Leveraging LLMs for Collaborative Ontology Engineering in Parkinson Disease Monitoring and Alerting

Georgios Bouchouras [a,*], Dimitrios Doumanas [a], Andreas Soularidis [a], Konstantinos Kotis [a,**] and George A. Vouros [b]

[a] *Intelligent Systems Laboratory, Department of Cultural Technology and Communication, University of the Aegean, Mytilene, 81100, Greece*
E-mails: cti23010@ct.aegean.gr, cti23009@ct.aegean.gr, soularidis@aegean.gr, kotis@aegean.gr
[b] *Artificial Intelligence Laboratory, Department Of Digital Systems, University of Piraeus, Piraeus, 18534, Greece*
E-mail: georgev@unipi.gr

Abstract. This paper explores the integration of Large Language Models (LLMs) in the engineering of a Parkinson's Disease (PD) monitoring and alerting ontology through four key methodologies: One Shot (OS) prompt techniques, Chain of Thought (CoT) prompts, X-HCOME, and SimX-HCOME+. The primary objective is to determine whether LLMs alone can create comprehensive ontologies and, if not, whether human-LLM collaboration can achieve this goal.Consequently, the paper assesses the effectiveness of LLMs in automated ontology development and the enhancement achieved through human-LLM collaboration. Initial ontology generation was performed using One Shot (OS) and Chain of Thought (CoT) prompts, demonstrating the capability of LLMs to autonomously construct ontologies for PD monitoring and alerting. However, these outputs were not comprehensive and required substantial human refinement to enhance their completeness and accuracy. X-HCOME, a hybrid ontology engineering approach that combines human expertise with LLM capabilities, showed significant improvements in ontology comprehensiveness. This methodology resulted in Ontologies that are very similar to those constructed by experts. Further experimentation with SimX-HCOME+, another hybrid methodology emphasizing continuous human supervision and iterative refinement, highlighted the importance of ongoing human involvement. This approach led to the creation of more comprehensive and accurate ontologies. Overall, the paper underscores the potential of human-LLM collaboration in advancing ontology engineering, particularly in complex domains like PD. The results suggest promising directions for future research, including the development of specialized GPT models for ontology construction.

Keywords: Ontology Engineering, LLMs, Parkinson Disease enhancement of ontology quality through the collaboration of human experts and LLMs

1. Introduction

The integration of LLMs (Large Language Models) with ontological frameworks is gaining prominence in the fields of knowledge representation (KR) and Artificial Intelligence (AI) [5, 7]. A noticeable trend is the use of LLMs for the construction, refinement, and mapping of ontologies, tasks traditionally performed and supervised by

[*] Corresponding author. E-mail: cti23010@ct.aegean.gr.
[**] Corresponding author. E-mail: kotis@aegean.gr.

human experts with in-depth domain and ontology engineering knowledge, as KR methods become more demanding [22]. Training LLMs on big data makes expert-level insights across domains more accessible and cost-effective. Moreover, while LLMs are getting more effective at engineering ontologies [9], their capabilities are significantly enhanced in the era of Neurosymbolic AI, i.e., combining the deep and varied knowledge of statistical AI with the semantic reasoning of symbolic AI [21].

Artificial intelligence is particularly significant in addressing complex health problems such as monitoring and alerting patients and doctors to Parkinson Disease (PD), the second most common neurodegenerative disease globally [8]. Despite extensive research, the nature of PD remains elusive, and current treatments offer only partial effectiveness [3]. In response, related ontologies have been developed to enhance understanding, monitoring, alerting, and treatment approaches. Specifically, the Wear4PDmove ontology [2, 24] has been recently developed with the aim of integrating 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' [2, 25]. This and associated ontologies facilitate the critical integration of domain-specific knowledge, making it easier to integrate and reason with health data and promoting PD treatment approaches.

Patients' PD monitoring and alerting requires flexible KR methods to effectively adapt to health changes. LLMs have demonstrated impressive abilities in handling large amounts of data and producing valuable insights from their near-real-time analysis. However, factors such as inadequate reasoning abilities and reliance on specialized health knowledge limit their use in monitoring PD and alerting patients. PD is a complex domain, with distinct contexts, subtle meaning variations, and disease-specific vocabularies. Effectively capturing and expressing this complex knowledge requires fine-tuning and training LLMs for the domain, demanding significant resources that are often unavailable or beyond the capacity of health and medical experts. Additionally, 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 exploits ontologies [23, 25]. However, maintaining these ontologies in this rapidly changing field of PD calls for constant effort and resources. Failure to update or refine the ontology may result in outdated information.This involves developing methods that streamline the ontology engineering process, making it more accessible and less resource-intensive.

Existing research has primarily focused on cooperation among participants, particularly domain experts collaborating with one another. However, real-time collaboration between humans and machines at various levels of participation in the development and improvement of ontologies using the OEM remains relatively underexplored. Notably, research has overlooked the extent of human involvement and the potential contribution of LLM assistance. Currently, many ontology engineers devote excessive time and resources to creating an initial ontology, known as the 'kick-off ontology,' but they often lack effective automated methods for further development and refinement. It is crucial to examine the varying levels of contributions from both humans and machines throughout an OEM to demonstrate the methodology's comprehensiveness and the diversity of results, while aiming to save time and resources.

