Using Generative Large Language Models for Hierarchical Relationship Prediction in Medical Ontologies

Hao Liu
School of Computing
Montclair State University
Montclair, USA
liuha@montclair.edu

Shuxin Zhou
Computer Science Dept.
New Jersey Institute of Tech
Newark, USA
sz23@njit.edu

Zhehuan Chen
Biomedical Engineering Dept.
Stevens Institute of Tech
Hoboken, USA
zchen28@stevens.edu

Yehoshua Perl
Computer Science Dept.
New Jersey Institute of Tech
Newark, USA
perl@njit.edu

Jiayin Wang
School of Computing
Montclair State University
Montclair, USA
wangji@montclair.edu

2024 IEEE 12th International Conference on Healthcare Informatics (ICHI) | 979-8-3503-8373-7/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICHI61247.2024.00040

***Abstract*—This study extends the exploration of ontology enrichment by evaluating the performance of various open-sourced Large Language Models (LLMs) on the task of predicting hierarchical relationships (IS-A) in medical ontologies including SNOMED CT Clinical Finding and Procedure hierarchies and the human Disease Ontology. With the previous finetuned BERT models for hierarchical relationship prediction as the baseline, we assessed eight open-source generative LLMs for the same task. We observed only three models, without finetuning, demonstrated comparable or superior performance compared to the baseline BERT-based models. The best performance model OpenChat achieved a macro average F1 score of 0.96 (0.95) on SNOMED CT *Clinical Finding* (*Procedure*) hierarchy, an increase over 7% from the baseline 0.89 (0.85). On human Disease Ontology, OpenChat excels with an F1 score of 0.91, outperforming the second-best performance model Vicuna (0.84). Notably, some LLMs prove unsuitable for hierarchical relationship prediction tasks or appliable for concept placement of medical ontologies. We also explored various prompt templates and ensemble techniques to uncover potential confounding factors in applying LLMs for IS-A relation predictions for medical ontologies.**

Keywords—Hieratical Relation Prediction, Large Language Models, Medical Ontology, Prompts Design, SNOMED CT

I. Introduction

Biomedical ontologies provided structured and formal representation of biomedical concepts and their interrelationships. The hierarchical relationship, also known as the "IS-A" relationship, is a fundamental concept within ontologies. It indicates that one concept is a subtype or subclass of another. In the context of biomedical ontologies, the IS-A relationship is crucial for organizing and structuring biomedical knowledge hierarchically. For example, the IS-A relationship would indicate that "Type 2 Diabetes" is a kind of or subclass of "Diabetes." As biomedical knowledge is constantly growing, the process of enriching medical ontologies with newly defined concepts, poses a multifaceted challenge in biomedical informatics. This process, specifically the placement of new concepts, demands more attention to hierarchical relationships. The intricate task of accurately placing new concepts within the existing ontology hierarchy is equivalent to determining if there should exist IS-A relationships between two concepts. Traditional approaches, often reliant on classifiers such as Snorocket [1] and HermiT [2], encounter limitations in handling the complexity and dynamic nature of evolving ontologies. The conventional reliance on classifiers, while providing a structured approach, falls short in cases where relationships are underspecified, or the ontology lacks the necessary granularity [3-5]. For example, SNOMED CT [6], as a widely used clinical terminology that is constantly updated with primitive concepts and underspecified relationships. The need for more efficient and accurate methodologies becomes evident, especially considering the time-consuming and error-prone nature of manually placing concepts.

A few studies have leveraged machine learning techniques for validate or identify missing IS-A relationships in SNOMED CT. Sun et al. investigated using deep learning models to aid in automatically validating IS-A relations suggested by nonlattice-based auditing approaches in SNOMED CT [7]. We also leveraged deep learning models, including Convolutional Neural Network [8] and BERT [9] to place new concepts in the SNOMED CT hierarchy using the new concepts' names. More recently, Hao et al. explored a logical definition-based approach to facilitate suggestion of new concepts for SNOMED CT [10].

Recently, Large Language Models (LLMs) have emerged as powerful tools in natural language processing and

2575-2634/24/$31.00 ©2024 IEEE
DOI 10.1109/ICHI61247.2024.00040

knowledge representation. Their ability to capture complex semantical patterns and relationships in data makes them promising candidates for addressing the challenges inherent in ontology curation. Pre-trained language models, such as BERT [11] or ELMo [12], are built using deep learning techniques, specifically a type of neural network architecture called transformer networks. These models were also applied in ontology curation to assist concept placement. In our previous work [13], we leveraged BERT to train a model to predict the IS-A relations between concepts in SNOMED CT, utilizing the next sentence prediction property of BERT.

As the field of NLP continues to evolve, LLMs represent a cornerstone in the pursuit of achieving human-level language understanding and generation. Generative LLMs have gained significant attention for their impressive language understanding and generation capabilities, influencing from natural language processing and content generation to customer support and education. LLMs are a type of artificial intelligence (AI) model designed to understand and generate human-like text. Generative LLMs such as Llama2 [14], Vicuna [15], Mistral [16], OpenChat [17], offer the potential to automate and enhance the process of predicting hierarchical relationships within ontologies.

In this study, we focus on evaluating the applicability of various LLMs to the task of predicting hierarchical relationships within the SNOMED CT. Recognizing the drawbacks of existing methods, we propose a comprehensive evaluation of LLMs' performance, including those without fine-tuning, challenging the conventional belief that extensive training is essential. Furthermore, we extended our investigation to Disease Ontology [18], expanding the scope of applicability and the versatility of LLMs in diverse medical ontology domains. To systematically assess and compare the performance of LLMs, we experiment with four different prompt templates and two kinds of ensemble techniques. This exploration aims to uncover potential confounding factors and optimize the utilization of LLMs for the critical task of predicting IS-A relationships. Through this research, we contribute valuable insights into the advantages and challenges of incorporating LLMs in ontology enrichment and concept placement in medical informatics.

