applied sciences

MDPI

Article

Fine-Tuning Large Language Models for Ontology Engineering: A Comparative Analysis of GPT-4 and Mistral

Dimitrios Doumanas [1,*], Andreas Soularidis [1], Dimitris Spiliotopoulos [2], Costas Vassilakis [3] and Konstantinos Kotis [1,*]

[1] Intelligent Systems Lab, Department of Cultural Technology and Communication, University of the Aegean, 81100 Mytilene, Greece; soularidis@aegean.gr

[2] Department of Management Science and Technology, University of the Peloponnese, 22100 Tripolis, Greece; dspiliot@uop.gr

[3] Department of Informatics and Telecommunications, University of the Peloponnese, 22100 Tripolis, Greece; costas@uop.gr

* Correspondence: cti23009@ct.aegean.gr (D.D.); kotis@aegean.gr (K.K.)

Abstract: Ontology engineering (OE) plays a critical role in modeling and managing structured knowledge across various domains. This study examines the performance of fine-tuned large language models (LLMs), specifically GPT-4 and Mistral 7B, in efficiently automating OE tasks. Foundational OE textbooks are used as the basis for dataset creation and for feeding the LLMs. The methodology involved segmenting texts into manageable chapters, generating question–answer pairs, and translating visual elements into description logic to curate fine-tuned datasets in JSONL format. This research aims to enhance the models' abilities to generate domain-specific ontologies, with hypotheses asserting that fine-tuned LLMs would outperform base models, and that domain-specific datasets would significantly improve their performance. Comparative experiments revealed that GPT-4 demonstrated superior accuracy and adherence to ontology syntax, albeit with higher computational costs. Conversely, Mistral 7B excelled in speed and cost efficiency but struggled with domain-specific tasks, often generating outputs that lacked syntactical precision and relevance. The presented results highlight the necessity of integrating domain-specific datasets to improve contextual understanding and practical utility in specialized applications, such as Search and Rescue (SAR) missions in wildfire incidents. Both models, despite their limitations, exhibited potential in understanding OE principles. However, their performance underscored the importance of aligning training data with domain-specific knowledge to emulate human expertise effectively. This study, based on and extending our previous work on the topic, concludes that fine-tuned LLMs with targeted datasets enhance their utility in OE, offering insights into improving future models for domain-specific applications. The findings advocate further exploration of hybrid solutions to balance accuracy and efficiency.

Keywords: large language models (LLMs) fine-tuning; ontology engineering (OE); domain-specific knowledge; search and rescue (SAR)

check for updates

Academic Editor: Jose María Alvarez Rodríguez

Received: 22 December 2024
Revised: 4 February 2025
Accepted: 13 February 2025
Published: 18 February 2025

Citation: Doumanas, D.; Soularidis, A.; Spiliotopoulos, D.; Vassilakis, C.; Kotis, K. Fine-Tuning Large Language Models for Ontology Engineering: A Comparative Analysis of GPT-4 and Mistral. *Appl. Sci.* **2025**, *15*, 2146. https://doi.org/10.3390/app15042146

1. Introduction

OE is fundamental to the organization and structuring of semantic knowledge, which is increasingly essential in domains such as artificial intelligence, knowledge management, and digital information systems. Ontologies, by defining concepts and relations between them, serve as the backbone for systems that rely on contextual understanding and reasoning. As the volume and complexity of information grows, traditional manual approaches to OE are proving insufficient. This has led to the exploration of advanced methodologies for OE (especially concerning ontology learning) by leveraging machine learning methods and natural language processing (NLP).

LLMs, such as GPT-4 and Mistral, represent a significant leap in NLP capabilities, enabling automation in tasks that previously required extensive human expertise. By processing vast datasets, LLMs can perform nuanced tasks such as text summarization, content generation, and knowledge extraction. However, their application in OE—a domain requiring deep understanding of structure, logic, and domain-specific knowledge—remains underexplored. Despite their general utility, pre-trained LLMs often fail to meet the precision and relevance required for tasks such as ontology generation, primarily due to a lack of domain-specific training.

Fine-tuning involves a detailed procedure where the model's parameters are adjusted using a smaller, highly specialized dataset after the initial broad pre-training on extensive data. This targeted training approach helps in refining the model's predictions, making it adept at handling specific types of queries and generating outputs that adhere to the specialized knowledge structures typical in OE.

The effectiveness of fine-tuning is evident in the model's enhanced ability to navigate the intricacies of domain-specific languages and terminologies, enabling it to produce outputs that are not only grammatically correct but also contextually aligned with the domain's requirements. By leveraging fine-tuned LLMs, professionals in OE can automate the generation of ontological structures, significantly reducing the time and effort involved in manual ontology construction and potentially increasing the scalability of knowledge management practices across various domains.

Fine-tuning LLMs for OE tasks represents a confluence of advanced AI capabilities with the specialized needs of knowledge management. This process underpins significant improvements in how models understand and interact with complex, domain-specific datasets, marking a pivotal step towards more intelligent and automated systems in ontology management.

The primary challenge addressed in this research lies in the limitations of pre-trained LLMs when applied to OE, creating a gap in their ability to generate outputs with low F1 scores, reflecting poor precision and recall in representing domain-specific concepts and properties (both object and data properties). This shortcoming underscores the importance of fine-tuning LLMs to enhance their understanding of OE principles and methodologies.

The objective of this study is to bridge this gap by fine-tuning LLMs on carefully curated datasets derived from foundational texts in OE. These datasets, encompassing question–answer pairs, description logic, and translations of visual representations, are designed to align with the conceptual and structural needs of OE. The study evaluates the impact of fine-tuning on the models' ability to generate high-quality, domain-specific ontologies. This research is driven by two key research questions:

1. Can fine-tuning LLMs for OE concepts significantly improve their performance, enabling them to generate outputs with higher accuracy and adherence to ontology syntax?
2. Is the incorporation of domain-specific datasets into the fine-tuning process able to enhance the practical utility of the generated ontologies, particularly in real-world applications?

To answer these questions, comparative experiments were conducted on GPT-4 and Mistral 7b, two leading LLMs with distinct capabilities. These experiments assessed the models' performance in generating ontologies, focusing on metrics such as precision, recall, and F1 score. The results provide valuable insights into the capabilities and limitations of LLMs in OE and offer practical recommendations for enhancing their utility in this specialized field.

In this study, we opted to apply fine-tuning directly to the LLMs without first employing retrieval-augmented generation (RAG) to address domain-specific OE tasks. This decision was driven by the need for precision and tailored adaptation to the highly specialized domain requirements. Fine-tuning integrates the domain's specific characteristics into the model's internal parameters, which enhances its ability to produce contextually precise and terminologically consistent outputs. This is particularly critical for tasks such as OE, where maintaining strict adherence to predefined logical structures and workflows is paramount. Conversely, while RAG excels at dynamically retrieving and incorporating external knowledge, it introduces additional complexity and potential latency. Moreover, the reliance of RAG on external, often heterogeneous, data sources could lead to inconsistencies, inaccuracies, or even irrelevance in outputs, especially if the external knowledge is not meticulously curated. By prioritizing fine-tuning, we ensured greater control over the model's behavior and outputs, optimizing for performance in a domain characterized by stable knowledge frameworks rather than dynamic informational updates. This approach aligns with established evidence that fine-tuning is more effective than retrieval-based methods for domains requiring precision, efficiency, and consistent adherence to structured frameworks, making it the logical choice for our objectives in OE.

In addition to contributing to the understanding of LLMs' application in OE, this study highlights the broader implications of integrating machine learning with NLP for domain-specific tasks. By addressing the challenges of domain knowledge representation, this research sets the stage for future advancements in automating complex knowledge-driven processes, offering a pathway towards more intelligent, context-aware systems in diverse fields.

Overall, the goals of the presented work are as follows:

- Enhance OE with Fine-Tuned LLMs: To demonstrate how fine-tuning LLMs can improve their performance in generating and understanding ontologies aligned with formal principles and domain-specific requirements.
- Develop Domain-Specific OE Training Datasets: To create and utilize a structured methodology for curating datasets based on foundational OE-related handbooks, including question–answer pairs and translated visual elements, ensuring alignment with OE standards.
- Provide a Comparative Analysis of LLMs: To evaluate and contrast the performance of GPT-4 and Mistral 7B in terms of accuracy, efficiency, cost, and their ability to handle domain-specific OE tasks, providing insights into their practical applicability.
- Apply Fine-Tuned Models to Real-World Scenarios: To explore the use of fine-tuned LLMs in practical applications, such as Search and Rescue (SAR) missions during wildfires, showcasing the potential utility of these models in critical, specialized fields.
- Identifying Limitations and Future Directions: To critically assess the strengths and weaknesses of the models, emphasizing the need for hybrid solutions and domain-specific training data to enhance their contextual understanding and emulate human expertise effectively.
- Pave the Way for AI-Enhanced OE Tools: To advocate advancements in OE by integrating LLMs with domain-specific insights, aiming to bridge the gap between theoretical principles and practical utility. The approach demonstrates how fine-tuned LLMs can automate key ontology development tasks, including concept extraction, relationship identification, and hierarchical structuring. By leveraging pre-trained models, we reduce the reliance on manual curation, allowing for a more scalable and efficient OE process. This advancement is particularly relevant for domains requiring management systems, where rapid adaptation to evolving technologies is crucial.

By addressing these goals, this paper aims to advance the field of OE by leveraging fine-tuned LLMs, like GPT-4 and Mistral 7B, to enhance the automation and accuracy of ontology generation. It seeks to bridge the gap between theoretical principles and real-world applications by developing domain-specific datasets and applying the models to practical scenarios, such as SAR missions during wildfires. Through a detailed comparative analysis, the study provides actionable insights, offering guidance for selecting and optimizing models for specialized tasks. By highlighting the importance of domain-specific training and exploring hybrid solutions, the paper sets a benchmark for integrating LLMs into complex, knowledge-driven domains, ultimately aiming to inspire further innovation and provide practical tools for researchers and practitioners.

While this study focuses on fine-tuning large language models (LLMs) for ontology engineering, it builds upon the foundation of existing research in automated ontology construction, knowledge representation, and semantic reasoning. A comprehensive review of prior work in these areas, including relevant methodologies and applications of LLMs in ontology engineering, is provided in Section 2. This ensures that our contributions are well grounded within the broader context of the existing literature.

The structure of the paper is as follows. Section 2 presents the related work. Section 3 describes the proposed approach. Section 4 presents the experiments and results, and Section 5 discusses the findings with respect to the research hypotheses. Finally, Section 6 summarizes the key findings of this study and identifies needs for future research.

2. Related Work

2.1. LLMs and OE

Zhang et al. [1] introduce a novel framework to streamline the process of OE (OE) through conversational interactions with a language model. Their study tackles the inherent complexities and resource-intensive nature of traditional OE, particularly in projects involving stakeholders from diverse backgrounds. OntoChat aims to mitigate systematic ambiguities and biases by facilitating requirement elicitation, analysis, and testing phases through a conversational AI interface. The framework enables users to create user stories and extract competency questions by interacting with a conversational agent, significantly reducing the manual effort typically required in these initial stages. The use of large language models (LLMs) in OntoChat allows for nuanced understanding and processing of natural language, enhancing the generation and refinement of ontology requirements. The paper details an evaluation of OntoChat using the Music Meta Ontology, demonstrating its effectiveness in improving the efficiency and accuracy of OE tasks. This approach not only streamlines the OE process but also promises to enhance the accessibility and adaptability of ontologies in various domains by incorporating advanced AI-driven methods.

Doumanas et al. [2] detail the evolution and application of OE methods in the context of large language models (LLMs), particularly focusing on human-to-machine centered methodologies. The paper discusses the gradual reduction in human involvement and corresponding increase in machine automation in the process of OE. This shift aims to leverage the efficiency of LLMs while retaining essential human oversight. The proposed spectrum of collaboration ranges from fully human-driven to completely LLM-driven OE methodologies. Through systematic experimentation across various levels of human and machine collaboration, the research evaluates the effectiveness of LLMs in creating ontologically sound structures with minimal human input. However, it is important to note that in this work, fine-tuning is not researched or evaluated; the study focuses solely on simple prompting techniques. The study provides a compelling look into how LLMs can significantly augment the OE process, making it faster and potentially more accurate while still needing human expertise for critical assessments and adjustments.

Garijo et al. [3] provide a detailed examination of how LLMs are applied in the domain of OE (OE). The paper categorizes various OE tasks that can benefit from LLMs, utilizing the Linked Open Terms (LOT) methodology as a framework for analysis. This research identifies the core phases of OE such as requirement specification, implementation, and maintenance, and highlights how LLMs are currently utilized within these areas. The authors discuss the diversity in task definitions and the need for standardized benchmarks and evaluation frameworks to measure LLM performance effectively in OE tasks. Through their analysis, Garijo et al. reveal significant gaps and suggest potential areas for further integration of LLMs to enhance the efficiency and effectiveness of OE processes.

Joachimiak et al. [4] discuss the development of the Artificial Intelligence Ontology (AIO), which organizes and defines AI concepts to support standardization and understanding in AI research. Developed at Lawrence Berkeley National Laboratory, the AIO employs LLMs to enhance manual curation, ensuring the ontology remains current with rapid advancements in AI technology. AIO is structured around six principal categories including Networks, Layers, and Bias, addressing both technical aspects and ethical considerations of AI. This approach not only facilitates modular composition of AI methodologies but also aids in navigating the ethical landscapes of AI applications, proving to be a vital resource for AI researchers and developers.

Saeedizade et al. [5] investigate the potential of LLMs to assist in the development of ontologies by generating OWL outputs directly from ontological requirements. The study explores several state-of-the-art models, utilizing a variety of prompting techniques such as Chain of Thoughts (CoT), Graph of Thoughts (GoT), and Decomposed Prompting to assess the ability of LLMs to produce sufficient quality OWL suggestions. The research demonstrates that GPT-4 is particularly capable of generating high-quality modeling suggestions, significantly outperforming other models in generating structurally and syntactically correct OWL files. This paper highlights the importance of carefully selected prompting techniques to leverage the capabilities of LLMs effectively in OE, suggesting a promising direction for automating this traditionally labor-intensive task.