This paper defines varying levels of human involvement in LLM-based/enhanced ontology engineering, corresponding to different OEMs. These levels range from minimal to moderate human involvement, allowing machines and humans to collaborate effectively. This transition moves from a human-centered to a more machine-centered OEM, with humans gradually transferring decision-making power to machines. This indicates an opportunity for developing new techniques to enhance ontologies, making them more time efficient and comprehensive. In this paper, the authors introduce experiments for ontology engineering, towards engineering an ontology in the challenging PD domain. They also focus on expanding the human-centered collaborative OEM (HCOME) [13] through LLM-based tasks, a concept proposed and evaluated as X-HCOME. The authors additionally utilize another extension, the simulated X-HCOME (SimX-HCOME+), to further enhance and evaluate the methodology. This extension features simulated environments to test the interaction between human and machine contributions under controlled conditions, providing deeper insights into the dynamics of collaborative OEM. The aim is to provide a novel OEM, including both humans and LLMs in the engineering of ontologies, with a focus on comprehensiveness and conciseness of conceptualizations, and the required level. The final product of this is an OEM constructing domain ontologies more effectively and with time efficiency than those used solely by humans or LLMs. The paper focuses on LLM-based collaborative OEM to create comprehensive PD ontologies and discusses findings and limitations of the LLM based collaborative process, identified from the experimental results.

Building upon previously published research [4] this paper studies several significant extensions: a) the implementation and evaluation of a new methodology for LLM-enhanced ontology engineering (SimX-HCOME+); b) the addition of a new capability of the proposed approach to convert a rule from natural language (NL) to Semantic Web Rule Language (SWRL); and c) a comparison of the highest LLM performance and the degree of human involvement, across all methodologies. These contributions aim to improve the comprehensiveness of the human-LLM generated ontology for PD monitoring and alerting.

This paper's organization is as follows: Section 2 presents related work on integrating LLMs into OEM; Section 3 describes the proposed research methodology and hypotheses; Section 4 presents the results of the conducted experiments; Section 5 presents a comprehensive evaluation comparing the performance of LLMs across various methodologies, highlighting the degree of human involvement in each approach; finally, section 6 discusses the results and draws 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 [18]. 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 excelling 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 [10]. 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 [17].

Lippolis et al. (2023) 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 [14]. Funk et al. (2023) investigates the capability of GPT3.5 (Generative Pre-trained Tranformer), 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 [9]. 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 [1]. 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 [16]. 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 [19]. 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 [12]. 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 [6]. 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 [15].

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 paper, is the integration of human expertise and LLMs capabilities in the process of OEM. This novel approach aims to harness the large corpus of knowledge, speed in shaping results 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 [26].

3. Research Methodology

This section presents experiments encompassing four 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 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 scenario, the model lacks multiple learning examples and must accomplish the task with minimal context. This is a straightforward approach where the model uses its pre-trained knowledge to infer the most likely answer. For the purposes of this paper, 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 designed to focus on a specific element of the overall task. By employing sequential prompting, the authors direct the language model to tackle each segment of the problem individually, thereby facilitating a cumulative build-up of information. Subsequently, in following experiments, hybrid OEMs are established, which integrates human expertise with the abilities of LLMs. This collaboration aims to elevate the comprehensiveness of the ontology within the PD monitoring and alerting framework. Figure 1 depicts a flowchart that outlines this four-phase experimental process. Initially, four LLMs independently develop an ontology with minimal human input (experiment 1). The process evolves into a more collaborative approach (Human and LLMs) (experiments 2-4). The authors compare the resulting ontologies against a gold standard ontology. In this paper, the Wear4PDmove [7, 8] is utilized as the gold standard ontology, and it will be referred to as such throughout the remainder of the paper.

Hypothesis 1: LLMs, when prompted with domain-specific queries, can autonomously develop a 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 corpus of domain knowledge, and construct ontologies using different prompts provided by humans. This hypothesis is tested in experiment 1, where LLMs are tasked with creating a PD monitoring and alerting ontology from ground zero, using domain-specific prompts. The effectiveness of LLMs in developing an accurate and relevant ontology is measured against the gold standard ontology.

Experiment 1: Initiating LLMs to develop the ontology. During the initial phase of the experiments, the LLMs will independently (minimum human-involvement) construct an ontology for PD monitoring and alerting 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 consistency with OOPS![2]and Protégé tools (Pellet)[3].

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

[2] https://oops.linkeddata.es.

[3] https://protege.stanford.edu.

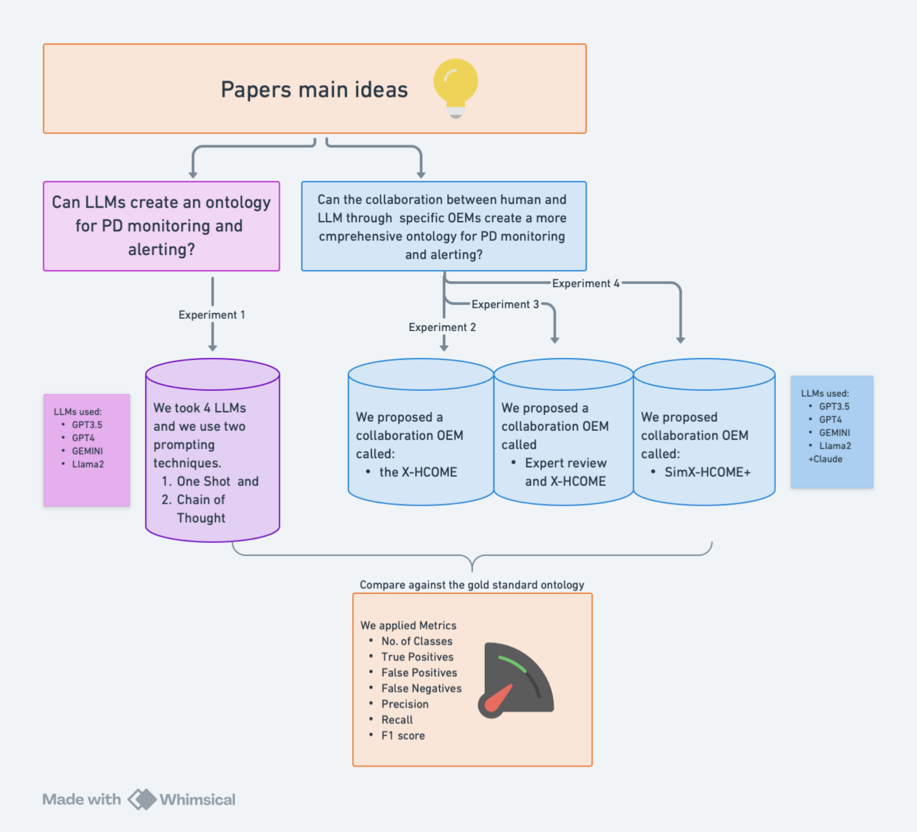
Papers main ideas
Can LLMs create an ontology for PD monitoring and alerting?
Can the collaboration between human and LLM through specific OEMs create a more cmprehensive ontology for PD monitoring and alerting?
Experiment 1
Experiment 2
Experiment 3
Experiment 4
LLMs used:
• GPT3.5
• GPT4
• GEMINI
• Llama2
We took 4 LLMs and we use two prompting techniques.
1. One Shot and
2. Chain of Thought
We proposed a collaboration OEM called:
• the X-HCOME
We proposed a collaboration OEM called
• Expert review and X-HCOME
We proposed collaboration OEM called:
• SimX-HCOME+
LLMs used:
• GPT3.5
• GPT4
• GEMINI
• Llama2
+Claude
Compare against the gold standard ontology
We applied Metrics
• No. of Classes
• True Positives
• False Positives
• False Negatives
• Precision
• Recall
• F1 score
Made with Whimsical

