LLMs for the Engineering of a Parkinson Disease Monitoring and Alerting Ontology

Georgios Bouchouras [1,*] †, Pavlos Bitilis [1†], Konstantinos Kotis[1†], and George A. Vouros [2†]

[1] *Intelligent Systems Lab, Dept. of Cultural Technology and Communication, University of the Aegean, Mytilene, 81100 Greece*

[2] *Artificial Intelligence Lab, Dept. Of Digital Systems, University of Piraeus, Piraeus, 18534 Greece.*

Abstract

This paper investigates the integration of Large Language Models (LLMs) in the engineering of a Parkinson's Disease (PD) monitoring and alerting ontology. The focus is on the ontology engineering methodology which combines the capabilities of LLMs and human expertise to develop more robust and comprehensive domain ontologies, faster than humans do alone. Evaluating models like ChatGPT-3.5, ChatGPT4, Gemini, and Llama2, this study explores various LLM based ontology engineering methods. The findings reveal that the proposed hybrid approach (both LLM and human involvement), namely X-HCOME, consistently excelled in class generation and F-1 score, indicating its efficiency in creating valid and comprehensive ontologies faster than humans do alone. The study underscores the potential of the combined LLMs and human intelligence to enrich PD domain knowledge and enhance expert-generated PD ontologies. In overall, the presented approach exemplifies a promising collaboration between machine capabilities and human expertise in developing ontologies for complex domains.

Keywords

Ontology Engineering, LLMs, Parkinson Disease, Human-LLM teaming.

1. Introduction

The integration of LLMs (Large Language Models) with ontological frameworks is gaining prominence in the field of knowledge Representation (KR) and Artificial Intelligence (AI) [1, 2]. As KR methods become more demanding, there is a noticeable trend towards the use of LLMs for the construction, refinement, and mapping of ontologies, tasks that have been traditionally performed and supervised by human experts with in-depth knowledge of the domain and of the engineering of ontologies [3]. Since LLMs are trained on big data, they are making expert-level insights across domains more accessible and cost-effective. Moreover, while LLMs are getting more effective at engineering ontologies, their capabilities are significantly enhanced in the

GeNeSy'24: First International Workshop on Generative Neuro-Symbolic Artificial Intelligence, co-located with ESWC 2024, May 26, 2024, Hersonissos, Crete, Greece.

*Corresponding author.

†These authors contributed equally.

cti23010@ct.aegean.gr (G.Bouchouras); pavlos.bitilis@aegean.gr (P.Bitilis); kotis@aegean.gr (K.Kotis); georgev@unipi.gr (G.Vouros);

0000-0003-0566-3615 (G.Bouchouras); 0000-0003-0548-6268 (P.Bitilis); 0000-0001-7838-9691 (K.Kotis); 0000-0001-5451-622X (G.Vouros)

era of Neurosymbolic AI, i.e., combining the deep and varied knowledge of statistical AI with the semantic reasoning of symbolic AI [4].

Neurosymbolic AI is particularly significant in addressing complex health problems such as the monitoring and alerting patients and doctors of Parkinson Disease (PD), the second most common neurodegenerative disease globally [5]. Despite extensive research, the nature of PD remains elusive, and current treatments offer only partial effectiveness [6]. In response, related ontologies have been developed to enhance understanding, monitoring and alerting, and treatment approaches. Specifically, the Wear4PDmove ontology [7, 8] has been recently developed with the aim to integrate heterogeneous sensor (movement) and personal health record (PHR) data, as a knowledge model used to interface/connect patients and doctors with smart devices and health applications. This ontology aims to semantically integrate heterogeneous data sources, such as dynamic/stream data from wearables and static/historic data from personal health records, to represent personal health knowledge in the form of a Personal Health Knowledge Graph (PHKG). Also, it supports health applications' reasoning capabilities for high-level event recognition in PD monitoring, such as identifying events like 'missing dose' or 'patient fall' [8, 9]. This and associated ontologies facilitate the critical integration of AI-driven tools and domain-specific knowledge, making it easier to integrate and reason with health data and promote creative PD treatment approaches.

PD monitoring and alerting of patients requires flexible KR methods to effectively adapt to their health changes. LLMs have shown impressive abilities in handling vast quantities of data and producing valuable insights from their near real-time analysis. Yet, their use in monitoring PD and alerting patients is limited by factors like inadequate reasoning abilities and reliance on specialized health knowledge. Health is a complicated domain, with distinct contexts, subtle meaning variations, and disease-specific vocabularies. To effectively capture and express this complex knowledge, it is necessary to fine-tune and train LLMs specifically for the domain, which can demand a significant number of resources that are not always available, or health/medical experts are not willing to provide for many different reasons. Also, healthcare ontologies now adhere to several standards and forms. The technical challenge, however, lies in the integration and reconciliation of information from many heterogeneous sources into a coherent ontology, while also ensuring interoperability. To achieve an efficient ontology development process within an ontology engineering methodology (OEM), LLMs must be able to navigate these disparities efficiently. Existing research on PD has already utilize ontologies [7, 10]. However, maintaining these ontologies in this rapidly changing field of PD, calls for constant effort and resources. Failure to update/refine the ontology may result in outdated information.

This study aims to investigate the possibilities of LLM-based collaborative OE (ontology engineer) to improve the speed and accuracy of PD knowledge representation. LLMs can efficiently analyze large volumes of health-related data, recognize patterns and semantic connections between them [11]. Human specialists contribute to ensuring the precision and domain-specific significance of the acquired knowledge. LLMs and humans, working together, can collaboratively engineer PD-related ontologies that efficiently support the monitoring and alerting of patients and doctors.