Mateiu et al. [6] explore the potential of fine-tuning GPT-3 to automate the translation of natural language sentences into Description Logic, specifically into OWL Functional Syntax. Their research involves developing a Protégé plugin that assists in both developing new ontologies and enriching existing ones by automatically translating domain-specific natural language into formal OWL axioms under human supervision. The authors fine-tuned the GPT-3 model using a dataset of 150 natural languages to OWL translation pairs, covering a variety of instances and relationships. This approach leverages the linguistic prowess of LLMs to streamline the OE process, potentially reducing the technical challenges and costs traditionally associated with manual ontology development. Their work demonstrates a significant advancement in applying LLMs to semantic web technologies, particularly in automating and facilitating the labor-intensive processes of OE.

Doumanas et al. [7] explore the application of LLMs to OE within the context of Search and Rescue (SAR) operations. Their research demonstrates how LLMs can be effectively utilized to automate the construction of ontologies, particularly for complex and dynamic scenarios like SAR missions in wildfire incidents. However, it is important to note that in this work, fine-tuning is not researched or evaluated; the study focuses solely on simple prompting techniques. The paper proposes a novel collaborative OE methodology that harnesses both human expertise and the advanced natural language processing capabilities of LLMs. This hybrid approach facilitates the rapid development of domain-specific ontologies, leveraging the LLMs' ability to process and structure large datasets into ontological knowledge, which is critical for enhancing situational awareness and decision-making in SAR operations. The integration of LLMs into OE represents a significant advancement in making these operations more efficient and effective by providing a structured knowledge framework that supports the dynamic requirements of SAR missions.

2.2. LLMs and Fine-Tuning

Gekhman et al. [8] investigate the effects of supervised fine-tuning of large language models (LLMs) on the introduction of new factual knowledge. They explore whether such fine-tuning leads to an increase in the generation of factually incorrect responses—a phenomenon known as hallucination. The study reveals that LLMs tend to acquire new facts at a slower rate during fine-tuning, compared to facts that align with their pre-existing knowledge. Moreover, the findings indicate a linear relationship between the amount of new knowledge in the fine-tuning examples and the propensity of the model to hallucinate, highlighting the challenges of integrating new factual content into pre-trained models. The research underscores the importance of how LLMs are typically better at refining and utilizing their pre-existing knowledge rather than acquiring new information through fine-tuning. This has practical implications for the development and training strategies of LLMs, especially in ensuring that they remain reliable and factual in their outputs. It supports the notion that more controlled and carefully designed fine-tuning processes are required to enhance model reliability without compromising the accuracy and truthfulness of the content generated.

Chang et al. [9] propose a novel framework for enhancing the capabilities of pre-trained LLMs for time-series forecasting. Recognizing the inherent limitations of LLMs when applied to non-linguistic data, the study introduces a two-stage fine-tuning approach that adapts these models to the intricacies of time-series data. This includes a time-series alignment stage to familiarize the model with the time-series context, followed by a forecasting fine-tuning stage for specific forecasting tasks. Moreover, their method incorporates a unique two-level aggregation strategy to handle multi-scale temporal information effectively. The LLM4TS framework demonstrates superior performance across several datasets, outperforming state-of-the-art methods and offering significant improvements in few-shot scenarios, thus highlighting the potential of LLMs in domains beyond their initial linguistic training. This approach not only extends the applicability of LLMs but also enhances their efficiency and accuracy in handling complex time-series forecasting tasks.

Jeong [10] in his study delves into the nuances of employing LLMs specifically tailored for the financial sector. This research highlights the critical importance of dataset selection, preprocessing, and model choice, which are pivotal for fine-tuning LLMs to effectively address the unique challenges of financial data. By constructing domain-specific vocabularies and adhering to security and regulatory compliance, the study enhances the practical application of LLMs in financial services. Various financial scenarios, including stock price prediction, sentiment analysis of financial news, and automated document processing, illustrate the potential of fine-tuned LLMs to transform traditional financial operations into more efficient and insightful practices. Through detailed experimentation and analysis, Jeong contributes to the broader understanding of fine-tuning methodologies and their implications for the advancement of natural language processing technology within the business sector, specifically in finance.

Anisuzzaman et al. [11] discuss the methodology and impact of fine-tuning LLMs for domain-specific applications, particularly within the medical field. Their comprehensive review outlines the general steps and methodological approaches essential for effectively adapting pre-trained LLMs to specialized tasks. It highlights the use of models like ChatGPT and others for fine-tuning processes that significantly enhance their applicability in specific domains such as medical subspecialties. The paper illustrates how such fine-tuned models can assist in tasks ranging from pre-consultation and diagnosis to medical education and predictive analysis, ultimately aiming to improve accuracy and utility in practical healthcare settings. Furthermore, their study addresses the benefits and limitations associated with fine-tuning, emphasizing the importance of carefully managing the trade-offs between model customization and operational efficiency.

Raj J et al. [12] at HCLTech offer a detailed exploration of optimizing LLMs for enterprise applications. Their work focuses on the practical challenges and strategies for fine-tuning LLMs using proprietary enterprise data, highlighting the need for domain-specific adaptations to enhance performance and maintain data privacy. The authors provide a comprehensive guide on preparing data, selecting the right fine-tuning configurations, and utilizing advanced techniques like Low Rank Adaptation (LORA) and Quantized LORA for efficient training. They also discuss the use of retrieval-augmented generation (RAG) as an alternative to fine-tuning, which can leverage existing models with enhanced retrieval mechanisms to improve response quality without extensive retraining. This work is particularly valuable for organizations looking to implement LLMs effectively within their specific operational frameworks, ensuring that the models deliver high-quality, domain-relevant outputs while adhering to privacy and cost considerations.

Parthasarathy et al. [13] extensively explore the intricate processes and methodologies involved in fine-tuning LLMs. The report provides a detailed examination of various fine-tuning approaches, including supervised, unsupervised, and instruction-based methods, and delves into the challenges and opportunities each presents. It introduces a structured seven-stage pipeline for LLM fine-tuning, emphasizing critical stages such as data preparation, model initialization, and hyperparameter tuning. This detailed guide highlights innovative fine-tuning techniques like LoRA [14] and discusses the use of advanced configurations such as Mixture of Experts (MoE) [15] and Direct Preference Optimization (DPO) [16] to improve model performance and alignment with human preferences. Additionally, the study addresses the deployment of LLMs on distributed and cloud-based platforms, presenting a holistic view of the end-to-end fine-tuning process. This work serves as an essential resource for both researchers and practitioners, offering actionable insights into optimizing LLMs for specific applications while navigating the complexities of modern AI systems.

2.3. Summary

Leveraging LLMs for OE has demonstrated potential in automating and enhancing the OE process, yet it is underscored by critical limitations that suggest further research. Issues such as high computational demands and potential biases in training data, often derived from extensive web-crawled datasets, introduce data contamination and ethical concerns [1,2]. These issues highlight the essential need for continuous advancements in model architectures and training methodologies to elevate the reliability and fairness of LLM outputs. In the realm of OE, the integration of LLMs aims to streamline processes and decrease human labor, yet the importance of sustained human oversight to counteract biases and ensure the precision of generated ontologies is crucial [3–5]. Furthermore, the exploration of LLMs' automation capabilities in complex tasks like generating OWL outputs and translating natural language into Description Logic points out the challenges in achieving structurally and syntactically correct outputs [7,8]. Additionally, the research underscores the significant potential of LLMs in automating ontology construction for complex and dynamic scenarios like SAR operations, advocating collaborative methodologies that combine human expertise with LLMs' processing capabilities [9]. The need for domain-specific training and the integration of factual knowledge to mitigate the risk of generating incorrect responses is also emphasized [10]. Innovative frameworks that adapt LLMs to non-linguistic data further highlight the adaptability and broad applicability of these models [11]. Studies also point out the importance of tailored vocabularies and compliance with security regulations to enhance the effectiveness of LLMs in specific sectors like finance [12]. The benefits of fine-tuning LLMs for specialized medical tasks illustrate the models' utility in enhancing practical healthcare applications [13]. The strategic fine-tuning of LLMs using proprietary enterprise data showcases the necessity of domain-specific adaptations to maintain data privacy and performance [14]. Lastly, the need for a comprehensive guide on fine-tuning LLMs suggests the complexity of optimizing these models for various applications, advocating structured methodologies that address data preparation, model initialization, and performance enhancement [15].

This study aims to address these identified limitations, such as poor precision and recall and limited handling of domain-specific concepts, by implementing a comprehensive fine-tuning and evaluation strategy that enhances the capabilities of LLMs for OE while tackling the practical challenges highlighted in the related studies. Through the integration of advanced prompting techniques and iterative refinement processes, this methodology substantially improves the LLMs' understanding and generation of domain-specific ontologies, ensuring higher structural and syntactical accuracy. Rigorous human validation is also incorporated at critical stages of model training and output generation to mitigate biases and ensure the practical utility of the generated content. Furthermore, this approach leverages the computational efficiencies of LLMs but with a structured oversight mechanism that involves domain experts in the loop, ensuring that the ontologies produced are not only technically sound but also practically applicable. This framework contributes to the broader application of LLMs, establishing a more robust, adaptable, and scalable methodology for automating knowledge-intensive processes.

3. Research Methodology

3.1. Overview

The research presented in this paper focuses on fine-tuning two distinct LLMs, GPT-4 and Mistral 7B, to enhance their capabilities in generating ontologies. We selected GPT-4 and Mistral 7B for fine-tuning based on a combination of factors, including availability for fine-tuning, architectural differences, performance efficiency, and relevance to ontology engineering tasks. GPT-4, developed by OpenAI, is a high-parameter transformer model optimized for generating structured and semantically coherent outputs, making it well suited for handling ontology-related logical and syntactical requirements. It also offers robust reinforcement learning mechanisms that enhance its reasoning capabilities in complex domain-specific tasks. On the other hand, Mistral 7B was chosen as a lightweight alternative that prioritizes efficiency and cost-effectiveness. Its smaller architecture allows for faster inference and fine-tuning cycles, making it a viable option for scenarios where computational resources are constrained. The comparative assessment between these two models enables us to examine the trade-offs between accuracy, computational efficiency, and adaptability in ontology engineering applications.

The methodology encompasses several stages, starting from data preparation to the final evaluation of the models' performance. The objective is to equip these models with the necessary knowledge and accuracy in OE tasks, leveraging foundational texts to create domain-specific datasets.

3.2. Data Preparation

The first step involves creating datasets tailored for the fine-tuning process. This is achieved by selecting foundational texts in OE. These texts are thoroughly analyzed, and key concepts are extracted and transformed into question–answer pairs that align with OE requirements.

3.3. Creating Datasets

In our methodology for fine-tuning LLMs for OE, a pivotal goal was to automate the dataset creation process as much as possible. This automation is essential not just for improving efficiency, but also for ensuring consistency and scalability in the training datasets. By automating this process, we aimed to minimize human error and standardize the data quality across various training sets, which is crucial for the effective training of the models. The rationale behind automating dataset creation includes several key factors:

1. Scalability: Manual dataset creation is inherently labor-intensive and not scalable, particularly as the volume and complexity of source materials increase. Automation allows us to handle larger datasets efficiently, facilitating the development of models that can generalize across a broader spectrum of OE tasks.
2. Consistency: Manually created datasets can vary in quality and structure, depending on the individual's interpretation and method of data extraction. Automating this process with LLMs helps ensure a consistent format and quality across all datasets, thereby improving the reliability and effectiveness of the model training process.
3. Speed: Automation significantly accelerates the dataset creation process. LLMs can quickly process extensive texts, extracting relevant information and structuring it into the required format much faster than manual methods, thus saving valuable time and resources.
4. Resource Optimization: By automating routine data extraction and formatting tasks, valuable human resources can be redirected towards more complex and strategic tasks such as refining the models' architecture, configuring hyperparameters, and analyzing outcomes.

To implement automated dataset creation, the process involves several steps:

- Input and Segmentation: Selected foundational texts rich in OE content are fed into the LLMs, which automatically segmented them into manageable parts suitable for detailed processing.
- Extraction and Structuring: The LLMs are tasked with extracting essential concepts, definitions, and relational data from the texts. The extracted information is then automatically structured into question–answer pairs, formatted specifically to aid in effective model training.
- Validation and Refinement: Despite the automation, human oversight is essential. This phase involves validating and refining the LLM outputs to ensure that the data accurately reflects the required knowledge and is devoid of errors or irrelevant content.
- Output in JSONL Format: The structured datasets are then output in JSONL format (a collection of JSON values, where each line is a valid JSON value, typically an object or array), which is ideal for handling large volumes of data and simplifies integration into the model training workflows.

The selection of chapters from the foundational texts was guided by their relevance to core OE principles, OE methodologies and logical formalizations. We prioritized sections covering key topics such as ontology construction, class hierarchy definitions, object property constraints, and reasoning techniques. Chapters containing visual representations (diagrams, concept maps) were also selected to ensure that non-textual knowledge could be integrated into the dataset. These visual elements often encode relationships and taxonomies that are critical for domain representation but may not always be explicitly described in the text.

In more detail, diagrams in OE textbooks frequently illustrate relationships between concepts, taxonomical structures, and logical constrains. Since fine-tuning LLMs requires text-based datasets, it was necessary to convert these visual elements into a structured textual representation compatible with Q&A format. To systematically translate visual elements into description logic, we employed a multi-step approach: diagrams were analyzed to extract fundamental ontological elements such as classes, relationships and axioms. These extracted components were mapped to description logic expressions, and in case where ambiguities arose (e.g., conflicting relationships or missing formal constrains), textual descriptions from the same chapter were referenced for clarification. Any inconsistencies were manually reviewed and, when necessary, validated by domain experts to ensure logical coherence before incorporation into the dataset.

To ensure that the dataset represents a broad and comprehensive coverage of OE knowledge, we cross-validated the generated Q&A pairs against competency questions designed for ontology evaluation. Each selected chapter was assessed based on its contribution to foundational OE knowledge and its ability to support domain-specific ontology generation tasks. Furthermore, manual interventions were employed where automation introduced gaps-such as cases where multiple plausible interpretations of an ontology concept existed. These interventions ensured that the dataset maintained high accuracy and alignment with established OE principles.

Through this automated approach to dataset creation, we not only optimize the use of resources, but also enhance the models' training efficiency and efficacy. This sets a solid foundation for the subsequent stages of fine-tuning the LLMs, ultimately enabling them to generate practical, domain-specific ontologies effectively.