Fig. 1. lowchart of a multi-phase experimentation assessing the construction and validation of ontologies using different methodologies created with AI-Whimsical ChatGPT, 2023[1]

3. Use metrics such as Precision, Recall, and the F-1-score (Table 1) to compare the LLM-generated ontology comprehensiveness against the gold standard ontology.

Table 1

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

Formulas	**Definitions**
Precision = $\frac{\text{True Positives}}{\text{True Positives}+\text{False Positives}}$	True Positives: classes correctly classified as positive in alignment with the 'gold standard' ontology (human judgment, alignment tool)
Recall = $\frac{\text{True Positives}}{\text{True Positives}+\text{False Negatives}}$	False Positives: classes incorrectly classified as positive in alignment with the 'gold standard' ontology
F-1 score = $2 \times \frac{\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}}$	False Negatives: classes that are incorrectly classified as negative despite being positive in the 'gold standard' ontology

The methods in this section and the results listed below, supported by supplementary material placed at a GitHub repository[4], focus on the complex process of creating ontologies for monitoring and alerting patients in PD. It is essential to clarify that the metrics presented in this paper are solely focused on the generated ontological classes and object properties. The validation involves both exact matching, where generated entities corresponded to entities in the gold standard ontology, and similarity matching, where entities are considered correct if they were semantically similar to the gold standard ones. Exact matching quantifies direct accuracies, while similarity matching captures the broader context and appropriateness of the generated entities. This approach aims to provide a comprehensive evaluation of the LLM's performance, capturing both direct accuracies and contextually appropriate approximations.

LLMs are initially given prompts with two methods. The one-shot prompting (OS) method provided the LLMs with a single, clear prompt that clearly 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 ontologies from a single standalone prompt. Along with this 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 as follows: 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 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.

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

This hypothesis is related to the second experimentation, specifies a new OEM called 'X-HCOME'. 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 human-centered ones) in the OE lifecycle. 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. The X-HCOME methodology, is 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. The experts play a crucial role in spotting variations and complexities that automated systems might miss, guaranteeing a technically sound, contextually rich, and real-world application-aligned ontology.

Experiment 2. The X-HCOME methodology that this paper presents 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. a) Define the aim and scope of the ontology: Explain the reasons for its development and the depth of the information it aims to encompass. b) Ontology Requirements: Enumerate the necessary knowledge that must be represented and explain its significance. c)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). d) Formulate specific questions (competency questions) in natural language that the ontology should be able to answer, as defined by knowledge workers.

[4] https://github.com/GiorgosBouh/Ontologies_by_LLMst.

2. (LLM): Construct a domain ontology using the input provided by the human, 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 [11].
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.

Hypothesis 3: Analyzing false positives and incorporating domain expert opinions, LLMs can identify relevant domain knowledge not included in the gold standard ontology.

Experiment 3. The expert review of the X-HCOME. To enhance the evaluation of the generated ontologies, the authors conducted an in-depth analysis of the false positives. Acting as domain experts, they assessed whether the LLMs provided domain knowledge that the gold standard ontology might have missed due to human bias or other limitations in the original engineering. The goal of this analysis was to determine if the LLM generated entities could be reclassified as true positives, even though they didn't match entities in the gold standard ontology, thereby enhancing the ontology. In this case, incorporating expert opinion was critical for expanding and enhancing the domain knowledge represented in the gold standard ontology. This method demonstrates an ever-changing way of thinking about ontology construction—a conversation between humans and machine intelligence that goes back and forth. By embracing this perspective, this experiment holds the promise of significantly advancing the field.

Hypothesis 4: Simulated collaboration between human experts and LLMs enhances ontology engineering by introducing a methodology where LLMs lead ontology development tasks within a controlled environment.

Experiment 4. This newly introduced methodology, named Simulated X-HCOME (SimX-HCOME+, allows LLMs to leverage their strengths in NLP and knowledge extraction to autonomously build ontologies under human supervision. The simulated environment is one where LLMs autonomously develop ontologies using NLP and knowledge extraction, under human expert supervision.This methodology takes the human-LLM collaboration/teaming a step further. The authors assume that this method would further integrate human and machine intelligence in ontology construction, potentially advancing the field by improving the efficiency and accuracy of ontology development.