This paper presents experiments with LLMs for PD ontology engineering. More important, in this paper, an extension of a human-centered collaborative OEM (HCOME) [12] with LLM-based tasks is propose and assessed (namely X-HCOME). The aim is to provide a novel OEM, including both humans and LLMs in the engineering of ontologies, with a focus on

speed, conceptualization, and human-assistance. The final product of this work will be an OEM more effective in knowledge representation than those used solely by humans or LLMs. The paper focuses on LLM-based collaborative OE to create comprehensive PD ontologies and discusses limitations identified from the experimental results.

The organization of this paper is as follows: Section 2 presents related work on integrating LLMs to OE; Section 3 describes the proposed research methodology; Section 4 presents the conducted experiment; Section 5 presents further experimentation; and finally, section 6. Discuss the results and draws the conclusions.

2. Related Work

Oksannen et al. (2021) developed an approach to derive product ontologies from textual reviews using BERT models. Their approach, which required minimum manual annotation, demonstrates increased precision and recall in comparison to established methods such as Text2Onto and COMET, signifying a noteworthy advancement in automatic ontology extraction [13]. The BERTMap, a tool designed for the visualization and analysis for Bidirectional Encoder Representations from Transformers by He et al. (2022), demonstrates the effectiveness of LLMs by exceling at ontology mapping (OM), especially in unsupervised and semi-supervised scenarios, surpassing current OM systems. It demonstrates the precision of LLMs in matching entities between knowledge graphs [14]. Ning et al. (2022), introduce a technique to extract factual information from LLMs by creating prompts for pairs of subjects and relations. They utilize an approach that incorporated pre-trained LLMs with prompt templates derived from web material and personal expertise. The authors identify effective prompts through a parameter selection technique and filter the generated entities to pinpoint reliable choices. They stress the significance of investigating parameter combinations, testing LLMs, and expanding research into different domains [15].

Lippolis et al. concentrate on harmonizing entities across ArtGraph and Wikidata. By combining traditional querying with LLMs, they achieve a high accuracy in entity alignment, showcasing the efficiency of LLMs in filling knowledge gaps in intricate databases [16]. Funk et al. (2023) investigates the capability of ChatGPT3.5 in creating concept hierarchies in several fields. Their method decreases mistakes and generates appropriate concept names, demonstrating the effectiveness of LLMs in the semi-automatic creation of ontologies. Studies on GPT4's abilities in structured intelligence within ontologies indicate its potential for groundbreaking progress. Their study emphasizes the importance of implementing controlled LLM integration in business environments through a collaborative framework. [17]. Biester et al. (2023) develops a technique that utilizes prompt ensembles to improve knowledge base development. When applied to models such as ChatGPT and Google BARD, they demonstrate notable enhancements in precision, recall, and F-1-score, highlighting the effectiveness of LLMs in improving knowledge bases [18]. Mountantonakis and Tzitzikas (2023) devise a technique to verify ChatGPT information by utilizing RDF Knowledge Graphs. They confirm the accuracy of 85.3% of ChatGPT facts, highlighting the significance of verification services in maintaining data precision [19]. Pan et al. (2023) suggests combining LLMs with KGs to improve reasoning skills. Their frameworks attempt to combine the benefits of both LLMs and KGs, resulting in enhanced data processing and reasoning abilities [20]. Joachimiac et al. (2023), used the Spindoctor approach, which employed LLMs to summarize gene sets, demonstrating the versatility of LLMs in analyzing intricate biological information. Their method showcased the effectiveness of LLMs in summarizing text specifically related to gene ontology [21]. The

SPIRES approach developed by Caufield et al. (2023) demonstrates the adaptability of LLMs in extracting information from unstructured texts in many fields. This zero-shot learning method does not require any model adjustment, demonstrating the wide range of applications of LLMs in various disciplines [22]. Mateiu et al. (2023) showcase the application of GPT3 in converting natural language words into ontology axioms. Their methodology facilitates ontology creation, enhancing accessibility and efficiency, demonstrating the effectiveness of LLMs in streamlining intricate ontology engineering processes [23].

However, the aforementioned studies primarily concentrate on the capabilities of LLMs in isolation or in comparison with traditional methods, often emphasizing automated or semi-automated processes. What remains less explored, and thus the focus of current study, is the symbiotic integration of both human expertise and LLMs in the process of OEM. This novel approach aims to harness the speed and computational efficiency of LLMs while simultaneously capitalizing on the complex understanding and conceptualization skills of human experts. Furthermore, it is reasonable to believe that the differences between LLMs have strengths and weaknesses that can help researchers and practitioners choose the best models for use in real-world entity resolution [24].

3. Research Methodology

The forthcoming section presents an experiment encompassing two distinct phases, focusing on the development and assessment of ontologies, with a special emphasis on classes. The initial phase involves generating an ontology for PD monitoring and alerting, mainly powered by the autonomous capabilities of LLMs. This process utilizes both 'One Shot' (OS) and 'Chain of Thought' (CoT) techniques. The OS method involves presenting a model with a single prompt and expecting it to produce a suitable response based only on this input. In a one-shot situation, the model is not provided with several examples for learning and must complete the task with little context. This is a straightforward approach where the model uses its pre-trained knowledge to infer the most likely answer. For this paper purposes, CoT refers to a methodological approach where the OS is segmented into two sequential prompts. This segmentation allows for a structured progression in the reasoning process, whereby each prompt is strategically designed to focus on a specific element of the overall task. By employing sequential prompting, we direct the language model to tackle each segment of the problem individually, thereby facilitating a cumulative build-up of information and reasoning. Subsequently, in the second phase, a hybrid OEM is established, which integrates human expertise with the abilities of LLMs. This collaboration aims to elevate the quality and practicality of the ontology within the PD monitoring and alerting framework. Figure 1 depicts a flowchart that outlines this two-phase experimental process. Initially, four LLMs independently develop an ontology with minimal human input (phase 1). The process evolves into a more collaborative approach (Human and LLMs) with the X-HCOME OEM (phase 2). The resulting ontologies are then compared against a gold standard ontology using various metrics. The process is further customized (further experimentation) through expert evaluations and refinement of the gold standard ontology.

