Spring 2024

Enhancing Environmental Health and Safety: Fine-Tuning Large Language Models for Domain-Specific Applications

Mohammad Adil Ansari

Enhancing Environmental Health and Safety: Fine-Tuning Large Language Models for Domain-Specific Applications

A Project

Presented to

The Faculty of the Department of Computer Science

San José State University

In Partial Fulfillment

of the Requirements for the Degree

Master of Science

by

Mohammad Adil Ansari

May 2024

The Designated Project Committee Approves the Project Titled

Enhancing Environmental Health and Safety: Fine-Tuning Large Language Models for Domain-Specific Applications

by

Mohammad Adil Ansari

APPROVED FOR THE DEPARTMENT OF COMPUTER SCIENCE

SAN JOSÉ STATE UNIVERSITY

May 2024

Dr. William Andreopoulos	Department of Computer Science
Dr. Genya Ishigaki	Department of Computer Science
Mr. Elton DSouza	Software Engineer, Google

ABSTRACT

Enhancing Environmental Health and Safety: Fine-Tuning Large Language Models for Domain-Specific Applications

by Mohammad Adil Ansari

This study aims to simplify Environmental Health and Safety (EHS) by leveraging the power of Large Language Models (LLMs). In this research, we focus on fine-tuning three LLMs — LLaMA, Mistral, and Falcon — using PEFT techniques such as QLoRA and SFT, to address domain-specific needs such as safety compliance, incident reporting, and knowledge dissemination. Our research methodology involves fine-tuning each LLM model on a custom dataset compiled from various regulatory agencies, supplemented by targeted web scraping and manual collection of questionnaires to capture and enrich the models with the latest regulations and guidelines. This study aims to compare the effectiveness of these fine-tuned models to identify the most effective model and fine-tuning techniques for specific EHS applications. We aim to integrate LLMs to support EHS practices and demonstrate a practical way of improving and automating EHS queries and reports to foster safer and more compliant workplace environments.

Keywords: Large Language Models, Environmental Health and Safety, LLaMA, Mistral, Falcon, Parameter-Efficient Fine-Tuning, QLoRA, Supervised Fine-Tuning, Safety compliance, Risk Assessment

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to my advisor, Dr. William Andreopoulos, for his invaluable guidance, patience, and support throughout this research project. His insights and suggestions were crucial in shaping both the direction and the execution of this study.

I am also immensely thankful to my committee members, Dr. Genya Ishigaki and Elton DSouza, for their valuable feedback and constructive critiques, which have significantly contributed to the enhancement of this research.

TABLE OF CONTENTS

CHAPTER

CHAPTER 1

Introduction

1.1 Introduction

Environmental Health and Safety (EHS) is one of the most essential cornerstones that focuses on encompassing the policies, standards, and practices designed to safeguard and protect the health and safety of people at work and the environment from hazards mainly caused by industrial operations. Due to the ever-evolving nature and complexity of these regulations navigating and managing these intricate changes requires continuous monitoring and adjustment to the existing processes which becomes a behemoth task for an EHS professional in the industry. Workers in the field may not be aware of the subtle or nuance changes in the regulations like when to use OSHA 300, 300-A, or 301 forms for incident reporting. There is a need for a tool that can simplify this. In recent years, due to rapid technological advancements, novel tools that have the potential to revolutionize the way EHS is managed and significantly improve the existing EHS processes have been developed. Among these innovations, artificial intelligence (AI) and more specifically Large Language Models (LLMs) such as LLaMA, Mistral, and Falcon are at the forefront due to their ability to understand the context and give results in natural and general-purpose ways. These models can generate user-friendly, plain-language explanations of complex safety procedures, making it easier for employees to understand and follow safety protocols. They can also be used to simulate real-world safety scenarios, answer employee questions, and provide personalized training experiences. Furthermore, they can even be used to provide recommendations for compliance adjustments helping organizations stay in line with the EHS regulations.

These models can be useful and have shown potential and capability to process and understand huge amounts of data which can be a lot of different formats be it structured or unstructured, this has proved to be instrumental in drawing inferences from complex documents and reports which is a common practice in EHS. There is an immense amount of potential to improve the EHS process and practices which will in turn help streamline the incident reporting and compliance aspect within EHS. This will help increase the accuracy of incident reporting that can not only be instrumental in correct data collection for analysis but also open avenues to automate the auditing of such information thereby greatly reducing the cost associated with such activities. These models can also be used for training purposes to simplify complex regulatory text in simple and understandable language thereby acting as a bridge to communicate specialized terminology to newcomers in the field of EHS. However, these general-purpose and normal LLMs lack the context to handle specific jargon such as *"STS (Standard Threshold Shift)"* and phrases such as *"Do I have to update the OSHA 301 Incident Reports?"* used in the EHS world. This gives rise to a need to have a customized LLM based on the generic ones so that not only we can harness the analytical power of these models but also adapt to the specific needs within EHS.

Therefore, this study aims to study ways to achieve customized LLM for specialized EHS tasks. For this, we will fine-tune three popular open-source LLM models on the custom EHS-specific dataset. We use state-of-the-art Parameter-Efficient Fine-Tuning (PEFT) methods such as Low-Rank Adaptation (LoRA)[1] and Supervised Fine-Tuning (SFT) to efficiently fine-tune the three general-purpose LLMs so that they can adapt to the needs of EHS and give responses in a particular format. We also investigate not only the technical avenues of LLM models in the EHS domain but also the practical feasibility of these models in actual real-world scenarios.

This study also investigates and highlights the transformative potential of LLMs in EHS process and management by providing a foundation for adapting general LLMs to handle complex and specialized EHS knowledge. We also identify existing gaps in incorporating LLMs into the current EHS practices and their future implications. We envision that soon, we can leverage these specialized models to enhance safety, minimize risks, and strengthen compliance outcomes, thereby moving towards a safer workplace environment and improved environmental protection strategy and outlook.

1.2 Motivation

The main motivation is driven by the pressing need to address the significant issues and substantial challenges faced by professionals who work in the field of Environmental Health and Safety (EHS). These EHS professionals face a constant battle with handling incident reporting in different and nuanced manufacturing settings, swiftly interpreting complex regulatory documents, and conducting rapid risk assessments to mitigate any potential issues in the auditing. Traditional EHS practices and management methods are riddled with manual, time-consuming, and archaic practices which include manually filing the incident reports using pen and paper, on-site in-person manual compliance checks, and manual on-call logging incidents. These methods not only provide an opportunity for an error to occur due to manual and human involvement but also are time-consuming and labor-intensive; making them a huge liability for a company that has to follow the EHS process. These manual tasks also overburden the system which is instrumental in workers and environmental safety. Moreover, the manual process might fail in case of rapidly evolving regulations and increasing volumes of data.

EHS professionals navigate a complex maze of city, county, state, and federal regulations, spread across different regulatory agencies, and an array of diverse codes and regulatory documents that encompass many different jurisdictions. This vast amount of regulatory information is nearly impossible to handle manually and is often time overwhelming for any resource-crunched EHS compliance unit within an organization.

Due to the sheer complexity of regulatory documentation, it requires significant time and resources to interpret and ensure compliance with all the regulations within the organization. Additionally, the dynamic nature of the regulations means that EHS professionals need to learn continuously and stay abreast with the latest amendments, a task that is critical, challenging, and burdensome.

In the EHS world incident reporting and risk assessment are equally challenging. It is crucial to report all safety incidents in a timely and accurate fashion to keep in line with legal and audit compliance, ensure the safety of all factory workers, and help prevent any similar incidents in the future. However, traditional methods of incident reporting are riddled with inaccuracies due to long manual cumbersome paperwork and slow ancient data entry processes that have not been changed in ages. This also causes a significant delay in response time and incorrect steps being taken when an actual incident occurs. Thus, to mitigate any potential hazard before any serious hazards occur, proactive risk assessments and a detailed preemptive strategy to prevent such incidents are essential. To draw such a plan a detailed analysis is required. We can chalk out a detailed strategy by automating and streamlining the data capture process to meticulously analyze and interpret the data.

In this context, Large Language Models (LLMs) emerge as a beacon of hope for EHS professionals. They offer a glimpse into a future where these critical yet cumbersome processes are streamlined and optimized. LLM models such as LLaMa, Mistral, and Falcon, with their deep learning capabilities, have significant potential in understanding human messages in text format, interpreting regulatory text, and presenting them in a lucid manner that can potentially revolutionize EHS practices. These models hold immense potential to significantly enhance the accuracy of interpreting regulatory text and reporting incidents. LLMs can ingest vast of EHS-related data, extract relevant key information, and generate insights much faster than traditional methods.

You

An employee claims he hurt himself at work. He takes it upon himself to take one or two days off using his own personal time. He does not provide us with any medical documentation to back up his being unable to work and claims he doesn't need to see a doctor, he just want to rest for a day or two. Do we need to record this as a days away from work case if we question that there is no medical evidence that he needed the time off?

ChatGPT

According to OSHA guidelines, if an employee misses work due to a work-related injury or illness and the absence extends beyond the day of injury, it should be recorded as a "days away from work" case, regardless of whether the employee uses paid time off or unpaid time off. However, there are some nuances to consider:

1. **Work-Relatedness**: If there is doubt about whether the injury or illness is work-related, it's essential to thoroughly investigate the circumstances surrounding the incident. If it's determined that the injury or illness is not work-related, then it may not need to be recorded as a days away from work case.
2. **Medical Documentation**: While OSHA doesn't explicitly require medical documentation for recordkeeping purposes, it's generally advisable to obtain medical documentation whenever possible to substantiate the need for time off due to an injury or illness. However, if an employee chooses not to seek medical treatment or provide documentation, it can complicate the recordkeeping process.
3. **Employer Discretion**: Ultimately, the decision to record an absence as a days away from work case may depend on the employer's policies and procedures, as well as any applicable state or local regulations. Some employers may choose to err on the side of caution and record the absence as a days away from work case if there is any doubt about its work-relatedness or the need for medical documentation.

It's essential to document the details of the incident, including the employee's report of injury, any investigations conducted, and any communications with the employee regarding the absence. Consult with your organization's HR department or legal counsel to ensure compliance with relevant regulations and company policies.

Figure 1: Respose of GPT 3.5 if asked an EHS related task

However, directly using these general-purpose LLMs comes with its own set of challenges and issues. We can't directly deploy these generic LLMs in the EHS domain, because they lack the domain-specific knowledge required to accurately interpret the

EHS regulations or generate precise information to capture the incidents. This exposes a significant limitation in the current usage of LLMs within the EHS domain. One way of eliminating this limitation is building our custom LLMs by fine-tuning existing models on EHS-domain-specific data.

In Fig. 1 we can see that when GPT 3.5 is prompted to give an EHS-related question it fails to answer the correct steps to be taken.

Therefore, the main motivation of our research is to bridge this gap by developing specialized LLMs tailored specifically for EHS applications. By powering the existing LLM models using fine-tuning techniques we can unlock the full potential of this technology and pave the way for a future that is more safer. This would also support EHS professionals in their day-to-day and would significantly lower the cost overhead for a lot of companies.

1.3 Problem Statement

The main issue that this research aims to address is the significant underutilization of Large Language Models (LLMs) within the Environmental Health and Safety (EHS) sector. Although a lot of industries have readily adopted them, LLMs haven't fully penetrated the EHS sector. The primary reason for this is the specialized nature of the EHS and the complex issues that come with it. The slow adoption is also due to a lack of specialized LLMs for this field that can easily understand the complex terminologies, regulatory frameworks, and compliance requirements. Such a model is also important as EHS requires a very high level of accuracy and reliability

Standard LLMs are trained on very broad and diverse datasets that span a variety of domains each with their own specialized and general linguistic requirements. While this is helpful when dealing with a wide range of topics these models lack the depth of understanding to interpret nuanced terminology within the EHS field. These models typically fail to accurately interpret the information when they are presented with the same sounding general phrases or words they can't differentiate but within the EHS context, they are miles apart. For example, terms like "exposure," "containment," and "mitigation" are similar in general usage but have specific interpretations in the EHS domain.

This limitation of general LLM models to effectively manage the linguistic and regulatory jargon limits their direct usage in the EHS domain. These models prove to be inadequate when navigating the convoluted path of compliance monitoring as they struggle to correctly identify and interpret relevant regulations. They also struggle to accurately interpret complex EHS documents due to the lack of domain-specific knowledge. This creates a huge gap between the capabilities of general-purpose LLMs and the actual implementation of these models within the EHS sector. The consequence of this is a missed opportunity to leverage the power of LLMs to streamline and enhance efficiency and accuracy in EHS management.

To bridge this gap and unlock the full potential of LLMs in EHS we focus on fine-tuning these models on EHS-specific datasets. The dataset consisted of regulatory documents, compliance guidelines, incident reports, and questionnaires collected from EHS personnel. The main aim is to compare the fine-tuned models to showcase how a specialized fine-tuned model can be used to accurately interpret complex EHS regulatory frameworks. This tuned model can help automate the regulatory compliance process, extract key information from complex EHS phrases, and facilitate an accurate way to report incidents. This will pave the way for a future where this can be used as a tool to improve management and operational efficiency, leading to a safer workplace and more sustainable environmental practices across industries.