SimX-HCOME+ introduces a simulated environment where LLMs take the lead in ontology development tasks, but under the supervision of human experts. Here LLMs leverage their capabilities in NLP and knowledge extraction to autonomously build ontologies. However, human supervision and intervention remain crucial. An iterative conversation between the three main roles Knowledge Worker (KW), Domain Expert (DE) and Knowledge Engineer (KE) is simulated. This approach is incorporating continuous ontology generation and refinement throughout the iterative process. It ensures that ontologies are produced at every step, allowing for more comprehensive and detailed results. Human experts play a more inclusive and active role, closely supervising and refining the ontologies generated by the LLMs at each iteration. This iterative refinement emphasizes the importance of human intervention, which can range from overseeing discussions between the three main roles (KW,DE, KE) to participating directly as one of the roles. The first prompt assigns the LLM both the initial role-playing simulation task and the specific OE role it will assume throughout the OE lifecycle. A supervisor who oversees the discussion between the three simulated roles and intervenes when necessary, or a human individual who assumes one of the three roles while allowing the machine to play the other two, can contribute in different ways.

During the OE lifecycle, the LLM user (human) feeds the LLM with related data, such as the aim and scope of the ontology, competency questions, and prompts the model to perform the LLM tasks defined in X-HCOME. These tasks involve constructive discussions among the three roles. When the collaborative and iterative execution of OE tasks ends, the final outcome (generated ontology) is delivered and evaluated.

The authors assessed the accuracy of LLM in properly identifying the number of entities, as well as its capability to transform rules from natural language (NL) to Semantic Web Rule Language (SWRL). The purpose of SWRL is to enable the creation and application of rules for reasoning about the relationships between entities within ontologies.The requested rule for the LLMs to generate and locate in the gold ontology was as follows: "If an observation indicates that there is bradykinesia of the upper limb (indicating slow movement) and this observation pertains to the property and the observation is made after medication dosing, then a notification should be sent indicating a <MissingDoseNotification> and this observation should be marked as a <PDpatientMissingDoseEventObservation>".

4. Results

4.1. Experiments 1 and 2 (OS, CoT and X-HCOME):

Ontological class definition consistency and syntactical correctness were observed in all LLM and collaborative generated ontologies, apart from the ones generated by Llama2 (OS, CoT and X-HCOME). The ontologies generated by Llama2 contained both syntactical errors and inconsistent definitions, which hindered its ability to produce 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.

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, Bard/Gemini OS had lower precision (8%) and recall (2%), identifying 13 classes, whereas Bard/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 accuracy but fewer classes. Conversely, OS tended to identify more classes 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%, a recall of 24%, and an F-1 score of 27%. Remarkably, the Bard/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. The Llama2 results indicated syntactical errors. 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 X-HCOME methodology performed better in all LLMs. This conclusion is drawn from the 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. The Bard/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.

As for the object properties, the F-1 score across all methods varied from 6% to 12% indicating low performance. For the ChatGPT3.5 CoT, ChatGPT3.5 OS, and BARD/Gemini OS methods, the F1 score was 0%. The ChatGPT3.5 X-HCOME method achieved an F1 score of 12%. ChatGPT4 CoT and OS both had an F1 score of 12%. The ChatGPT4 X-HCOME method demonstrated a lower F1 score of 6%. The BARD/Gemini CoT method also had an F1 score of 6%, while BARD/Gemini X-HCOME showed an F1 score of 12%. Llama2 CoT achieved an F1 score of 12%, and Llama2 OS had an F1 score of 6%. Finally, the Llama2 X-HCOME method resulted in an F1 score of 0%. The complete table of results for object properties is in the GitHub repository[5]

[5] https://github.com/GiorgosBouh/Ontologies_by_LLMst.

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%
Bard/Gemini CoT	8	5	3	36	63%	12%	20%
Bard/Gemini OS	13	1	12	40	8%	2%	4%
Bard/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%

4.2. Experiment 3 (Expert review of the X-HCOME):

The ChatGPT3.5 CoT and OS methods have comparable results, with the CoT method showing slightly higher precision but equal recall and an 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). Specifically, for ChatGPT3.5, X-HCOME significantly improved metrics for classes, achieving 92% precision, 56% recall, and a 70% F-1 score, compared to lower scores for CoT and OS. ChatGPT4 showed similar trends, with X-HCOME achieving 88% precision, 71% recall, and a 78% F-1 score. Bard/Gemini's X-HCOME method excelled, showing no false positives and a high true positive rate, with an F-1 score of 110%, though CoT and OS methods lagged. Values above 100% suggests that the true positives reported exceed the actual positives in the gold ontology. Llama2 had high precision but low recall and struggled with consistent ontology creation. Significantly, the X-HCOME method for both ChatGPT3.5 and ChatGPT4 demonstrated a marked improvement in precision and recall, notably reducing false positives after expert review. The Bard/Gemini X-HCOME method stood out with exceptional precision and recall, indicating no false positives and a high rate of true positives. However, Bard/Gemini's CoT and OS methods lagged considerably behind in these metrics. Llama2's CoT and OS methods achieved high precision but low recall. Notably, Llama2 failed to create a consistent ontology without errors, which is a critical aspect of OE. In summary, the X-HCOME method demonstrated superior performance across all LLMs, including ChatGPT3.5, ChatGPT4, and Bard/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, Bard/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 humans alone might overlook. Patients who have undergone surgical interventions such as deep brain stimulation may significantly alter their medication regimens. 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 who have reduced or switched medications. Also, in patients experiencing significant rigidity, a missed dose of medication can lead to rapid symptom exacerbations. The alert system can be calibrated to be more sensitive and prompt in these cases, ensuring quick notification of a missed dose to prevent the 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 patient care and intervention.