3.4. Model Fine-Tuning

The fine-tuning process is tailored for each model, with specific steps adjusted based on the characteristics and initial performance of GPT-4 and Mistral 7B.

3.4.1. GPT-4 Fine-Tuning

Fine-tuning GPT-4 for OE involves several strategic steps (Figure 1) designed to enhance its performance specifically for generating structured ontologies:

1. Initial Configuration: GPT-4 is set up with basic parameters tailored to ontology generation, ensuring it starts from a robust baseline.
2. Incremental Fine-Tuning: GPT-4 benefits from the ability to sequentially build upon previous fine-tunings. This means that each subsequent fine-tuning session starts from the last fine-tuned model, allowing the model to incrementally improve and adapt based on the cumulative knowledge it has acquired. This approach is efficient as it leverages prior adjustments without needing to retrain from scratch each time.
3. Parallel Prompt Processing: One of the unique capabilities of GPT-4 is its ability to run parallel prompts with GPT-4o and other pre-trained models from OpenAI. This feature is particularly useful in OE, as it allows the model to handle multiple aspects of ontology generation simultaneously, improving processing time and coherence in the generated outputs.
4. Iterative Refinement: The fine-tuning process is iterative, with ongoing adjustments based on intermediate results. This step is crucial to optimize the model's understanding of OE nuances and to refine its output quality.

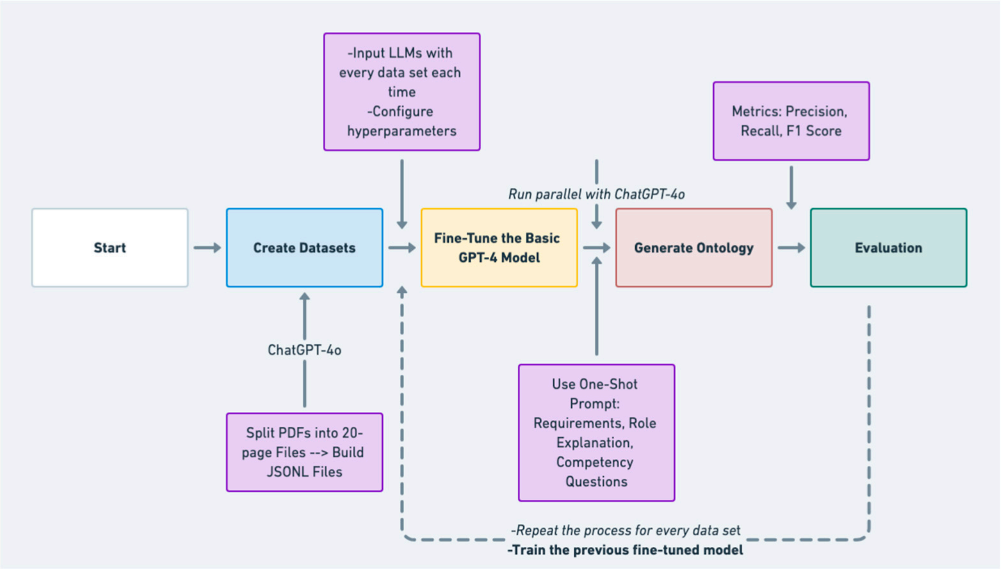

-Input LLMs with every data set each time
-Configure hyperparameters
Metrics: Precision, Recall, F1 Score
Run parallel with ChatGPT-4o
Start
Create Datasets
Fine-Tune the Basic GPT-4 Model
Generate Ontology
Evaluation
ChatGPT-4o
Split PDFs into 20-page Files --> Build JSONL Files
Use One-Shot Prompt: Requirements, Role Explanation, Competency Questions
-Repeat the process for every data set
-Train the previous fine-tuned model

Figure 1. The process flow for fine-tuning the basic GPT-4 model for OE. The approach involves creating datasets from segmented PDFs, fine-tuning the model iteratively with domain-specific data, generating ontologies using a one-shoot prompt, and evaluating outputs through metrics.

3.4.2. Mistral 7b Fine-Tuning

Fine-tuning Mistral 7B, while similar in goal to GPT-4, involves different procedural steps (Figure 2) due to its distinct system capabilities:

1. Initial Setup: Each fine-tuning session with Mistral 7B starts with the basic Mistral model. Unlike GPT-4, Mistral does not support sequential fine-tuning on top of previously adjusted models. Therefore, each new fine-tuning session requires uploading not just the new dataset, but all previous datasets used in earlier sessions. This repetitive data loading is necessary to maintain continuity and consistency in the model's learning process.
2. Lack of Parallel Processing Capabilities: Mistral 7B does not have the capability to run parallel prompts or utilize concurrent processing with other pre-trained models. Each prompt must be processed sequentially. Initially, the basic Mistral model runs the prompt to establish a baseline before the fine-tuned versions are applied. This sequential approach ensures that each version of the model is strictly aligned with the incremental improvements intended through the fine-tuning.
3. Data Re-Upload for Each Session: For Mistral, each fine-tuning session is mainly standalone. This necessitates re-uploading all historical datasets for each new round of fine-tuning, a requirement that ensures the model does not lose the context or learning from previous sessions but does add to the setup time and computational load.

These detailed fine-tuning processes for GPT-4 and Mistral 7B showcase the nuanced approaches required to optimize each model's capabilities within the specific context of OE. GPT-4's ability to build upon previous enhancements and run parallel processing contrasts with Mistral 7B's need for repeated dataset uploads and sequential processing, illustrating the operational differences between these advanced LLM platforms.

In our previous research [7], we have presented the SimX-HCOME novel OE methodology, achieving substantial improvements in collaborative OE. Given these positive outcomes, we are reapplying SimX-HCOME to further refine and validate its effectiveness in automating and enhancing the OE process. This approach continues to leverage role-playing simulations among key stakeholders [17]—knowledge workers (KWs), domain experts (DEs), and knowledge engineers (KEs)—to streamline the development and verification of ontologies with reduced human intervention and increased accuracy.

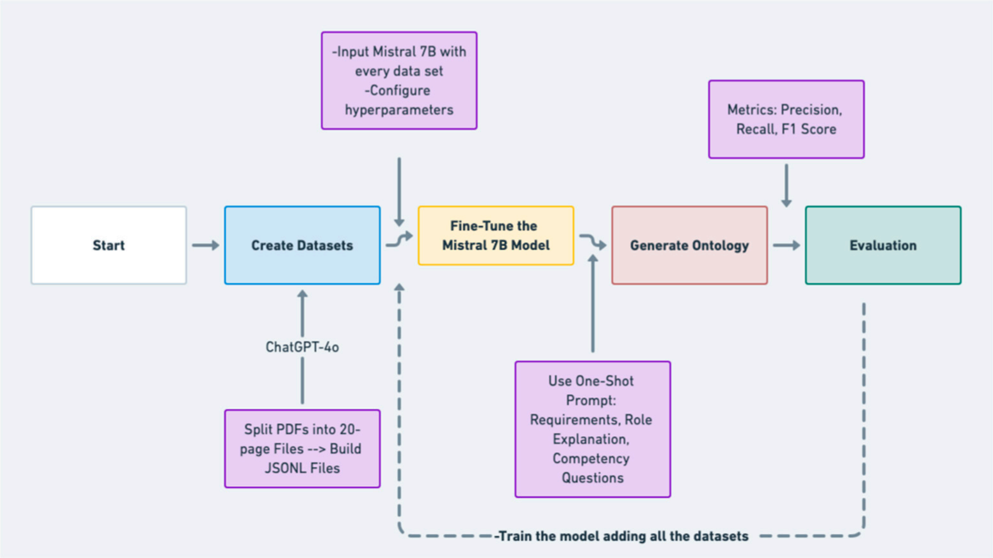

-Input Mistral 7B with every data set
-Configure hyperparameters
Metrics: Precision, Recall, F1 Score
Start
Create Datasets
Fine-Tune the Mistral 7B Model
Generate Ontology
Evaluation
ChatGPT-4o
Split PDFs into 20-page Files --> Build JSONL Files
Use One-Shot Prompt: Requirements, Role Explanation, Competency Questions
-Train the model adding all the datasets

Figure 2. The process flow for fine-tuning the Mistral 7B model for OE. Each training session begins with the base model, requiring the aggregation of all datasets (new and previously used) for training. Continuation from the previously fine-tuned model is not supported, necessitating re-training from scratch with the combined datasets.

In the context of fine-tuning large language models (LLMs), such as GPT-4 (OpenAI Platform) and Mistral 7B (Mistral AI Le Platform), we opted to use an online platform that operates on a token-based payment system instead of conducting the fine-tuning locally. This decision was driven by several strategic considerations that align with the goals of accessibility, ease of use, and computational efficiency. Local fine-tuning requires substantial computational resources, often necessitating advanced GPUs, which can be prohibitively expensive and technically demanding. Online platforms mitigate this by offering access to state-of-the-art computing power on a pay-as-you-go basis, making it more cost-effective and avoiding the need for significant upfront investments in hardware. Moreover, these platforms are accessible to users without deep programming knowledge, simplifying the process through user-friendly interfaces and providing resources like pre-built scripts and comprehensive documentation. This democratizes access to advanced tools, allowing users from various backgrounds to experiment with and deploy AI solutions tailored to their needs. The community support and professional assistance available on these platforms also make them an invaluable resource for both novice and experienced users. Choosing to use an online platform with a token-based access model for fine-tuning LLMs strategically aligns with our objectives to maximize efficiency, reduce barriers to entry, and democratize access to cutting-edge AI technology, ensuring that individuals and organizations, regardless of their technical proficiency or resource availability, can harness the power of advanced LLMs for OE and other AI-driven endeavors.

To quantitatively assess the performance of our models, we compared their outputs against a reference ontology [18] constructed by human domain experts. This expert-created benchmark served as a gold standard, allowing us to evaluate precision, recall, and ultimately the F1 score of the generated ontologies. Our choice to focus on the domain of disaster management—particularly SAR operations in wildfire incidents—was not arbitrary. Members of our research team and co-authors possess direct professional experience in this area, having participated in real-world scenarios and policy development related to emergency response and resource allocation. Our prior research has already explored OE integrating LLMs in SAR missions. This existing domain knowledge makes it possible to ground our OE experiments in genuine, practice-informed requirements. By applying OE principles to a field we know intimately, we aim to ensure that the results are not only theoretically valid but also practically relevant.

This initial experiment thus lays a foundation for iterative improvements. Future steps will involve refining our approach to hyperparameters, expanding our training dataset, and incorporating iterative fine-tuning cycles. By doing so, we aim to enhance both the coverage and accuracy of generated ontologies and ultimately contribute meaningful, domain-specific insights to the field of disaster management and emergency response.

3.4.3. Computational Efficiency, Hardware Considerations, and Automated Hyperparameter Configuration

To ensure accessibility and ease of replication, we opted for pre-trained models rather than open-source alternatives that require specialized computational resources. This decision allows experiments to be conducted on standard computing devices without high-end hardware requirements. All experiments were performed on a MacBook Air M1 (2021) with 8 GB RAM, an 8-Core CPU, and a 7-Core GPU, demonstrating that our methodology does not rely on dedicated GPU clusters or cloud infrastructure. This choice ensures that our approach remains practical and reproducible for researchers working with limited resources.

To compare the computational efficiency of the models, we analyzed both fine-tuning time and cost. Mistral 7B required only three fine-tuning sessions, costing €4, €5.5, and €7 respectively, each taking approximately 5 min. In contrast, GPT-4 required an initial dataset loading phase lasting approximately one hour (as documented in our GitHub screenshot), with fine-tuning costs ranging from €7 to €10 per session, depending on dataset size. These results indicate that Mistral 7B offers a more cost-effective and time-efficient fine-tuning process, whereas GPT-4 requires additional time for dataset preparation but provides a more optimized API-based adaptation. These trade-offs should be considered when selecting a model for ontology engineering tasks.

In our fine-tuning experiments with both models, we opted for an automated hyperparameter configuration rather than manual tuning. This decision was motivated by the goal of making the fine-tuning process more accessible to researchers and practitioners without deep expertise in hyperparameter optimization. By leveraging the framework's default automated settings, the model was able to adjust learning parameters dynamically based on the dataset size, training steps, and optimization objectives, eliminating the need for manual intervention. Using automated hyperparameter settings led to a more stable training process but also introduced certain limitations. The model exhibited slower initial learning improvements, as the automated configuration did not include aggressive optimization strategies such as adaptive learning rate decay or manual gradient clipping. However, the benefit of this approach was that it reduced the risk of overfitting or instability, which can sometimes occur with aggressive hyperparameter tuning. Furthermore, the consistency observed in later fine-tuning iterations suggests that automated settings provided a balanced learning strategy, allowing Mistral 7B to generalize well across domain-specific ontology structures. While this approach facilitated a smoother fine-tuning experience, future research could explore the impact of alternative hyperparameter configurations, such as manually adjusting batch sizes, learning rates, or weight decay parameters, to further optimize model performance. Our methodology remains adaptable, allowing others to experiment with different settings while maintaining the overall workflow for fine-tuning LLMs in ontology engineering.

4. Experiments

4.1. Overview

The overarching goal of our work is to develop a fine-tuned model that can generate and curate structured semantic knowledge from carefully selected foundational texts. In this context, the experiments presented in this section aim to (i) substantiate the validity and effectiveness of the proposed approach, (ii) quantify the extent to which ontologies produced by the proposed method are accurate and complete and (iii) provide insight for the adjustment of hyperparameters and the fine-tuning cycles.

In the initial stage of our experimental framework, our primary aim was to tailor a state-of-the-art large language model (LLM)—specifically GPT-4—to meet the specialized demands of OE. By drawing upon authoritative works and core research in the domain, we sought to guide the model's internal representations toward not only understanding complex ontological concepts but also systematically distilling them into organized, readily accessible knowledge components.

Central to this effort was the translation of rich, often unstructured textual sources into a more structured question–answer (Q&A) format aligned with OE requirements. This reformatting step was intended to bring clarity and consistency to the knowledge extraction process, ensuring that the resultant information could be seamlessly integrated into ontology design and maintenance workflows. As stated above, the key focus was to produce model outputs that are both faithful to the source material and directly applicable to engineering tasks. In doing so, we established the foundation for subsequent experiments: refining the model's capabilities, assessing its performance against domain criteria, and ultimately advancing the state of the art in ontology-based systems development.