II. Background

A. Medical & Clinical Ontology

1) ***SNOMED CT***. SNOMED CT (Systematized Nomenclature of Medicine Clinical Terms) is a comprehensive and globally recognized clinical terminology, with standardized medical concepts of clinical information. SNOMED CT facilitates interoperability in healthcare data exchange and supports precise and standardized communication among healthcare professionals [19]. Its hierarchical structure organizes clinical terms into a network of relationships, allowing for nuanced representation of medical knowledge. This standardized terminology forms the backbone of various health information systems and significantly contributes to improving the quality and efficiency of healthcare delivery. Its rich network of IS-A hierarchical relationships not only organizes clinical terms but also provides a nuanced portrayal of the hierarchical connections between concepts. SNOMED CT contains more than 357,000 health care concepts with unique meanings and formal logic-based definitions organized into hierarchies, of which the largest two hierarchies are Clinical Findings and Procedure.

2) ***Disease Ontology***. Disease Ontology serves as a valuable resource in organizing and classifying information related to human diseases [18]. It provides a structured framework for representing disease entities and their interconnections. By capturing the relationships between diseases and linking them to relevant biomedical knowledge, Disease Ontology contributes to a better understanding of the complex landscape of human health and diseases. The IS-A hierarchy in Disease Ontology serves as a structured framework that enhances our comprehension of the hierarchical connections between different disease entities. Its utility extends to diverse applications, including biomedical research, clinical decision support, and the integration of disease information in bioinformatics workflows.

B. Large Language Models (LLMs)

In recent years, the field of natural language processing (NLP) has witnessed a paradigm shift propelled by the advent of Large Language Models (LLMs) [20, 21]. These models, characterized by their vast scale and immense number of parameters, have demonstrated unprecedented capabilities in summarizing and generating human-like text. The significance of LLMs lies in their ability to learn intricate patterns and representations from massive datasets through unsupervised pre-training. During this pre-training phase, these models acquire a deep knowledge of the complexities of language, capturing syntactic, semantic, and contextual nuances. Transfer learning is a key principle underpinning the success of LLMs. By initially training on diverse and unlabeled text corpora, these models gain a generalized knowledge of language that can be leveraged for downstream tasks, such as text completion, sentiment analysis, question answering, language translation, and more. The transferability of knowledge acquired during pre-training contributes to the efficiency and effectiveness of LLMs in real-world applications. In this work, we experimented with eight open sourced LLM models:

1) Llama2-7b & Llama2-13b. Llama 2 is an open-source language model from Meta AI that outperforms other open-source language models on many benchmarks, including reasoning, coding, proficiency, and knowledge tests [14]. Llama2-7b and -13b are 7-billion and 13-billion parameter models fine-tuned for chat completions.

2) WizardLM-13B V1.2. WizardLM [22] is trained from Llama2-13b and it is designed to follow complex instructions and generate coherent and fluent text in response to various inputs. It is fine-tuned on AI-evolved instructions using the

Evol+ approach. The model is pre-trained on a large corpus of text data and fine-tuned on the Llama2 dataset to generate high-quality responses to complex instructions.

3) Vicuna-13b-v1.5. Vicuna-13B [15] is an open-source conversational model trained using the LLaMa-13B model. It is fine-tuned with user-shared conversations gathered from SharedGPT. Preliminary evaluations indicate that Vicuna-13B achieves a quality exceeding 90% of ChatGPT and Google Bard. It has outperformed other models such as LLaMa and Alpaca in over 90% of instances.

4) Xwin-LM-13B-V0.2. Xwin-LM [23], built-upon on the Llama2 base models, employed open-source alignment technologies including supervised fine-tuning (SFT), reward models (RM), reject sampling, reinforcement learning from human feedback (RLHF), for training large language models. It achieved good performance on the AlpacaEval benchmark.

5) OpenChat_3.5. OpenChat [17] is an open-source language model with mixed-quality data, consisting of a small amount of expert data mixed with a large proportion of sub-optimal data, without any preference labels. It employed the C(conditioned)-RLFT, which regards different data sources as coarse-grained reward labels and learned a class-conditioned policy to leverage complementary data quality information. The OpenChat-13b model fine-tuned with C-RLFT achieves the highest average performance among all 13b open-source language models on three standard benchmarks. The code, data, and models are publicly available at https://github.com/imoneoi/openchat.

6) Mistra-7B-Instruct-v0.1. Mistral [16] model leveraged grouped-query attention (GQA) for faster inference, coupled with sliding window attention (SWA) to effectively handle sequences of arbitrary length with a reduced inference cost. The Mistral 7B outperforms the Llama2-13B model across all evaluated benchmarks, and Llama-34B in reasoning, mathematics, and code generation. The Mistral-7B-Instruct is a fine-tuned model to follow instructions that surpasses the Llama2-13B-Chat model both on human and automated benchmarks.

7) Zephyr-7b-beta. Zephyr [24] is a series of language models that are trained to act as helpful assistants. Zephyr-7B-β is the second model in the series and is a fine-tuned version of Mistral-7B-v0.1 that was trained on a mix of publicly available, synthetic datasets using Direct Preference Optimization (DPO).