As for the object properties, the F-1 score across all LLMs varied from 6% to 84%. Specifically, for the ChatGPT3.5 CoT and ChatGPT3.5 OS methods, the F-1 score was 0%, indicating no effective prediction. The ChatGPT3.5 X-HCOME method achieved an F-1 score of 60%. ChatGPT4 CoT and ChatGPT4 OS both had an F1 score of 12%. The ChatGPT4 X-HCOME method demonstrated a higher F-1 score of 64%. The Bard/Gemini CoT method had an F-1 score of 6%, while Bard/Gemini OS showed an F-1 score of 0%. Bard/Gemini X-HCOME displayed an F1 score of 84%. Llama2 CoT achieved an F1-score of 13%, and Llama2 OS had an F1 score of 7%. Lastly, the Llama2 X-HCOME method resulted in an F1 score of 0%. All the related metric for object properties are presented at the GitHub repository[6]

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%
Bard/Gemini CoT	8	5	3	36	63%	12%	20%
Bard/Gemini OS	13	1	12	40	8%	2%	4%
Bard/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%

4.3. Experiment 4 (SimX-HCOME+):

For SimX-HCOME, the evaluation criteria include ontology reusability, consistency (using Pellet Reasoner), syntactical errors, and whether the ontology can be opened by Protege. ChatGPT4, ChatGPT3.5, and Claude all achieved ontology reusability, consistency without syntactical errors, and could be opened by Protege. Gemini, while reusing ontology and being editable by Protege, had syntactical errors. Finally, all the developed ontologies were validated with OOPS!, identifying only one minor pitfall (pitfall P36-URI, file extension) during the experimental process.

The evaluation metrics of the SimX-HCOME+ generated ontologies in the PD domain reveal varying performances among the methods used (table 4). ChatGPT4 identified 17 classes, with 9 true positives, 8 false positives, and 32 false negatives, resulting in a precision of 52%, recall of 21%, and an F-1 score of 31%. ChatGPT3.5 identified 21 classes, with 14 true positives, 7 false positives, and 27 false negatives, achieving a precision of 66%, recall of 34%, and an F-1 score of 45%. Gemini identified 22 classes, with 15 true positives, 7 false positives, and 26 false negatives, yielding a precision of 68%, recall of 36%, and an F-1 score of 48%. Lastly, Claude identified 24 classes, with 12 true positives, 12 false positives, and 29 false negatives, resulting in a precision of 50%, recall of 29%, and an F-1 score of 37%. These results highlight that Gemini performed the best in terms of F-1 score, indicating a relatively balanced precision and recall among the evaluated methods. Finally, the authors evaluated object properties, but due to space limitations, these results are not presented here. The results obtained for object properties, with this method, were less than optimal, as evidenced by the observed low F-1 scores as presented in the GitHub. Specifically, the F-1 scores for object properties for the ChatGPT-4 method, the F-1 score was 4.75%. The ChatGPT3.5 method achieved an F1 score of 4.25%. Both the Gemini and Claude methods displayed an F-1 score of 5%.The full results for the object properties are in the GitHub repository[7].

[6] https://github.com/GiorgosBouh/Ontologies_by_LLMs.

[7] https://github.com/GiorgosBouh/Ontologies_by_LLMs.

Table 4

Evaluation metrics on SimX-HCOME+ generated ontologies in PD domain (classes).

Method	Number of Classes	True Positives	False Positives	False Negatives	Precision	Recall	F-1 Score
Gold ontology	41						
ChatGPT-4	17	9	8	32	52%	21%	31%
ChatGPT-3.5	21	14	7	27	66%	34%	45%
Gemini	22	15	7	26	68%	36%	48%
Claude	24	12	12	29	50%	29%	37%

Regarding the SWRL rules, while all LLMs except Gemini were able to generate the correct SWRL format, only a small number of logical atoms were detected, resulting in low performance and metrics. Among them, Claude had slightly better results (table 5).

Table 5

Evaluation metrics on SimX-HCOME+ generated ontologies in PD domain (NL2SWRL) with SC: syntactical comparison and LC: Logical Comparison.

Method	Number of Atoms	True Positives SC	True Positives LC	False Positives SC	False Positives LC	False Negatives SC	False Negatives LC	Precision SC (%)	Precision LC (%)	Recall SC (%)	Recall LC (%)	F-1 Score SC	F-1 Score LC
Gold ontology	8												
ChatGPT-4	13	0	3	13	10	8	5	0	23	0	27	0%	13%
ChatGPT-3.5	17	1	3	16	14	7	5	5	17	12.5	3	1%	11%
Gemini	0	0	0	0	0	0	0	0	0	0	0	0%	0%
Claude	12	0	5	12	7	8	3	0	41.6	0	28.4	0%	20%

5. Levels of Human Involvement Across Different Methodological Approaches in OE

All the methodological approaches introduced in this paper align with distinct levels of human-machine collaboration, forming a spectrum from human-centered to LLM-centered collaborative ontology engineering. Within this spectrum, this sectionc presents each methodology proposed. The authors arbitrarily allocated the degrees of human involvement to assess and compare the impact of varying levels of human participation on the ontology engineering process. The authors created a scale from 1 to 5 (with respect to LLMs participation) to assess the different levels of human involvement (table 6). This arbitrary assignment enables controlled analysis, guaranteeing consistent interpretation of the results across various methodological approaches. This approach facilitates a better understanding of the role human expertise plays in ontology engineering, providing valuable insights into the effectiveness of human-LLM collaboration in creating high-quality ontologies.