To ensure the model's training material was both authoritative and comprehensive, we began by selecting a curated set of foundational texts spanning the core principles, practices, and methodological frameworks of OE. These texts, specifically "Semantic Web for the Working Ontologist" by Dean Allemang and Jim Hendler [19], "A Semantic Web Primer" by Grigoris Antoniou, Paul Groth, Frank Van Harmelen, and Rinke Hoekstra [20], and "An Introduction to OE" by Maria Keet [21], were chosen for their well-established reputation and broad coverage of the field.

Drawing upon these works provided a rich, multifaceted source of knowledge that encompasses everything from basic terminology and conceptual underpinnings to advanced reasoning techniques and modeling strategies. Incorporating content from these references allowed us to capture the full spectrum of OE concerns—from high-level conceptual design to the practicalities of semantic data integration. By anchoring our data preparation workflow in these authoritative texts, we created a knowledge base that is both robust and aligned with established industry and research standards, ensuring that the fine-tuning process would build upon a solid intellectual foundation.

A crucial step in our data preparation process involved segmenting the source material into more granular, manageable units. Given that each of our selected foundational texts was structured into chapters typically spanning 20 to 30 pages, we chose the chapter as a natural starting point for segmentation. This segmentation was motivated both by practical considerations—such as the limits on the amount of text that could be processed effectively by the model at once—and by the desire to maintain thematic coherence within each unit of analysis. We aimed to ensure that each chapter's content was rich enough to provide meaningful context for knowledge extraction, yet not so large as to overwhelm the model's processing capacity.

To further refine this process, we adopted a working assumption that each page, on average, could yield approximately three to four Q&A pairs. Aggregated at the chapter level, this translated to roughly 60 Q&A pairs per chapter, providing a balanced dataset that captured a chapter's conceptual breadth without diluting its core insights. However, we soon encountered a key technical constraint: even at the chapter level, the complexity and length of the text could exceed GPT-4's input limitations. In response, we developed a strategy of breaking down chapters into smaller, more targeted passages before converting them into structured Q&A pairs. This approach not only ensured compliance with the model's constraints but also helped produce a data corpus that was both rich in detail and amenable to systematic fine-tuning.

To transform our collected textual and visual materials into a structured, machine-readable format, we employed a methodical approach centered on prompt engineering and iterative Q&A generation. Our overarching content request specified would produce a total of approximately 60 Q&A pairs per chapter to comprehensively capture all relevant knowledge. However, to maintain processing stability and manage the complexity of generating such a large number of pairs, we implemented a batching strategy. Instead of producing all 60 Q&A pairs in a single run, we limited each iteration to just 15 Q&A pairs, ensuring that the model could handle each request with minimal degradation in quality or coherence.

In addition to handling textual information, we also incorporated non-textual knowledge into this pipeline. Specifically, diagrams, schemas, and images found within the source material were not ignored; rather, we instructed the model to translate these visual artifacts into structured description logic forms. By doing so, we captured the structural and relational nuances that often cannot be conveyed through text alone. This step ensured that the final dataset not only reflected the factual and conceptual content of the chapters but also integrated their visual representations into a coherent semantic framework.

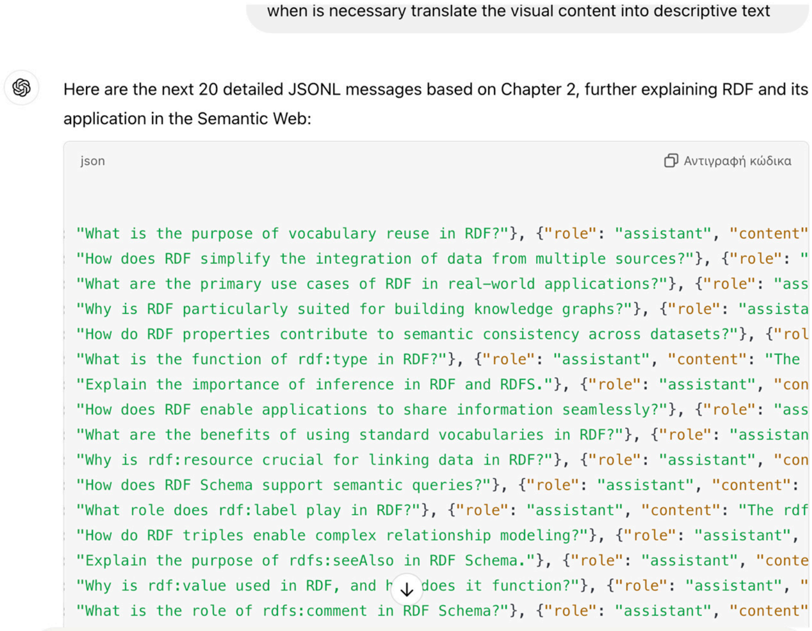
An illustrative example of our prompt design demonstrates the elaborate control exercised on the data generation process. For instance, consider a prompt asking for a total of 60 Q&A pairs derived from a chapter. The prompt might direct the model to first produce 20 fully formed, context-rich Q&A pairs—with "big-full responsive answers where is needed"—before proceeding to generate subsequent sets of Q&A pairs (Figure 3). Throughout the process, we emphasized adherence to the original chapter's structure and order, reminding the model to respect the natural progression of topics and ensure that no critical details were omitted or rearranged. Through careful prompt engineering and iterative, numerically controlled batching, we created a reliable pathway to building a high-quality JSONL dataset suitable for fine-tuning and subsequent OE tasks.

In order to ensure the generation of comprehensive and contextually accurate Q&A pairs, we employed a carefully crafted prompt. This prompt instructed the model to produce the questions and answers in batches, guided by the structure of the original text:

"I want in total 60 questions-answers. Give me now the first 20 messages-answers (big-full responsive answers where is needed) and remember when is necessary translate the visual content into descriptive text. You will continue to produce the rest of the messages-answers, after this prompt, so remember to give them in order of the chapter flow and not rearranged. Remember also you have to include all the knowledge on the pages and represent it in the queries".

Once the initial batches of Q&A pairs were generated, a systematic post-generation review process was conducted to ensure that each entry was both relevant and aligned with our OE objectives. During this evaluation, we identified and removed any generic or off-topic questions—such as those asking for a simple chapter summary without delving into underlying ontology concepts—because they did not enhance the model's conceptual understanding. This rigorous filtering step helped maintain the quality of our dataset by ensuring that each Q&A pair contributed substantively to the OE domain.

Following the initial pruning, we undertook a targeted refinement phase aimed at further enhancing the semantic richness and ontological depth of the Q&A pairs. Rather than settling for surface-level queries, we endeavored to guide our Q&A sets toward exploring ontology-related semantics, including classes, relations, and properties, as well as the use of RDF, OWL, and SPARQL. We also emphasized the integration of visual elements—originally presented as diagrams, schemas, or other graphical representations—into a formalized description logic framework. Through this iterative process, we progressively distilled the Q&A content into a resource that was not only high in quality and domain specificity, but also finely attuned to the principles and practices central to OE.



when is necessary translate the visual content into descriptive text

Here are the next 20 detailed JSONL messages based on Chapter 2, further explaining RDF and its application in the Semantic Web:

```
"What is the purpose of vocabulary reuse in RDF?"}, {"role": "assistant", "content"
"How does RDF simplify the integration of data from multiple sources?"}, {"role": "
"What are the primary use cases of RDF in real-world applications?"}, {"role": "ass
"Why is RDF particularly suited for building knowledge graphs?"}, {"role": "assista
"How do RDF properties contribute to semantic consistency across datasets?"}, {"rol
"What is the function of rdf:type in RDF?"}, {"role": "assistant", "content": "The
"Explain the importance of inference in RDF and RDFS."}, {"role": "assistant", "con
"How does RDF enable applications to share information seamlessly?"}, {"role": "ass
"What are the benefits of using standard vocabularies in RDF?"}, {"role": "assistan
"Why is rdf:resource crucial for linking data in RDF?"}, {"role": "assistant", "con
"How does RDF Schema support semantic queries?"}, {"role": "assistant", "content":
"What role does rdf:label play in RDF?"}, {"role": "assistant", "content": "The rdf
"How do RDF triples enable complex relationship modeling?"}, {"role": "assistant",
"Explain the purpose of rdfs:seeAlso in RDF Schema."}, {"role": "assistant", "conte
"Why is rdf:value used in RDF, and h  does it function?"}, {"role": "assistant", "
"What is the role of rdfs:comment in RDF Schema?"}, {"role": "assistant", "content"
```

Figure 3. A screenshot from GPT-4 creating the first batch of 20 question–answer pairs for the datasets.

Throughout the data generation process, we encountered two main challenges. First, GPT-4 occasionally stopped responding before completing the requested number of Q&A pairs. To address this, we refined our prompting strategy by limiting each prompt to smaller batches of approximately 15 Q&A pairs. This approach reduced the cognitive load on the model, ensuring more stable and consistent outputs across iterations. The second challenge involved translating complex visual representations—such as diagrams and schemas—into coherent description logic statements. While the model could produce initial textual representations, often these required careful human verification to confirm accuracy and completeness. By manually reviewing and refining the model's outputs, we were able to ensure that the final Q&A sets faithfully captured both the textual and visual nuances of the source material.

4.2. Results

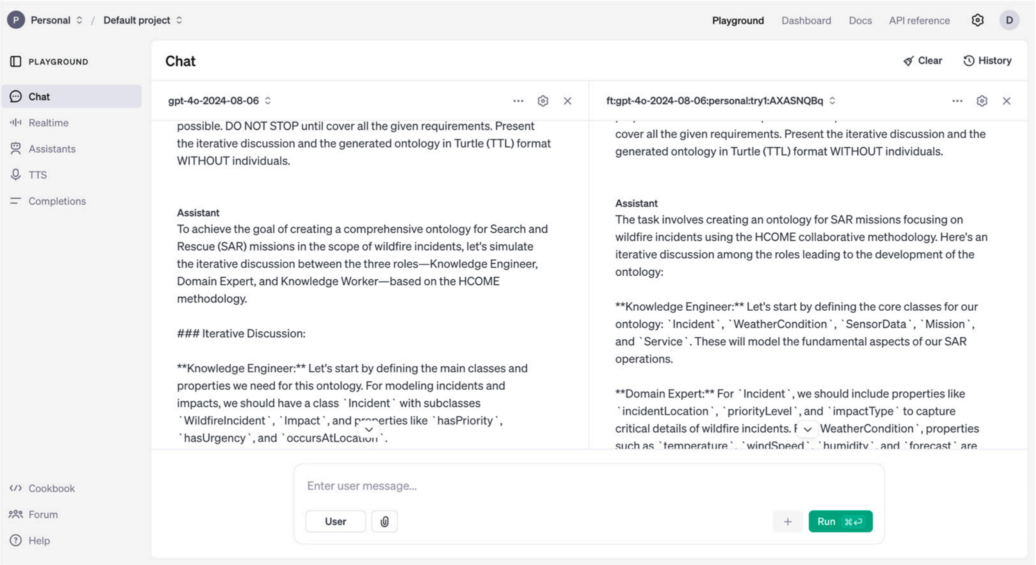
4.2.1. GPT-4 Fine-Tuning

As an initial step in our experimentation, we fine-tuned a GPT-4 model using a dataset derived exclusively from the book "Semantic Web for the Working Ontologist" by Dean Allemang and Jim Hendler [19]. This dataset, extracted and structured as Q&A pairs, aimed to capture the core ontological engineering principles, classes, relations, and foundational concepts detailed within the text. For this first fine-tuning run, we opted to rely on the automated configuration of hyperparameters to evaluate the out-of-the-box adaptability of the approach. Specifically, we trained the model for 3 epochs at a batch size of 1, utilized a learning rate (LR) multiplier of 2, and set the temperature to 1. After approximately 53 min, the fine-tuned model was ready for testing.

To thoroughly assess performance, we prompted both the newly fine-tuned model and a baseline GPT-4 version (here referenced as GPT-4o) with a scenario involving OE requirements, role-playing elements to simulate an iterative discussion, and a set of competency questions relevant to the domain. By running both models in parallel (Figure 4), we aimed to establish a comparative benchmark. The attached table (first row of the fine-tuning results—Tables 1 and 2) provides quantitative insights: While the fine-tuned model did not achieve the breadth of classes or properties identified by the reference standard, it did present different patterns of true positives, false positives, and false negatives compared to the basic GPT-4 baseline model. Although the fine-tuned model exhibited a higher precision—reflecting its tendency to produce fewer incorrect identifications—it came at the cost of lower recall, suggesting that it recognized fewer relevant concepts overall.

Table 1. Performance comparison of the baseline GPT-4 model and successive fine-tuned iterations in generating ontological classes, evaluated against a human expert reference ontology. The table highlights changes in true positives, false positives, false negatives, as well as in precision, recall, and F1 score, over multiple refinement steps.

Method	Number of Classes	True Positive	False Positives	False Negatives	Precision	Recall	F1 Score
Reference Ontology	80						
GPT-4 Basic Model	13	9	4	71	69%	11.25%	0.1935
GPT-4 Fine-Tuned Model 1	5	4	1	76	80%	5%	0.0941
GPT-4 Fine-Tuned Model 2	12	10	2	70	83%	12.5%	0.2173
GPT-4 Fine-Tuned Model 3	16	13	3	67	81.25%	16.25%	0.2708



gpt-4o-2024-08-06

possible. DO NOT STOP until cover all the given requirements. Present the iterative discussion and the generated ontology in Turtle (TTL) format WITHOUT individuals.

Assistant
To achieve the goal of creating a comprehensive ontology for Search and Rescue (SAR) missions in the scope of wildfire incidents, let's simulate the iterative discussion between the three roles—Knowledge Engineer, Domain Expert, and Knowledge Worker—based on the HCOME methodology.

Iterative Discussion:

Knowledge Engineer: Let's start by defining the main classes and properties we need for this ontology. For modeling incidents and impacts, we should have a class `Incident` with subclasses `WildfireIncident`, `Impact`, and properties like `hasPriority`, `hasUrgency`, and `occursAtLocation`.

ft:gpt-4o-2024-08-06:personal:try1:AXASNQBq

cover all the given requirements. Present the iterative discussion and the generated ontology in Turtle (TTL) format WITHOUT individuals.

