Applying Large language models to capture impact assessment from AI incident reports in order to gain insight into relationships and correlations between impact categorisation ontologies

Patrick O'Callaghan

A Dissertation

Presented to the University of Dublin, Trinity College

in partial fulfillment of the requirements for the degree of

Master of Computer Science

Supervisor: Dr. David Lewis

April 2024

Declaration

I, the undersigned, declare that this work has not previously been submitted as an exercise for a degree at this, or any other University, and that unless otherwise stated, is my own work.

I have read and I understand the plagiarism provisions in the General Regulations of the University Calendar for the current year, found at http://www.tcd.ie/calendar.

I have completed the Online Tutorial in avoiding plagiarism ‘Ready, Steady, Write’, located at http://tcd-ie.libguides.com/plagiarism/ready-steady-write.

Patrick O’Callaghan
April 15, 2024

Applying Large language models to capture impact assessment from AI incident reports in order to gain insight into relationships and correlations between impact categorisation ontologies

The abilities of Large Language Models has grown exponentially in recent years. With a heightened focus on safety concerns with this technology, there has been a greater emphasis placed on categorizing high risk AI systems, in order to better understand their risks and dangers. This categorization is done through incident report annotation. By applying semantic models, namely impact categorisation ontologies, to these incident reports, stakeholders can better understand these risks and dangers.

This research aims to leverage large language models to optimize the process of capturing impact assessments from publicly available AI incident reports. This will be done through the application of large language models to the task of extracting AI incident report data into existing impact categorization ontologies.

By examining the annotated incident reports, conclusions can be drawn about the ability of large language models to perform this annotation task that is traditionally time-intensive, as manual annotation is the current standard method. As well as this, conclusions can be drawn regarding correlations and relationships between the different impact categorisation ontologies used, which can lead to further work being done on standardizing the impact categorization methodology in the future.

Acknowledgements

I would like to thank my supervisor Professor David Lewis whose input and guidance throughout this dissertation was invaluable to me.

I would also like to thank my family for their continued support and encouragement throughout this research, as well as over the course of my academic career.

Table of Contents

Chapter 1 Introduction

Chapter 2 State of the Art

Chapter 3 Design

Chapter 4 Implementation

Chapter 5 Evaluation

Chapter 6 Conclusions and Future work

References

Chapter 1

Introduction

This chapter shall introduce the reader to the dissertation work. It begins with the motivation which highlights the importance and relevance of the work. Following sections include both the research question and the research objectives, with the final section summarizing the structure of the report.

1.1 Motivation

As Artificial Intelligence (AI) improves, its impact on our society grows dramatically. It is vital to understand the potential risks that this growth may have. AI incidents, such as privacy breaches, systematic biases or safety hazards can cause significant damage to both individuals and society as a whole. In order to effectively mitigate these risks, it is important to understand them. The importance of recording and analyzing AI incidents is key to a successful mitigation of risk. Efforts to collect information on AI incidents have thus become more important. These efforts have taken the form of the collection of large databases of media reports of AI incidents such as the AI, Algorithmic and Automation Incidents and Controversies (AIAAIC)[1] .and the AI Incident Database (AIID) [2]

Semantic models that categorize risk and impact such as the Vocabulary of AI Risks (VAIR) or the Common Impact Data Standard (CIDS), are commonly used to annotate these incidents in order to standardize and categorize them. This research

[1] AIAAIC Repository https://www.aiaaic.org/

[2] AI Incident Database https://incidentdatabase.ai/

aims to understand how the current state of the art Large Language models (LLMs) like openAI's GPT-4 [3] and Anthropic's Clause 3[4] can be leveraged to aid the process of applying these ontologies to AI incidents, as well as to examine how these ontologies correlate to each other. This may uncover valuable insights into the relationships between similar AI incidents and risks, and the correlations between different ontologies.