Table 6

Levels of Human Involvement Across Different Methodological Approaches in Ontology Engineering

Methodological Approach	OS	CoT	SimX-HCOME+	X-HCOME	Expert Review X-HCOME
Level of Human Involvement	1	2	3	4	5

The results indicate that the Expert Review X-HCOME Bard/Gemini model achieved the highest F-1 score, exceeding 100% and nearing 110% with a human involvement level of 5. Sim-X-HCOME+ Gemini also showed a relatively high F-1 score, around 47,6%, with a human involvement level of 3. X-HCOME Bard/Gemini had an F1 score slightly above 40% with the maximum human involvement level of 4. CoT Gemini and OS ChatGPT4 showed lower F-1 scores, around 20%, with lower human involvement levels of 2 and 1 respectively (Figure 2). These findings point to a positive relationship between the models' F-1 scores and the level of human engagement, with greater human involvement typically translating into higher performance.

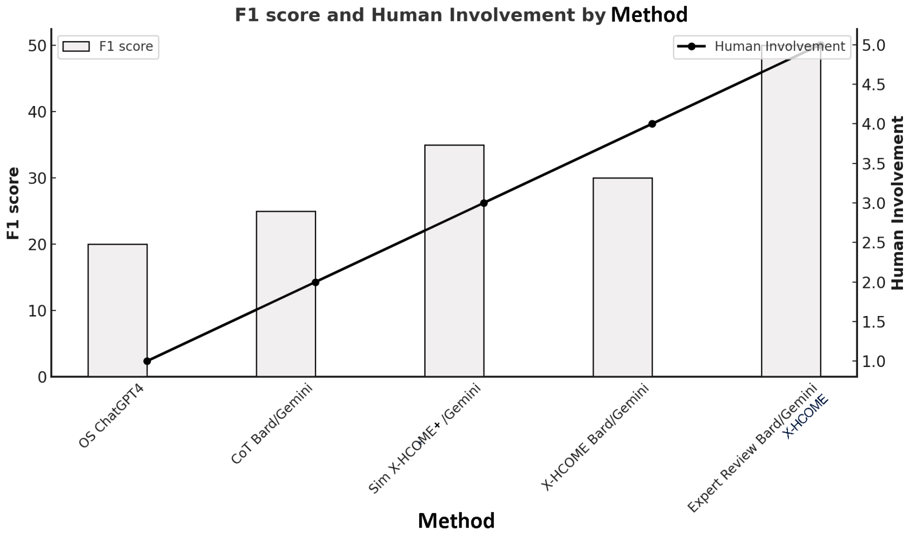

Fig. 2. The graph compares the highest F1 scores from various LLM methods and the degree of human involvement in the PD domain. The x-axis represents the different methods: CoT Gemini, OS ChatGPT4, XHCOME Bard/Gemini, Expert Review X-HCOME Bard/Gemini, and Sim X-HCOME+ Gemini. The left y-axis shows the F1 Score, while the right y-axis indicates the degree of human involvement, measured on a scale from 1 (minimum) to 5 (maximum)

6. Discussion

The results presented in this paper partially confirme the first 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, and data). While LLMs demonstrated the capability to construct an ontology, the comprehensiveness of this ontology did not fully align with the authors expectations. LLMs efficiently acquired knowledge from big data repositories and generated ontologies using various prompting engineering techniques, but 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 to achieve comprehensive knowledge representation in specific domains like PD monitoring and alerting of patients.

On the other hand, the second hypothesis, which stated that combining human expertise with LLM capabilities improves the developed ontology's quality and comprehensiveness, was confirmed for PD monitoring and alerting of patients. The current paper demonstrates that the X-HCOME methodology, which is enhanced by the capabilities of LLMs, provides a robust approach for developing comprehensive ontologies in the PD domain.

Regarding the third hypothesis the results showed that expert revision can improve ontology generation, especially when it comes to reducing false positives. This is especially clear in the large improvements seen in precision and F-1 scores. This kind of collaboration not only improves the structure and usefulness of the ontologies that are made, but it also finds new information and ideas that add to the domain-specific data and help the representation of knowledge keep changing.

Also, concerning the papers fourth hypothesis, the experiments conducted using the SimX-HCOME+ methodology further illustrate the importance of human-LLM collaboration. Experiments highlighted the importance of human involvement in the OE lifecycle, as demonstrated through the iterative discussions and refinements by Knowledge Workers (KW), Domain Experts (DE), and Knowledge Engineers (KE).The inclusion of human experts in the iterative ontology generation and refinement process ensures that the ontologies produced are more comprehensive and detailed. The simulated environment facilitated continuous ontology development, where human experts provided oversight and direct participation, ensuring that the ontologies remained relevant and comprehensive at each step. However, regarding the transformation of NL to SWRL, the method did not fully manage to generate the SWRL rule, presenting a significant challenge for future experiments. This limitation indicates a critical area for improvement and suggests that future research should focus on enhancing LLMs' ability to handle SWRL rule generation effectively.

The conclusions of the paper emphasizes that the three collaborative methods—X-HCOME, SimX-HCOME, and expert review of the X-HCOME—significantly enhance the comprehensiveness and time efficiency of ontology development. By integrating human expertise with the capabilities of LLMs, these methodologies address the limitations of LLMs in generating comprehensive ontologies independently. The X-HCOME approach showed notable improvements in ontology quality through human-LLM collaboration, while the SimX-HCOME methodology demonstrated the benefits of iterative refinement in a simulated environment. Additionally, the expert review of X-HCOME further improved precision and reduced false positives, highlighting the critical role of human oversight. Moreover, the synergy between human expertise and advanced LLMs in OEM holds enormous potential for future developments. It paves the way for more comprehensive knowledge representation systems that can significantly contribute to the advancement of various fields, especially in complex areas like PD. The Expert Review Bard/Gemini model, with the highest F-1 score and significant human involvement, further highlights the critical role of human oversight in achieving high-quality outcomes. There is a strong link between human involvement and methodology performance, suggesting that experts need to be involved to ensure the creation of comprehensive ontologies