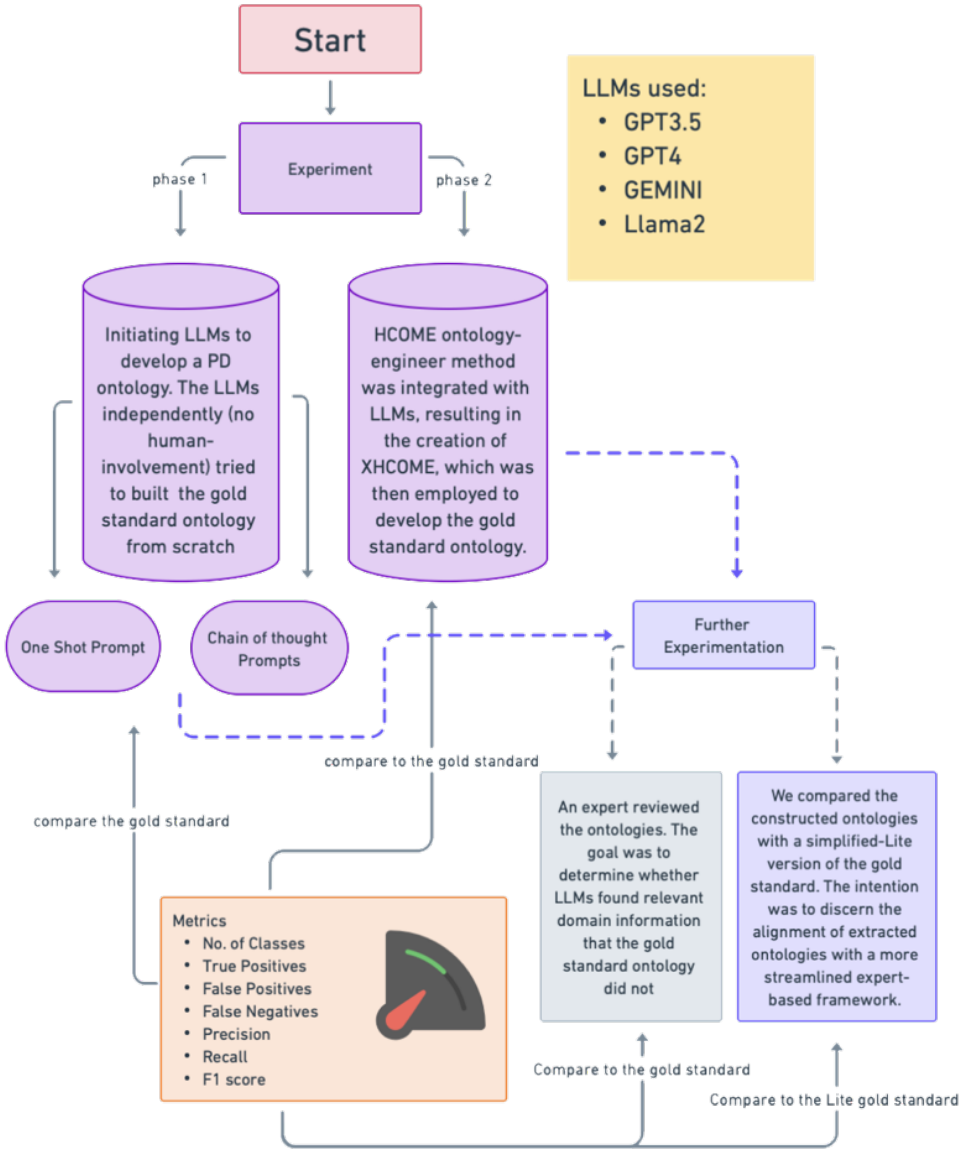

Figure 1. Flowchart of a multi-phase experimentation assessing the construction and validation of ontologies using different methodologies (created with AI-Whimsical ChatGPT, 2023[2]).

To fulfill the study's objective, the following will be conducted: a) an examination of the LLMs attempting to construct ontologies in with minimal human intervention and b), an examination of the X-HCOME methodology in OE and its evaluation by comparing the quality of LLM-generated ontologies with human-generated ones. The X-HCOME methodology is an extension of the Human-Centered Collaborative Ontology Engineering methodology (HCOME) [12]. This extension concerns the inclusion of LLM-based tasks (along with the human-centered

[2] *OpenAI. 2023. "Whimsical Diagrams." ChatGPT Functionality. OpenAI. https://openai.com/chatgpt.*

ones) in the OE lifecycle. This study aims to show that ontologies that are collaboratively engineered by humans (knowledge engineers, knowledge workers, domain experts, etc.) and machines (LLMs) are of higher quality than ontologies that are created by humans or LLMs alone. A secondary goal is to support the hypothesis that working along with LLMs, humans can complete ontology engineering tasks (and consequently, the OE lifecycle) much faster i.e., from several days or weeks to hours. The proposed research methodology is driven by two specific hypotheses. These hypotheses drive the experimental phases carried out to assess the efficacy of the proposed approach.

Hypothesis 1: LLMs, when prompted with domain-specific queries, can autonomously develop a coherent and comprehensive ontology, as it is in the case of PD monitoring and alerting ontology. LLMs have the ability to extract domain knowledge efficiently from their extensive data repositories, and construct ontologies using different prompts engineered by human-user of the LLM.

- This hypothesis is tested in Phase 1 of our experiments, where LLMs are tasked with creating a PD patients' monitoring and alerting patients ontology from ground zero, using domain-specific prompts. The effectiveness of LLMs in developing an accurate and relevant ontology is measured against a gold standard -expert-generated ontology. In this study, the Wear4PDmove [7, 8] is utilized as the gold standard ontology, and it will be referred to as such throughout the remainder of the study.
    - **Phase 1**: Initiating LLMs to develop the ontology. During the initial phase of the experiments, the LLMs will independently (no human-involvement) reconstruct the Wear4PDmove ontology from scratch. This phase comprises the following steps:

        1. LLMs construct an ontology in Turtle format. The ontology represents various aspects of PD patient care, including monitoring, alerting, patients' health record and healthcare team coordination.
        2. Validate the ontology by assessing its accuracy and coherence with OOPS![3] and Protégé[4] tools (Pellet).
        3. Use metrics such as Precision, Recall, and the F-1-score (Table 1) to compare the LLM-generated ontology with the gold standard ontology created by human experts.