1.4 Research Objectives

The research achieves the following objectives:

- Employ Parameter-Efficient Fine-Tuning (PEFT) techniques such as Low-Rank Adaptation (LoRA)[1] and Supervised Fine-Tuning (SFT) to fine-tune three open-source Large Language Models (LLMs)
- Construct a custom dataset for Environment, Health, and Safety (EHS) by scraping regulatory websites, manually collecting data from various EHS books and anonymous answers from EHS professionals in the industry
- Compare the performance of different LLMs on various performance metrics to determine the most efficient model and provide a foundation for future work
- Test the fine-tuned models on various real-world EHS scenarios to demonstrate the ability to employ fine-tuned LLMs in the EHS domain.

By achieving these objectives, the study aims to bridge the gap between theoretical possibility and actual implementation of fine-tuned models in the EHS domain. The objective of this study is to transform the field of EHS management and open doors to more usage of such models to improve the practices in the compliance and safety monitoring domains.

CHAPTER 2

Related Works

2.1 Large Language Models

Large Language Models (LLMs) have made remarkable strides in the area of natural language processing and artificial intelligence. The below background work portrays various types of models, such as GPT4, Palm2, Claude, Cohere, Falcon, and Meta LLaMA and similarly it encompasses a variety of models. An LLM can simply be defined as, "a deep neural network model which has been trained on large amounts of data, such as books, code, articles, and websites, to learn the underlying patterns and relationships in the language that it was trained for. By doing so, the proposed model is able to generate unique and particularly correct content such as grammatically correct sentences and paragraphs that closely resembles human language or code that is syntactically correct" [2].

Author Ipek Ozkaya[2], explains that LLM are generally correct regarding the generation of generating grammatically correct sentences but that may not be accurate all the time. She explains that, "The probabilistic and randomized selection of the "next token" in constructing the outputs on one hand gives the end user the impressions of correctness and style, on the other hand may result in mistakes" [2].

Author Jiaxi Cui et al.,[3] explains in the proposed paper that have shown potential to change or revolutionize the NLP (Natural Language Processing) work. Various GPT's that is the Generative Pre-trained transformers in multiple fields are making strides such BloombergGPT and FinGPT.

In the proposed system the authors have introduced, "a method that combines vector database retrieval with keyword retrieval to effectively reduce the inaccuracy of relying solely on vector database retrieval" [3]. In the proposed approach the authors have made a special attention to enhance the capability of LLM and overcome any shortcomings and errors that are present in the current systems they have tried to, “further optime the issue of model hallucinations at the model level and improve the problem-solving capabilities of large models” [3].

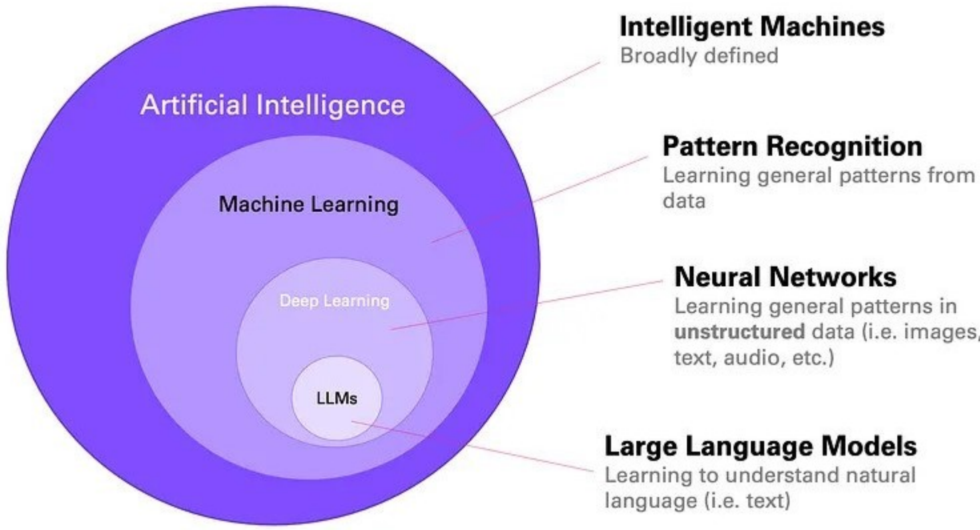

Figure 2: Large Learning Models with respect to AI [4]

2.2 Pre-trained Models with Domain-specific Fine-tuning

Author Yi Yang et al.,[5] explains a LLM specifically fine tuned for financial domain, the author emphasizes that, “that a high-quality domain specific LLM can be tuned using a small set of carefully verified instructions on a well-trained foundation model, that is consistent with the Superficial Alignment Hypothesis” [5]. The authors explains that it is built upon LLaMA-65B and fine-tuned with a diverse set of financial instructions, aiming to provide helpful investment-related response. The authors also emphasize on various research insights for tuning the model.

Zhixiao Qi et al.,[6] explains that incremental pre-training model is used to integrate domain specific knowledge from scanned documents which is not usually used or included in the base LLM's. The author mentions, "Some models also incorporate knowledge bases without the need for pre-training" [6]. This can be due to the base already contains the domain specific knowledge. Hence, "the authors built a model or specifically a large language model for food testing. where a significant amount of data in the domain exists in Scanning format for domain standard documents" [6].

Author Ou Zheng et al.,[7] explains how LLM can be used in transportation and its safety. The authors put the emphasis on how LLM are fine tuned in every sector but transportation sector is not well equipped with these models. To address this challenge, the authors, "introduce TrafficSafetyGPT, a novel LLAMA-based model, which has undergone supervised fine-tuning using TrafficSafety-2K dataset which has human labels from government produced guiding books and ChatGPT-generated instruction-output pairs" [7]. The authors explain how the TrafficSafetyGPT outperforms the LLM which are not specialized. This model also claims to produce almost accurate traffic safety responses. This is achieved by fine tuning the model with TrafficSafety-2K dataset. The performance enhancement can also be credited due to its specialized focus tuning and therefore a increase in efficiency is observed in the response.

Author Yuhao Dan et al.,[8] explains how LLM can be used in education and how a personalized experience can be provided to students, teachers and teaching professionals to enhance the learnings. The authors focus on, "strengthening educational functions such as open question answering, essay assessment, Socratic teaching, and emotional support based on the existing basic LLMs" [8]. The authors achieve this by fine tuning and pre-training the corpus of educational materials by doing so they propose a system where education can be thought in an intelligent way by the working professionals to the students for education.

Author Le Xiao et al.,[9] proposes a system where LLM can be used in the generation and creation of news using evolutionary fine tuning. The authors propose, “a new way or process for news summary generation as well as news curation using LLM with powerful natural language understanding and generative capabilities. They use LLM to extract multiple structured event patterns from the events contained in news paragraphs, evolve the event pattern population with genetic algorithm, and select the most adaptive event pattern to input into the LLM to generate news summaries” [9]. The proposed approach signifies the importance of fine tuning for generation of news using LLM. Authors experimental report shows the improvement of news generation and improvement in quality of news summary reliability.

Author Teodoro Baldazzi et al,[10] explains how the fine tuning a model does not always yield a proper result and generally tuning is used to diversify goals in LLM. They explain, “Task specificity should go hand in hand with domain orientation” [10]. They also emphasize on the fact that data that is publicly available or the data that is available over the web is fine-tuned which clearly ignores the domain experience and business level definitions. The authors propose, “A novel neurosymbolic architecture that leverages the power of ontological reasoning to build task- and domain-specific corpora for LLM fine-tuning” [10].

Authors Ishika Singh et al.,[11] explains in their proposed paper the importance of LLM in creation of task planning and how the robot can be used to automate the next task and enhance its performance using the large language model. The authors propose, “a programmatic LLM prompt structure that enables plan generation functional across situated environments, robot capabilities, and tasks” [11]. The authors aim to provide prompt like instructions to the robot to perform a particular task. The identification of and an object to perform the desired task is achieved via the LLM model. The proposed system is achieved through structure constrains and generations constrains for the model to work in a desired manner.

Author Shengcheng Yu et al., [12] explains in their proposed approach, the use of LLM in the testing of software application and their importance. The authors explain in their findings that LLM can be useful in testing where migration of software takes place or the development takes place in diverse platforms. Running automation test cases through LLM can be very useful and provide a great insight after the test report are generated. Automation testing can be achieved with great efficiency when processed trough LLM. The authors also propose, "the application of LLMs in cross-app migration, where it generates test scripts across different applications and software environments based on existing scripts" [12].

Authors Weirui Kuang et al.,[13] explains in the proposed system that streamlining the processing of dataset and fine-tuning it resulted in performance increment and the authors also proposed of implementing the fine-tuning algorithms in programming the interfaces so that it can result in increased performance results. The main the of the proposed approach is that Federated Learning or (FL) in LLM can yield promising results and also help in research and development it also significantly reduces the risk of compliance risk. So therefore, fine-tuning LLM with module specific or task specific activity enhance the efficiency notably.

Authors Zhixin Lai et al.,[14] proposes a system of fine-tuning LLM for the detection of fake text generation. The authors explain the importance of implementing and enhancing the use LLM in text generation to counter fake news and its potential risks related to it. They implemented this by testing five different transformer-based model in both in-distribution as well as out-distribution model. They claimed it better helped them to understand their performance and also increase their efficiency. The LLM based text generated can cause implication and can be used for immoral deeds. The authors achieve this by, "combining the individual classifiers models using adaptive ensemble algorithms, which improved the average accuracy significantly from 91.8% to 99.2% on an in-distribution test set and from 62.9% to 72.5% on an out-of-distribution test set" [14]. The results not only show the increase in accuracy but also significantly increases the test result outcome.

Author Yaqi Zhang et al.,[15] presents a unique fine-tuned motion generator (MotionGPT) or simply Motion pre-trained generator. The main aim of the proposed system is to generate motion answers. The proposed system claims, "human motion generation model with multimodal control signals by tuning a mere 0.4% of LLM parameters" [15]. The system has unique text-to-motion generation ability where it can solve tasks using a unified model. The authors fine tune the LLM with Low Rank Adaptation or (LoRA) to generate the motion of the texts provided. The decoding of the motion of humans are achieved by "Vector Quantized-Variational Autoencoder (VQ-VAE)". The proposed system achieves an almost identical text-to-motion human output by tuning a mere 0.4% parameter of LLM, and also claims to be the first multimodal control signal to generate human motions.

Authors SeungHeon Doh et al.,[16] proposes a fine-tuned system for text-to-music retrieval. The proposed approach aims to retrieve music tracks based on their genre and keyword searched. To address this concerns the authors specifically utilized various music datasets and which contained the music tag and caption and a knowledge graph of various artists. They achieved this by, "providing an improved Text-to-Music Retrieval model, denoted as TTMR++, which utilizes rich text descriptions generated with a finetuned large language model and metadata" [16]. This aims to retrieve the exact music based on the descriptive query provided and fine-tuning the LLM to achieve this results in enhanced and accurate results.

CHAPTER 3

Dataset

3.1 Data Collection

Specialized data which encompasses all the nuances and complexities of the EHS is critical when fine-tuning general LLMs for EHS-specific tasks. As there was no existing dataset available to fine-tune the LLM model we had to make the dataset from scratch. This dataset is critical in providing accurate insights into the complex EHS standard. We employed different strategies to curate data so that we could encompass a variety of different topics without any bias.

The first step we took in collecting the dataset was to identify the data sources from where we could get correct and accurate information in the EHS domain. After going through different government regulatory agencies website we narrowed our focus on Occupational Safety and Health Administration (OSHA) website [17]. We selected this particular regulatory agency due to its exhaustive and detailed resources on EHS regulations, frameworks, and preventive measures.

The next step in the process was to identify sources of information that are relevant to fine-tuning our general-purpose LLM models. To get the ground truth and domain-specific knowledge we scrapped the recordkeeping page of OSHA [18] and created a manually labeled dataset from the OSHA Field Safety and Health Manual [19]. This department guidebook and record-keeping encompasses all the accurate, complex, and nuanced information on various regulations, preventive measures, and policies within EHS. Additionally, these resources also have examples of real-world scenarios which is particularly useful in giving context to the final tuned LLM model. Furthermore, the information contained in the book also gives a holistic view of the complex jargon, guidelines, and codes specific to EHS. Consequently, when LLMs are trained on such data they will be able to offer more insights and recommendations for the government codes and regulations in the EHS domain.

OSHA recordkeeping FAQ Website [18]: The key component of this website is that it contains various real-life examples and scenarios that might be possible within the EHS domain. It is a standard FAQ where a user can go and search from all the available entries but the main issue comes with the accurate searching and the technical jargon in the answers. This data also contains a variety of user cases with age, and other information to accurately mimic a real-life scenario.