III. Method

The process of this project is depicted in Fig. 1, comprising three pivotal sections: *Concept Pair Extraction*, *Prompt Generation*, and *LLM Prediction*. In the Concept Pair Extraction phase, we extracted IS-A connected concept pairs as positives and from SNOMED CT's Clinical Finding and Procedure hierarchies and Disease Ontology. The same number of concept pairs that are not connected by IS-A relation (denoted as ***unrelated*** concept pair) were randomly generated as negatives. Subsequently, in the Prompt Generation section, we introduced a component-based template design, leveraging strategically positioned placeholders for concepts, context prompt (task emphasis) and few-shots prompt (examples). These placeholders allowed for the generation of specific prompts tailored to individual positive and negative concept pairs. The templates, ranging from a simple baseline to nuanced ontology hierarchy considerations, provided a diverse set of stimuli for LLMs. Finally, in the LLM Prediction phase, eight prominent models were systematically fed with the generated prompts, evaluating their correctness in predicting IS-A relations. The prediction results were compared with true labels for each concept pair to compute precision, recall and F1, as the metrics to evaluate models' performance.

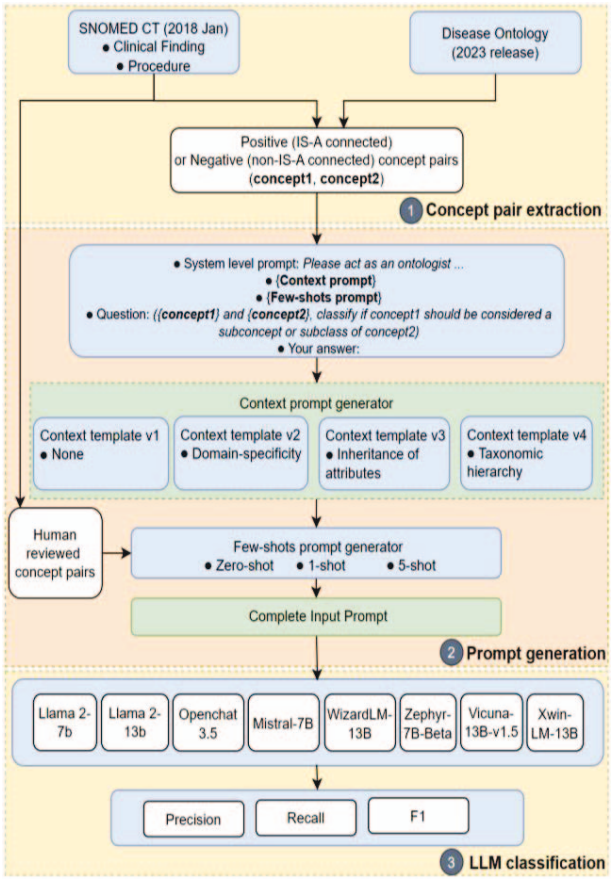

Fig. 1. Experiment Workflow

A. Dataset Selection and Preparation

The choice of an appropriate dataset is paramount in evaluating the performance of Large Language Models (LLMs) in predicting IS-A relationships within medical ontologies. We used the same test dataset used in [13], which were new concepts added to the Clinical Finding hierarchy and Procedure hierarchy in the SNOMED CT 2018 release. For the Clinical Finding hierarchy, a total of 8,574 concept pairs were used for evaluation, which contained 4,287 IS-A

connected concept pairs as positives and 4,287 unrelated concept pairs as negatives. Similarly, the Procedure hierarchy, comprising 3,908 pairs, was crafted with an equal distribution of 1,954 positives and 1,954 negatives. The evaluation also includes Disease Ontology, which includes 1500 concept pairs of positives and 1500 negatives. This meticulous curation ensured that our experimental setup represented a diverse and comprehensive spectrum of IS-A relationships within medical ontologies.

TABLE I. INPUT PROMPT TEMPLATE

Input Prompt with Component Placeholders
"*Please act as an ontologist who can assist with ontology design and knowledge representation for medical/clinical domains.* *The task is that given two concepts, ["Concept_1"] and ["Concept_2"], determine if Concept_1 should be considered a subconcept or subclass of Concept_2.* *{**Context Prompt**}* *If yes, then respond with "YES". Otherwise, respond with "NO".* *Note that your answer must start with a YES or NO answer, then provide reasoning or justifications for your response in less than five sentences, taking into account any relevant domain knowledge, ontological principles, and relationships observed between the concepts.* *Here are some examples:* *{**Few_shots_text**}* *Here is your question:* *With ["{**concept1**}"] as Concept_1 and ["{**concept2**}"] as Concept_2, determine if Concept_1 should be considered a subconcept or subclass of Concept_2.* *Your answer:*"

B. Input Prompt Design

The input prompt to LLMs consists of four sub-prompts (shown in Table 1): 1) *System-level prompt*; 2) *Context prompt*; 3) *Few-shots prompt*; and 4) *Question prompt* with hint words "Your answer" to indicate the starting point for LLMs' generation.

1) ***System-level prompt*** describes the role of a model ("act as an ontologist"), and the high-level description of the task ("*determine if Concept_1 should be considered a sub-concept or subclass of Concept_2*").

2) ***Context prompt*** is varied with 4 different template versions (Table II) with different considerations: version 1 template is plain with no context information; version 2 template focuses on Domain-specificity about the two concepts; version 3 template highlighted the Inheritance of attributes between two concepts of a pair; version 4 template emphasizes taxonomic hierarchical connection between two concepts. Other details of context prompt design are elaborated in section *C. Context Prompt Template Design*.

3) ***Few-shots prompt*** was used to provide examples for LLMs to learn from. We tested three types of configurations including zero-shot, one-shot and five-shots. Details of few-shot examples were elaborated in section *E. Zero-shot Vs. Multi-shots*.