Table 1: Summary of metrics for classes evaluation. This table presents the formulas for Precision, Recall, and the F-1-score, along with their definitions.

Formulas	Definitions
Precision = True Positives / (True Positives + False Positives)	True Positives: classes correctly classified as positive in alignment with the 'gold standard' ontology,
Recall = True Positives / (True Positives + False Negatives)	False Positives: classes incorrectly classified as positive in alignment with the "gold standard' ontology

[3] *https://oops.linkeddata.es*
[4] *https://protege.stanford.edu*

F-1 score = 2 * (Precision * Recall) / (Precision + Recall)	False Negatives: classes that are incorrectly classified as negative despite being positive in the 'gold standard' ontology

Hypothesis 2: The combination of human expertise and LLM capabilities enhances the quality and applicability of the developed ontology, as it is in the case of PD monitoring and alerting ontology.

- This hypothesis is related to Phase 2 experimentation, where the X-HCOME methodology is deployed. It assesses how the collaboration between humans and LLMs contributes to refining and validating the ontology, ensuring its relevance and accuracy e.g., in the case of PD monitoring and alerting patients.
    - **Phase 2**. The X-HCOME methodology presented in this paper involves a number of steps assigned to either human experts or LLMs in an alternating manner during the OE process. These steps are:

1. (Human): Define prompts and provide LLMs with the specified data.[5]
    - Define the aim and scope of the ontology: Explain the reasons for its development and the depth of the information it aims to encompass.
    - Ontology Requirements: Enumerate the necessary knowledge that must be represented and explain its significance.
    - Integrate data from PD cases. This data was specifically asked for from the LLM to give a full and accurate picture of the condition (i.e. make sure that PD tremor is properly represented in the ontology).
    - Formulate specific questions (competency questions) in natural language that the ontology should be able to answer, as defined by knowledge workers.
2. (LLM): Construct a domain ontology using the input provided previously, in specific syntax e.g., Turtle . This is a fully automated task performed by the LLM, asking it to act as an ontology engineer and a domain expert.
3. (Human): Compare the LLM-generated ontology with existing gold standard (or widely accepted) ontologies. This is a human based comparison performed either manually or assisted by ontology alignment-mapping tools e.g., LogMap [25].
4. (LLM): Perform a machine-based comparison of LLM-generated ontology against the gold standard ontology. This is a fully automated comparison of the two ontologies, asking LLM to act as an ontology engineer using an OM tool such as LogMap.
5. (Human): Develop a revised domain ontology by combining an existing ontology with the one generated by the LLM.
6. (LLM): Repeat step 4 (LLM-based evaluation of the developed ontology).
7. (Human): Evaluate the revised/refined ontology using OE tools. This step includes a comprehensive assessment of the engineered ontology to confirm that it fulfills the particular requirements and attains the intended level of validity.

4. Methodology Assessment Through Experiment

The results described in this section, supported by supplementary material placed at a GitHub repository [5], focus on the complex process of creating ontologies for monitoring and alerting patients in PD. The conducted experimentation progresses through the two distinct phases presented in Section 3. This experiment evaluates the proposed research methodology by comparing the ontologies generated in the experiment with the gold standard ontology. It is essential to clarify that the metrics presented in this paper solely focused on the generated ontological classes. The validation involves both exact matching, where generated classes corresponded to entities in the gold standard ontology, and similarity matching, where classes were considered correct if they were semantically similar to the gold standard classes. This dual approach ensures a comprehensive evaluation of the LLM's performance, capturing both direct accuracies and contextually appropriate approximations. While our study did include calculations for object properties, unfortunately, due to space limitations, they were not included in this paper. Having said that, the results obtained for object properties were less than optimal, as evidenced by the observed low F1 scores as presented in the GitHub repository [6].

Ontological class definitions consistency and syntactical correctness were observed in all LLM and hybrid generated ontologies, apart from the ones generated by Llama2 (OS, CoT and X-HCOME)[7]. Llama2-generated ontologies included both syntactical errors and inconsistent definitions, and thus it failed to generate a valid ontology. Also, all the developed ontologies were validated with OOPS!, identifying only one minor pitfall (pitfall P36-URI, file extension) during the experimental process[7].

Phase 1 experimentation. LLMs are initially given prompts with two methods. One-shot prompting (OS): with this method, the LLMs were given a single, clear prompt that stated the aim and scope of the gold standard ontology without any additional information or background. The goal was to test LLMs' initial response effectiveness by generating accurate and relevant ontology from a single standalone prompt. Along with thus test, a focus on minimal human effort was given.

The following paragraph provides an example of an OS prompt:

> "*Act as an Ontology Engineer, I need to generate an ontology about Parkinson disease monitoring and alerting patients. The aim of the ontology is to collect movement data of Parkinson disease patients through wearable sensors, analyze them in a way that enables the understanding (uncover) of their semantics, and use these semantics to semantically annotate the data for interoperability and interlinkage with other related data. You will reuse other related ontologies about neurodegenerative diseases. In the process, you should focus on modeling different aspects of PD, such as disease severity, movement patterns of activities of daily living and gait. Give the output in TTL format.*"

Chain-of-Thought prompting (CoT): The CoT prompting method, which breaks down the OS prompt into two distinct prompts. The following paragraph provides an example of an CoT prompt:

> *Prompt 1: "Act as an Ontology Engineer, I need to generate an ontology about Parkinson disease monitoring and alerting patients. The aim of the ontology is to collect movement*

---

[5] *https://github.com/GiorgosBouh/Ontologies_by_LLMs*
[6] *https://github.com/GiorgosBouh/Ontologies_by_LLMs*
[7] *https://oops.linkeddata.es/catalogue.jsp*

data of Parkinson disease patients through wearable sensors, analyze them in a way that enables the understanding (uncover) of their semantics, and use these semantics to semantically annotate the data for interoperability and interlinkage with other related data."