The key component of this dataset is that it is already present in a question-answer format which is adept when fine-tuning LLMs. Thus, by using this dataset we save time and effort on cleaning and pre-processing the dataset to get it ready for fine-tuning. Additionally, exposure to real-life use cases and applications will help our final LLM model to understand user perspective and context to give an output in a simple lucid manner that can be understood by all the stakeholders in the project. Lastly, this website covers various aspects of incident reporting, injuries, and regulatory information which is a very valuable resource for answering queries for specific and nuances scenarios.

Fig. 3 shows a snapshot of the OSHA recordkeeping FAQ highlighting different real-world scenarios. This ensures that the final tuned model is not only able to handle procedural guidelines and regulatory frameworks but also a variety of case studies and best practices in EHS management.

OSHA Field Safety and Health Manual [19]: This book is published by the government regulatory agency OSHA and it contains a detailed description of the various procedures to be followed in case of an incident, reporting of incidents and further precautions to prevent any such incident in the future. This manual acts as

Q: **A company has been notified that an employee may have been exposed to TB while overseas while working. Is this a recordable case?**

A: No. Employers are only required to record injuries and illnesses if they occur within the geographic coverage of the OSH Act. The Occupational Safety and Health Act, and therefore the OSHA Recordkeeping Regulation, apply only within the jurisdictional boundaries of the United States as defined in Section 4(a) of the Act. If the exposure had occurred in the United States, the case would be recordable if that employee subsequently developed a tuberculosis infection, as evidenced by a positive skin test or diagnosis by a physician or other licensed health care professional.

Source: OSHA e-correspondence FAQ ID: 578

Q: **A company is located with multiple buildings within one complex. Can I produce one log for the complex or do I need to have one log for each building?**

A: You may complete one log for the complex. Generally, there should be one log per "establishment". An establishment is defined as a single physical location and can include campus and complex type locations.

Source: OSHA e-correspondence FAQ ID: 142

Q: **A corporation pays a licensed professional to offer massage on-site as an employee benefit. The employee chooses by their own to see the on-site professional. Because the work the employee does is physical, he benefits from the soft tissue therapy and uses the service as needed. When should an employee's condition be categorized as recordable upon receipt of the massage?**

A: Massage therapy is considered first aid and is included as item M in the rule's first aid list. Message therapy does not make a case recordable.

Source: OSHA e-correspondence FAQ ID: 322

Q: **A doctor places an employee on restricted duty but the employer can not accommodate the restrictions and, as a result, the employee is losing time. Are the lost days recorded as lost time or restricted duty?**

A: You must count the days as days away from work. To count days as restricted days, restricted work activity must be made available to the employee.

Source: OSHA e-correspondence FAQ ID: 201

Q: **A dust particle is removed from the eye using a cotton swab, but prescription eye drops are given to prevent the possibility of infection. Do the eye drops make the case recordable even though they were only precautionary?**

A: Yes, the case is recordable. Any use of RX medicine in treating a work-related injury or illness is considered medical treatment regardless of purpose prescribed.

Source: OSHA e-correspondence FAQ ID: 171

Q: **A health care worker was exposed to pertussis and was given an antibiotic by a medical provider as a preventative measure. We know that the mere exposure is not recordable. However, does the fact that the antibiotic was given make it recordable?**

A: If an employee is exposed to pertussis (or other communicable diseases) and does NOT exhibit any symptoms of the disease, the administration of prophylactic antibiotics does not make the case recordable. Please note if the employee shows ANY signs or symptoms of the disease and is given antibiotics, then the case is recordable.

Source: OSHA e-correspondence FAQ ID: 545

Figure 3: Snapshot of OSHA recordkeeping FAQ webpage

an official guideline when dealing with hazardous chemicals, safety procedures to be followed, emergency procedures to be followed in case of spill or incident, and step-by-step methods to be followed to dispose of such harmful chemicals. Additionally, it also contains information about the legal framework, federal and state regulations, rights of the employees, and the responsibilities of the employers for the safety of workers.

The information provided in this manual is crucial to understanding the vocabulary and terminology used in the EHS sector. The data extracted from this book will help the final fine-tuned model better understand the various jargon used in the industry. It will also provide a contextual understanding of the question being asked to the model. Moreover, it also contains information covering various real-world scenarios which will help the model learn and answer similar questions with accuracy.

The manual is divided into several chapters and sections which encompass all the scenarios in an orderly fashion. It also contains appendices that give further detail on the procedure for logging incidents, conducting inspections, and enforcing the rules and regulations.

However, books only have limited knowledge and may not consider all of the real-life situations. This might prove to be insufficient when tuning a general LLM model. To make sure that our final model can understand human queries and can answer them in such a way that is being understood by humans we need to add some data from humans as well. For this, we shared a Google form anonymously with EHS professionals in the industry to get their feedback. This ensured that the final dataset contained data not only from the government regulatory agencies but also from people who face issues and maybe the end users of such a system. The reason for adding human-generated data was to prevent any bias in the system and add nuances that may not be covered with only textual data. By simulating real-world questions the model will learn and understand specific queries that EHS personnel face in their day-to-day activities, enhancing its pragmatic application. Moreover, using information directly from government sources ensures that the model's response will always be in line with the legal framework laid out by the various agencies.

3.2 Dataset Construction

The dataset is the most vital piece of the puzzle when fine-tuning an LLM model. To ensure that we have the correct data that covers all the scenarios and aspects of the EHS domain we have used a variety of sources to collect the data. Once we have finalized the data source, extracting that data poses another challenge. Extracting and constructing data manually is tedious and time-consuming, it may also lead to incorrect data being added to the dataset which might hamper the performance of the model. To automate this task we employ the method of scraping data from the website.

For scraping the OSHA FAQ webpage we wrote a custom logic in Python, manually copy-pasting the data is humanly impossible. The first step in this process was to analyze the website DOM layout and figure out an algorithm to accurately and efficiently scrape all the data. Once we finalized the algorithm and carved out a common logic to get data from the DOM elements we then proceeded to look for an inbuilt library that would help us achieve this task. We choose Beautiful Soup due to its inbuilt support for parsing HTML and XML documents.

The next method we employed was getting and labeling data from the OSHA handbook. This handbook covers a variety of different aspects of EHS scenarios which is critical for a well-balanced dataset. To ensure that we extract all the different use cases from this journal, we leveraged the power of the AI PDF tool in the ChatGPT store. This tool not only helped us capture a plethora of complex real-world use cases but also gave the response in the question-answer format, thus saving us the overhead of actually cleaning and modeling data in a proper format. Apart from using this tool, we also manually labeled data from this book, to ensure that we incorporate topics, scenarios, jargon, codes, and other intricate details that this tool may have missed.

Lastly, we contacted EHS professionals in the industry to get their queries, and questions that they face in their day-to-day activities. We then collected the questions and answers responses in different formats and consolidated them into one CSV. The main motivation and reason for employing this method was to ensure that the tuned model understands the complexities and nuances of real-world problems plaguing the EHS personnel.

Let's discuss each method in detail and look into the exact methodology, tools used, and how the data is consolidated.

Method 1: Web Scraping

Tools and Technology Used: Python, Chromium, Beautiful Soup, a Python library for parsing HTML and XML documents.

Process:

In this method, we first set up the Python environment and installed the Beautiful Soup. As shown in Fig. 4 we wrote a custom logic to scrape the DOM elements and using Beautiful Soup we systematically navigated and scrapped data from each page of the website. We also identified the relevant sections and the URL from where we needed to scrape the data. We then executed this Python script to fetch the data and then transform them into a question-and-answer pair to finally store them in a CSV file with two columns.

```python
import requests
from bs4 import BeautifulSoup
import csv
import time

def scrape_page(url, writer):
    headers = {
        'User-Agent': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/96.0.4664.110 Safari/537.36'
    }

    response = requests.get(url, headers=headers)

    if response.status_code == 200:
        soup = BeautifulSoup(response.text, 'html.parser')
        question_answers = soup.select('div.recordkeeping-qna')

        # Loop through all available questions and answers
        for qa in question_answers:
            question = qa.select_one('div.views-field.views-field-title strong').get_text(strip=True)
            answer = qa.select_one('div.views-field.views-field-body p').get_text(strip=True)

            # Write the data to CSV
            writer.writerow([question, answer])

        # Find the URL of the next page if available
        next_page_link = soup.select_one('li.page-item.pager__item.pager__item--next a')
        if next_page_link:
            page_number = next_page_link['href']
            next_page_url = "https://www.osha.gov/recordkeeping/faq-search" + page_number
            time.sleep(3)  # Adding a delay to respect website limitations
            scrape_page(next_page_url, writer)

# Start scraping from the initial page
initial_url = 'https://www.osha.gov/recordkeeping/faq-search?page=0'

# Open a CSV file to write the data
csv_filename = 'qa_data.csv'
with open(csv_filename, 'w', newline='', encoding='utf-8') as csvfile:
    csv_writer = csv.writer(csvfile)
    csv_writer.writerow(['Question', 'Answer'])  # Write header

    scrape_page(initial_url, csv_writer)

print(f"Scraping complete. Data saved to {csv_filename}.")
```

Figure 4: Function to scrape webpage

Method 2: Using AI Tool and Manual Labelling

Tool and Technology Used: AI PDF tool

Process:

To ensure that we cover every aspect of the data in the government-issued manual, we leveraged the power of the AI PDF tool with the GPT store to help us analyze the PDF and give us the relevant question-answer pairs. To further strengthen our dataset we also manually labeled the data directly from the journal, more specifically the appendix which contains the complex jargon, code, and other definitions of phrases used in the EHS world. We consolidated the data into a single CSV file that we will use for further processing and fine-tuning of our LLM model. Fig. 5 shows the step-by-step process of how we generated the CSV file for tuning our LLM.

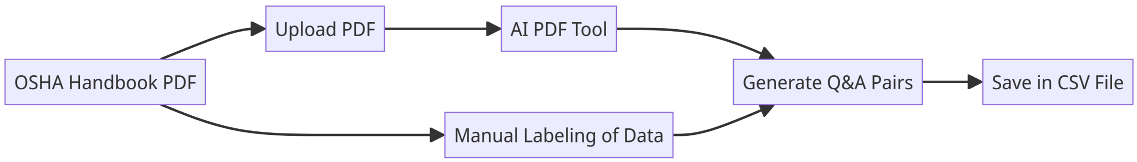

Figure 5: Steps to generate data using Manual Labeling and AI PDF Tool

Method 3: Collecting data from EHS personnel

Tools and Technology Used: Google Docs, Google Forms, Python

Process:

For the model to understand the daily linguistic language of the EHS personnel it is important to include data from the human perspective as well. For this, we reached out to EHS professionals in the industry to collect questionnaires they encounter in their day-to-day work. We took surveys and collected the data in a common Google Doc file. We also sent out Google Forms to collect answers to some of the questions. After collecting the raw data we wrote a simple Python program to consolidate all the data in a single CSV file.

3.3 Data Pre-processing

Large Language Models require data in a particular format for fine-tuning. After we have collected the data we need to pre-process it so that it is in the correct format. In the data collection steps, we have cumulated the data in CSV or text format using different methods. Almost all of our data is in the question-answer pair, we wrote a Python script to cumulate the data into one main CSV file.

After consolidating all the data in one CSV file with two columns namely question and answer, we wrote the code to get the data in the format required for fine-tuning. In this code, we wrote the logic to read the CSV file containing the scraped, manually labeled, and AI PDF-generated data to convert it into a dataset of prompt and response. Each prompt represented the corresponding question in the CSV file and the response is the ground truth. The response not only makes the model more adept at answering real-world and nuanced questions related to EHS but also helps the model learn exactly in which format to answer the queries or questions when prompted after the fine-tuning is done.

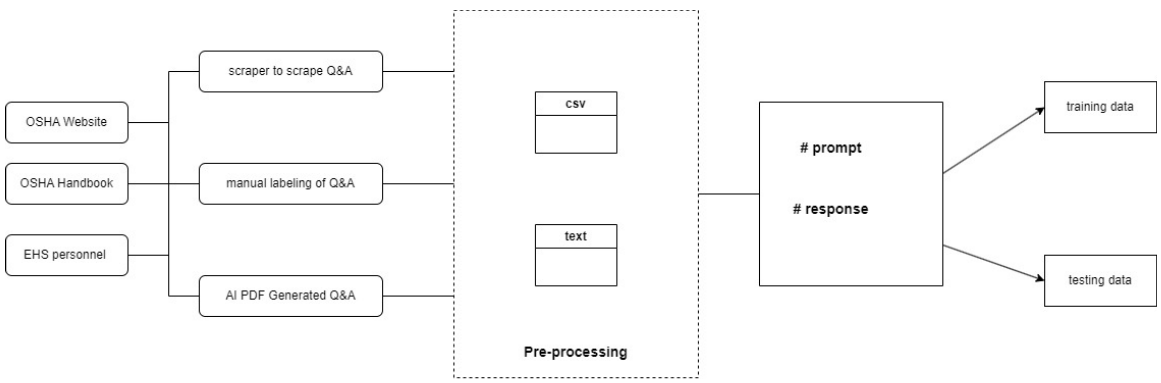

Figure 6: Steps to generate data using Manual Labeling and AI PDF Tool

Fig. 6 shows us exactly how we curated the data from different sources namely the OSHA website, OSHA handbook, and data from EHS personnel. We then employed different methodologies to get this data from the source like manually labeling the data, scraping the data, and using AI-PDF to generate the question-answer pairs. After this, we store the data either in CSV or text format; and then consolidate it in a common CSV file. Thereafter, we write the code to transform it into the format required for fine-tuning and then split it into training and testing sets.