4) ***Question prompt*** specifies the task for LLMs and embedded concept placeholders, denoted as {concept1} and {concept2}. These placeholders were strategically positioned to be replaced with specific concepts from the positive and negative pairs. For instance, in a positive concept pair (*Deep partial thickness burn of perineum*, *Dermatosis of perineum*), the placeholders would be substituted with the actual concepts, creating a specific prompt for the LLMs:

"With ["Deep partial thickness burn of perineum"] as Concept_1 and ["Dermatosis of perineum"] as Concept_2, determine if Concept_1 should be considered a sub-concept or subclass of Concept_2."

C. Context Prompt Template Design

The essence of our method lies in the careful design of context prompt templates to encapsulate the complexity of IS-A relationships. Four templates with different semantic priority were crafted to probe different perspectives in ontology curation. The first template, conceived as a simple and abstract baseline, aimed to establish a foundational understanding. Subsequent templates delved into nuanced dimensions, namely "Domain-specificity," "Inheritance of attributes," and "Taxonomic hierarchy." These templates were not merely linguistic constructs; rather they were carefully engineered to capture the intricacies of hierarchical relationships, fostering a rich and dynamic interaction with the LLMs under scrutiny.

TABLE II. FOUR TEMPLATES OF THE CONTEXT PROMPTS

No.	Context Prompt Templates
T1	**Semantic:** Simple
	Prompt: N/A
T2	**Semantic:** Domain-specificity
	Prompt: "To make this determination, please define the two concepts and analyze the context and domain relevance of the concepts: Is Concept_1 a narrower or more specific concept within the domain of Concept_2?"
T3	**Semantic**: Inheritance of attributes
	Prompt: "To make this determination, please define the two concepts and assess whether Concept 1 naturally inherits or extends the attributes and properties of Concept 2."
T4	**Semantic**: Taxonomic hierarchy
	Prompt: "To make this determination, please define the two concepts and examine the overall structure and organization of the ontology. Does the inclusion of Concept_1 as a child concept of Concept 2 align with the existing hierarchy?"

The "Domain-specificity" template sought to investigate the impact of contextual specificity in eliciting accurate predictions, while the "Inheritance of attributes" template aimed to capture the inheritance patterns that define IS-A

relationships within medical ontologies. The "Taxonomic hierarchy" template, on the other hand, probed the hierarchical structure of the ontologies, challenging the models to discern and comprehend the taxonomy underlying IS-A relationships. Each template was meticulously designed to serve as a lens, allowing us to discern the models' proficiency in understanding and navigating different dimensions of hierarchical relationships in medical ontologies.

D. Model Selection and Experimentation

Eight prominent LLMs were selected for experimentation, each chosen for its unique architecture and capabilities. The lineup included Llama2-7b, Llama2-13b, Vicuna-13b-v1.5, WizardLM-13B-V1.2, Xwin-LM-13B-V0.2, OpenChat_3.5, Zephyr-7b-beta, and Mistral-7B-Instruct-v0.1. This diverse set of models brought forth a rich array of linguistic and semantic processing capabilities, making our investigation robust and comprehensive.

In our experimental setup, each model was meticulously fed with the four distinct templates, each embedded with positive and negative concept pairs. This approach allowed us to systematically evaluate how different LLMs responded to variations in prompt structure and content. The templates served as dynamic stimuli, challenging the models to interpret and predict IS-A relationships across different linguistic and semantic contexts.

E. Zero-shot Vs. Multi-shots

To explore the adaptability of the LLMs to situations with minimal training examples, we introduced zero-shot and few-shot scenarios. In the zero-shot scenario, models were presented with IS-A relationships without any prior training on those specific instances. In the few-shot scenarios (1 example and 5 examples), the models were exposed to a selected number of human-reviewed IS-A relationships, simulating scenarios where only a handful of examples were available. A total of 15 human-reviewed IS-A connected concept pairs (with 5 pairs each corresponding to T2, T3, T4 in *Context Prompt Templates*) and 5 negative pairs. This setup allowed us to assess the generalization capabilities of the LLMs and their performance under varying degrees of training data availability.

F. Two-Dimensional Majority Vote

1) Template Layer. Recognizing the potential advantages of breaking down the task into subtasks, we introduced a Majority Vote Ensemble. This ensemble approach involved aggregating the models' prediction results from templates T2, T3, and T4. The motivation behind this ensemble was to assess whether subtask breakdowns in the prompt variants contributed to improved accuracy in identifying correct answers. By comparing the ensemble results with those obtained using only template T1, we aimed to unravel the effectiveness of partitioning the task into subtasks and exploring the impact on predictive accuracy.

2) Model Layer. After applying Majority Vote among templates T2, T3 and T4, we additionally used the majority vote method for the model level. Specifically, we identified the top-performing three models and produced majority vote results based on their respective template levels. Subsequently, we conducted a majority vote among these three selected models as Figure 2 demonstrates.