Prompt 2: "You will reuse other related ontologies about neurodegenerative diseases. In the process, you should focus on modeling different aspects of PD, such as disease severity, movement patterns of activities of daily living and gait. Give the output in TTL format."

The first prompt cover the role and aim and scope of the ontology and is crucial as it sets the foundation for the ontology. The second prompt deals with the processing and utilization of the data collected as per the framework set up in the first prompt.

Phase 2 experimentation. Subsequently, we have developed and evaluated the X-HCOME methodology, a novel approach in OE, that seamlessly integrates the expertise of human experts (domain and ontology engineer) with the computational power of LLMs in domain knowledge acquisition and ontology engineering. At each stage of this iterative process, human domain experts critically examine and provide feedback on the ontologies generated by the LLMs. This collaborative working and human-machine teaming is central to the X-HCOME methodology, as it allows for the integration of expert knowledge and insights with the advanced data processing capabilities of LLMs. The experts' contributions are pivotal in identifying variations and complexities that might be overlooked by automated systems, ensuring that the resulting ontology is not only technically sound but also contextually rich and aligned with real-world applications.

Following is a presentation of the two phases' findings. Based on the data provided in Table 2, the chatGPT3.5 OS method identified 5 classes but had relatively low accuracy (Precision 40%, Recall 5%, F-1 score 9%). ChatGPT3.5 CoT achieved higher precision (67%) with limited recall (5%), identifying only 3 classes. ChatGPT4 OS improved, identifying 9 classes (Precision 56%, Recall 12%, F-1 score 20%), while ChatGPT4 CoT showed further enhancement with 6 classes (Precision 67%, Recall 10%, F-1 score 17%). Conversely, GEMINI OS had lower precision (8%) and recall (2%), identifying 13 classes, whereas GEMINI CoT identified 8 classes with better precision (63%) and recall (12%), mirroring ChatGPT4 OS's performance. To summarize, the CoT method generally returned higher precision than the OS method, indicating more accurate but fewer classes. Conversely, OS tended to identify more classes but with lower precision, suggesting a broader but less accurate approach to class identification. While CoT focused on the quality of classifications, OS emphasized quantity, leading to differences in their overall effectiveness in ontology creation.

For the X-HCOME method, the ChatGPT3.5 X-HCOME generated 25 classes with a Precision of 40%, a Recall of 24%, and an F-1 score of 30%, balancing the number of classes identified and accuracy. The ChatGPT4 X-HCOME generated 33 classes but with lower precision, reflected in a Precision of 30%, Recall of 24%, and an F-1 score of 27%. Remarkably, the GEMINI X-HCOME method produced the highest number of classes (50) with a Precision of 38%, a Recall of 46%, and an F-1 score of 42%, showcasing the best recall rate among the methods.

Syntactical errors were indicated by the Llama2 results. However, it is noted that its CoT and OS methods showed high Precision but were limited in overall performance due to the restricted number of classes identified.

Overall, the performance of the X-HCOME methodology was superior in all LLMs. This conclusion is drawn from its consistently higher number of classes identified and the overall

better F-1 score when compared to the other methods (OS and CoT) for each LLM. GEMINI X-HCOME method appeared to be the most effective overall in the context of ontology creation. It produced the highest number of classes (50) and achieved the best recall rate (46%) among all the methods tested. Additionally, its F-1 score of 42% was the highest, suggesting a relatively better balance between precision and recall compared to other methodologies. The F-1 score for the object properties across all LLMs varied from 6% to 12%.[8]

Table 2. Comparative evaluation of methodologies used for ontology creation against the gold standard ontology.

Method	Number of Classes	True Positives	False Positives	False Negatives	Precision	Recall	F-1 score
Gold-ontology	41						
ChatGPT3.5 CoT	3	2	1	39	67%	5%	9%
ChatGPT3.5 OS	5	2	3	39	40%	5%	9%
ChatGPT3.5 X-HCOME	25	10	15	31	40%	24%	30%
ChatGPT4 CoT	6	4	2	37	67%	10%	17%
ChatGPT4 OS	9	5	4	36	56%	12%	20%
ChatGPT4 X-HCOME	33	10	23	31	30%	24%	27%
GEMINI CoT	8	5	3	36	63%	12%	20%
GEMINI OS	13	1	12	40	8%	2%	4%
GEMINI X-HCOME	50	19	31	22	38%	46%	42%
Llama2 CoT	3	3	0	38	100%	7%	14%
Llama2 OS	2	2	0	39	100%	5%	9%
Llama2 X-HCOME	32	4	28	37	13%	10%	11%

5. Further Experimentation

To better evaluate the generated ontologies, we further analyzed the results obtained for False Positives, serving as a domain experts, checking whether LLMs have discovered relevant domain knowledge that the gold standard ontology has not included (incomplete engineering due to human bias or other reasons). This analysis aimed to understand whether the generated classes, despite not matching entities within the gold standard ontology, could be reclassified as true positives, potentially improving the ontology. The integration of expert opinion in this case was crucial for expanding and enhancing the domain knowledge represented in the gold standard ontology. This method shows an ever-changing way of thinking about ontology

[8] *https://github.com/GiorgosBouh/Ontologies_by_LLMs*

construction—as a conversation between human and machine intelligence that goes back and forth. By embracing this perspective, this experiment holds the promise of significantly advancing the field.