After the data is collected, transformed, pre-processed, and then split into training and testing sets we can use this to fine-tune our base or foundation models using the PEFT tuning techniques namely LoRA, QLora, and SFT. The only change when getting the data ready for tuning would be the format in which the string for the prompt is generated. Thus, all the models will use the same data but how they are prompted would be different.

CHAPTER 4

Methodology

In this section, we will discuss the foundation models being used for fine-tuning. Additionally, we will look into the fine-tuning techniques used to build our final model, particularly Parameter-Efficient Fine-Tuning (PEFT), and the overall architecture of the system.

4.1 Foundation Models

Large Language Models (LLMs) can understand the context and can respond in a human-like way because they are trained on a huge amount of data. They can generate text and answer questions by employing a technique known as next-token prediction. When we fine-tune a model based on a dataset of a particular domain (for this study we are using EHS data) they can predict tokens in a way that aligns with the characteristics of the dataset that they were fine-tuned on. This makes the model effective in handling specialized tasks and responding to queries specific to the domain. Moreover, this also ensures that the model accurately generates texts for particular nuance scenarios.

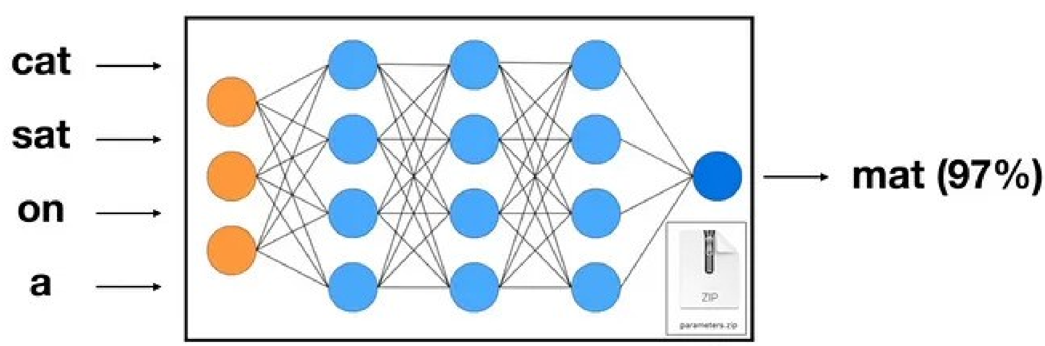

Figure 7: Next Token Prediction [20]

For this study, we choose three foundation models namely Llama 2 7b, Mistral

7b, and Falcon 7b. As our dataset consists of question-answer pairs and we are testing our models based on different EHS scenarios we used a specialized fine-tuned model of these foundation models. The reason for the selection of the 7b variant of these models was the limited computing power and less time required for fine-tuning.

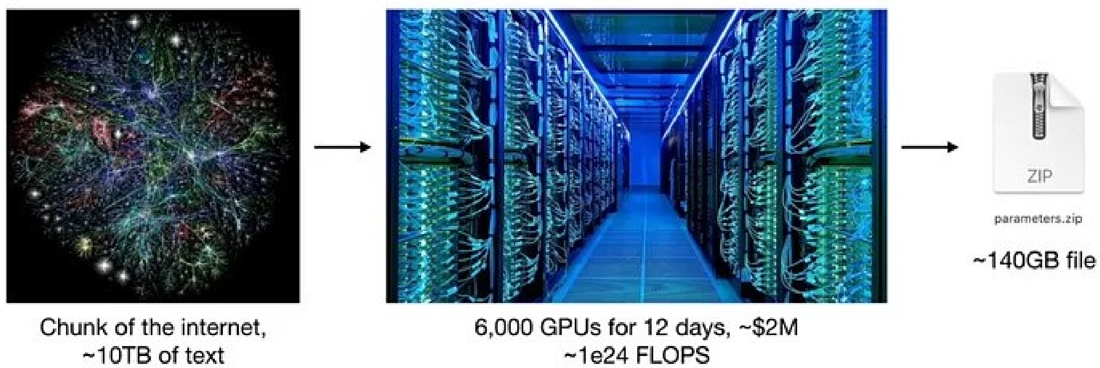

Figure 8: Llama 2 70b [20]

- **Llama 2 7b (Llama-2-7b-chat-hf)**

This model is fine-tuned to handle chats in the English language. As the name suggests this model is trained on 7 billion parameters. For fine-tuning this model to need to format the dataset to include INST and <<SYS>> tags. After the model is fine-tuned we prompt it using the same format to get the domain-specific result.

- **Mistral 7b (Mistral-7B-Instruct-v0.1)**

This is a specialized model on top of Mistral-7B-v0.1 for conversation. It is fine-tuned on a variety of publicly available conversation datasets giving it contextual awareness of everyday language. To leverage fine-tuning in this model we need to construct the prompt surrounded by [INST] and [/INST] tokens.

- **Falcon 7b (falcon-7b-instruct)**

This is the instruct model of Falcon 7b intended to be used in conversational scenarios.

We use this model so that it can better respond to the questions presented by the EHS professionals.

After we finalize what foundation models we will be using, we can then move on to fine-tuning these foundation models.

4.2 Fine-tuning

Fine-tuning is a process through which a general-purpose Large Language Model (LLM) can be modified for a specific task. In other words, we can enhance the understanding of the model for a specific domain. LLMs are trained on a wide range of data and may lack information about specific information or tasks, they are more suited for a broad range of tasks and basic contextual relationships. For our study, we focus on making three LLM models focus on EHS tasks and terminologies. Thus, by fine-tuning these models we can tailor them to resolve and answer queries and questions related to the EHS domain.

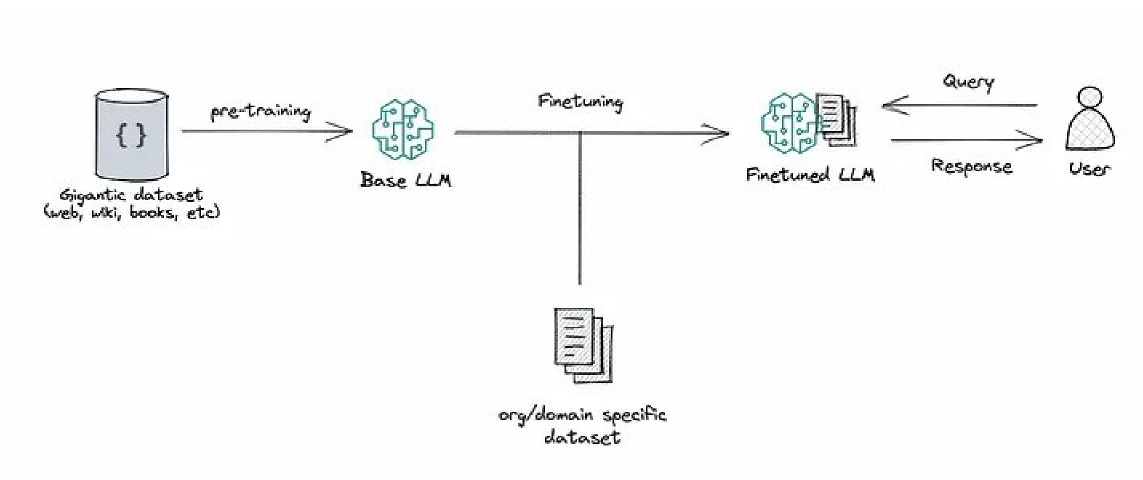

Figure 9: Fine-Tuning an LLM model [21]

In Fig. 9 can see a high-level overview of how fine-tuning works. First, the base LLM is trained on a gigantic dataset that contains a plethora of general data from technical documents, code, and reports, to internet dialogues and scientific manuals.

This helps the LLM model get a holistic view of things and helps it generate coherent text when it is prompted. The primary goal at this stage is to have a model that understands fundamental aspects of the language and responds accordingly. Now from Fig. 9, we can see that if we want it to respond to queries and questionnaires specific to a domain in our case the EHS domain we need to fine-tune it based on a domain-specific dataset. In this study, we employ PEFT and SFT fine-tuning techniques to customize the base LLM for EHS-specific queries. Once the fine-tuning is completed the user in our case an EHS personnel can query this fine-tuned LLM to get an EHS-relevant response.

4.2.1 Parameter-Efficient Fine-Tuning (PEFT)

Parameter-Efficient Fine-Tuning (PEFT) is a method used to adapt an existing pre-trained, for a particular domain or task. The main advantage of using PEFT techniques is that the model need not be re-trained for the given specialized scenario. PEFT leverages the idea of modifying a small number of model parameters, saving a lot of computational time and resources. This method is particularly useful when dealing with Large Language Models (LLMs), as re-training and full fine-tuning of such models requires a mammoth amount of computational cost and storage.

In this study, we make use of a PEFT technique known as QLoRA a slight modification or extension of LoRA. When we use PEFT methods, only a few MBs of storage are required for each downstream dataset, contrary to full-tuning when around 40GB of data storage is required for each downstream dataset. Although, a fraction of data storage is required when compared to full-tuning the models that are fine-tuned using PEFT techniques have similar accuracy.

4.2.1.1 Low-Rank Adaptation (LoRA)

LoRA or Low-Rank Adaptation of Large Language Models is a PEFT technique to train LLMs efficiency for a specific domain. When we use LoRA we change the underlying parameters of the model. The main technique that LoRA used is to introduce a low-rank decomposition matrix into each layer of the model [1]. Fig. 10 shows how this technique injects a trainable rank decomposition matrix into each layer of the transformer. This technique significantly reduces trainable parameters for downstream tasks.

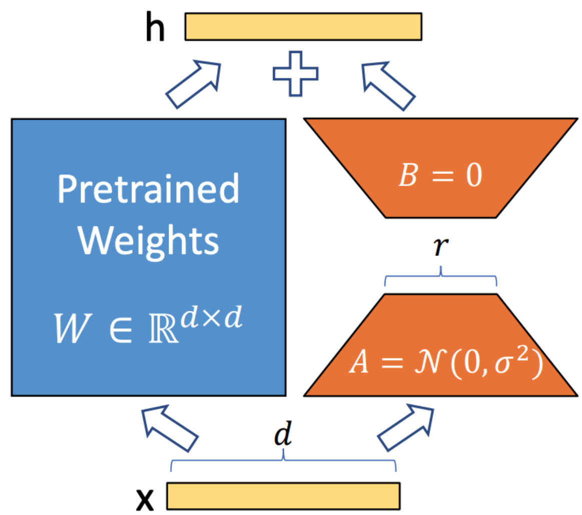

Figure 10: LoRA Fine-tuning [1]

To reduce the computational cost rather than modifying the weight matrix W in each layer, LoRA creates two smaller matrices, A and B whose matrix product roughly has the same dimension as the modification to W. As shown in Fig. 10 if W is a dxd matrix then A is a dxr matrix and B is a rxd matrix where r is rank and is much smaller than d. Using this method we need not modify the large matrix W instead we compute the values of A and B; and take a linear project and add it to the model's pre-trained weights. The main advantage of using the rank-decomposition matrix is that we have to train a minimum number of parameters.

$$W_{\mathtt{ft}} = W_{\mathtt{pt}} + \underbrace{\Delta W = W_{\mathtt{pt}} + \overbrace{AB}^{\text{Rank Decomposition Matrix}}}_{\text{Approximation}}$$

$$\text{where} \quad W_{\mathtt{ft}}, W_{\mathtt{pt}}, \Delta W, AB \in \mathbb{R}^{d \times d}$$

$$\text{and} \quad \underbrace{A \in \mathbb{R}^{d \times r}, B \in \mathbb{R}^{r \times d}}_{\text{Low Rank}}$$

Figure 11: LoRA uses Rank Decomposition Matrix during fine-tuning [22]

Using LoRA we ensure that we need to update very few trainable parameters, this significantly reduces the computation cost and also makes the fine-tuning much faster.

4.2.1.2 QLoRA

Quantized Low-Rank Adaptation or QLoRA is one of the most popular extensions of the LoRA. At a very high level, QLoRA uses model quantization during fine-tuning to reduce memory requirements. More specifically, QLoRA works by quantizing the weights of the pre-trained model to a 4-bit format which is typically stored in a 32-bit format[23]. This method significantly reduced the memory footprint required to tune LLM. Using QLoRA it is even possible to fine-tune models on a single GPU. In hindsight, QLoRA reduces the training speed to save memory. This reduces the memory requirement, making it practical to tune models even on less powerful hardware.