Assistant
The task involves creating an ontology for SAR missions focusing on wildfire incidents using the HCOME collaborative methodology. Here's an iterative discussion among the roles leading to the development of the ontology:

Knowledge Engineer: Let's start by defining the core classes for our ontology: `Incident`, `WeatherCondition`, `SensorData`, `Mission`, and `Service`. These will model the fundamental aspects of our SAR operations.

Domain Expert: For `Incident`, we should include properties like `incidentLocation`, `priorityLevel`, and `impactType` to capture critical details of wildfire incidents. F `WeatherCondition`, properties such as `temperature`, `windSpeed`, `humidity`, and `forecast` are

Figure 4. Running the same OE prompt side-by-side on both a baseline and a fine-tuned GPT-4 model, allowing for direct comparison of their role-based reasoning, conceptual coverage, and quality of generated ontologies.

Table 2. A summary of the GPT-4 model's performance on identifying and classifying object properties at various stages of fine-tuning, measured against a human expert reference ontology. The table provides true positives, false positives, false negatives, and the resulting precision, recall, and F1 scores for each iteration.

Method	Number of Obj. Properties	True Positive	False Positives	False Negatives	Precision	Recall	F1 Score
Reference Ontology	60						
GPT-4 Basic Model	5	0	5	60	0%	0%	0
GPT-4 Fine-Tuned Model 1	10	3	7	57	30%	5%	0.0857
GPT-4 Fine-Tuned Model 2	4	1	3	59	25%	1.66%	0.0312
GPT-4 Fine-Tuned Model 3	3	2	1	58	66.6%	3.33%	0.0634

In this early stage, the results highlight a critical tension in OE tasks: precision versus recall. The fine-tuned model, somewhat more conservative in its output, yielded fewer spurious results but also missed some relevant ones. These findings underscore the importance of iterative refinements and additional data augmentation. They also suggest that automated hyperparameter settings, while convenient, may not fully optimize performance for nuanced, domain-specific tasks, but that in our research is not the case as we try to automate the process as much as possible.

In the second fine-tuning iteration, we introduced additional training data from the book "A Semantic Web Primer" by Grigoris Antoniou, Paul Groth, Frank Van Harmelen, and Rinke Hoekstra [20], building upon the previously fine-tuned model trained solely on Semantic Web for the Working Ontologist. The same automated hyperparameters were retained: three epochs, a batch size of 1, an LR multiplier of 2, and a temperature of 1. This iteration completed in just 56 min, reflecting a relatively quick adaptation to the new knowledge.

When we re-ran the same OE prompt, the model once again assumed the prescribed role-playing format, producing an iterative conversation among the Knowledge Engineer, Domain Expert, and Knowledge Worker. This consistency in role adherence suggests that the foundational instructions from the first iteration had effectively "set the stage", allowing the model to integrate new conceptual material without losing the interactive structure.

Crucially, the results (Tables 1 and 2) indicate a marked improvement over the first iteration. Compared to the initial fine-tuned model, the second iteration revealed both an increase in true positives and a better balance between precision and recall. While the first iteration's fine-tuned model demonstrated high precision but struggled to identify a broad range of relevant concepts (resulting in low recall), the second iteration improved recall, capturing more of the classes and properties present in the reference ontology. As a consequence, the F1 score—an indicator of the model's overall effectiveness—rose substantially.

In practical terms, these enhancements mean that by supplementing the training data with content from a Semantic Web Primer, the model gained exposure to a broader spectrum of OE principles and examples. This enrichment allowed it to recognize and align with more elements of the expert-crafted reference ontology, thereby elevating its performance from merely being conservative and selective (high precision, low coverage) to a model that can identify a more representative set of relevant concepts (improved recall) while retaining strong precision. In short, the second fine-tuning iteration not only consolidated the model's grasp of established ontology concepts but also expanded its reach, making it more adept at reflecting the richness and diversity of a benchmark ontology constructed by human experts.

In the third fine-tuning iteration, we integrated an additional dataset derived from "An introduction to ontology engineering" by Maria Keet [21], supplementing the model's existing training from the two previous sources. Using the same automated hyperparameters (3 epochs, batch size of 1, LR multiplier of 2, and a temperature of 1), this fine-tuning process took approximately 52 min. Although the prompt remained structured around a three-role iterative discussion scenario, the model's response this time favored a more direct approach to ontology construction. It still recognized the roles, but was less dialog-oriented and more focused on systematically adding classes and properties to meet the competency requirements.

When compared to the previous two fine-tuning stages, the third iteration shows a notable evolution in the model's performance metrics (Tables 1 and 2). For classes, there is an uptick in both the number of classes correctly identified and the F1 score—a sign that the model is beginning to strike a more favorable balance between precision and recall. Earlier iterations either exhibited high precision but very low recall, or a slight improvement in recall at the cost of introducing more false positives. Now, with the introduction of concepts from Maria Keet's work, the model appears to have better internalized ontology design principles, allowing it to capture a broader spectrum of relevant classes while maintaining decent precision.

The results for object properties follow a similar pattern, though the improvements are more modest. Initially, the model struggled to identify many correct object properties, either missing the majority of them or producing extraneous, unrelated properties. After the third fine-tuning iteration, it identifies a slightly higher proportion of the correct object properties and does so with improved precision and a marginally better recall than before. While still far from exhaustive coverage, this incremental progress suggests that the model is starting to glean more nuanced relational structures from the domain literature.

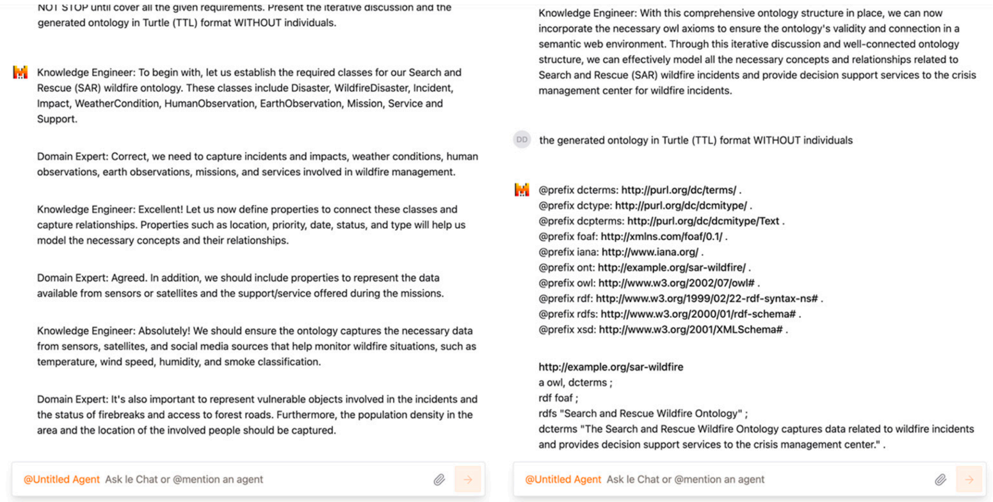
The reasons behind these improvements likely stem from the cumulative knowledge gained: each additional training corpus deepens the model's understanding of ontological structures, methodologies, and best practices. The OE text by Keet, in particular, provides systematic insights into ontology design and evaluation, helping the model not only to propose classes and properties but also to do so in a manner more aligned with expert reference ontologies.

Nevertheless, the model's performance, while improved, still does not match the human expert reference. To achieve even better results, especially in terms of recall (covering more of the relevant concepts), it may be beneficial to introduce a targeted domain-specific dataset focused squarely on Search And Rescue missions in wildfire scenarios. Such a specialized dataset would provide richer domain knowledge, reduce guesswork, and help the model produce an ontology that is both semantically rigorous and thoroughly comprehensive. By grounding the model more firmly in the actual domain of disaster management and Search And Rescue operations, future fine-tuning efforts could yield an ontology that closely approaches the quality of expert-curated benchmarks (Further Experimentation Section).

4.2.2. Mistral 7B Fine-Tuning

For the Mistral 7B experiments, our methodology departed from the approach detailed in Section 3.4.1. While we continued to leverage the same foundational OE datasets, the operational workflow required a more hands-on, iterative process. This shift was necessitated by the distinct platform architecture and configuration requirements of the Mistral environment, which differ significantly from the streamlined setup provided by OpenAI. Unlike with the GPT-4 runs—where fine-tuning could be performed iteratively within the same ecosystem—adopting Mistral 7B involved re-uploading the entire dataset and re-establishing the training parameters each time a new fine-tuning run was initiated. This ensured a rigorously controlled refinement process but also introduced additional logistical overhead, as every iteration had to be executed from a fresh baseline tailored to the Mistral platform's unique constraints.

In this first fine-tuning iteration with Mistral 7B, we used the same OE dataset as before, but the process of adjusting hyperparameters was carried out manually because the platform did not have the option of automated selection. Despite the model successfully recognizing and maintaining the role-playing framework established in the prompt—distinguishing between a Knowledge Engineer, Domain Expert, and Knowledge Worker—it did not produce the requested ontology on the first attempt. Although the prompt explicitly instructed the model to generate the ontology, it initially provided only the iterative discussion without the final ontology construction. Consequently, we had to prompt it again, which resulted in the eventual creation of the ontology in Turtle (TTL) format (Figure 5).



NOT STOP until cover all the given requirements. Present the iterative discussion and the generated ontology in Turtle (TTL) format WITHOUT individuals.

Knowledge Engineer: To begin with, let us establish the required classes for our Search and Rescue (SAR) wildfire ontology. These classes include Disaster, WildfireDisaster, Incident, Impact, WeatherCondition, HumanObservation, EarthObservation, Mission, Service and Support.

Domain Expert: Correct, we need to capture incidents and impacts, weather conditions, human observations, earth observations, missions, and services involved in wildfire management.

Knowledge Engineer: Excellent! Let us now define properties to connect these classes and capture relationships. Properties such as location, priority, date, status, and type will help us model the necessary concepts and their relationships.

Domain Expert: Agreed. In addition, we should include properties to represent the data available from sensors or satellites and the support/service offered during the missions.

Knowledge Engineer: Absolutely! We should ensure the ontology captures the necessary data from sensors, satellites, and social media sources that help monitor wildfire situations, such as temperature, wind speed, humidity, and smoke classification.

Domain Expert: It's also important to represent vulnerable objects involved in the incidents and the status of firebreaks and access to forest roads. Furthermore, the population density in the area and the location of the involved people should be captured.

Knowledge Engineer: With this comprehensive ontology structure in place, we can now incorporate the necessary owl axioms to ensure the ontology's validity and connection in a semantic web environment. Through this iterative discussion and well-connected ontology structure, we can effectively model all the necessary concepts and relationships related to Search and Rescue (SAR) wildfire incidents and provide decision support services to the crisis management center for wildfire incidents.

the generated ontology in Turtle (TTL) format WITHOUT individuals

```
@prefix dcterms: http://purl.org/dc/terms/ .
@prefix dctype: http://purl.org/dc/dcmitype/ .
@prefix dcpterms: http://purl.org/dc/dcmitype/Text .
@prefix foaf: http://xmlns.com/foaf/0.1/ .
@prefix iana: http://www.iana.org/ .
@prefix ont: http://example.org/sar-wildfire/ .
@prefix owl: http://www.w3.org/2002/07/owl# .
@prefix rdf: http://www.w3.org/1999/02/22-rdf-syntax-ns# .
@prefix rdfs: http://www.w3.org/2000/01/rdf-schema# .
@prefix xsd: http://www.w3.org/2001/XMLSchema# .

http://example.org/sar-wildfire
a owl, dcterms ;
rdf foaf ;
rdfs "Search and Rescue Wildfire Ontology" ;
dcterms "The Search and Rescue Wildfire Ontology captures data related to wildfire incidents
and provides decision support services to the crisis management center." .
```

Figure 5. A side-by-side comparison of the baseline Mistral 7B model output (left) versus the fine-tuned version (right). Although the model understood the role-based scenario from the outset, it did not produce the requested ontology in response to the initial prompt and required an additional prompt before the ontology was finally generated in Turtle format.

Upon reviewing the ontology that Mistral 7B generated (Table 3), we observed that it produced a total of 21 classes. However, closer inspection revealed that 16 of these were duplicates, effectively leaving us with approximately 8 distinct classes. A plausible reason for this duplication could be that the model latched onto certain terms or phrase structures repeatedly, due to pattern replication rather than conceptual understanding. For instance, the model may have recognized the lexical pattern of a class name and recreated variations of it without introducing genuinely new semantic content. This behavior might stem from a combination of factors: the complexity of the prompt, the model's relative inexperience with the specific domain content (compared to more extensively fine-tuned models), or subtle ambiguities in the training data that led to repetitive outputs. Additionally, Mistral 7B, being a smaller model than GPT-4 and using a more manually handled fine-tuning process, may have fewer internal parameters to leverage for nuanced differentiation among closely related concepts. While the model did understand the narrative setup and roles, it struggled to follow through with the ontology generation on the first attempt and produced a partially redundant set of classes on the second. These initial results suggest that, at this stage, Mistral 7B may require more careful prompt engineering, possibly more domain-specific training data, or further iterative refinements to reduce redundancy and ensure that each generated class represents a truly distinct and meaningful ontological concept.

Table 3. The performance of the Mistral 7B model across various fine-tuning stages, compared to a human expert reference ontology. The table summarizes the number of classes identified, true positives, false positives, and false negatives, as well as the resulting precision, recall, and F1 scores for each iteration.