The ChatGPT3.5 CoT and OS methods had comparable results, with the CoT method showing slightly higher precision but equal recall and F-1 score as OS. For ChatGPT4, both CoT and OS showed similar trends, with CoT slightly outperforming OS in precision and recall (table 3).

Significantly, the X-HCOME method for both ChatGPT3.5 and ChatGPT4 displayed a marked improvement in precision and recall, notably reducing false positives after expert review. The GEMINI X-HCOME method stood out with exceptional precision and recall, indicating no false positives and a high rate of true positives. However, GEMINI's CoT and OS methods lagged considerably behind in these metrics. Llama2's CoT and OS methods achieved high precision but lower recall. Notably, Llama2 failed to create a consistent ontology without errors, which is a critical aspect in OE. In summary, the X-HCOME method demonstrated superior performance across all LLMs, including ChatGPT3.5, ChatGPT4, and GEMINI, particularly after human expert intervention. This methodology proved more effective in accurately classifying classes with minimal false positives, highlighting its robustness and efficiency in ontology creation tasks. Post-revision, X-HCOME emerges as a highly effective method for ontology generation, balancing class creation with accuracy. For instance, GEMINI X-HCOME generated classes like "*Surgical Intervention*," "*Rigidity*," and "*Cognitive Impairment*", that were absent in the gold standard ontology. This fact underscores its ability to uncover comprehensive knowledge in PD monitoring/alerting that experts alone might overlook. For patients who have undergone surgical interventions like deep brain stimulation, medication regimens may be altered significantly. The alert system needs to be adaptable to reflect these changes. To avoid false alerts about missed doses, the system should account for post-surgical patients reduced or different medication. Also, in patients experiencing significant rigidity, a missed dose of medication can lead to rapid symptom exacerbation. The alert system can be calibrated to be more sensitive and prompting in these cases, ensuring quick notification of a missed dose to prevent worsening of rigidity. Patients with more severe rigidity might receive early or more frequent reminders to take their medication to maintain optimal symptom control. Lastly cognitive impairment can make it challenging for patients to remember their medication schedules. In such cases, the alert system can include more robust, frequent, and clear reminders, possibly using different modalities (like visual or auditory cues) to ensure the patient is aware of the missed dose. Classes like these enhance the ontology's utility in developing sophisticated PD monitoring and alerting systems, ensuring a more rounded approach to patientcare and intervention.

Finally, the F1 score for the object attributes across all LLMs varied from 6% to 84%.[9]

[9] *https://github.com/GiorgosBouh/Ontologies_by_LLMs*

Table 3. Comparative evaluation of ontology creation methods' post expert review on False Positives.

Method	Number of Classes	True Positives	False Positives	False Negatives	Precision	Recall	F-1 score
Gold-ontology	41						
ChatGPT3.5 CoT	3	2	1	39	67%	5%	9%
ChatGPT3.5 OS	5	2	3	39	40%	5%	9%
ChatGPT3.5 X-HCOME	25	23	2	18	92%	56%	70%
ChatGPT4 CoT	6	4	2	37	67%	10%	17%
ChatGPT4 OS	9	5	4	36	56%	12%	20%
ChatGPT4 X-HCOME	33	29	4	12	88%	71%	78%
GEMINI CoT	8	5	3	36	63%	12%	20%
GEMINI OS	13	1	12	40	8%	2%	4%
GEMINI X-HCOME	50	50	0	-9	100%	122%	110%
Llama2 CoT	3	3	0	38	100%	7%	14%
Llama2 OS	2	2	0	39	100%	5%	9%
Llama2 X-HCOME	32	26	6	15	81%	63%	71%

Lastly, an additional experiment was carried out to assess the efficacy of the proposed approach after the X-HCOME methodology was applied. This involved using a modified version of the gold standard ontology, thereby altering the ground truth of the experiments in a controlled manner. We have removed the imported ontologies from the gold standard ontology in order to create a simplified/light version of it. Specifically we removed the SOSA[10], the DAHCC[11] and the PMDO[12] ontologies. This "light" ontology excluded certain complexities found in the original (Wear4PDmove), enabling a focused comparison with a ground truth constructed solely by experts. The intention was to discern the alignment of LLM-extracted ontologies with a more streamlined expert-based conceptualization of the domain. Also, comparing the above methodologies to a "light" expert-based ground truth (ontology) facilitates a more direct evaluation of the LLMs' performance in capturing the essential elements of PD monitoring and alerting without extraneous informative details. This comparison can highlight the LLMs' effectiveness in essential knowledge capture and representation. To assess the accuracy and consistency of the constructed ontologies compared to this version of gold standard ontology, we have employed the metrics mentioned previously.

As seen in Table 4, while the ChatGPT3.5 and ChatGPT4 methods with CoT and OS approaches showed varying levels of success, their X-HCOME counterparts showed better F-1 score, indicating a better balance of precision and recall. Notably, GEMINI X-HCOME achieved the highest F-1 score of 36%, significantly outperforming other methods. This suggests that the X-HCOME method is particularly effective in achieving a balance between accuracy and comprehensiveness in ontology creation tasks.

[10] *http://www.w3.org/ns/sosa/*
[11] *https://dahcc.idlab.ugent.be/Ontology/SensorsAndWearables/*
[12] *http://www.case.edu/PMDO*

This indicates the X-HCOME method's enhanced ability to identify a broader range of relevant classes, showcasing its overall superiority in ontology creation tasks. The F1 score for the object attributes across all LLMs varied from 6% to 24%.[13]

Table 4. Comparative evaluation of methods used for ontology generation against the simplified-/light version of the gold standard ontology.