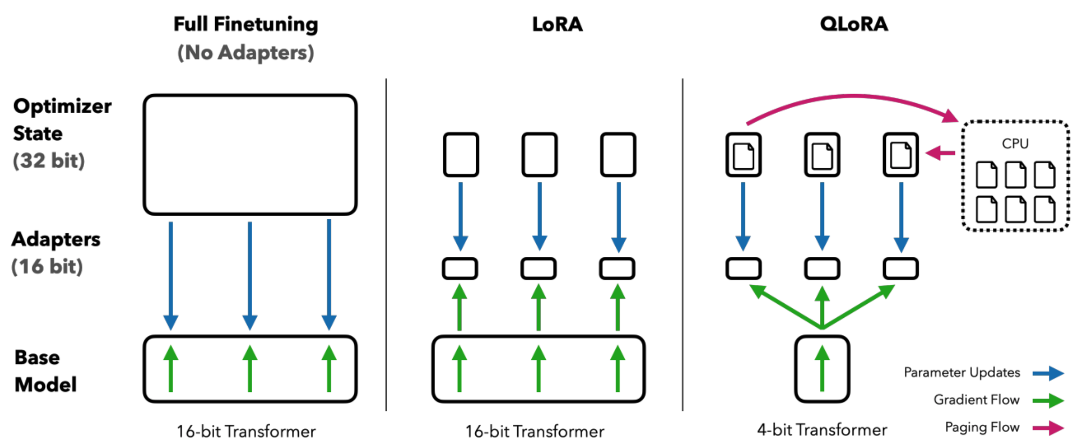

Figure 12: Finetuning vs LoRA vs QLoRA [23]

4.2.2 Supervised fine-tuning (SFT)

Supervised fine-tuning, involves adapting a pre-trained Language Model (LLM) to a specific downstream task using labeled data [24]. In supervised fine-tuning, the finetuning data is collected from a set of responses validated beforehand. That's the main difference to unsupervised techniques, where data is not validated beforehand. While LLM training is (usually) unsupervised, Finetuning is (usually) supervised.

During supervised fine-tuning, the pre-trained LLM is fine-tuned on this labeled dataset using supervised learning techniques. The model's weights are adjusted based on the gradients derived from the task-specific loss, which measures the difference between the LLM's predictions and the ground truth labels.

The supervised fine-tuning process allows the model to learn task-specific patterns and nuances present in the labeled data. By adapting its parameters according to the

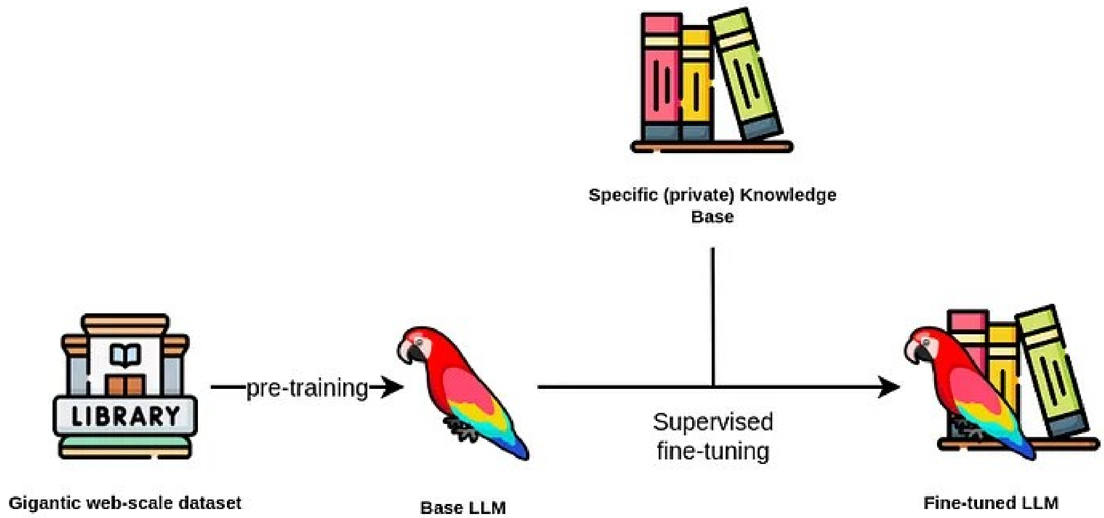

Figure 13: Supervised Fine-tuning on LLM [24]

specific data distribution and task requirements, the model becomes specialized in performing well on the target task.

4.3 Architecture

In this study, we have fine-tuned three foundation models on the custom-curated EHS dataset. After we have collected data from various sources and divided it into training and testing sets, we are ready to start with the fine-tuning process. In Fig. 14 we can see that we are feeding the foundation model the tuning data and the fine-tuning technique to adapt it so that it can answer domain-specific queries - for fine-tuning we have used QLoRA and SFT. The advantage of using these PEFT techniques is the low computational cost and time.

We started with loading the dataset which was stored in the Google drive, after mounting the drive and loading the training data. We then proceeded to make sure that the data was formatted in the required format for fine-tuning. Different LLM models require different prompt formats for fine-tuning. The base model Mistral-7B-Instruct-v0.1 requires the prompt surrounded by [INST] and [/INST] tokens whereas

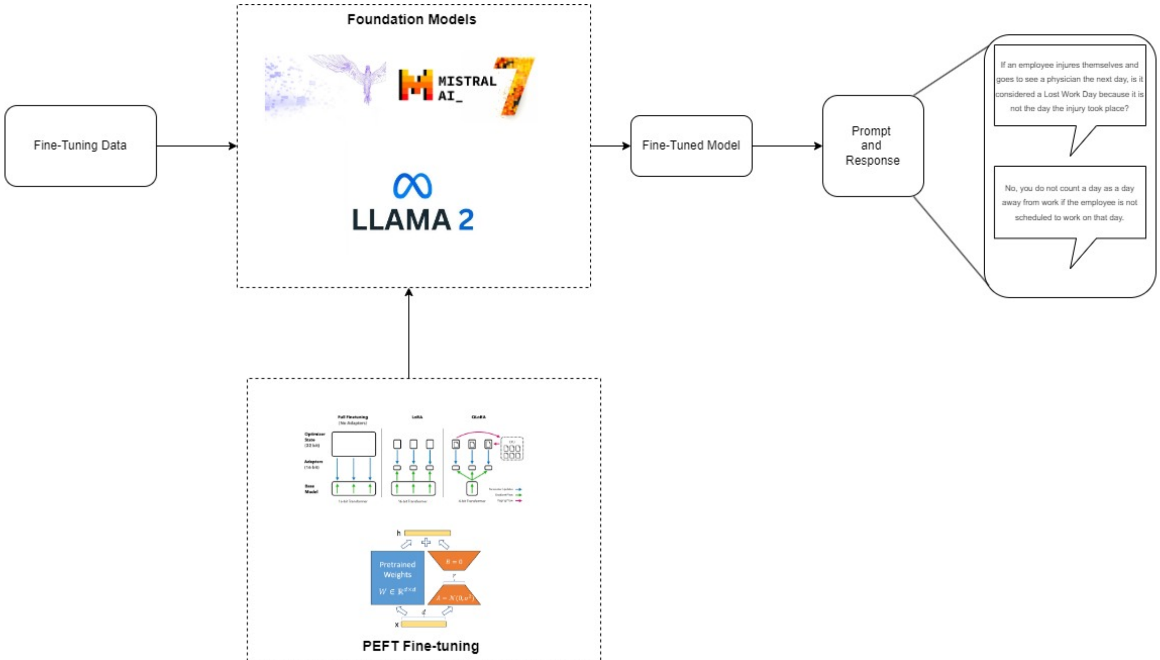

Figure 14: Architecture diagram for fine-tuning Llama 7b, Mistral 7b and Falcon 7b

Llama-2-7b-chat-hf requires the dataset to include INST and <<SYS>> tags. After this, we load the base model and set up the bitsandbytes config; once this is done we then load the tokenizer and set up the LoRA config. After this step, we set up the LoRA config and load the trainer. We then pass the arguments to the SFT trainer and start the training. Thereafter, we save the model and push it to Hub, now we can infer the model to give the fine-tuned response.

We have the same architecture and steps for fine-tuning for all three foundation models used in this study. By employing PEFT techniques we reduce the computation demand or memory usage. We used Google Colab's V100 GPU for fine-tuning all our models and saved the model to Hugging Face Hub. After the fine-tuning was done we curated a list of questions covering a wide array of topics within EHS for the testing. We then employed different evaluation metrics to compare the performance of all three models. Apart from this we anonymously surveyed EHS personnel on the model's response to mitigate any bias in the metric evaluation. This also ensured that the tuned model could further be extended in real-world applications.

CHAPTER 5

Experimental Results

To evaluate the results of our fine-tuned model we employed different evaluation metrics namely ROUGE, BERT, and an evaluation by EHS professionals. For covering all the aspects of EHS we picked up different questions covering a variety of different nuances within EHS. Using this testing data we were able to conduct an extensive and exhaustive evaluation of EHS-related tasks.

5.1 Metrics

For converge we selected ROUGE and BERT scores, to ensure that we are not restricted only by scores of standard evaluations we also evaluated the generated text by anonymously surveying EHS personnel and getting their response.

ROUGE (Recall-Oriented Understudy for Gisting Evaluation): is a method or algorithm for evaluating the performance of text summarization as well as machine translation. The way it works is by comparing and evaluating the information coverage of the summary generated by the machine against a set of reference passages usually written by a human. ROUGE is useful as helps us gauge how good the generated text is against the given reference text. ROUGE consists of several metrics, for our study, we have captured ROUGE-1, ROUGE-2, and ROUGE-L.

ROUGE-1: It measures the overlap of unigrams (individual words) between the machine-generated summary and the reference summary. It is a straightforward indicator of how many words the generated summary shares with the reference, ignoring their order. ROUGE-1 is computed in terms of recall, precision, and F1-score.

ROUGE-2: It evaluates the overlap of bigrams (sequences of two consecutive words) between the generated and reference summaries. By considering pairs of adjacent words, ROUGE-2 provides insight into the syntactic structure and order of words, making it a more stringent measure compared to ROUGE-1.

ROUGE-L: This metric measures the longest common subsequence (LCS) between the generated text and the reference texts. A subsequence is a sequence that appears in the same relative order but not necessarily consecutively. ROUGE-L is particularly valuable for evaluating the fluency and the order of the text, as it does not require consecutive word matches but respects the order of words.

ROUGE score is widely used in text summarization tasks as it provides a way to objectively assess the quality of machine-generated summaries compared to reference summaries. It takes into account the overlap of n-grams, which helps in capturing the important content of the summary. ROUGE score is also flexible as it allows the use of different n-gram lengths based on the task requirements.

Significance of ROUGE score:

ROUGE score is widely used in text summarization tasks as it provides a way to objectively assess the quality of machine-generated summaries compared to reference summaries. It takes into account the overlap of n-grams, which helps in capturing the important content of the summary. ROUGE score is also flexible as it allows the use of different n-gram lengths based on the task requirements.

BERTScore: BERTScore is an advanced metric for evaluating text generation, utilizing the contextual embeddings provided by BERT (Bidirectional Encoder Representations from Transformers). Unlike traditional metrics such as BLEU or ROUGE that rely on exact word overlap, BERTScore computes the semantic similarity between words in the generated text and the reference text using the cosine similarity of their BERT embeddings. This method captures the semantic nuances of the language by considering the context in which each word is used, allowing for a more precise evaluation of text quality.

To implement BERTScore, embeddings for each word in both the generated and reference texts are extracted using a pre-trained BERT model. The metric then calculates the cosine similarities between each token in the generated text and each token in the reference text, forming a matrix of similarity scores. The highest similarity scores for each token are used to compute the Precision, Recall, and F1 scores of the generated text relative to the reference text. Precision measures how much of the content in the generated text is relevant to the reference, while Recall assesses how much of the reference content is captured in the generated text. The F1 score combines both to provide a balanced evaluation.

Significance of BERTScore:

The introduction of BERTScore represents a significant advancement in the field of natural language processing, particularly in the evaluation of text generation tasks such as machine translation, summarization, and content generation. BERTScore's method of using semantic similarity instead of relying solely on exact word matches allows it to better assess the quality of text in a way that reflects human judgment more closely. This is crucial in applications where the precise meaning and context of words are more important than their mere presence in the text.

Manual Evaluation: In this case, we manually evaluated the responses generated by each model. We selected a variety of questions covering various aspects of the EHS domain, We then reached out to an EHS professional in the industry to anonymously score the response. This ensures that the response generated by the different fine-tuned models is understood by industry professionals. We employed a method where for each answer the user will give a score between 0-10 where 0 means not all relevant and 10 means that it is the exact answer that the user is looking for.

5.2 Results

We saved the responses generated by each model to a file in a question-answer format. After, saving the results for all the models, we wrote functions to evaluate ROUGE and BERTScore. We then used the data in the saved file to manually evaluate the responses.