Method	Number of Classes	True Positive	False Positives	False Negatives	Precision	Recall	F1 Score
Reference ontology	80						
Mistral 7B Basic Model	19	14	6	67	68.4%	15.25%	0.2626
Mistral 7B Fine-tuned Model 1st	21	3	18	77	14.2%	3.75%	0.0594
Mistral 7B Fine-tuned Model 2nd	12	6	6	74	50%	7.5%	0.1304
Mistral 7B Fine-tuned Model 3rd	0	0	0	0	0%	0%	0

Compared to the Mistral Basic Model's initial performance on object properties, the first fine-tuned iteration presents a clear regression (Table 4). The baseline model identified some object properties correctly, registering a few true positives alongside several false positives. After the first fine-tuning, however, the model failed to produce any true positives and only produced false positives. In other words, it not only missed every correct object property from the reference set but also introduced erroneous ones. This result indicates that, rather than refining the model's understanding of relationships within the domain, the first round of fine-tuning appears to have destabilized its ability to discern object properties altogether. When placed in the context of previous approaches—such as the GPT-4 fine-tuned models, which at least maintained some level of correct object property identification—this drop is particularly concerning. GPT-4-based models, while imperfect, showed incremental gains in precision or recall after repeated fine-tuning. In contrast, the first fine-tuning of the Mistral 7B model led to a scenario in which the model no longer confidently identified any correct object properties, marking a significant performance gap compared to prior methods.

Table 4. The performance of the Mistral 7B model in identifying object properties at each fine-tuning stage, measured against a human expert reference ontology. The table details the progression (or regression) in terms of true positives, false positives, false negatives, and derived metrics (precision, recall, F1 score).

Method	Number of Obj. Properties	True Positive	False Positives	False Negatives	Precision	Recall	F1 Score
Reference ontology	60						
Mistral 7B Basic Model	17	4	13	56	23.5%	6.66%	0.1038
Mistral 7B Fine-tuned Model 1st	4	0	4	60	0%	0%	0
Mistral 7B Fine-tuned Model 2nd	0	0	0	0	0%	0%	0
Mistral 7B Fine-tuned Model 3rd	0	0	0	0	0%	0%	0

In this second fine-tuning iteration using Mistral 7B, we incorporated both the original dataset and the new one, effectively building upon the model that had already been exposed to the OE patterns from the first dataset. This cumulative training provided a stronger foundation of domain knowledge and formatting expectations. As a result, unlike in the first iteration, the model succeeded in generating the requested ontology on the first attempt, without needing a follow-up prompt.

Qualitatively, the model's response showed a more seamless integration of the role-based narrative and the final goal—producing a well-structured ontology (Table 3). The ontology's classes were introduced more coherently, reflecting a better conceptual mapping of key domain elements. While the quantitative metrics (precision, recall, and F1 score) may not have improved dramatically when compared to the previous iteration, the ability of the model to immediately produce a functioning ontology suggests it was better aligned with the task requirements. This indicates that the additional dataset and iterative fine-tuning have begun to guide the model toward a more stable and contextually appropriate representation of the ontology, making the second fine-tuning a noteworthy qualitative improvement over the first.

The second fine-tuning iteration did not improve upon the shortcomings observed after the first (Table 4). In fact, it eliminated all attempts at identifying object properties entirely—no true positives, but also no false positives. While removing false positives could be seen as the model becoming more conservative, the lack of any true positives or meaningful predictions indicates that the model has essentially "given up" on identifying object properties. This outcome might suggest that the additional training data or instructions did not help the model internalize the relationships between classes and their corresponding object properties. Instead of correcting course, the second fine-tuning iteration pushed the model toward a stance of total non-commitment. Comparing this to previous results, including the GPT-4 fine-tuned models, makes the problem stark. GPT-4 iterations typically attempted to refine their object property recognition, and even if they struggled, they did not default to a state of complete inactivity. The Mistral 7B model's second iteration shows no productive engagement with the domain's relational structures, falling well short of both its own baseline performance and GPT-4's incremental improvements in similar stages of refinement.

By the time we reached the third fine-tuning iteration with Mistral 7B, its performance in class identification had deteriorated to the point of producing no meaningful outputs at all (Table 3). This result starkly contrasts with the reference ontology's comprehensive structure of 80 classes and falls well short of any progress achieved in earlier attempts. In fact, while the first fine-tuning run offered some semblance of engagement—albeit with duplicates and low precision—and the second iteration at least attempted to identify a handful of correct concepts, the third yielded a complete absence of true positives. Compared to GPT-4's iterative refinements, which, despite challenges, consistently showed incremental gains or at least balanced trade-offs in precision and recall, Mistral 7B's final iteration shows no such adaptability. The model's stagnation could stem from multiple factors: its smaller size and capacity relative to GPT-4, the manual and more cumbersome re-upload process required for each new fine-tuning stage, and possible misalignments between the prompt instructions, data structure, and Mistral's internal representation capabilities. Taken together, these issues suggest that the model was neither fully assimilating the domain knowledge from the training sets nor adjusting its latent space to accurately reflect the reference ontology's complexity, ultimately resulting in disappointingly null outputs by the end of the third fine-tuning round.

Furthermore, the model's performance on object properties remained stagnant at zero across all relevant metrics: no true positives, no false positives, and no false negatives (Table 4). This flatline suggests that the model has not recovered or improved its internal representation of object properties. Instead, it remains in a neutral state of producing no object property identifications at all. In comparison, previous models—particularly the GPT-4-based ones—demonstrated some capacity for adaptation over multiple fine-tuning rounds, either increasing their precision or expanding their recall. The persistent zero-performance state in the Mistral model's third iteration emphasizes a critical failure to benefit from iterative training, and it underscores the challenge of aligning Mistral 7B with the complex relational reasoning required in OE tasks. While GPT-4 fine-tuning rounds often yielded at least incremental gains or trade-offs (such as improved precision at the cost of reduced recall), the Mistral 7B model's repeated failure to identify any object properties indicates that its training regime and data exposure might need substantial revision.

4.2.3. Error Analysis

To provide a deeper analysis of model performance, we examined errors in the generated ontologies to understand recurring limitations and potential challenged. Errors were primarily observed in three areas: (1) generating invalid axioms, where models produced statements that were either syntactically incorrect or logically inconsistent; (2) mismatched relationships, where object properties were incorrectly assigned, leading to incorrect domain and range specifications; and (3) poorly structured hierarchies, where models misclassified subclasses under incorrect parent concepts, affecting the overall ontology structure.

One notable example of invalid axiom generation was observed in Mistral 7B's handling of disjoint class constraints. It classified Firefighter as a subclass of FirstResponder, correctly indicating that all firefighters are first responders. However,, it then correctly classified Firefighter as a subclass of "not FirstResponder", implying that firefighters are explicitly not first responders. This contradiction creates an inconsistency in the ontology, as an entity cannot belong to both class and its negation. Such errors disrupt logical reasoning and can cause failures in automated inference systems. A correct ontology should either maintain Firefighter as a subclass of FirstResponder without contradiction or correctly define disjointness when necessary. Mistral 7B also incorrectly assigned the isAssignedTo relationship between firetrucks and firefighters. The model stated that a firetruck is assigned to a firefighter, which is incorrect. In reality, it is firefighters who are assigned to firetrucks, not the other way around. This error misrepresents the logical structure of the ontology and could lead to incorrect reasoning in applications. The correct relationship should define firefighters as the subject of the assignment, with firetrucks as the object. Poorly structured hierarchies were frequently observed in cases where models had to infer subclass relationships. For example, Mistral 7B placed "EmergencyResponseTeam" as a subclass of "FirstAidKit", which is a category error. GPT-4, on the other hand, successfully maintained the integrity of subclass structures but tended to create overly granular subclass divisions when unnecessary, leading to ontology bloat.

These observations highlight the different strengths and weaknesses of each model and reinforce the necessity of structured fine-tuning with domain-specific datasets to improve the logical coherence of ontology outputs.

4.3. Discussion

4.3.1. Quantitative Evaluation Metrics

The comparison between GPT-4 and Mistral 7B for generating domain-specific ontologies through fine-tuning reveals significant contrasts in their capabilities and highlights areas requiring further refinement. Both models were fine-tuned using foundational OE texts, and their performance was evaluated based on precision, recall, and F1 scores compared to a benchmark ontology crafted by human experts. The results from Section 4.2 illustrate the differing abilities of the two models to integrate and represent complex ontological knowledge.

GPT-4 demonstrated a notable capacity for incorporating foundational knowledge and iteratively improving its outputs through successive fine-tuning. Although the initial results were modest, each fine-tuning iteration led to incremental improvements in precision, such as reaching 83% in the second iteration, indicating better accuracy in identifying relevant classes. However, its recall metrics, such as 12.5% in the same iteration, reflect persistent challenges in capturing the breadth of the benchmark ontology. More specifically, for class identification, the precision of the basic model starts at 69% and improves to 83% by the second fine-tuned model, an increase of approximately 20.3%. In the final iteration, precision stabilizes at 81.25%, confirming a consistent upward trend. However, recall progresses more slowly, increasing from 11.25% in the basic model to 16.25% in the third fine-tuned model, a relative improvement of 44.4%. This discrepancy between precision and recall indicates that while GPT-4 becomes increasingly accurate, it still struggles to identify a broader set of relevant classes, reflected in its F1 score, which improves from 0.1935 in the basic model to 0.2708 in the final iteration—an overall improvement of approximately 39.9%. In the evaluation of object properties, GPT-4 begins with no true positives in its basic form, yielding a precision and recall of 0%. However, fine-tuning progressively improves precision, which reaches 66.6% by the third iteration. Despite this, recall remains extremely low, improving only marginally from 0% to 3.33%. As a result, the F1 score, which starts at 0, achieves only 0.0634 by the final iteration. GPT-4 consistently displayed a capacity to adapt and refine its understanding of the domain, effectively leveraging the prompts and incorporating iterative feedback.

In contrast, Mistral 7B faced more pronounced limitations. While the first fine-tuning iteration demonstrated the model's ability to generate classes, the presence of duplicate classes and a relatively low precision of 14.2% highlighted significant issues with internal consistency. Subsequent iterations showed slight improvements in precision and recall but remained far behind GPT-4 in both metrics. By the third iteration, Mistral 7B completely failed to produce meaningful outputs, with precision and recall dropping to 0%, indicating a complete breakdown in its ability to align with the benchmark ontology. These results underscore the inherent architectural and computational constraints of Mistral 7B compared to GPT-4. More specifically, for class identification, the basic model achieves a precision of 68.4% and a recall of 15.25%, resulting in an F1 score of 0.2626. However, the first fine-tuned iteration shows a sharp decline in performance, with precision dropping to 14.2%, a decrease of 79.2%, while recall plummets to 3.75%, reflecting a 75% reduction. Although precision recovers to 50% in the second fine-tuning iteration, recall remains low at 7.5%, and the final iteration fails entirely, producing no true positives and reducing recall to 0%. Despite an improvement in precision to 80% in the third iteration, this result is misleading, as it arises in the absence of any identified classes, effectively nullifying practical utility. In the case of object properties, Mistral 7B exhibits similar trends. The basic model begins with a precision of 23.5% and a recall of 6.66%, yielding an F1 score of 0.1038. However, fine-tuning results in complete failure across all iterations. The first fine-tuned model identifies no true positives, leading to a precision and recall of 0%. Subsequent iterations show no recovery, with precision and recall remaining at 0%. Even though the final iteration reports a precision of 80%, this figure is deceptive, as the absence of any true positives means the model fails to provide any meaningful object property identification.

The disparity in performance between the two models can be attributed to several factors. First, GPT-4's larger parameter space and advanced architecture equipped it to handle the logical reasoning and complexity required for OE tasks. Conversely, Mistral 7B's smaller size limited its ability to generalize and represent nuanced relationships effectively. Second, the fine-tuning methodologies differed significantly. GPT-4 employed an iterative approach, building upon previous fine-tuned models, which facilitated cumulative learning and enhanced performance. Mistral 7B, due to platform constraints, required re-uploading datasets for each iteration, disrupting continuity and hindering knowledge retention. Third, while GPT-4 consistently understood and executed the role-playing prompts designed to guide the ontology generation process, Mistral 7B struggled with prompt comprehension and produced incomplete or redundant outputs.

Despite GPT-4's superior performance, both models fell short of fully replicating the benchmark ontology. This highlights a critical limitation in their ability to generalize domain-specific knowledge solely from foundational texts. For instance, GPT-4 and Mistral 7B struggled with object properties, with the former achieving a maximum precision of 66.6% but low recall, while the latter failed to produce any meaningful outputs in later iterations. These shortcomings suggest a gap in the models' ability to integrate complex, domain-specific relationships and semantic structures.

To address these challenges, further experimentation is essential. Incorporating real-world domain-specific datasets, such as those derived from Search and Rescue (SAR) missions, could significantly enhance the models' contextual understanding and enable them to generate more robust and realistic ontologies. Such datasets would provide a richer foundation for fine-tuning, allowing the models to align more closely with practical requirements. Additionally, expanding the scope and depth of competency questions could help capture intricate relationships and guide the models toward improved precision and recall. This would not only refine their outputs but also make them more relevant to real-world applications.

In conclusion, while GPT-4 demonstrated greater potential for OE tasks compared to Mistral 7B, both models exhibited limitations that underscore the need for more comprehensive datasets and refined methodologies. Incorporating domain-specific data and leveraging iterative, feedback-driven approaches would likely yield more robust ontologies and bridge the gap between automated generation and expert-crafted benchmarks. These advancements could establish large language models as indispensable tools for domain-specific knowledge representation, ultimately enhancing their utility in OE and related fields.

4.3.2. Qualitative Analysis

Beyond numerical evaluation metrics, it is essential to analyze the ontologies' structural integrity, semantic coherence, and practical usability. The qualitative assessment provides deeper insight into the generated ontologies' logical consistency, hierarchy formation, and alignment with expected domain-specific knowledge.

1. Structural soundness: one of the key factors in assessing ontology quality is the correct hierarchical arrangement of concepts. GPT-4 consistently maintained structured and logically aligned hierarchies, preserving correct superclass-subclass (subsumption) relationships in most cases. However, it occasionally overgeneralized categories, grouping semantically distinct concepts under broader umbrella terms, requiring manual refinement. In contrast, Mistral 7B, particularly in early iterations, exhibited misclassifications and redundancy issues, such as assigning multiple conflicting parent classes to a single entity. However, with additional domain-specific fine-tuning, Mistral 7B showed significant improvement in class hierarchy consistency, correcting misplaced classifications and better differentiating subclass relationships.
2. Semantic coherence: the accuracy of relationships and constraints between entities is another important aspect of ontology quality. Mistral 7B struggled in initial fine-tuning iterations with incorrect domain and range assignments for object properties, often misaligning relationships (e.g., classifying an entity to an incorrect category). For example, it initially misclassified "EmergencyResponder" as both a subclass of

"MedicalPersonnel" and "Firefighter", leading to inconsistencies. After incorporating domain-specific datasets, the model demonstrated improved understanding, correctly placing "EmergencyResponder" as a broader category encompassing both roles without conflicting subclass assignments. GPT-4, on the other hand, rarely misclassified relationships but occasionally omitted necessary inverse relations, requiring manual correction.