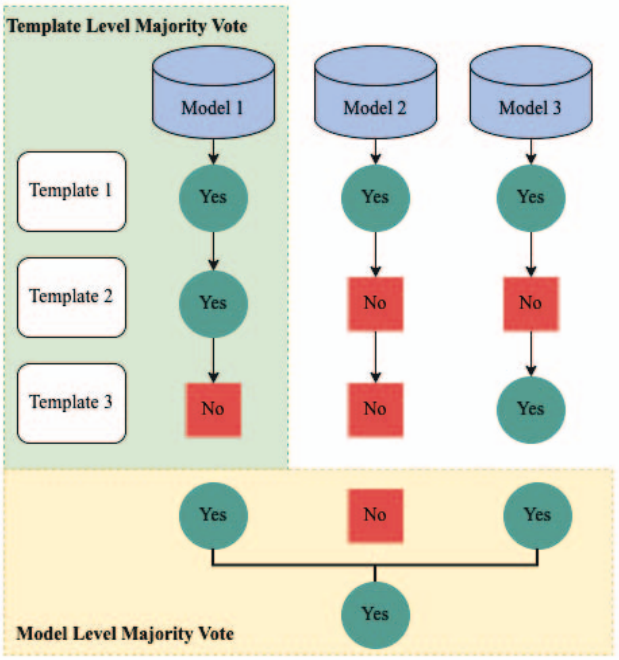

Fig. 2. Example of Two-Dimensional Majority Vote

G. Evaluation Metrics

Each model predicts a "YES" or "NO" label based on the input prompt for a given concept pair. This label is compared with the ground truth label for this concept pair. The true label for two concepts that are connected by IS-A relation (positive) is "YES." The true label for two concepts that are unrelated (negative) is "No." The model's prediction is correct if its prediction is equal to the true label, otherwise the model is wrong. Precision, recall, and F1 score were chosen as the primary metrics to provide a nuanced and comprehensive assessment. Precision gauged the accuracy of positive predictions, recall measured the models' ability to capture all positive instances, and F1 (1), the harmonic means of precision and recall, provided a balanced evaluation considering both false positives and false negatives. The macro average F1 score (2) is computed to evaluate the performance of individual models. In this study, the macro average F1 score is the mean of F1 score of IS-A concept pairs ($F1_{positive}$) and F1 score unrelated concept pairs ($F1_{negative}$).

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (1)$$

$$Macro\ Avg\ F1 = \frac{F1_{positive} \times F1_{nagative}}{2} \quad (2)$$

H. Experiment configuration

For our experiments, we fixed the temperature which refers to the "softness" of the probability distribution at 0.1. Specifically, a higher temperature value will cause the higher randomness of the output, resulting in more "innovative" feedback. On the other hand, a lower temperature value, closer to 0 will make the output more deterministic, leading to responses that are more likely according to the model's training [25]. The experiment was implemented using FastChat [26], and all LLMs were loaded from the HuggingFace [27] platform. Intel(R) Xeon(R) CPU E5-4627 v4 with 40 cores processor, 4 NVIDIA A100 GPUs, and a memory size of 40G were used in this work.

IV. Result

We reported the model's performance on testing datasets selected from the *Clinical Finding* hierarchy and the *Procedure* hierarchy of SNOMED CT, and the human Disease Ontology (DO), and their performance under different context prompt templates and few-shot training strategy settings.

In Table III, we compared the baseline model BERT with eight models using context prompt template T2, as described in the Method section. Openchat-3.5 demonstrated optimal performance across all three datasets, with average macro F1 scores consistently exceeding 0.90. The second-best performing model was Vicuna-13b-v1.5. While Zephyr-7b-beta exhibited notable accuracy in the context of SNOMED CT's two hierarchies, its performance was less favorable when applied to Disease Ontology. On average, OpenChat-3.5, Vicuna-13b-v1.5 and Zephyr-7b-beta outperformed the baseline BERT model on the two SNOMED CT datasets.

TABLE III. F1 SCORE (MACRO) OF MODELS' PERFORMANCE ON CLINICAL FINDING (CF) HIERARCHY, PROCEDURE (PROC) HIERARCHY AND HUMAN DISEASE ONTOLOGY (DO) DATASET

Model	Dataset		
	CF	***PROC***	***DO***
OpenChat-3.5	**0.96**	**0.95**	**0.91**
Vicuna-13b-v1.5	0.89	0.90	0.84
BERT*	0.89	0.85	–
Zephyr-7b-beta	0.88	0.92	0.64
Llama-2-13b-chat	0.85	0.85	0.58
Llama-2-7b-chat	0.33	0.33	0.33
Mistral-7B-Instruct-v0.1	0.33	0.33	0.33
WizardLM-13B-V1.2	0.33	0.33	0.33
Xwin-LM-13B-V0.2	0.30	0.37	0.33

* BERT model's performance reported from [13].

In Table IV, we showed the model performances under four different context prompt templates. Among the eight Large Language Models (LLMs), three models including Llama-2-7b-chat-hf, Mistral-7B-Instruct-v0.1, and WizardLM-13B-V1.2 were insensitivities to prompt variations, with no changes in F1 scores. Notably, for Xwin-LM-13B-V0.2, the enriched context prompts T2, T3, and T3 resulted in a decline in performance compared with T1. Using template T2, OpenChat-3.5 (0.96), Vicunna-13b-v1.5 (0.89), and Zephyr-7b-beta (0.88) achieved the top-3 F1 scores. Although prompt template T3 achieved optimal F1 score (0.89) for the Vicuna model, the improvement compared with T2 is trivial (0.01). T2 demonstrates a high potential as the most instructive prompt among 4 templates.

TABLE IV. F1 SCORE (MACRO) OF EIGHT MODELS QUERIED BY FOUR DIFFERENT TEMPLATES ON CLINICAL FINDING DATASET

Model	Template			
	T1	***T2***	***T3***	***T4***
Llama-2-7b-chat	0.33	0.33	0.33	0.33
Llama-2-13b-chat	0.47	0.85	0.67	0.69
Mistral-7B-Instruct-v0.1	0.33	0.33	0.33	0.33
OpenChat-3.5	**0.93**	**0.96**	**0.94**	**0.96**
Vicuna-13b-v1.5	0.82	0.89	0.90	0.63
WizardLM-13B-V1.2	0.33	0.33	0.33	0.33
Xwin-LM-13B-V0.2	0.63	0.30	0.33	0.33
Zephyr-7b-beta	0.39	0.88	0.64	0.38

We also observed that a model's performance varied with the few-shot strategies embedded in the input prompt. As illustrated in Table V, three models (Mistral-7B-Instruct-v0.1, WizardLM-13B-V1.2, Xwin-LM-13B-V0.2) exhibited significant improvements when trained with one-shot and five-shot strategies. It is interesting that these three models did not perform well with the original settings and were insensitive to the context prompt templates; but were able to "learn" from the training shots (examples) to achieve better performance. In contrast, models such as OpenChat, Vicuna, Zephyr, and Llama-2-13b performed well without training shots, but their performance degraded with additional training shots.