5.2.1 ROUGE

We calculate the ROUGE-1, ROUGE-2 and ROUGE-L for all the models and then take an average of all the ROUGE results for each model.

```
rouge scores for mistral
rouge1: Score(precision=0.8333333333333334, recall=0.20202020202020202, fmeasure=0.3252032520325203)
rouge1: Score(precision=0.19117647058823528, recall=0.20634920634920634, fmeasure=0.19847328244274806)
rouge1: Score(precision=0.5526315789473685, recall=0.3888888888888889, fmeasure=0.4565217391304348)
rouge1: Score(precision=1.0, recall=0.34375, fmeasure=0.5116279069767442)
rouge1: Score(precision=0.5806451612903226, recall=0.5454545454545454, fmeasure=0.5625)
rouge1: Score(precision=0.5, recall=0.53125, fmeasure=0.5151515151515151)
rouge1: Score(precision=0.9565217391304348, recall=0.5789473684210527, fmeasure=0.721311475409836)
rouge1: Score(precision=0.4603174603174603, recall=0.6904761904761905, fmeasure=0.5523809523809524)
rouge1: Score(precision=0.43103448275862066, recall=0.6578947368421053, fmeasure=0.5208333333333334)
Score function: calculate_rouge1
Average score: 0.519062771173187

rouge2: Score(precision=0.391304347826087, recall=0.09183673469387756, fmeasure=0.1487603305785124)
rouge2: Score(precision=0.014925373134328358, recall=0.016129032258064516, fmeasure=0.015503875968992248)
rouge2: Score(precision=0.16216216216216217, recall=0.11320754716981132, fmeasure=0.13333333333333336)
rouge2: Score(precision=1.0, recall=0.3225806451612903, fmeasure=0.4878048780487805)
rouge2: Score(precision=0.26666666666666666, recall=0.25, fmeasure=0.2580645161290323)
rouge2: Score(precision=0.21212121212121213, recall=0.22580645161290322, fmeasure=0.21875)
rouge2: Score(precision=0.8636363636363636, recall=0.5135135135135135, fmeasure=0.6440677966101694)
rouge2: Score(precision=0.22580645161290322, recall=0.34146341463414637, fmeasure=0.27184466019417475)
rouge2: Score(precision=0.19298245614035087, recall=0.2972972972972973, fmeasure=0.23404255319148934)
Score function: calculate_rouge2
Average score: 0.29309672643316526

rougeL: Score(precision=0.7083333333333334, recall=0.1717171717171717, fmeasure=0.2764227642276423)
rougeL: Score(precision=0.08823529411764706, recall=0.09523809523809523, fmeasure=0.09160305343511449)
rougeL: Score(precision=0.39473684210526316, recall=0.2777777777777778, fmeasure=0.32608695652173914)
rougeL: Score(precision=1.0, recall=0.34375, fmeasure=0.5116279069767442)
rougeL: Score(precision=0.3548387096774194, recall=0.3333333333333333, fmeasure=0.34375)
rougeL: Score(precision=0.38235294117647056, recall=0.40625, fmeasure=0.393939393939394)
rougeL: Score(precision=0.9565217391304348, recall=0.5789473684210527, fmeasure=0.721311475409836)
rougeL: Score(precision=0.36507936507936506, recall=0.5476190476190477, fmeasure=0.43809523809523804)
rougeL: Score(precision=0.20689655172413793, recall=0.3157894736842105, fmeasure=0.25)
Score function: calculate_rougeL
Average score: 0.4029723641755729
```

Figure 15: ROUGE Scores - Mistral 7b

In Fig. 15 we can see the ROUGE score of all the test questions and the final average ROUGE score for Mistral 7b

```
rouge scores for llama
rouge1: Score(precision=0.8225806451612904, recall=0.5151515151515151, fmeasure=0.6335403726708074)
rouge1: Score(precision=0.14473684210526316, recall=0.1746031746031746, fmeasure=0.15827338129496402)
rouge1: Score(precision=0.2857142857142857, recall=0.3333333333333333, fmeasure=0.30769230769230765)
rouge1: Score(precision=0.4262295081967213, recall=0.8125, fmeasure=0.5591397849462365)
rouge1: Score(precision=0.3333333333333333, recall=0.5454545454545454, fmeasure=0.41379310344827586)
rouge1: Score(precision=0.6428571428571429, recall=0.5625, fmeasure=0.6000000000000001)
rouge1: Score(precision=0.6585365853658537, recall=0.7105263157894737, fmeasure=0.6835443037974684)
rouge1: Score(precision=0.2823529411764706, recall=0.5714285714285714, fmeasure=0.3779527559055118)
rouge1: Score(precision=0.6521739130434783, recall=0.39473684210526316, fmeasure=0.49180327868852464)
Score function: calculate_rouge1
Average score: 0.48498106604680785

rouge2: Score(precision=0.5737704918032787, recall=0.35714285714285715, fmeasure=0.44025157232704404)
rouge2: Score(precision=0.0, recall=0.0, fmeasure=0.0)
rouge2: Score(precision=0.08064516129032258, recall=0.09433962264150944, fmeasure=0.08695652173913043)
rouge2: Score(precision=0.25, recall=0.4838709677419355, fmeasure=0.32967032967032966)
rouge2: Score(precision=0.16981132075471697, recall=0.28125, fmeasure=0.21176470588235294)
rouge2: Score(precision=0.2222222222222222, recall=0.1935483870967742, fmeasure=0.20689655172413793)
rouge2: Score(precision=0.525, recall=0.5675675675675675, fmeasure=0.5454545454545455)
rouge2: Score(precision=0.14285714285714285, recall=0.2926829268292683, fmeasure=0.192)
rouge2: Score(precision=0.36363636363636365, recall=0.2162162162162162, fmeasure=0.2711864406779661)
Score function: calculate_rouge2
Average score: 0.26291636723243267

rougeL: Score(precision=0.5806451612903226, recall=0.36363636363636365, fmeasure=0.4472049689440994)
rougeL: Score(precision=0.09210526315789473, recall=0.1111111111111111, fmeasure=0.10071942446043165)
rougeL: Score(precision=0.25396825396825395, recall=0.2962962962962963, fmeasure=0.2735042735042735)
rougeL: Score(precision=0.32786885245901637, recall=0.625, fmeasure=0.43010752688172044)
rougeL: Score(precision=0.25925925925925924, recall=0.42424242424242425, fmeasure=0.3218390804597701)
rougeL: Score(precision=0.4642857142857143, recall=0.40625, fmeasure=0.43333333333333333)
rougeL: Score(precision=0.4878048780487805, recall=0.5263157894736842, fmeasure=0.5063291139240507)
rougeL: Score(precision=0.2, recall=0.40476190476190477, fmeasure=0.2677165354330708)
rougeL: Score(precision=0.391304347826087, recall=0.23684210526315788, fmeasure=0.2950819672131147)
Score function: calculate_rougeL
Average score: 0.3528716277494124
```

Figure 16: ROUGE Scores - Llama 7b

In Fig. 16 we can see the ROUGE score of all the test questions and the final average ROUGE score for Llama 7b

In Fig. 17 we can see the ROUGE score of all the test questions and the final average ROUGE score for Falcon 7b

From the results we can see that Mistral 7b has the highest ROUGE-1 and ROUGE-L score, whereas Falcon 7b has the highest ROUGE-2 out of the three. Overall, based on the ROUGE score of all three models we can conclude that Mistral 7b has the best performance.

```
rouge scores for falcon
rouge1: Score(precision=0.9473684210526315, recall=0.18181818181818182, fmeasure=0.3050847457627119)
rouge1: Score(precision=0.14285714285714285, recall=0.015873015873015872, fmeasure=0.028571428571428567)
rouge1: Score(precision=0.3780487804878049, recall=0.5740740740740741, fmeasure=0.45588235294117646)
rouge1: Score(precision=1.0, recall=0.34375, fmeasure=0.5116279069767442)
rouge1: Score(precision=0.7619047619047619, recall=0.48484848484848486, fmeasure=0.5925925925925926)
rouge1: Score(precision=0.8333333333333334, recall=0.3125, fmeasure=0.45454545454545453)
rouge1: Score(precision=1.0, recall=0.34210526315789475, fmeasure=0.5098039215686275)
rouge1: Score(precision=0.8076923076923077, recall=0.5, fmeasure=0.6176470588235294)
rouge1: Score(precision=0.8181818181818182, recall=0.23684210526315788, fmeasure=0.3673469387755102)
Score function: calculate_rouge1
Average score: 0.5009000033741624

rouge2: Score(precision=0.5555555555555556, recall=0.10204081632653061, fmeasure=0.1724137931034483)
rouge2: Score(precision=0.0, recall=0.0, fmeasure=0.0)
rouge2: Score(precision=0.1728395061728395, recall=0.2641509433962264, fmeasure=0.208955223880597)
rouge2: Score(precision=1.0, recall=0.3225806451612903, fmeasure=0.4878048780487805)
rouge2: Score(precision=0.45, recall=0.28125, fmeasure=0.34615384615384615)
rouge2: Score(precision=0.5454545454545454, recall=0.1935483870967742, fmeasure=0.2857142857142857)
rouge2: Score(precision=1.0, recall=0.32432432432432434, fmeasure=0.489795918367347)
rouge2: Score(precision=0.36, recall=0.21951219512195122, fmeasure=0.27272727272727276)
rouge2: Score(precision=0.5, recall=0.13513513513513514, fmeasure=0.21276595744680854)
Score function: calculate_rouge2
Average score: 0.3297304899699096

rougeL: Score(precision=0.7894736842105263, recall=0.15151515151515152, fmeasure=0.2542372881355932)
rougeL: Score(precision=0.14285714285714285, recall=0.015873015873015872, fmeasure=0.028571428571428567)
rougeL: Score(precision=0.25609756097560976, recall=0.3888888888888889, fmeasure=0.3088235294117647)
rougeL: Score(precision=1.0, recall=0.34375, fmeasure=0.5116279069767442)
rougeL: Score(precision=0.5714285714285714, recall=0.36363636363636365, fmeasure=0.4444444444444444)
rougeL: Score(precision=0.75, recall=0.28125, fmeasure=0.4090909090909091)
rougeL: Score(precision=0.9230769230769231, recall=0.3157894736842105, fmeasure=0.47058823529411764)
rougeL: Score(precision=0.6538461538461539, recall=0.40476190476190477, fmeasure=0.5)
rougeL: Score(precision=0.8181818181818182, recall=0.23684210526315788, fmeasure=0.3673469387755102)
Score function: calculate_rougeL
Average score: 0.4334073866259241
```

Figure 17: ROUGE Scores - Falcon 7b

5.2.2 BERT

```
bert scores for mistral
Score function: calculate_bert_score
Average score: 0.9075832190337003
```

Figure 18: BERT Scores - Mistral 7b

Based on the average BERTScore of all three models we can conclude that Mistral 7b is the best among the three models.

```
bert scores for llama
Score function: calculate_bert_score
Average score: 0.902605935379311
```

Figure 19: BERT Scores - Llama 7b

```
bert scores for falcon
Score function: calculate_bert_score
Average score: 0.9072098599539863
```

Figure 20: BERT Scores - Falcon 7b

METRIC	Mistral 7b	Llama 7b	Falcon 7b
ROUGE-1	0.52	0.48	0.5
ROUGE-2	0.29	0.26	0.32
ROUGE-L	0.40	0.35	0.43
BERTScore	0.91	0.90	0.90

Table 1: Result Summary

Table 1 shows the summary of ROUGE and BERTScore for all the three fine-tuned model

5.2.3 Manual Evaluation

For manually we not only evaluated the fine-tuned model but also manually generated the data from GPT 3.5, we then manually evaluated the responses to each question. We scored them between 0-10 based on their relevance to the answer from a human perspective. Table 2. shows the scores for each model for all the fine-tuned models and GPT 3.5.

Question Number	**Mistral 7b**	**Llama 7b**	**Falcon 7b**	**GPT 3.5**
Q1	9	9	9	4
Q2	9	9	9	8
Q3	10	8	9	4
Q4	9	9	9	7
Q5	9	9	9	7
Q6	9	9	9	8
Q7	9	9	9	8
Q8	9	8	9	8
Q9	10	10	10	9
Average	9.22	8.8	9.11	7

Table 2: Human Evaluation score for each question

From Table 2 we can see that GPT 3.5 performs very poorly in nuanced and real-life questions like Q1 and Q3 which is *"If an employee injures themselves and goes to see a physician the next day, is it considered a Lost Work Day because it is not the day the injury took place?"* and *"If an employee receives a bee sting (or any* *other bug bite) at work and has an allergic reaction requiring prescription medication and/or days off work, is this considered a recordable injury?"* respectively.