3. Usability and practical implementation: a key consideration in evaluating ontology quality is its real-world applicability. GPT-4-generated ontologies were more structured and required minimal post-processing, making them suitable for immediate integration into knowledge-based systems. However, they sometimes lacked specificity, necessitating additional manual refinement. Mistral 7B required more adjustments in earlier iterations, but after targeted fine-tuning, its outputs became increasingly reliable, showcasing improved adaptability to domain-specific needs. This suggests that while GPT-4 provides a strong initial structure, Mistral 7B has the potential to generate highly customized ontologies when trained with more targeted data.

The evolution of Mistral 7B's ontology quality through iterative fine-tuning highlights the impact of dataset quality, competency question expansion, and domain adaptation. These refinements allowed the model to generate ontologies that better align with expert-defined structures. The comparison between GPT-4 and Mistral 7B demonstrates that while pre-trained models can produce well-formed structures, domain-specific fine-tuning plays a crucial role in refining their logical consistency and contextual accuracy.

4.3.3. Expert Evaluation on Ontology Quality

Beyond quantitative evaluation metrics, expert assessment plays a critical role in validating the logical coherence and practical applicability of generated ontologies. To further understand the strengths and weaknesses of the models, domain experts were consulted to assess the conceptual accuracy of class definitions and relationships, the logical consistency of generated axioms, and the overall utility of the ontology for real-world applications.

Experts noted that GPT-4 generally produced well-structured ontologies, but some inconsistencies appeared in domain-specific adaptations, requiring manual corrections. For example, in a generated ontology for Search and Rescue (SAR) operations, GPT-4 correctly classified IncidentCommander as a subclass of EmergencyResponder but mistakenly included it under both OperationalUnit and CommandUnit without clear distinction. This overlap had to be manually refined to maintain logical consistency.

Mistral 7B, on the other hand, was found to be more computationally efficient but required substantial post-processing due to missing or misclassified relationships. In some cases, it omitted necessary transitive relationships. For example, the model failed to infer that if a Firetruck isAssignedTo Station1 and Station1 isPartOf ResponseZoneA, then the Firetruck should be implicitly associated with ResponseZoneA. These inference gaps were commonly identified in expert reviews.

To ensure the reliability of expert assessments, multiple evaluators reviewed the ontologies independently. While formal inter-rater reliability calculations were not conducted, the consistency of independent expert feedback suggests a high level of agreement in assessing logical coherence and applicability. Experts particularly emphasized that while fine-tuned LLMs significantly aid ontology engineering, human validation remains essential to refine structure, resolve ambiguities, and correct logical inconsistencies.

4.4. Extended Experimentation

To enhance the robustness of the models and their ability to generate domain-specific ontologies, we expanded our experimentation by integrating two additional, domain-specific, datasets. These datasets were designed with the expertise of the authoring team, under their capacity as domain experts, and aimed to address the unique requirements of SAR operations during wildfire incidents. The need for domain-specific datasets beyond basic OE knowledge emerged from the findings of previous experiments, where the results fell short of representing the full capabilities of the large language models (LLMs) and their fine-tuning potential. While the models demonstrated some ability to process foundational OE concepts, their outputs lacked the contextual richness and applicability needed for complex, real-world scenarios. This highlighted the importance of integrating targeted, domain-specific datasets to boost the models' performance and enable them to generate more robust and nuanced ontologies.

The first dataset builds on the original set of 18 competency questions derived from the reference ontology. Recognizing gaps in coverage and the need for more granular representation of wildfire SAR operations, we subdivided the competency questions into subsections reflective of specific operational contexts as followed: Weather-Related Questions, Incident-Related Questions, Data from Human and Earth Observations, Missions and Services, Population and Community Impact, Environmental Impact and Rehabilitation, Technology and Modeling, Safety and Protocols and some Additional Questions. This structured expansion was guided by critical areas of knowledge that must be represented for effective SAR decision-making. By integrating domain expertise, we extended the dataset to 48 competency questions, ensuring comprehensive coverage of SAR ontology requirements. These competency questions were meticulously answered based on our domain expertise, forming a knowledge-rich dataset. The finalized Q&A pairs were uploaded into GPT-4 for JSONL conversion, ensuring consistency and accessibility for fine-tuning purposes. This expanded dataset aimed to provide nuanced insights and a broader perspective on SAR operations, enriching the model's training data and enabling it to represent complex, domain-specific scenarios accurately.

The second dataset was constructed from incident report templates used by the Greek Fire Department during actual wildfire incidents. These templates encapsulate critical information recorded during operations, such as resource allocation, mission timelines, environmental conditions, and outcomes. These real-world data were translated into a textual format and subsequently processed through GPT-4 to generate Q&A pairs in JSONL format. This dataset brings an operationally grounded perspective to the model, incorporating the terminology, procedural details, and contextual nuances encountered in SAR missions. By aligning the training data with practical applications, the dataset serves to bridge the gap between theoretical ontology principles and real-world use cases.

To evaluate the impact of these datasets, we fine-tuned the models in the configurations that each of them requires: on GPT-4 we followed the Separate Integration. Each dataset (expanded competency questions and incident reports) was uploaded independently to the last trained model (3rd fine-tuned) for fine-tuning. On Mistral 7B we comply with Sequential Integration. Following the sequence used in previous experiments, we fine-tuned the models first with the expanded competency questions dataset, followed by the incident report dataset. This method simulated a cumulative knowledge-building process, where the foundational ontology knowledge was augmented with real-world operational data.

Initial experiments suggest that incorporating these datasets leads to a noticeable improvement in the models' ability to generate robust, contextually accurate ontologies. The competency questions provided a theoretical framework, while the incident reports added operational depth, enabling the models to represent both strategic and tactical dimensions of SAR operations effectively. Further evaluation of precision, recall, and F1 scores, as well as qualitative assessments of generated ontologies, is ongoing to quantify these improvements. These datasets highlight the critical role of domain-specific knowledge in enhancing the utility of large language models for OE tasks, particularly in specialized applications like wildfire SAR missions.

4.4.1. GPT4

The new fine-tuned GPT-4 model demonstrated exceptional performance in generating a robust SAR ontology. By incorporating domain-specific datasets such as the expanded competency questions and wildfire incident reports, GPT-4 effectively utilized its role-playing capabilities to structure and generate a cohesive ontology. The model generated an ontology that aligns closely with practical SAR requirements, including detailed classifications of classes, object properties, and data properties (Tables 5 and 6). As seen in the ontology metrics, GPT-4 produced a structure with 72 distinct classes, 32 object properties, and 7 data properties, effectively capturing both theoretical and operational aspects. The model also achieved logical coherence, as evidenced by the 284 logical axioms and the robust use of subclass relationships. These results highlight GPT-4's ability to translate domain-specific knowledge into comprehensive and structured ontologies, showcasing its potential as a reliable tool for OE in specialized fields like wildfire SAR operations.

Table 5. A comparison of class generation metrics between the fine-tuned GPT-4 and Mistral 7B models after incorporating domain-specific datasets and their basic models, highlighting significant improvements in recall and F1 scores compared to previous experiments limited to OE training data.

Method	Number of Classes	True Positive	False Positives	False Negatives	Precision	Recall	F1 Score
Reference	80						
GPT-4 Basic Model	13	9	4	71	69%	11.25%	0.1935
Mistral 7B Basic Model	19	14	6	67	68.4%	15.25%	0.2626
New Fine-Tuned GPT-4	72	32	40	48	44.4%	40%	0.4210
New Fine-Tuned Mistral 7B	65	39	26	41	60%	48.75%	0.5379

Table 6. A comparison of object property generation metrics between the fine-tuned GPT-4 and Mistral 7B models after incorporating domain-specific datasets and their basic models, emphasizing significant improvements in recall and F1 scores over previous experiments relying solely on OE training data.

Method	Number of Obj. Properties	True Positive	False Positives	False Negatives	Precision	Recall	F1 Score
Reference	60						
GPT-4 Basic Model	5	0	5	60	0%	0%	0
Mistral 7B Basic Model	17	4	13	56	23.5%	6.66%	0.1038
New Fine-Tuned GPT-4	32	14	18	46	43.75%	23.33%	0.3043
New Fine-Tuned Mistral 7B	51	27	24	33	52.94%	45%	0.4864

More specifically, for class identification, the GPT-4 basic model achieved a precision of 69%, a recall of 11.25%, and an F1 score of 0.1935. In comparison, the new fine-tuned GPT-4 model demonstrated a precision of 44.4%, showing a 35.6% decrease, but its recall improved dramatically to 40%, marking a 255.5% improvement. This significant recall increase contributed to the F1 score rising to 0.4210, an overall improvement of 117.5%. For object property identification, the GPT-4 basic model started with 0% precision and 0% recall, producing an F1 score of 0. After fine-tuning, the new GPT-4 model achieved a precision of 43.75% and a recall of 23.33%, representing significant improvements over the baseline. The F1 score increased to 0.3043, highlighting the model's enhanced ability to identify object properties and maintain a reasonable balance between precision and recall. In summary, the new fine-tuned GPT-4 model demonstrates substantial gains over the basic model, particularly in recall, which shows remarkable improvement across both tasks. Despite a decrease in precision for class identification, the overall F1 score improvements confirm the effectiveness of the fine-tuning process.

4.4.2. Mistral 7B

The fine-tuned Mistral 7B model exhibited an impressive leap in performance during the generation of domain-specific ontologies for SAR operations. Unlike the earlier experiments, where Mistral 7B struggled compared to GPT-4, this iteration not only matched but surpassed GPT-4 in both depth and contextual accuracy. Leveraging domain-specific data, the model elevated the interaction among the three roles based on the HCOME OE methodology [17]. The iterative discussions, enriched by the model's ability to simulate nuanced domain knowledge, produced more insightful and well-rounded outputs, with enhanced clarity and purpose in defining and linking key concepts relevant to wildfire SAR scenarios. The ontology metrics further validate this success. Mistral 7B generated an ontology comprising 65 classes and 51 object properties, surpassing GPT-4 in capturing relationships and semantic depth (Tables 5 and 6). More specifically, for class identification, the Mistral 7B basic model achieved a precision of 68.4%, a recall of 15.25%, and an F1 score of 0.2626. After fine-tuning, precision decreased to 60%, reflecting a slight 12.3% decrease, but recall improved significantly to 48.75%, corresponding to an improvement equal to 219.7%. As a result, the F1 score rose to 0.5379, marking an overall improvement of 104.8%. This highlights a substantial enhancement in the model's ability to identify relevant classes, despite the slight decline in precision. For object property identification, the basic Mistral 7B model achieved a precision of 23.5%, a recall of 6.66%, and an F1 score of 0.1038. The new fine-tuned model improved precision to 52.94%, a 125.3% increase, and recall to 45%, reflecting an exceptional 575.7% improvement. This significant recall gain was the main contributor to the F1 score rising to 0.4864, an overall improvement of 368.4%.

Notable additions included classes and properties that detailed relationships between incidents, services, and environmental observations. The model introduced richer subclass hierarchies and properties, reflecting its improved grasp of the domain-specific complexities. This performance underscores the model's ability to semantically integrate knowledge, connecting heterogeneous data sources such as sensor observations and real-time weather conditions, and translating these into meaningful ontological structures. The fine-tuned Mistral 7B also excelled in leveraging discussions to optimize the ontology iteratively. The elevated level of role-play discussions facilitated not just the coverage of competency questions but also the generation of more context-aware and operationally useful ontologies. Mistral's ability to encapsulate operationally grounded and semantically rich outputs signifies its potential as a cost-efficient alternative to larger models, capable of yielding high-quality domain-specific ontologies tailored to real-world SAR challenges.

The performance of Mistral 7B exhibited significant inconsistencies across different fine-tuning sessions. Initial training attempts (as presented in Section 4.2.2, Tables 3 and 4) displayed notable deficiencies, such as class duplication, misclassification of object properties, and inconsistencies in relationship assignments. These issues primarily stemmed from early-stage hyperparameter settings, dataset preprocessing inconsistencies, and limitations in the initial prompt structures guiding the model. However, later iterations (Section 4.4.2, Tables 5 and 6) demonstrated a marked improvement, even surpassing GPT-4 in certain cases.

Several key factors contributed to this transition from suboptimal to high-performance results:

1. Data Preprocessing Adjustments: Early dataset formatting included overlapping or redundant entity descriptions, leading to class duplication and property assignment errors. By refining the dataset through more structured entity extraction and deduplication techniques, later training runs exhibited more consistent outputs.
2. Incorporation of Domain-Specific Datasets: A significant improvement was observed when ontology fine-tuning was expanded to include domain-specific datasets, which provided the model with richer contextual knowledge and helped mitigate early-stage errors in entity classification and relationship structuring. By grounding Mistral 7B in more domain-relevant examples, it was able to generate more coherent ontologies with improved logical consistency.
3. Expansion of Competency Questions: The number and complexity of competency questions used for model evaluation were increased in later iterations, providing a broader test set that reinforced correct interpretations of ontology structures. This iterative reinforcement led to improved model accuracy in reasoning-based ontology generation.
4. Iterative Fine-Tuning and Model Adaptation: The observed performance shift also highlights the cumulative effect of iterative fine-tuning, where each session builds upon prior refinements. While early fine-tuning attempts exposed model weaknesses, successive sessions reinforced correct classifications, improving coherence in ontology structuring.

These findings suggest that fine-tuning LLMs for ontology engineering requires a dynamic and adaptive strategy. Performance can fluctuate significantly based on domain-specific datasets and the iterative refinement process. This highlights the importance of continuous optimization and specific domain data to maximize model effectiveness in domain-specific knowledge representation.

5. Discussion

The results from Tables 5 and 6 underscore the transformative impact of incorporating domain-specific datasets into the fine-tuning process of large language models (LLMs), particularly in comparison to the earlier experiments outlined in Tables 1–4. These domain-specific datasets, focusing on wildfire SAR operations, allowed both GPT-4 and Mistral 7B to achieve significantly better precision, recall, and F1 scores, demonstrating their enhanced ability to represent domain-specific knowledge comprehensively.