TABLE V. F1 SCORES (MACRO) OF EIGHT MODELS WITH THREE DIFFERENT SHOTS CONFIGURATIONS (DATASET: CF; TEMPLATE: T2)

Model	Shots		
	0-shot	***1-shot***	***5-shot***
Llama-2-7b-chat	0.33	0.19↓	0.34↑
Llama-2-13b-chat	0.85	0.53↓	0.22↓
Mistral-7B-Instruct-v0.1	0.33	0.65↑	0.67↑
OpenChat-3.5	0.96	0.93↓	0.91↓
Vicuna-13b-v1.5	0.89	0.68↓	0.89
WizardLM-13B-V1.2	0.33	0.71↑	0.83↑
Xwin-LM-13B-V0.2	0.30	0.59↑	0.73↑
Zephyr-7b-beta	0.88	0.84↓	0.81↓

In addition, we implemented a two-dimensional majority vote as a way of enhancing performance. The outcomes of the

template-level majority vote with the average of the F1 scores for zero-shot, one-shot, and five-shot configurations are presented in Table VI. Notably, 7 out of the 8 models demonstrated improved performance with the majority vote strategy. The combination of training shots with the template-level majority vote strategy proved effective in elevating the accuracy of the model's predictions.

TABLE VI. F1 SCORE (MACRO) OF ZERO-, ONE-, AND FIVE-SHOTS WITH TEMPLATE T1 AND MAJORITY VOTE OF T2, T3 AND T4

Model	T1	Majority Vote (T2, T3, T4)
Llama-2-7b-chat-hf	0.27	0.41
Llama-2-13b-chat-hf	0.42	0.65
Mistral-7B-Instruct-v0.1	0.53	0.54
OpenChat-3.5	**0.92**	**0.94**
Vicuna-13b-v1.5	0.82	0.82
WizardLM-13B-V1.2	0.51	0.60
Xwin-LM-13B-V0.2	0.43	0.54
Zephyr-7b-beta	0.68	0.77

We also experimented with the majority vote at the model level. The performance comparison between individual model and the ensemble of the three models is shown in Figure 3. It is evident that OpenChat outperformed the majority votes of three models (OpenChat, Vicuna and Zephyr) across all three datasets. In other words, the combination of multiple models' predictions did not lead to more correct results. Therefore, employing a majority vote across multiple models is not an effective strategy for improving prediction accuracy.

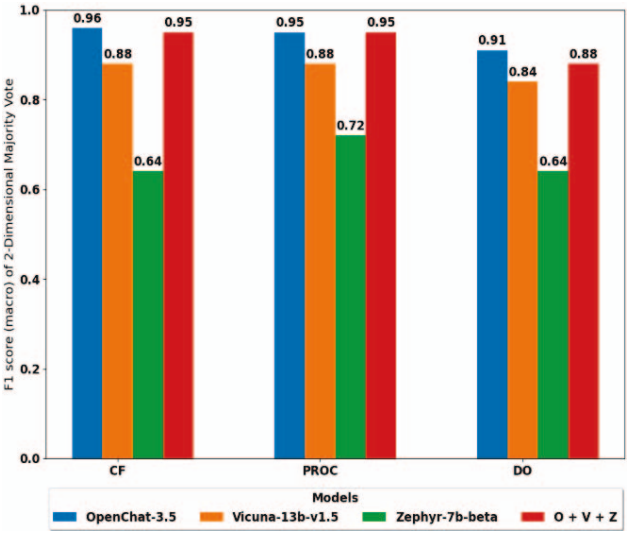

Fig. 3. Macro F1 Score of Three different Shot Scenarios with 2-dimentional majority votes (Blue, Orange and Green bars are the template level majority vote results of 3 best performed model: OpenChat, Vicuna and Zephyr, Red bar is the result produced by the Model-level Majority Vote of these 3 models)

Table VII presents several examples of prediction by OpenChat with context prompt template T2 and zero-shot strategy. It demonstrates accurate predictions of the IS-A relation between concepts "*Sequela of non-traumatic intracranial subarachnoid hemorrhage*" and "*Dysphagia due to and following non-traumatic subarachnoid hemorrhage*". The explanation of this prediction is precise and logical as follows "*[Dysphagia due to and following non-traumatic subarachnoid hemorrhage] is a specific [Sequela of a non-traumatic intracranial subarachnoid hemorrhage], which falls under the broader category of sequelae of disorders classified by disorder-system. The relationship between the two concepts is that Concept_1 is a specific instance of Concept_2.*" For the true negative example (non-IS-A concept pair), OpenChat provided the correct and informative response: "*[Ecthyma caused by staphylococcus aureus] is a bacterial skin infection, while [Dementia] is a neurological disorder affecting cognitive function. They belong to different medical domains and are not related in a subconcept/subclass relationship."*

TABLE VII. PREDICTION EXAMPLES BY OPENCHAT WITH TEMPLATE T2 (DATASET: CF AND PROC)

Metrics	Concept 1	Concept 2
TP	Sequela of non-traumatic intracranial subarachnoid hemorrhage	Dysphagia due to and following non-traumatic subarachnoid hemorrhage
TN	Dementia	Ecthyma caused by staphylococcus aureus
FP*	Cerebral ischemia	Lacunar ataxic hemiparesis of right nondominant side
FN*	Urge incontinence due to prolapse of female genital organ	Mixed incontinence due to prolapse of female genital organ

*FP: Concept 2 is not subclass of Concept 1, but falsely predicted they have IS-A relation; FN: Concept 2 is actual the subclass of Concept 1, but falsely predicted they don't have IS-A relation.