Question	If an employee injures themselves and goes to see a physician the next day, is it considered a Lost Work Day because it is not the day the injury took place?
GPT 3.5	It depends on the specific policies of the employer and the regulations in the jurisdiction where the incident occurred. In some cases, if the injury is directly related to work duties, the day the employee seeks medical attention could be considered a lost work day. It's best to consult with HR or legal experts for precise guidance.
Mistral 7b	No, you do not count a day as a day away from work if the employee is not scheduled to work on that day.
Llama 7b	No. If the employee goes to see a physician the next day, it is not considered a lost work day. The day the injury occurred is the day that is counted. If the employee goes to see a physician the next day, it is not a lost work day. If the employee is restricted from work, it is a lost work day.
Falcon 7b	No. A day away from work is counted only for the day on which the employee is not working.

Table 3: Answer generated by GPT 3.5, Fine-tuned Mistal 7b, Llama 7b and Falcon 7b

Table 3 shows the response for Q1 and we can see that GPT 3.5 was not able to understand the context and give a proper response, while the fine-tuned models were able to accurately give the correct response to the question.

Fig 21 shows a graph for the average score for human evaluation for Mistral 7b, Llama 7b, Falcon 7b, and GPT 3.5. From Table 2 and Fig 21 we can conclude that even though GPT 3.5 can correctly answer some questions related to EHS it fails to give correct responses for real-life examples.

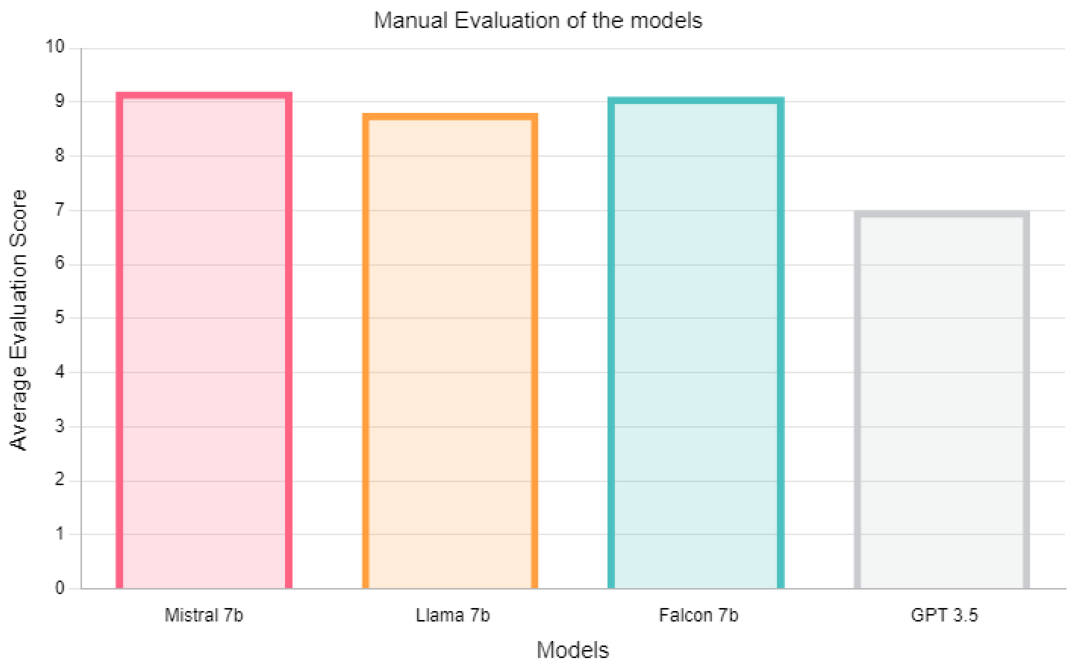

Figure 21: Average manual evaluation score for different fine-tuned models

CHAPTER 6

Conclusion and Future Works

In this paper, we fine-tuned three foundation models namely Mistral 7b, Llama 7b, and Falcon 7b for domain-specific EHS tasks. The data for fine-tuning was collected from various sources like the OSHA faq website and OSHA handbook manual using different techniques such as web-scraping and manual annotation. We then employed Parameter-Efficient Fine-Tuning (PEFT) techniques such as Quantized Low-Rank Adapters(QLoRA) and Supervised Fine-tuning (SFT) to fine-tune all three base models. The fine-tuned models are then compared using metrics like ROUGE and BERTScore. To compare it with the existing LLM models we did a manual evaluation of all three models with GPT 3.5. From the manual evaluation results, we can conclude that GPT 3.5 fails to answer real-life nuance EHS questions because of a lack of domain context. Apart from this all the foundation models used for fine-tuning are open-source and the performance can be further improved with a larger dataset and using base models with a higher number of parameters like 13b and 70b.

In the future, we can extend the fine-tuned models to be used as a chatbot to train and onboard new employees in the EHS team while also simplifying the whole EHS process. They can also be integrated into the existing incident reporting software to automate the recording of incidents and help organizations stay in line with the EHS regulations

LIST OF REFERENCES

[1] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, and W. Chen, ``Lora: Low-rank adaptation of large language models,'' *CoRR*, vol. abs/2106.09685, 2021. [Online]. Available: https://arxiv.org/abs/2106.09685

[2] I. Ozkaya, ``Application of large language models to software engineering tasks: Opportunities, risks, and implications,'' *IEEE Software*, vol. 40, no. 3, pp. 4--8, 2023.

[3] J. Cui, Z. Li, Y. Yan, B. Chen, and L. Yuan, ``Chatlaw: Open-source legal large language model with integrated external knowledge bases,'' 2023.

[4] Andreas Stöffelbauer. Accessed: April 17, 2024. [Online]. Available: https://medium.com/data-science-at-microsoft/how-large-language-models-work-91c362f5b78f

[5] Y. Yang, Y. Tang, and K. Y. Tam, ``Investlm: A large language model for investment using financial domain instruction tuning,'' 2023.

[6] Z. Qi, Y. Yu, M. Tu, J. Tan, and Y. Huang, ``Foodgpt: A large language model in food testing domain with incremental pre-training and knowledge graph prompt,'' 2023.

[7] O. Zheng, M. Abdel-Aty, D. Wang, C. Wang, and S. Ding, ``Trafficsafetygpt: Tuning a pre-trained large language model to a domain-specific expert in transportation safety,'' 2023.

[8] Y. Dan, Z. Lei, Y. Gu, Y. Li, J. Yin, J. Lin, L. Ye, Z. Tie, Y. Zhou, Y. Wang, A. Zhou, Z. Zhou, Q. Chen, J. Zhou, L. He, and X. Qiu, ``Educhat: A large-scale language model-based chatbot system for intelligent education,'' 2023.

[9] L. Xiao and X. Chen, ``Enhancing llm with evolutionary fine tuning for news summary generation,'' 2023.

[10] T. Baldazzi, L. Bellomarini, S. Ceri, A. Colombo, A. Gentili, and E. Sallinger, ``Fine-tuning large enterprise language models via ontological reasoning,'' 2023.

[11] I. Singh, V. Blukis, A. Mousavian, A. Goyal, D. Xu, J. Tremblay, D. Fox, J. Thomason, and A. Garg, ``Progprompt: Generating situated robot task plans using large language models,'' in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 11 523--11 530.

[12] S. Yu, C. Fang, Y. Ling, C. Wu, and Z. Chen, "Llm for test script generation and migration: Challenges, capabilities, and opportunities," in *2023 IEEE 23rd International Conference on Software Quality, Reliability, and Security (QRS)*, 2023, pp. 206--217.

[13] W. Kuang, B. Qian, Z. Li, D. Chen, D. Gao, X. Pan, Y. Xie, Y. Li, B. Ding, and J. Zhou, "Federatedscope-llm: A comprehensive package for fine-tuning large language models in federated learning," 2023.

[14] Z. Lai, X. Zhang, and S. Chen, "Adaptive ensembles of fine-tuned transformers for llm-generated text detection," 2024.

[15] Y. Zhang, D. Huang, B. Liu, S. Tang, Y. Lu, L. Chen, L. Bai, Q. Chu, N. Yu, and W. Ouyang, "Motiongpt: Finetuned llms are general-purpose motion generators," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 7, pp. 7368--7376, Mar. 2024. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/28567

[16] S. Doh, M. Lee, D. Jeong, and J. Nam, "Enriching music descriptions with a finetuned-llm and metadata for text-to-music retrieval," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 826--830.

[17] Occupational Safety and Health Administration. "Workplace safety and health." Accessed: April 17, 2024. [Online]. Available: https://www.osha.gov/

[18] Occupational Safety and Health Administration. "Osha recordkeeping faq." Accessed: April 17, 2024. [Online]. Available: https://www.osha.gov/recordkeeping/faq-search

[19] Occupational Safety and Health Administration, *OSHA Field Safety and Health Manual*. Occupational Safety and Health Administration, 2020. [Online]. Available: https://www.osha.gov/sites/default/files/enforcement/directives/ADM_04-00-003.pdf

[20] Ali Mobarekati. Accessed: April 17, 2024. [Online]. Available: https://medium.com/@codersama/fine-tuning-mistral-7b-in-google-colab-with-qlora-complete-guide-60e12d437cca

[21] Heiko Hotz. Accessed: April 17, 2024. [Online]. Available: https://towardsdatascience.com/rag-vs-finetuning-which-is-the-best-tool-to-boost-your-llm-application-94654b1eaba7

[22] Accessed: April 17, 2024. [Online]. Available: https://cameronrwolfe.substack.com/p/easily-train-a-specialized-llm-peft

[23] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "Qlora: Efficient finetuning of quantized llms," 2023.

[24] Jose J. Martinez. Accessed: April 17, 2024. [Online]. Available: https://medium.com/mantisnlp/supervised-fine-tuning-customizing-llms-a2c1edbf22c3

APPENDIX A

Responses for EHS Questions

A.1 Fine-tuned Mistral 7b


```json
[
  {
    "q": "If an employee injures themselves and goes to see a physician the next day, is it considered a Lost Work Day because it is not the day the injury took place?",
    "a": "No, you do not count a day as a day away from work if the employee is not scheduled to work on that day."
  },
  {
    "q": "If our company has multiple offices in multiple states does each office have to maintain an OSHA 300 Log or should one log for all locations be kept at the Corporate office only?",
    "a": "Each individual establishment must maintain a separate OSHA 300 Log. An establishment is defined as a single physical location where business is conducted or where services or industrial operations are performed. Multiple physical locations that are located on one site or are otherwise related to each other are considered one establishment if they are owned or operated by the same entity and are located on the same site."
  },
  {
    "q": "If an employee receives a bee sting (or any other bug bite) at work and has an allergic reaction requiring prescription medication and/or days off work, is this considered a recordable injury?",
    "a": "No, an allergic reaction to a bee sting or other insect bite is not considered an injury for OSHA recordkeeping purposes if the employee is not otherwise restricted from work or required to transfer or modify their job."
  },
  {
    "q": "Will OSHA accept an electronic signature on the 300A Summary?",
    "a": "Yes, OSHA will accept an electronic signature on the 300A Summary."
  },
  {
    "q": "We are a hospital. Under our umbrella, there are numerous physician clinics and practices. My understanding is the physician clinics are not required to maintain the 300 log. Is this correct?",
    "a": "Yes, you are correct. The OSHA recordkeeping requirements apply to the establishment, not the parent company. If the physician clinics are separate establishments, they must maintain their own OSHA 300 Logs."
  },
  {
    "q": "Our business operates two production facilities (appx 40 miles apart). Should we be keeping a separate 300 log for both facilities?",
    "a": "You must keep a separate OSHA 300 Log for each establishment that is expected to be in operation for one year or longer. An establishment is a separate physical location where business is conducted."
  },
  {
    "q": "May a third party submit data for an establishment or firm?",
    "a": "Yes, a third party may submit data for an establishment or firm. However, the employer is responsible for the accuracy of the data."
  },
  {
    "q": "If the injured or ill worker produces fewer goods or services than he or she would have produced prior to the injury or illness but otherwise perform all of the routine functions of his or her work, is the case considered a restricted work case?",
    "a": "No. A case is considered restricted work if the employee is not able to perform all of the routine functions of his or her job, or if he or she is able to perform all of the routine functions of his or her job but is restricted from working the full day or full shift that he or she would otherwise have worked."
  },
  {
    "q": "If an employer has no recordable cases for the year, is an OSHA 300-A, Annual Summary, still required to be completed, certified and posted?",
    "a": "Yes, even if the employer has no recordable injuries or illnesses, the employer must complete the OSHA 300 Log, keep it at the establishment, and make it available for inspection. The employer must also prepare an annual summary, the OSHA 300-A, and post it in a conspicuous place or places where notices to employees are customarily posted."
  }
]
```