However, collaborative methods like X-HCOME ans SimX-HCOME+ may contain inherent biases in LLMs due to their training with unfair or biased algorithms and data, as well as biases resulting from the opinions and experiences of specific domain experts. These biases may affect the validity and correctness of the knowledge that comes from LLMs. The results of experiments suggest that ontologies generated by LLMs using a well-defined collaborative OEM may have the potential to be comparable to those created solely by humans. This indicates the importance of considering hybrid methodologies in OE, which enable collaboration between humans and machines, potentially enhancing efficiency in knowledge-based tasks for both parties involved. Another challenge with this paper is that it may have oversimplified the process of building an ontology by focusing too much on the classes and object properties created as key indicators to compare ontology-building approaches (OS, CoT, and X-HCOME, SimX-HCOME+). This perspective may have led to an oversight of other crucial aspects, such as data properties and diverse axioms. These entities are essential for crafting a comprehensive ontology. Unfortunately, this research did not thoroughly investigate these aspects, revealing a potential gap in developing a comprehensive and detailed ontology. At last, in the collaborative methodologies OEMs, human evaluation is an integral part of the process, which might raise questions about the necessity of comparing the generated ontologies with a gold standard ontology at the end, and the purpose of the metrics used. In any case, the comparison with the gold standard ontology and the use of metrics are essential for validating, benchmarking, and improving the collaborative OEMs, ensuring that the collaborative efforts of humans and LLMs yield high-quality ontologies.

The promising results of the collaborative OEMs (X-HCOME, expert review of the X-HCOME and SimX-HCOME+) in this paper suggest their potential in further research efforts in LLM-enhanced OE, yet they also underscore the need for significant refinement and enhancement before they can be considered revolutionary OEMs. Given the complexities of ontology engineering in general, these methodologies require further development to create comprehensive and accurate ontologies. Future research could benefit from investigating the adaptability and effectiveness of the collaborative ontology engineering approach in diverse healthcare contexts, such as chronic disease management, mental health interventions, and personalized medicine. Expanding the application of these OEMs to a broader range of healthcare domains could provide valuable insights into their versatility and potential impact on knowledge representation systems across various medical specialties Additionally, extensive practice with these methodologies by ontology engineers and domain experts across various fields is essential to fully harness their capabilities and adapt them effectively to diverse knowledge areas. The OEMs proposed lie in a narrow spectrum focused on PD monitoring and alerting, which may limit the generalizability to other domains. Regarding future work, it would be intriguing to explore the development of a specialized GPT model that is tailored specifically for ontology construction, utilizing the X-HCOME and SimX-HCOME+ methodologies. This could involve tuning a GPT on datasets that are representative of ontology structures and concepts, aligned with the principles and techniques of the collaborative OEMs introduced in the current paper. 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 collaborative OEMs. As OE continues to evolve, the integration of these methodologies will play a pivotal role in shaping the future of knowledge representation, offering new possibilities for innovation and improvement in various domains.

References

[1] F. Biester, D.D. Gaudio and M. Abdelaal, Enhancing Knowledge Base Construction from Pre-trained Language Models using Prompt Ensembles, *CEUR Workshop Proceedings* **3577** (2023). https://ceur-ws.org/Vol-3577/paper4.pdf.

[2] P. Bitilis, N. Zafeiropoulos, A. Koletis and K. Kotis, 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. ISBN 9798350318067. doi:10.1109/IISA59645.2023.10345958.

[3] U. Bonuccelli and R. Ceravolo, The safety of dopamine agonists in the treatment of Parkinson's disease, *Expert Opinion on Drug Safety* **7**(2) (2008), 111–127. doi:10.1517/14740338.7.2.111.

[4] Bouchouras Georgios , Bitilis Pavlos , Kotis Konstantinos and V.A. George., LLMs for the Engineering of a Parkinson Disease Monitoring and Alerting Ontology, in: *GeNeSy2024 Workshop hosted by Extended Semantic Web Conference 2024*, 2024. https://sites.google.com/view/genesy2024/program.

[5] T.B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D.M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. Mccandlish, A. Radford, I. Sutskever and D. Amodei, Language Models are Few-Shot Learners, *Advances in Neural Information Processing Systems* **33** (2020), 1877–1901. https://commoncrawl.org/the-data/.

[6] J.H. Caufield, H. Hegde, V. Emonet, N.L. Harris, M.P. Joachimiak, N. Matentzoglu, H. Kim, S.A.T. Moxon, J.T. Reese, M.A. Haendel, P.N. Robinson and C.J. Mungall, Structured prompt interrogation and recursive extraction of semantics (SPIRES): A method for populating knowledge bases using zero-shot learning (2023), 1–19. http://arxiv.org/abs/2304.02711.

[7] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H.W. Chung, C. Sutton, S. Gehrmann, P. Schuh, K. Shi, S. Tsvyashchenko, J. Maynez, A. Rao, P. Barnes, Y. Tay, N. Shazeer, V. Prabhakaran, E. Reif, N. Du, B. Hutchinson, R. Pope, J. Bradbury, J. Austin, M. Isard, G. Gur-Ari, P. Yin, T. Duke, A. Levskaya, S. Ghemawat, S. Dev, H. Michalewski, X. Garcia, V. Misra, K. Robinson, L. Fedus, D. Zhou, D. Ippolito, D. Luan, H. Lim, B. Zoph, A. Spiridonov, R. Sepassi, D. Dohan, S. Agrawal, M. Omernick, A.M. Dai, T.S. Pillai, M. Pellat, A. Lewkowycz, E. Moreira, R. Child, O. Polozov, K. Lee, Z. Zhou, X. Wang, B. Saeta, M. Diaz, O. Firat, M. Catasta, J. Wei, K. Meier-Hellstern, D. Eck, J. Dean, S. Petrov and N. Fiedel, PaLM: Scaling Language Modeling with Pathways (2022). http://arxiv.org/abs/2204.02311.