Method	Number of Classes	True Positives	False Positives	False Negatives	Precision	Recall	F-1 score
Simplified-Lite Gold standard ontology	27						
ChatGPT3.5 CoT	3	2	1	25	67%	7%	13%
ChatGPT3.5 OS	5	3	2	24	60%	11%	19%
ChatGPT3.5 X-HCOME	25	5	20	22	20%	19%	19%
ChatGPT4 CoT	9	3	6	24	33%	11%	17%
ChatGPT4 OS	9	2	7	25	22%	7%	11%
ChatGPT4 X-HCOME	33	6	27	21	18%	22%	20%
GEMINI CoT	9	2	7	25	22%	7%	11%
GEMINI OS	14	1	13	26	7%	4%	5%
GEMINI X-HCOME	50	14	36	13	28%	52%	36%
Llama2 CoT	3	0	3	27	0%	0%	0%
Llama2 OS	2	1	1	26	50%	4%	7%
Llama2 X-HCOME	34	3	31	24	9%	11%	10%

6. Discussion

The research study presented in this paper partially confirmed our initial hypothesis that LLMs can autonomously develop an ontology for PD monitoring and alerting patients when provided with domain-specific input (aim, scope, requirements, competency questions, data). While LLMs demonstrated the capability to construct an ontology, the comprehensiveness of these ontologies did not fully align with our expectations. LLMs have efficiently acquired knowledge from big data repositories and generated ontologies using various prompting engineering techniques, yet the resulting ontologies were not as comprehensive as anticipated. This suggests that while LLMs are effective in ontology creation, their output still requires further refinement

[13] *https://github.com/GiorgosBouh/Ontologies_by_LLMs*

to achieve comprehensive knowledge representation in specific domains like PD monitoring and alerting of patients.

On the other hand, our second hypothesis, which stated that combining human expertise with LLM capabilities improves the developed ontology's quality and applicability was confirmed for PD monitoring and alerting of patients. Our study demonstrated that the X-HCOME methodology, which is enhanced by the capabilities of LLMs, is a robust approach for developing quality ontologies in the PD domain. This methodology not only enhances the structural integrity of ontologies but also enriches them with a more extensive range of knowledge, ensuring their vitality and relevance to contemporary needs, while also showcasing notable time efficiency. Moreover, the collaboration between human expertise and advanced LLMs in OE holds great potential for future developments. It paves the way for more intelligent, adaptive, and comprehensive knowledge representation systems that can significantly contribute to the advancement of various fields, especially in complex areas like healthcare. Through expert revision, particularly evident in the significant improvements seen in precision and F-1 scores, our findings underscore the value of expert intervention in enhancing ontology generation, particularly in mitigating false positives. Notably, the X-HCOME method demonstrated excellence post-revision, showcasing its potential for ontology refinement.

However, biases such as interpretation bias resulting from the opinions and experiences of specific domain experts, as well as biases inherent in LLMs due to their training with unfair or biased algorithms and data, may be present in hybrid methods such as X-HCOME. These biases might affect how valid and correct the knowledge that comes from LLMs is. The results of experiments suggest that ontologies generated by LLMs using a well-defined collaborative OE methodology may have the potential to be comparable to those created solely by humans. This indicates the importance of considering hybrid approaches in OE, which enable collaboration between humans and machines, potentially enhancing efficiency in knowledge-based tasks for both parties involved. Moreover, another limitation of the current study is that it might have oversimplified the ontology-building process by using the number of classes generated as a crucial metric to evaluate ontology-building methodologies (OS, CoT, and X-HCOME). This perspective may have led to an oversight of other crucial aspects such as data/object properties and diverse axioms. These entities are essential for crafting a rich and expansive ontology. Unfortunately, they were not thoroughly investigated in this research, indicating a potential gap in fully realizing a comprehensive and detailed ontology development. While object properties were also calculated in the current, details of these findings are available in the associated GitHub repository[14].

The promising results of X-HCOME in our study suggest its potential, yet they also underscore the need for significant refinement and enhancement before it can be considered a revolutionary methodology in OE. Given the complexities of ontology construction, X-HCOME requires further development for comprehensive and accurate ontology creation. Additionally, extensive practice with this methodology by ontology engineers and domain experts across various domains is essential to fully harness its capabilities and adapt it effectively to diverse knowledge areas.

---

[14] *https://github.com/GiorgosBouh/Ontologies_by_LLMs*

Regarding future work, it would be intriguing to explore the development of a specialized GPT (Generative Pre-trained Transformer) model that is tailored specifically for ontology construction, utilizing the X-HCOME methodology. This could involve training a GPT on datasets that are representative of ontology structures and concepts, aligned with the principles and techniques of the X-HCOME approach. Such an attempt would not only harness the advanced capabilities of GPTs in understanding and generating complex language patterns but also integrate the methodological strengths of X-HCOME. As OE continues to evolve, the integration of methodologies like X-HCOME will play a pivotal role in shaping the future of knowledge representation, offering new possibilities for innovation and improvement in various domains.