The relevancy of this research extends across multiple stakeholders and domains. Companies currently developing AI systems need to understand the landscape of AI incidents in order to deploy safe and fair systems. Understanding AI risk categorization can allow these companies to preemptively mitigate risks that may occur from their AI systems by studying similar risks from other systems or across other domains. With the EU AI act recently coming into law, regulators and policymakers have a heightened interest in AI risk categorization. Understanding and optimizing effective AI risk categorization would allow these regulators and policymakers to focus their attention on studying the causes and possible mitigations for risks that are deemed to have the highest severity, or that appear most frequently, thus improving the regulation of AI systems.

As AI integration increases across different domains such as healthcare, finance and law enforcement, the understanding of how AI incidents relate to each other, as well as how their various ontological categorisations correlate becomes more relevant. This integration gives the various stakeholders involved the opportunity to compare similar risks across different jurisdictions, industries or organizations. This can be an effective way for stakeholders to understand how risks are handled, and how to effectively mitigate risks stemming from their own integration by learning from others.

[3] GPT-4 https://chat.openai.com/

[4] Claude 3 https://claude.ai/

1.2 Research Question

With AI regulation coming to the EU through the new AI Act, there is a need more than ever for regulatory assessment and the impact that it has on new AI systems. This research has two main aims: to investigate how LLMs can be used to aid the process of to capture impact assessments in order to categorize AI incidents using existing ontologies and to analyze the relationships and correlations between the existing ontologies used to categorize AI incidents. As a result, the research question for this dissertation is as follows:

To what extent can existing Large language models and ontologies be leveraged to capture impact assessment from AI incident reports in order to gain insight into relationships and correlations between impact categorisation ontologies.

1.3 Research Objectives

From the research question defined in 1.2, there are three main points. The first being the extent to which existing LLMs can be used to capture impact assessment from AI incident reports This leads to the first research objective below, which aims to develop a methodology to capture impact assessments from AI incidents using LLMs.

The second point is to how LLMs can be used to annotate these impact categorisation ontologies using the impact assessments captured. This leads to the second research objective, which aims to develop a methodology for using an LLM to annotate these impact assessments using impact categorisation ontologies.

The third point is the insight into relationships and correlations between different impact categorisation ontologies. This is the reason for the third research objective, to analyze the annotated impact categorization ontologies to try and gain insight into the relationships and correlations between them.

In summary, below are the research objectives of this dissertation:

1. To develop and evaluate a methodology for using LLMs like GPT-4 and Claude 3 to capture impact assessments from AI incidents.

2. To develop a methodology for using LLMs to annotate these impact assessments using impact categorisation ontologies.

3. To analyze the annotated impact categorization ontologies in order to gain insight into the relationships and correlations between them.

1.4 Structure and Contents

The following chapters will cover the State of the Art, Design, Implementation, Evaluation and Conclusions & Future Work. State of the Art will cover the background context of this work, as well as closely related works. The Design chapter will cover the research focus, design choices and design overview. Implementation will cover the data preprocessing, LLM choices, prompt engineering, ontology preparation and selection, and the correlation analysis process. The Evaluation chapter will cover the evaluation framework used, and presents the results using this framework. The Conclusions & Future Work chapter will conclude this dissertation by reflecting on the work completed, and possible future adaptations.

Chapter 2
State of the Art

This chapter aims to look at background topics related to the area of research as well as related works. The aim here is to give context for the research of this dissertation, and to give the reader a deeper understanding of the relevance of the work in the area of research. By presenting the existing literature it gives insight into current gaps, as well as providing the grounds for a deeper understanding of the research topic for the reader.

2.1 Background

Over the last few years there has been an increasing interest in the area of artificial intelligence, especially considering the paradigm shift of Large Language Models (LLM) like GPT-4 and Claude-3. The need for regulations of Artificial Intelligence systems has in turn been accelerated at an unprecedented rate. Many new frameworks have been proposed across the globe, with the European AI act garnering attention as the world's first AI law[5] . This new AI Act stands to be the inaugural legislation in the field of artificial intelligence, with the added importance of covering the entire European Union. (5) In an environment where regulatory compliance is becoming more important, and AI associated risks are becoming more prevalent every day, the need for an analysis of ways to categorize incidents and mitigate future risks is growing. Here, these incident reports will be discussed in depth, along with the AI Act that applies to them. As well as this, the relevant semantic models and ontologies will be discussed in order to understand the current state of the art in this area.

