] L. N. DeLong, R. F. Mir, J. D. Fleuriot, Neurosymbolic AI for Reasoning over Knowledge Graphs: A Survey, 2024. URL: <http://arxiv.org/abs/2302.07200>, arXiv:2302.07200 [cs, stat].
- [3] F. J. Ekaputra, M. Llugiqi, M. Sabou, A. Ekelhart, H. Paulheim, A. Breit, A. Revenko, L. Waltersdorfer, K. E. Farfar, S. Auer, Describing and Organizing Semantic Web and Machine Learning Systems in the SWeMLS-KG, 2023. URL: <http://arxiv.org/abs/2303.15113>, arXiv:2303.15113 [cs].
- [4] L.-P. Meyer, C. Stadler, J. Frey, N. Radtke, K. Junghanns, R. Meissner, G. Dziwis, K. Bulert, M. Martin, LLM-assisted Knowledge Graph Engineering: Experiments with ChatGPT, 2023. URL: <http://arxiv.org/abs/2307.06917>, arXiv:2307.06917 [cs].
- [5] E. Marin, D. Perino, R. Di Pietro, Serverless Computing: A Security Perspective, 2022. URL: <http://arxiv.org/abs/2107.03832>. doi:10.48550/arXiv.2107.03832, arXiv:2107.03832 [cs].
- [6] T. Çöplü, M. Loedi, A. Bendiken, M. Makohin, J. J. Bouw, S. Cobb, A Performance Evaluation of a Quantized Large Language Model on Various Smartphones, 2023. URL: <https://arxiv.org/abs/2312.12472>. arXiv:2312.12472.
- [7] Q. Guo, Z. Jin, X. Qiu, W. Zhang, D. Wipf, Z. Zhang, CycleGT: Unsupervised Graph-to-Text and Text-to-Graph Generation via Cycle Training, 2020. URL: <http://arxiv.org/abs/2006.04702>. doi:10.48550/arXiv.2006.04702, arXiv:2006.04702 [cs].
- [8] X. Li, F. Yin, Z. Sun, X. Li, A. Yuan, D. Chai, M. Zhou, J. Li, Entity-Relation Extraction as Multi-Turn Question Answering, 2019. URL: <http://arxiv.org/abs/1905.05529>. doi:10.48550/arXiv.1905.05529, arXiv:1905.05529 [cs].
- [9] Y. Xu, L. Fu, Z. Lin, J. Qi, X. Wang, INFINITY: A Simple Yet Effective Unsupervised Framework for Graph-Text Mutual Conversion, 2022. URL: <http://arxiv.org/abs/2209.10754>. doi:10.48550/arXiv.2209.10754, arXiv:2209.10754 [cs].
- [10] R. Anantharangachar, S. Ramani, S. Rajagopalan, Ontology Guided Information Extraction from Unstructured Text, International journal of Web & Semantic Technology 4 (2013) 19–36. URL: <http://arxiv.org/abs/1302.1335>. doi:10.5121/ijwest.2013.4102, arXiv:1302.1335 [cs].

- [11] T. Çöplü, A. Bendiken, A. Skomorokhov, E. Bateiko, S. Cobb, J. J. Bouw, Prompt-Time Symbolic Knowledge Capture with Large Language Models, 2024. URL: <http://arxiv.org/abs/2402.00414>. doi:10.48550/arXiv.2402.00414, arXiv:2402.00414 [cs].
- [12] H. B. Giglou, J. D’Souza, S. Auer, LLMs4OL: Large Language Models for Ontology Learning, 2023. URL: <http://arxiv.org/abs/2307.16648>. doi:10.48550/arXiv.2307.16648, arXiv:2307.16648 [cs, math].
- [13] P. Mateiu, A. Groza, Ontology engineering with Large Language Models, 2023. URL: <http://arxiv.org/abs/2307.16699>, arXiv:2307.16699 [cs].
- [14] K. Lu, X. Pan, K. Song, H. Zhang, D. Yu, J. Chen, PIVOINE: Instruction Tuning for Open-world Information Extraction, 2023. URL: <http://arxiv.org/abs/2305.14898>. doi:10.48550/arXiv.2305.14898, arXiv:2305.14898 [cs].
- [15] N. Mihindukulasooriya, S. Tiwari, C. F. Enguix, K. Lata, Text2KGBench: A Benchmark for Ontology-Driven Knowledge Graph Generation from Text, 2023. URL: <http://arxiv.org/abs/2308.02357>. doi:10.48550/arXiv.2308.02357, arXiv:2308.02357 [cs].
- [16] M. Funk, S. Hosemann, J. C. Jung, C. Lutz, Towards Ontology Construction with Language Models, 2023. URL: <http://arxiv.org/abs/2309.09898>. doi:10.48550/arXiv.2309.09898, arXiv:2309.09898 [cs].
- [17] T. Çöplü, A. Bendiken, A. Skomorokhov, E. Bateiko, S. Cobb, Prompt-Time Ontology-Driven Symbolic Knowledge Capture with Large Language Models, 2024. URL: <https://arxiv.org/abs/2405.14012>. arXiv:2405.14012.
- [18] A. Hannun, J. Digani, A. Katharopoulos, R. Collobert, MLX: Efficient and flexible machine learning on Apple silicon, <https://github.com/ml-explore>, 2023.
- [19] T. Dettmers, A. Pagnoni, A. Holtzman, L. Zettlemoyer, QLoRA: Efficient Finetuning of Quantized LLMs, 2023. URL: <http://arxiv.org/abs/2305.14314>. doi:10.48550/arXiv.2305.14314, arXiv:2305.14314 [cs].
- [20] E. Beeching, C. Fourrier, N. Habib, S. Han, N. Lambert, N. Rajani, O. Sanseviero, L. Tunstall, T. Wolf, Open LLM Leaderboard, https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard, 2023.
- [21] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. I. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, W. E. Sayed, Mistral 7B, 2023. URL: <http://arxiv.org/abs/2310.06825>. doi:10.48550/arXiv.2310.06825, arXiv:2310.06825 [cs].
- [22] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, T. Scialom, Llama 2: Open foundation and fine-tuned chat models, 2023. arXiv:2307.09288.
- [23] A. Bendiken, KNOW: A Real-World Ontology for Knowledge Capture with Large Language Models, 2024. URL: <https://arxiv.org/abs/2405.19877>. arXiv:2405.19877.

- [24] R. V. Guha, D. Brickley, S. Macbeth, Schema.org: evolution of structured data on the web, *Communications of the ACM* 59 (2016) 44–51. URL: <https://doi.org/10.1145/2844544>. doi:10.1145/2844544.
- [25] D. B. Lenat, R. V. Guha, *Building Large Knowledge-Based Systems; Representation and Inference in the Cyc Project*, 1st ed., Addison-Wesley Longman Publishing Co., Inc., USA, 1989.

A. Online Resources

Please refer to the paper's corresponding GitHub repository at <https://github.com/HaltiaAI/paper-OGODCKC>