References

1. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., Mccandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language Models are Few-Shot Learners. Adv Neural Inf Process Syst. 33, 1877–1901 (2020).
2. Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H.W., Sutton, C., Gehrmann, S., Schuh, P., Shi, K., Tsvyashchenko, S., Maynez, J., Rao, A., Barnes, P., Tay, Y., Shazeer, N., Prabhakaran, V., Reif, E., Du, N., Hutchinson, B., Pope, R., Bradbury, J., Austin, J., Isard, M., Gur-Ari, G., Yin, P., Duke, T., Levskaya, A., Ghemawat, S., Dev, S., Michalewski, H., Garcia, X., Misra, V., Robinson, K., Fedus, L., Zhou, D., Ippolito, D., Luan, D., Lim, H., Zoph, B., Spiridonov, A., Sepassi, R., Dohan, D., Agrawal, S., Omernick, M., Dai, A.M., Pillai, T.S., Pellat, M., Lewkowycz, A., Moreira, E., Child, R., Polozov, O., Lee, K., Zhou, Z., Wang, X., Saeta, B., Diaz, M., Firat, O., Catasta, M., Wei, J., Meier-Hellstern, K., Eck, D., Dean, J., Petrov, S., Fiedel, N.: PaLM: Scaling Language Modeling with Pathways. (2022). https://doi.org/https://doi.org/10.48550/arXiv.2204.02311.
3. Uschold, M., Gruninger, M.: Ontologies: Principles, methods and applications. Knowledge Engineering Review. 11, 93–136 (1996). https://doi.org/10.1017/s0269888900007797.
4. Sheth, A., Roy, K., Gaur, M.: Neurosymbolic AI -- Why, What, and How. d, (2023).
5. Corrà, M.F., Vila-Chã, N., Sardoeira, A., Hansen, C., Sousa, A.P., Reis, I., Sambayeta, F., Damásio, J., Calejo, M., Schicketmueller, A., Laranjinha, I., Salgado, P., Taipa, R., Magalhães, R., Correia, M., Maetzler, W., Maia, L.F.: Peripheral neuropathy in Parkinson's disease: prevalence and functional impact on gait and balance. Brain. 146, 225–236 (2023). https://doi.org/10.1093/BRAIN/AWAC026.
6. Bonuccelli, U., Ceravolo, R.: The safety of dopamine agonists in the treatment of Parkinson's disease. Expert Opin Drug Saf. 7, 111–127 (2008). https://doi.org/10.1517/14740338.7.2.111.
7. Zafeiropolos, N., Bitilis, P., Kotis, K.: Wear4pdmove: An Ontology for Knowledge-Based Personalized Health Monitoring of PD Patients. CEUR Workshop Proc. 3632, 4 (2023).
8. Bitilis, P., Zafeiropoulos, N., Koletis, A., Kotis, K.: Uncovering the Semantics of PD Patients' Movement Data Collected via off-the-shelf Wearables. In: 14th International

Conference on Information, Intelligence, Systems and Applications, IISA 2023 (2023). https://doi.org/10.1109/IISA59645.2023.10345958.
9. Zafeiropoulos, N., Bitilis, P., Tsekouras, G.E., Kotis, K.: Graph Neural Networks for Parkinson's Disease Monitoring and Alerting. Sensors (Basel). 23, 8936 (2023). https://doi.org/10.3390/s23218936.
10. Younesi, E., Malhotra, A., Gündel, M., Scordis, P., Kodamullil, A.T., Page, M., Müller, B., Springstubbe, S., Wüllner, U., Scheller, D., Hofmann-Apitius, M.: PDON: Parkinson's disease ontology for representation and modeling of the Parkinson's disease knowledge domain. Theor Biol Med Model. 12, (2015). https://doi.org/10.1186/S12976-015-0017-Y.
11. Sachdeva, N., Coleman, B., Kang, W.-C., Ni, J., Hong, L., Chi, E.H., Caverlee, J., McAuley, J., Cheng, D.Z.: How to Train Data-Efficient LLMs. (2024).
12. Kotis, K., Vouros, G.A.: Human-centered ontology engineering: The HCOME methodology. Knowl Inf Syst. 10, 109–131 (2006). https://doi.org/10.1007/s10115-005-0227-4.
13. Oksanen, J., Cocarascu, O., Toni, F.: Automatic Product Ontology Extraction from Textual Reviews. Association for Computing Machinery (2021).
14. He, Y., Chen, J., Antonyrajah, D., Horrocks, I.: BERTMap: A BERT-Based Ontology Alignment System. Proceedings of the 36th AAAI Conference on Artificial Intelligence, AAAI 2022. 36BERTMap:, 5684–5691 (2022). https://doi.org/10.1609/aaai.v36i5.20510.
15. Ning, X., Celebi, R.: Knowledge Base Construction from Pre-trained Language Models by Prompt learning, https://ceur-ws.org/Vol-3274/paper4.pdf, (2022).
16. Lippolis, A.S., Klironomos, A., Milon-Flores, D.F., Zheng, H., Jouglar, A., Norouzi, E., Hogan, A.: Enhancing Entity Alignment Between Wikidata and ArtGraph Using LLMs. CEUR Workshop Proc. 3540, (2023).
17. Funk, M., Hosemann, S., Jung, J.C., Lutz, C.: Towards Ontology Construction with Language Models. CEUR Workshop Proc. 3577, (2023).
18. Biester, F., Gaudio, D. Del, Abdelaal, M.: Enhancing Knowledge Base Construction from Pre-trained Language Models using Prompt Ensembles. CEUR Workshop Proc. 3577, (2023).
19. Mountantonakis, M., Tzitzikas, Y.: Validating ChatGPT Facts through RDF Knowledge Graphs and Sentence Similarity. (2023).
20. Pan, S., Luo, L., Wang, Y., Chen, C., Wang, J., Wu, X.: Unifying Large Language Models and Knowledge Graphs: A Roadmap. 14, 1–29 (2023).
21. Joachimiak, M.P., Caufield, J.H., Harris, N.L., Kim, H., Mungall, C.J.: Gene Set Summarization using Large Language Models. (2023).
22. Caufield, J.H., Hegde, H., Emonet, V., Harris, N.L., Joachimiak, M.P., Matentzoglu, N., Kim, H., Moxon, S.A.T., Reese, J.T., Haendel, M.A., Robinson, P.N., Mungall, C.J.: Structured prompt interrogation and recursive extraction of semantics (SPIRES): A method for populating knowledge bases using zero-shot learning. 1–19 (2023).
23. Mateiu, P., Groza, A.: Ontology engineering with Large Language Models. (2023).
24. Zeakis, A., Papadakis, G., Skoutas, D., Koubarakis, M.: Pre-trained Embeddings for Entity Resolution: An Experimental Analysis. Proceedings of the VLDB Endowment. 16, 2225–2238 (2023). https://doi.org/10.14778/3598581.3598594.
25. Jiménez-Ruiz, E., Cuenca Grau, B.: LogMap: Logic-based and scalable ontology matching. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial

Intelligence and Lecture Notes in Bioinformatics). 7031 LNCS, 273–288 (2011). https://doi.org/10.1007/978-3-642-25073-6_18.