For false predictions, while the outcome may not align with expectations, the accompanying explanations remain acceptable to some degree. For example, the false positive case "*Cerebral ischemia*" and "*Lacunar ataxic hemiparesis of right nondominant side*" are not related as "IS-A", but OpenChat considered the second one is the sub-concept of first one with reasons: "*[Lacunar ataxic hemiparesis of the right nondominant side] is a specific type of brain damage caused by ischemia, which is a lack of blood supply to the brain. Therefore, Concept_1 can be considered a subconcept or subclass of Concept_2.*" For the false negative example, OpenChat predicted that "*Urge incontinence due to prolapse of female genital organ*" is not the subconcept of "*Mixed incontinence due to prolapse of female genital organ*" because "*[Mixed incontinence due to prolapse of female genital organ] is a type of incontinence that involves both stress and urge incontinence, while [Urge incontinence due to prolapse of female genital organ] is a specific type of incontinence that is characterized by a sudden, strong urge to*

urinate. These are two distinct types of incontinence, and one is not a subtype of the other." Although the explanation asserted that they are not IS-A related which they are in SNOMED CT, OpenChat accurately described the meaning of these two concepts, and the incorrect prediction may stem from the absent information that "Mixed incontinence" is modeled as a subtype of "Urge incontinence" for this specific scenario in SNOMED CT.

V. Discussion

Our comprehensive exploration into the application of Large Language Models (LLMs) in predicting IS-A relationships within medical ontologies, specifically the Disease Ontology (DO), Clinical Finding and Procedure hierarchies of SNOMED CT, has unearthed valuable insights into the capabilities and potential applications of these advanced language models.

The findings reveal a nuanced interplay between LLMs and the complex semantic relationships inherent in medical ontologies. Across the diverse set of prompts and models employed in our study, certain LLMs, including OpenChat, Vicuna, and Zephyr, demonstrated remarkable performance even without the need for finetuning. This challenges traditional approaches and opens avenues for leveraging pre-trained models in ontology enrichment tasks.

Prompt template variations played a pivotal role in influencing the models' predictive accuracy. The introduction of concept placeholders provided a dynamic and specific prompt for each instance, contributing to the granularity of predictions. Our ensemble approach, incorporating subtask breakdowns, showcased promising potential in enhancing predictive accuracy, and underscoring the significance of task decomposition in certain scenarios.

The success of LLMs in predicting IS-A relationships within medical ontologies holds promising implications for healthcare informatics. These models could assist ontology curation and periodic updates by automating the placement of concepts, reducing the manual effort required by curators. The ability to generalize to zero-shot and few-shot scenarios opens avenues for rapid integration of new concepts into ontologies with minimal training data.

Furthermore, the findings pave the way for improved clinical decision support systems. The accurate prediction of hierarchical relationships enables more precise and context-aware clinical information retrieval. This can enhance the quality and efficiency of healthcare delivery, supporting clinicians in decision-making processes.

Limitations: The application of Large Language Models (LLMs) in predicting IS-A relationships within medical ontologies encounters multifaceted challenges. One prominent limitation lies in the semantic ambiguity present in concept naming rather than general English context. For instance, when introducing a new concept like "Red Spotted Fever," the term's multiple interpretations, ranging from a specific clinical finding to a broader disease category, can confound LLMs, potentially leading to misplacement within the ontology hierarchy [4]. Another significant challenge arises from contextual sensitivity and specificity. Prompt templates designed to explore relationships might struggle in contexts where nuanced clinical distinctions are crucial. For example, differentiating between "Chronic Pain" and "Acute Pain" demands both linguistic comprehension and clinical understanding, posing challenges for LLMs to capture these nuances accurately. Moreover, LLMs exhibit limitations in their understanding of clinical intricacies, such as the handling of rare or novel concepts. LLMs may lack sufficient context for accurate predictions. The introduction of highly specialized medical conditions, absent in the training data, underscores the limitations in adapting to emerging medical concepts. These limitations collectively emphasize the need for ongoing refinement and careful consideration when employing LLMs in the intricate domain of medical ontology curation.

Future Work: Future research should explore methods to incorporate biomedical ontologies and domain-specific datasets during pre-training or fine-tuning to enhance the LLMs' understanding of medical concepts and relationships. Experiment of training top performed models with large training datasets, may prone the efficiency of the suggested pretraining. We plan to use the "IS-A" related concepts pairs in this study as the training dataset to finetune to the top three performance models for improving their prediction accuracy. Additionally, we plan to assess the performance of various proprietary LLMs, such as OpenAI's ChatGPT-4 and Anthropic's Claude on this task and compare it with other open-source generative LLMs. We will also explore dynamic prompt generation strategies that adapt to the evolving nature of medical ontologies. This involves creating prompt templates that dynamically adjust based on the specific domains of concepts and semantic relationships that need to be curated in the ontologies, providing LLMs with a more adaptive and context-aware task description. Future work could also delve into interdisciplinary collaborations involving computer scientists, healthcare professionals, and ontology experts. This collaborative approach could facilitate the development of hybrid systems that combine the strengths of LLMs with the nuanced expertise of domain specialists, addressing the challenges posed by the intricate nature of medical ontologies.