Figure A.22: Screenshot of the response for Fine-tuned Mistral 7b

A.2 Fine-tuned Llama 7b


```json
[
  {
    "q": "If an employee injures themselves and goes to see a physician the next day, is it considered a Lost Work Day because it is not the day the injury took place?",
    "a": "No. If the employee goes to see a physician the next day, it is not considered a lost work day. The day the injury occurred is the day that is counted. If the employee goes to see a physician the next day, it is not a lost work day. If the employee is restricted from work, it is a lost work day."
  },
  {
    "q": "If our company has multiple offices in multiple states does each office have to maintain an OSHA 300 Log or should one log for all locations be kept at the Corporate office only?",
    "a": "Each establishment must maintain its own OSHA 300 Log. If you have multiple establishments in multiple states, each establishment must maintain its own log. You may use a single form to record information for all of your establishments, but each establishment must have its own copy of the form. You may also use a computer program to print out the information from the single form and attach it to each establishment's copy of the form."
  },
  {
    "q": "If an employee receives a bee sting (or any other bug bite) at work and has an allergic reaction requiring prescription medication and/or days off work, is this considered a recordable injury?",
    "a": "Yes. If the employee requires prescription medication or days off work, the case is recordable. If the employee does not require prescription medication or days off work, the case is not recordable. If the employee requires only over-the-counter medication, the case is not recordable. If the employee requires prescription medication but does not require days off work, the case is recordable."
  },
  {
    "q": "Will OSHA accept an electronic signature on the 300A Summary?",
    "a": "Yes, OSHA will accept electronic signatures on the 300A Summary. You may use an electronic signature tool such as Adobe Acrobat or other software to create the signature. You may also use a digital signature tool such as a digital certificate or a digital signature service. You must ensure that the signature is secure and cannot be altered or tampered with."
  },
  {
    "q": "We are a hospital. Under our umbrella, there are numerous physician clinics and practices. My understanding is the physician clinics are not required to maintain the 300 log. Is this correct?",
    "a": "Yes, physician clinics and practices are not required to maintain the OSHA 300 Log. However, if the physician clinic or practice is a separate business entity, it must maintain its own OSHA 300 Log. If the physician clinic or practice is part of the hospital, it must maintain the hospital's OSHA 300 Log."
  },
  {
    "q": "Our business operates two production facilities (appx 40 miles apart). Should we be keeping a separate 300 log for both facilities?",
    "a": "Yes. You must keep a separate 300 log for each establishment. If you have more than one establishment, you must keep a separate 300 log for each establishment."
  },
  {
    "q": "May a third party submit data for an establishment or firm?",
    "a": "Yes, a third party may submit data for an establishment or firm. However, the third party must have access to the establishment or firm's OSHA 300 Log and must be authorized by the establishment or firm to submit the data."
  },
  {
    "q": "If the injured or ill worker produces fewer goods or services than he or she would have produced prior to the injury or illness but otherwise perform all of the routine functions of his or her work, is the case considered a restricted work case?",
    "a": "Yes. If the injured or ill worker produces fewer goods or services than he or she would have produced prior to the injury or illness but otherwise performs all of the routine functions of his or her work, the case is considered a restricted work case. The case would be classified as a restricted work case if the worker is restricted in the kind of work he or she can perform or the work he or she can perform is limited in some other way."
  },
  {
    "q": "If an employer has no recordable cases for the year, is an OSHA 300-A, Annual Summary, still required to be completed, certified and posted?",
    "a": "Yes. Even if you have no recordable cases, you must still complete the OSHA 300-A, Annual Summary, certify it, and post it."
  }
]
```


Figure A.23: Screenshot of the response for Fine-tuned Llama 7b

A.3 Fine-tuned Falcon 7b


```
[
  {
    "q": "If an employee injures themselves and goes to see a physician the next day, is it considered a Lost Work Day because it is not the day the injury took place?",
    "a": "No. A day away from work is counted only for the day on which the employee is not working."
  },
  {
    "q": "If our company has multiple offices in multiple states does each office have to maintain an OSHA 300 Log or should one log for all locations be kept at the Corporate office only?",
    "a": "Each establishment must keep a separate log."
  },
  {
    "q": "If an employee receives a bee sting (or any other bug bite) at work and has an allergic reaction requiring prescription medication and/or days off work, is this considered a recordable injury?",
    "a": "Yes, this is a recordable case. An injury or illness is work-related if an event or exposure in the work environment either caused or contributed to the resulting condition or significantly
aggravated a pre-existing injury or illness. Work-relatedness is presumed for injuries and illnesses resulting from events or exposures occurring in the work environment, unless an exception in \u00a7 1904
.5(b)(2) specifically applies. There is no exception for insect bites or stings in section 1904.5(b)(2)."
  },
  {
    "q": "Will OSHA accept an electronic signature on the 300A Summary?",
    "a": "Yes, OSHA will accept an electronic signature on the 300A Summary."
  },
  {
    "q": "We are a hospital. Under our umbrella, there are numerous physician clinics and practices. My understanding is the physician clinics are not required to maintain the 300 log. Is this correct?",
    "a": "The OSHA recordkeeping requirements apply to all workplaces that are covered by the OSH Act. This includes physician clinics and practices."
  },
  {
    "q": "Our business operates two production facilities (appx 40 miles apart). Should we be keeping a separate 300 log for both facilities?",
    "a": "Yes, you must keep a separate OSHA 300 Log for each establishment."
  },
  {
    "q": "May a third party submit data for an establishment or firm?",
    "a": "Yes, a third party may submit the data for an establishment or firm."
  },
  {
    "q": "If the injured or ill worker produces fewer goods or services than he or she would have produced prior to the injury or illness but otherwise perform all of the routine functions of his or her
work, is the case considered a restricted work case?",
    "a": "No. A case is considered a restricted work case if the worker is restricted from performing one or more routine functions of his or her job."
  },
  {
    "q": "If an employer has no recordable cases for the year, is an OSHA 300-A, Annual Summary, still required to be completed, certified and posted?",
    "a": "Yes, the employer must still complete and post the annual summary."
  }
]
```


Figure A.24: Screenshot of the response for Fine-tuned Falcon 7b

A.4 GPT 3.5

M **You**

If an employee injures themselves and goes to see a physician the next day, is it considered a Lost Work Day because it is not the day the injury took place?

ChatGPT

It depends on the specific policies of the employer and the regulations in the jurisdiction where the incident occurred. In some cases, if the injury is directly related to work duties, the day the employee seeks medical attention could be considered a lost work day. It's best to consult with HR or legal experts for precise guidance.

You

If our company has multiple offices in multiple states does each office have to maintain an OSHA 300 Log or should one log for all locations be kept at the Corporate office only?

ChatGPT

Each office should maintain its own OSHA 300 Log. According to OSHA (Occupational Safety and Health Administration) regulations, employers with more than ten employees are required to keep a record of serious work-related injuries and illnesses. If your company has multiple offices in different states, each location should maintain its own OSHA 300 Log to accurately track incidents and ensure compliance with OSHA regulations specific to each state. While it might be beneficial to have a centralized system for recordkeeping purposes, it's essential to maintain separate logs for each location to accurately reflect the safety performance and incidents at each site.

You

If an employee receives a bee sting (or any other bug bite) at work and has an allergic reaction requiring prescription medication and/or days off work, is this considered a recordable injury?

ChatGPT

Whether a bee sting or bug bite incident at work is considered a recordable injury depends on various factors, including the severity of the reaction, the specific regulations of the jurisdiction, and the company's internal policies.

You

Will OSHA accept an electronic signature on the 300A Summary?

ChatGPT

As of my last update in January 2022, OSHA (Occupational Safety and Health Administration) does not explicitly accept or reject electronic signatures on the Form 300A Summary. However, OSHA regulations do allow electronic recordkeeping and signatures for certain documents, but it's essential to ensure compliance with all relevant regulations and guidelines.

It's advisable to consult the most recent OSHA guidelines or contact OSHA directly to confirm whether electronic signatures are currently accepted for the Form 300A Summary or if there have been any updates to their policies regarding electronic signatures since my last update.

You

We are a hospital. Under our umbrella, there are numerous physician clinics and practices. My understanding is the physician clinics are not required to maintain the 300 log. Is this correct?

ChatGPT

As of my last update in January 2022, the requirement to maintain a Form 300 log, also known as the OSHA Form 300, pertains to certain types of employers and workplaces. Generally, the OSHA Form 300 log is required for most employers with more than 10 employees, unless they fall into an exempted industry or category.

Physician clinics and practices may or may not be required to maintain the OSHA Form 300 log, depending on various factors such as the number of employees, the nature of the work conducted, and the jurisdiction's specific regulations. It's essential to consult with legal counsel or an occupational safety professional to determine whether your physician clinics are subject to this requirement.

Additionally, OSHA regulations and requirements may have changed since my last update, so it's always a good idea to check for any updates or changes to compliance obligations.

You

Our business operates two production facilities (appx 40 miles apart). Should we be keeping a separate 300 log for both facilities?

ChatGPT

The decision to maintain separate 300 logs for each production facility depends on various factors, including legal requirements, organizational policies, and practical considerations. Here are some points to consider:

1. **Legal Requirements**: Occupational Safety and Health Administration (OSHA) regulations in the United States require certain employers to maintain injury and illness records, including the OSHA Form 300 Log, which documents workplace injuries and illnesses. Depending on the jurisdiction and industry, regulations may vary regarding whether each facility needs its own log.
2. **Operational Differences**: If the two production facilities have distinct operations, hazards, workforce compositions, or safety programs, it may be beneficial to maintain separate logs to accurately track and analyze workplace incidents specific to each location.
3. **Consistency in Recordkeeping**: Maintaining separate logs can help ensure that each facility's safety performance is accurately documented and analyzed without potential confounding factors from another location.
4. **Centralized Reporting**: On the other hand, consolidating records into a single log could streamline administrative processes, simplify reporting, and provide a comprehensive overview of the organization's safety performance.
5. **Data Analysis**: Consider how you plan to use the data from the 300 logs. Separate logs may allow for more detailed analysis of trends and patterns at each facility, whereas a combined log may facilitate broader comparisons and identification of company-wide trends.
6. **Resource Allocation**: Assess the resources available for recordkeeping and analysis. Maintaining separate logs may require additional administrative efforts and resources, whereas a single log could be more manageable and efficient, particularly for smaller organizations.

Ultimately, it's essential to weigh these factors and consult legal counsel or regulatory authorities to ensure compliance with applicable laws and regulations. Additionally, consider engaging with relevant stakeholders within your organization to determine the most practical and effective approach for your specific circumstances.

You

May a third party submit data for an establishment or firm?

ChatGPT

Yes, in many cases, third parties can submit data for an establishment or firm, depending on the context and the policies in place. For example:

1. **Surveys and Feedback:** Companies often use third-party survey companies to collect feedback from customers or clients. These third parties then submit the data they collect to the company for analysis.
2. **Data Aggregation Services:** Some firms utilize third-party data aggregation services to gather information from various sources and compile it into a usable format for analysis or reporting.
3. **Government Reporting:** In certain industries, third-party consultants or specialists may assist establishments or firms in compiling and submitting data required for regulatory compliance or government reporting.
4. **Vendor Management:** Companies may allow third-party vendors or suppliers to directly input data related to their products or services into the company's systems.

However, it's essential to ensure that any third-party data submission is done securely and in compliance with relevant privacy regulations and internal data governance policies. Additionally, firms should have mechanisms in place to verify the accuracy and integrity of the data submitted by third parties.

You

If the injured or ill worker produces fewer goods or services than he or she would have produced prior to the injury or illness but otherwise perform all of the routine functions of his or her work, is the case considered a restricted work case?

ChatGPT

Yes, typically, if an injured or ill worker is able to perform all routine functions of their work but at a reduced productivity level due to the injury or illness, it would still be considered a restricted work case. In such situations, the worker's ability to perform their job is restricted, albeit not completely, which would classify the case as restricted work.

You

If an employer has no recordable cases for the year, is an OSHA 300-A, Annual Summary, still required to be completed, certified and posted?

ChatGPT

Even if there are no recordable cases for the year, employers subject to OSHA recordkeeping requirements must still complete, certify, and post the OSHA Form 300-A, the Summary of Work-Related Injuries and Illnesses. This form summarizes the total number of job-related injuries and illnesses that occurred during the year and must be posted in a conspicuous location where employees can easily see it. It provides transparency about the workplace safety performance for both employees and potential visitors to the workplace.

Figure A.25: Screenshot of the response for GPT 3.5

"You are an AI trained to assist with Environmental, Health, and Safety (EHS) regulations and compliance. Your role is to provide accurate, detailed, and up-to-date information on EHS standards, practices, and legislation. You should offer guidance that is in strict adherence to the relevant laws and industry standards, such as those set by OSHA, EPA, and other regulatory bodies. When responding to queries, your answers must be clear, concise, and practical, providing actionable advice whenever possible. You should consider the context of each question, whether it pertains to workplace safety, environmental protection, health standards, or risk management. In cases where specific regulatory details are requested, provide precise references to the relevant sections of laws or guidelines."

"If an employee injures themselves and goes to see a physician the next day, is it considered a Lost Work Day because it is not the day the injury took place?"

Injuries resulting in lost workdays are typically determined based on the day of the injury, not the day the employee seeks medical treatment. If an employee injures themselves on a particular day, that day is typically counted as a lost workday if the injury prevents them from returning to work for the remainder of their scheduled shift. Whether the employee seeks medical attention on the same day or the next day does not typically affect the classification of the injury as a lost workday.

Figure A.26: Screenshot of the response for GPT 3.5 with additional context