In the earlier experiments (Table 1), GPT-4 exhibited limited recall and F1 scores even after iterative fine-tuning on basic OE knowledge. The baseline GPT-4 model achieved a precision of 69% but suffered from a recall as low as 11.25% and an F1 score of 0.1935. Similarly, Mistral 7B struggled with basic datasets, often failing to generate distinct classes without duplication, with minimal improvements across iterations.

With the introduction of domain-specific datasets, both models demonstrated significant performance gains in class generation. As shown in Table 5:

- GPT-4 improved its recall to 40%, doubling its performance compared to prior experiments. The addition of SAR-specific knowledge enriched the model's understanding, reflected in its ability to generate 72 distinct classes while maintaining a precision of 44.4%.

- Mistral 7B, previously outperformed by GPT-4 in class generation, surpassed GPT-4 in this iteration. It achieved a recall of 48.75% and an F1 score of 0.5379, emphasizing its ability to better integrate domain-specific insights into a semantically rich ontology with 65 classes. This represents a marked improvement over its earlier iterations (Table 3), where precision and recall were as low as 14.2% and 3.75%, respectively.

Object property generation results also highlight the value of domain-specific datasets (Table 6). Previously (Table 2), both GPT-4 and Mistral 7B struggled significantly with object property identification, with Mistral 7B failing to generate any true positives in later iterations and GPT-4 showing limited recall improvements.

- GPT-4, in its fine-tuned version with SAR datasets, generated 32 object properties, achieving a precision of 43.75% and an F1 score of 0.3043. Although this was a significant improvement over its earlier attempts (maximum recall of 3.33% in Table 2), it still lagged behind Mistral 7B in representing relational knowledge effectively.
- Mistral 7B, benefiting from sequential fine-tuning with domain-specific datasets, excelled in this category. It generated 51 object properties, achieving a recall of 45% and an F1 score of 0.4864. This improvement reflects the model's ability to map complex relationships, such as those between incidents, resources, and environmental factors, a crucial aspect of wildfire SAR operations.

The comparative analysis across these experiments underscores the critical role of domain-specific datasets in OE. While basic OE knowledge provides a foundational understanding, it falls short of enabling LLMs to address real-world, complex domains effectively. The SAR-specific datasets bridged this gap, enriching the models' knowledge bases with operationally grounded concepts and relationships.

The inclusion of expanded competency questions and incident report templates allowed the models to achieve the following:

- Better address the complexity of domain knowledge, including subclass hierarchies and property relationships.
- Improve recall, particularly in identifying classes and properties relevant to wildfire SAR operations.
- Enhance F1 scores, indicating a balanced ability to reduce false positives while capturing more true positives.

These results suggest that fine-tuning LLMs with targeted, domain-specific data are pivotal for achieving robust knowledge representation. Domain-specific datasets not only enable the models to better emulate human expert ontologies but also highlight their ability to capture operational nuances, such as environmental impacts, mission timelines, and resource allocation.

Furthermore, the performance disparity between Mistral 7B and GPT-4 in this context demonstrates that smaller models like Mistral, when equipped with rich domain-specific data, can outperform larger models trained on general datasets. This finding is particularly valuable for applications where computational resources or budget constraints favor smaller models.

Further experimentation reinforces the importance of aligning training datasets with practical, domain-specific requirements. By integrating such datasets, LLMs can achieve higher precision and recall, bridging the gap between theoretical OE and real-world applications. These insights pave the way for more refined methodologies and hybrid approaches to optimize knowledge representation in specialized fields like wildfire SAR operations.

6. Conclusions

This research paper explores the fine-tuning of LLMs, specifically GPT-4 and Mistral 7B, for OE tasks. Through comprehensive experiments, the impact of general OE training and domain-specific fine-tuning on the models' ability to generate robust and practical ontologies was evaluated. The results validated the identified research questions and revealed critical insights into the role of targeted data in enhancing the models' performance.

The first research question (RQ1) related to the fine-tuning of LLMs on general OE concepts and the improvement in their performance was partially validated. While fine-tuning improved the baseline capabilities of both GPT-4 and Mistral 7B, the results highlighted a few limitations. GPT-4 demonstrated incremental improvements, achieving a precision of 81.25% and a modest recall increase from 11.25% to 16.25% in OWL class generation after three iterations. Mistral 7B, however, struggled with basic OE datasets, showing limited progress and regression in key metrics, with a recall of only 15.25% in preliminary experiments. These findings suggest that while foundational OE training is essential, it alone cannot shape models for the nuanced requirements of domain-specific knowledge representation.

The second research question (RQ2)—that incorporating domain-specific datasets would enhance the practical utility of generated ontologies—was strongly validated. The addition of SAR-specific datasets, including expanded competency questions and real-world wildfire incident reports, significantly improved the models' performance. GPT-4's recall for class generation increased to 40%, with an F1 score of 0.4210, while its recall for object properties rose to 23.33%. More notably, Mistral 7B, which had previously lagged, surpassed GPT-4 in both class and object property metrics. It achieved a recall of 48.75% and an F1 score of 0.5379 for classes, and a recall of 45% and an F1 score of 0.4864 for object properties. These results underscore the transformative impact of domain-specific data in enabling models to accurately represent complex and operationally grounded knowledge structures.

The key findings from this study highlight the necessity of combining domain-specific datasets and fine-tuning methodologies for maximizing LLM performance in specialized tasks. The integration of SAR-specific data enriched the models' understanding of key relationships and hierarchies, enabling the generation of semantically rigorous and practically relevant ontologies. Mistral 7B's performance in these experiments was particularly noteworthy, as it not only overcame its earlier struggles but also outperformed GPT-4 in recall and F1 scores for both classes and object properties. This demonstrates that smaller, cost-efficient models can achieve competitive results when trained on targeted data.

This paper also revealed the importance of aligning training data with real-world applications. Earlier experiments using general OE data showed limited recall and low F1 scores, indicating a lack of contextual understanding. In contrast, domain-specific fine-tuning improved both models' ability to generate ontologies that aligned with practical needs, including detailed representations of SAR concepts such as weather impacts, resource allocation, and incident timelines. These improvements emphasize that domain-specific datasets bridge the gap between theoretical ontology principles and operational utility. In conclusion, this research confirms the critical role of fine-tuning with domain-specific datasets in enhancing the utilization of LLMs for OE. While general OE training lays the groundwork, it is insufficient for generating robust, application-ready ontologies. Domain-specific datasets enable models to achieve significantly higher recall and F1 scores, reflecting better contextual understanding and practical relevance. The findings demonstrate the potential of LLMs, particularly when fine-tuned with targeted data, to advance OE and serve as indispensable tools for knowledge representation in specialized fields such as wildfire SAR operations. These insights pave the way for future studies to refine fine-tuning methodologies and extend the applicability of LLMs across diverse domains.

While this study demonstrates the feasibility of fine-tuning LLMs for ontology engineering in SAR contexts, an important next step is the integration of these ontologies into operational decision support systems. In practical applications, ontology-driven reasoning can be embedded into SAR platforms to enhance incident classification, automate information retrieval, and support knowledge-based decision-making. For example, ontology-enhanced knowledge graphs could improve real-time situational awareness by dynamically linking incident reports, geographical data, and response protocols to assist emergency teams.

Future improvements could involve incorporating retrieval-augmented generation (RAG) strategies and active learning approaches to continuously refine the model based on real-time domain expert feedback. This would allow the system to adapt to new terminologies, evolving threat landscapes, and emerging response protocols in SAR environments. Such an approach could ensure that the model remains aligned with the operational needs of first responders while improving its generalization capabilities across different emergency scenarios.

Furthermore, we recognize the need for additional details regarding data preparation steps, computational efficiency, and real-world deployment scenarios. Future research could explore a fully operational prototype of an ontology-powered SAR assistant, systematically evaluating scalability, response accuracy, and integration challenges in real-world settings. These enhancements would further solidify the role of LLMs in ontology-driven decision support at scale.

The experiments detailed (source files, screenshots, ttl files and chat conversations) are saved in the following GitHub repository: https://github.com/dimitrisdoumanas19/Fine-tuning-LLMS.git (accessed on 7 December 2024).

Author Contributions: Conceptualization, D.D., A.S. and K.K.; Methodology, D.D., A.S. and K.K.; Writing—original draft, D.D.; Writing—review & editing, D.D., D.S., C.V. and K.K.; Supervision, D.S., C.V. and K.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are openly available in https://github.com/dimitrisdoumanas19/Fine-tuning-LLMS.git, accessed on 7 December 2024.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Zhang, B.; Carriero, V.A.; Schreiberhuber, K.; Tsaneva, S.; González, L.S.; Kim, J.; de Berardinis, J. OntoChat: A Framework for Conversational Ontology Engineering using Language Models. *arXiv* **2024**, arXiv:2403.05921. [CrossRef]
2. Doumanas, D.; Soularidis, A.; Kotis, K.; Vouros, G. Integrating LLMs in the Engineering of a SAR Ontology. In *Artificial Intelligence Applications and Innovations*; Maglogiannis, I., Iliadis, L., Macintyre, J., Avlonitis, M., Papaleonidas, A., Eds.; Springer Nature: Cham, Switzerland, 2024; pp. 360–374. [CrossRef]
3. Garijo, D.; Poveda-Villalón, M.; Amador-Domínguez, E.; Wang, Z.; García-Castro, R.; Corcho, O. LLMs for Ontology Engineering: A Landscape of Tasks and Benchmarking Challenges. In Proceedings of the International Semantic Web Conference ISWC2024, Baltimore, MD, USA, 11–15 November 2024.
4. Joachimiak, M.P.; Miller, M.A.; Caufield, J.H.; Ly, R.; Harris, N.L.; Tritt, A.; Mungall, C.J.; Bouchard, K.E. The Artificial Intelligence Ontology: LLM-assisted construction of AI concept hierarchies. *arXiv* **2024**, arXiv:2404.03044.

5. Saeedizade, M.J.; Blomqvist, E. Navigating Ontology Development with Large Language Models. In *The Semantic Web*; Lecture Notes in Computer Science; Peñuela, A.M., Dimou, A., Troncy, R., Hartig, O., Acosta, M., Alam, M., Paulheim, H., Lisena, P., Eds.; Springer Nature: Cham, Switzerland, 2024; Volume 14664, pp. 143–161. [CrossRef]
6. Mateiu, P.; Groza, A. Ontology engineering with Large Language Models. In Proceedings of the 2023 25th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC), Nancy, France, 11–14 September 2023.
7. Doumanas, D.; Bouchouras, G.; Soularidis, A.; Kotis, K.; Vouros, G. From Human- to LLM-centered Collaborative Ontology Engineering. Large Language Models (LLMs) and Ontologies. *Appl. Ontol.* **2024**, 15705838241305067.
8. Gekhman, Z.; Yona, G.; Aharoni, R.; Eyal, M.; Feder, A.; Reichart, R.; Herzig, J. Does Fine-Tuning LLMs on New Knowledge Encourage Hallucinations? *arXiv* **2024**, arXiv:2405.05904. [CrossRef]
9. Chang, C.; Wang, W.-Y.; Peng, W.-C.; Chen, T.-F. LLM4TS: Aligning Pre-Trained LLMs as Data-Efficient Time-Series Forecasters. *arXiv* **2023**, arXiv:2308.08469. [CrossRef]
10. Jeong, C. Fine-tuning and Utilization Methods of Domain-specific LLMs. *arXiv* **2024**, arXiv:2401.02981.
11. Anisuzzaman, D.M.; Malins, J.G.; Friedman, P.A.; Attia, Z.I. Fine-Tuning LLMs for Specialized Use Cases. *Mayo Clin. Proc. Digit. Health* **2024**, *3*, 100184. [CrossRef]
12. J, M.R.; VM, K.; Warrier, H.; Gupta, Y. Fine Tuning LLM for Enterprise: Practical Guidelines and Recommendations. *arXiv* **2024**, arXiv:2404.10779. [CrossRef]
13. Parthasarathy, V.B.; Zafar, A.; Khan, A.; Shahid, A. The Ultimate Guide to Fine-Tuning LLMs from Basics to Breakthroughs: An Exhaustive Review of Technologies, Research, Best Practices, Applied Research Challenges and Opportunities. *arXiv* **2024**, arXiv:2408.13296. [CrossRef]
14. Pathak, A.; Shree, O.; Agarwal, M.; Sarkar, S.D.; Tiwary, A. Performance Analysis of LoRA Finetuning Llama-2. In Proceedings of the 2023 7th International Conference on Electronics, Materials Engineering & Nano-Technology (IEMENTech), Kolkata, India, 18–20 December 2023; pp. 1–4. [CrossRef]
15. Nguyen, H.D.; Chamroukhi, F. Practical and theoretical aspects of mixture-of-experts modeling: An overview. *WIREs Data Min. Knowl. Discov.* **2018**, *8*, e1246. [CrossRef]
16. Bernardelle, P.; Demartini, G. Optimizing LLMs with Direct Preferences: A Data Efficiency Perspective. In Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region, Tokyo, Japan, 9–12 December 2024; pp. 236–240. [CrossRef]
17. Kotis, K.; Vouros, G.A. Human-centered ontology engineering: The HCOME methodology. *Knowl. Inf. Syst.* **2006**, *10*, 109–131. [CrossRef]
18. Masa, P.; Meditskos, G.; Kintzios, S.; Vrochidis, S.; Kompatsiaris, I. Ontology-based Modelling and Reasoning for Forest Fire Emergencies in Resilient Societies. In Proceedings of the 12th Hellenic Conference on Artificial Intelligence, Corfu, Greece, 7–9 September 2022; pp. 1–9. [CrossRef]
19. Allemang, D.; Hendler, J. *Semantic Web for the Working Ontologist*; Elsevier: Amsterdam, The Netherlands, 2011. [CrossRef]
20. Antoniou, G.; Groth, P.; van Harmelen, F.; Hoekstra, R. A Semantic Web Primer. In *Cooperative Information Systems*, 3rd ed.; MIT Press: Cambridge, MA, USA, 2012.
21. Keet, C.M. *An Introduction to Ontology Engineering*; College Publications: London, UK, 2018.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.