VI. Conclusion

We presented a rigorous investigation into the application of Large Language Models (LLMs) in predicting IS-A relationships within medical ontologies. The strategic exploration of diverse prompt templates with concept placeholders, model variations, few-shots examples, and model ensembles has provided valuable insights into the potential applications of LLMs in ontology curation. Our findings revealed the potential of LLMs, particularly OpenChat, Vicuna, and Zephyr, in predicting hierarchical relationships without finetuning. We found that strategic use of the prompts to LLMs allowed for dynamic and specific

prompts tailored to each concept pair, contributing to the accuracy and granularity of IS-A relationship predictions.

ACKNOWLEDGMENT

This research project was supported by the U.S. National Science Foundation Grant No. 2018575.

REFERENCES

[1] A. Metke-Jimenez and M. Lawley, "Snorocket 2.0: Concrete Domains and Concurrent Classification," *ORE,* vol. 1015, pp. 32-38, 2013.

[2] R. D. Shearer, B. Motik, and I. Horrocks, "Hermit: A highly-efficient OWL reasoner," in *Owled*, 2008, vol. 432, p. 91.

[3] L. Luo, J. Feng, H. Yu, and J. Wang, "Automatic structuring of ontology terms based on lexical granularity and machine learning: Algorithm development and validation," *JMIR medical informatics,* vol. 8, no. 11, p. e22333, 2020.

[4] H. Liu, S. Carini, Z. Chen, S. P. Hey, I. Sim, and C. Weng, "Ontology-based categorization of clinical studies by their conditions," *Journal of Biomedical Informatics,* vol. 135, p. 104235, 2022.

[5] M. K. H. Dehkordi *et al.*, "Using annotation for computerized support for fast skimming of cardiology electronic health record notes," in *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2023, pp. 4043-4050: IEEE.

[6] SNOMED International (2024, 01/28). *Systematised nomenclature of medicine clinical terms (SNOMED CT)*. Available: https://www.snomed.org/

[7] Q. Sun, G.-Q. Zhang, W. Zhu, and L. Cui, "Validating auto-suggested changes for SNOMED CT in non-lattice subgraphs using relational machine learning," *Studies in health technology informatics,* 2019.

[8] L. Zheng, H. Liu, Y. Perl, and J. Geller, "Training a convolutional neural network with terminology summarization data improves SNOMED CT enrichment," in *AMIA Annual Symposium Proceedings*, 2019, vol. 2019, p. 972: American Medical Informatics Association.

[9] H. Liu, Y. Perl, and J. Geller, "Transfer learning from BERT to support insertion of new concepts into SNOMED CT," in *AMIA Annual Symposium Proceedings*, 2019, vol. 2019, p. 1129: American Medical Informatics Association.

[10] X. Hao, R. Abeysinghe, F. Zheng, and L. Cui, "Leveraging non-lattice subgraphs for suggestion of new concepts for SNOMED CT," in *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2021, pp. 1805-1812.

[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:.04805,* 2018.

[12] M. E. Peters, W. Ammar, C. Bhagavatula, and R. Power, "Semi-supervised sequence tagging with bidirectional language models," *arXiv preprint arXiv:.00108,* 2017.

[13] H. Liu, Y. Perl, and J. Geller, "Concept placement using BERT trained by transforming and summarizing biomedical ontology structure," *Journal of Biomedical Informatics,* vol. 112, p. 103607, 2020/12/01/ 2020.

[14] H. Touvron *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:.09288,* 2023.

[15] W.-L. Chiang *et al.*, "Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality," *https://vicuna.lmsys.org,* 2023.

[16] A. Q. Jiang *et al.*, "Mistral 7B," *arXiv preprint arXiv:.06825,* 2023.

[17] G. Wang, S. Cheng, X. Zhan, X. Li, S. Song, and Y. Liu, "Openchat: Advancing open-source language models with mixed-quality data," *arXiv preprint arXiv:.11235,* 2023.

[18] L. M. Schriml *et al.*, "Disease Ontology: a backbone for disease semantic integration," *Nucleic acids research,* vol. 40, no. D1, pp. D940-D946, 2012.

[19] E. Chang and J. Mostafa, "The use of SNOMED CT, 2013-2020: a literature review," *Journal of the American Medical Informatics Association,* vol. 28, no. 9, pp. 2017-2026, 2021.

[20] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, "Large language models are zero-shot reasoners," *Advances in neural information processing systems,* vol. 35, pp. 22199-22213, 2022.

[21] A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, L. Gutierrez, T. F. Tan, and D. S. W. Ting, "Large language models in medicine," *Nature Medicine,* vol. 29, no. 8, pp. 1930-1940, 2023/08/01 2023.

[22] C. Xu *et al.*, "Wizardlm: Empowering large language models to follow complex instructions," *arXiv preprint arXiv:.12244,* 2023.

[23] Xwin-LM Team (2023, 01/28). *Xwin-LM.* Available: https://github.com/Xwin-LM/Xwin-LM

[24] L. Tunstall *et al.*, "Zephyr: Direct distillation of lm alignment," *arXiv preprint arXiv:.16944,* 2023.

[25] F. F. Xu, U. Alon, G. Neubig, and V. J. Hellendoorn, "A systematic evaluation of large language models of code," in *Proceedings of the 6th ACM SIGPLAN International Symposium on Machine Programming*, 2022, pp. 1-10.

[26] L. Zheng *et al.*, "Judging LLM-as-a-judge with MT-Bench and Chatbot Arena," *arXiv preprint arXiv:.05685,* 2023.

[27] T. Wolf *et al.*, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, 2020, pp. 38-45.