[8] M.F. Corrà, N. Vila-Chã, A. Sardoeira, C. Hansen, A.P. Sousa, I. Reis, F. Sambayeta, J. Damásio, M. Calejo, A. Schicketmueller, I. Laranjinha, P. Salgado, R. Taipa, R. Magalhães, M. Correia, W. Maetzler and L.F. Maia, Peripheral neuropathy in Parkinson's disease: prevalence and functional impact on gait and balance, *Brain* **146**(1) (2023), 225–236. doi:10.1093/BRAIN/AWAC026.

[9] M. Funk, S. Hosemann, J.C. Jung and C. Lutz, Towards Ontology Construction with Language Models, *CEUR Workshop Proceedings* **3577** (2023). http://arxiv.org/abs/2309.09898.

[10] Y. He, J. Chen, D. Antonyrajah and I. Horrocks, BERTMap: A BERT-Based Ontology Alignment System, *Proceedings of the 36th AAAI Conference on Artificial Intelligence, AAAI 2022* **36BERTMap:** (2022), 5684–5691. ISBN 1577358767. doi:10.1609/aaai.v36i5.20510.

[11] E. Jiménez-Ruiz and B. Cuenca Grau, 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**(PART 1) (2011), 273–288. ISBN 9783642250729. doi:10.1007/978-3-642-25073-6$_1$8.

[12] M.P. Joachimiak, J.H. Caufield, N.L. Harris, H. Kim and C.J. Mungall, Gene Set Summarization using Large Language Models, *ArXiv* **13**(61) (2009), 4. ISBN 15410617. /pmc/articles/PMC10246080//pmc/articles/PMC10246080/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC10246080/.

[13] K. Kotis and G.A. Vouros, Human-centered ontology engineering: The HCOME methodology, *Knowledge and Information Systems* **10**(1) (2006), 109–131. doi:10.1007/s10115-005-0227-4.

[14] A.S. Lippolis, A. Klironomos, D.F. Milon-Flores, H. Zheng, A. Jouglar, E. Norouzi and A. Hogan, Enhancing Entity Alignment Between Wikidata and ArtGraph Using LLMs, *CEUR Workshop Proceedings* **3540** (2023). https://ceur-ws.org/Vol-3540/paper7.pdf.

[15] P. Mateiu and A. Groza, Ontology engineering with Large Language Models, *Proceedings - 2023 25th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing, SYNASC 2023* (2023), 226–229. ISBN 9798350394122. doi:10.1109/SYNASC61333.2023.00038. https://arxiv.org/abs/2307.16699v1.

[16] M. Mountantonakis and Y. Tzitzikas, Real-Time Validation of ChatGPT facts using RDF Knowledge Graphs, *CEUR Workshop Proceedings* **3632** (2023), 0–5.

[17] X. Ning and R. Celebi, Knowledge Base Construction from Pre-trained Language Models by Prompt learning, Vol. 3274, 2022, pp. 46–54. ISSN 16130073. https://ceur-ws.org/Vol-3274/paper4.pdf.

[18] J. Oksanen, O. Cocarascu and F. Toni, *Automatic Product Ontology Extraction from Textual Reviews*, Vol. 1, Association for Computing Machinery, 2021. http://arxiv.org/abs/2105.10966.

[19] S. Pan, L. Luo, Y. Wang, C. Chen, J. Wang and X. Wu, Unifying Large Language Models and Knowledge Graphs: A Roadmap **14**(8) (2023), 1–29. http://arxiv.org/abs/2306.08302.

[20] N. Sachdeva, B. Coleman, W.-C. Kang, J. Ni, L. Hong, E.H. Chi, J. Caverlee, J. McAuley and D.Z. Cheng, How to Train Data-Efficient LLMs (2024). https://arxiv.org/abs/2402.09668v1.

[21] A. Sheth, K. Roy and M. Gaur, Neurosymbolic AI – Why, What, and How **d** (2023). http://arxiv.org/abs/2305.00813.

[22] M. Uschold and M. Gruninger, Ontologies: principles, methods and applications, *The Knowledge Engineering Review* **11**(2) (1996), 93–136. doi:10.1017/S0269888900007797. https://www.cambridge.org/core/journals/knowledge-engineering-review/article/abs/ontologies-principles-methods-and-applications/2443E0A8E5D81A144D8C611EF20043E6.

[23] E. Younesi, A. Malhotra, M. Gündel, P. Scordis, A.T. Kodamullil, M. Page, B. Müller, S. Springstubbe, U. Wüllner, D. Scheller and M. Hofmann-Apitius, PDON: Parkinson's disease ontology for representation and modeling of the Parkinson's disease knowledge domain, *Theoretical Biology Medical Modelling* **12**(1) (2015). doi:10.1186/S12976-015-0017-Y. /pmc/articles/PMC4580356//pmc/articles/PMC4580356/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC4580356/.

[24] N. Zafeiropolos, P. Bitilis and K. Kotis, Wear4pdmove: An Ontology for Knowledge-Based Personalized Health Monitoring of PD Patients, *CEUR Workshop Proceedings* **3632** (2023), 4.

[25] N. Zafeiropoulos, P. Bitilis, G.E. Tsekouras and K. Kotis, Graph Neural Networks for Parkinson's Disease Monitoring and Alerting, *Sensors (Basel, Switzerland)* **23**(21) (2023), 8936. doi:10.3390/s23218936. /pmc/articles/PMC10648881//pmc/articles/PMC10648881/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC10648881/.

[26] A. Zeakis, G. Papadakis, D. Skoutas and M. Koubarakis, Pre-trained Embeddings for Entity Resolution: An Experimental Analysis, *Proceedings of the VLDB Endowment* **16**(9) (2023), 2225–2238. doi:10.14778/3598581.3598594.