[5] EU AI Act https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_EN.pdf

2.1.1 Incident Reports

Artificial Intelligence systems have inherent risks due to the nature of the technology. These risks manifest in various ways, such as biases in decision making, safety control failure and breaches in cybersecurity. As the development of AI systems continues despite these issues, a need for systematic documentation and analysis of AI incidents has emerged. AI incident reports are the solution for this. These incident reports play an important role in the understanding and mitigation of AI risks. These reports document the failures, biases and other risks that are commonplace for AI systems and serve as a valuable resource for the various stakeholders in the improvement of AI safety.

Databases such as the AI, Algorithmic and Automation Incidents and Controversies (AIAAIC) repository and the AI Incident Database (AIID) have been founded to catalog these incidents (16). These serve as centralized repositories for AI incident reports that are invaluable for developers, researchers and regulators. These repositories categorize a wide range of AI incidents, mainly sourced from news articles. Currently there is a lack of standardized AI incident reporting. Organizations like OECD are working on developing a common framework for AI incident reporting to address this gap. The global AI Incidents Monitor (AIM) intends to collect incidents in real-time from publically available resources.[6] Until this source becomes publicly available for use, news article based repositories like the AIAAIC repository serve as a valuable source of incident reports.

2.1.2 AIAAIC repository

The AI, Algorithmic and Automation Incident and Controversy (AIAAIC) repository is an open-source and independent database which collects information about incidents related to AI. It was launched in 2019, and has over 1000 entries occurring from 2012 to present. The repository serves as a valuable tool for accessing structured and verifiable incident reports. The content of the repository is available to

[6] AI Incidents Monitor https://oecd.ai/en/ai-principles

use, copy, redistribute and adapt under a CC BY-SA 4.0 license.[7] This repository of incidents will be used in this work as the main source of incident reports, as it is currently the state of the art in the area.

2.1.3 The EU AI Act

The EU AI Act stands as the world's first Artificial Intelligence law. This piece of legislation intends to set the precedent for AI system risk mitigation in the EU. (24) The act takes a risk-based categorization approach, classifying AI systems into four categories: unacceptable risk, high-risk, limited risk and minimal risk. The Act imposes strict obligations on both developers and deployers of high-risk AI systems in order to mitigate the risks involved. Systems categorized as high-risk must undergo outlined assessments before being available for use in the EU. (10). The Act aims to regulate these AI systems comprehensively based on their risks and impacts, with an aim to promote trustworthy AI innovation.

2.2 Generative Artificial intelligence

Generative Artificial Intelligence has come to the forefront of Artificial Intelligence recently, referring to systems which generate new information rather than regurgitating information that the system was trained on.(7) It is this new form of artificial intelligence that demands new regulations due to its limitless potential. These tools have the ability to analyze many different types of data such as text, images, video and audio.(14) The current state of the art is changing daily, with new tools constantly pushing the leading edge of the technology out further. For the purpose of this work, the main generative artificial intelligence tool focused on is the large language model (LLM).

[7] AIAAIC Terms of Use: https://www.aiaaic.org/terms

2.2.1 Large Language Models (LLMs)

The Large Language Model is a generative artificial intelligence tool that uses deep learning techniques and a massive volume of data to understand and generate text. They are based on the transformer architecture which uses attention mechanisms to process text. Multiple layers of neural networks are used with huge numbers of parameters that are then fine tuned during model training. (11) In order to train these layers of huge networks, self-supervised learning is used on a vast corpora of text. During this training, the model learns to predict the most likely next word in a sequence of words, which leads to the model having the ability to generate text while seeming to understand context, patterns, grammar and semantics. (17)

2.2.2 Benchmarking LLM Performance

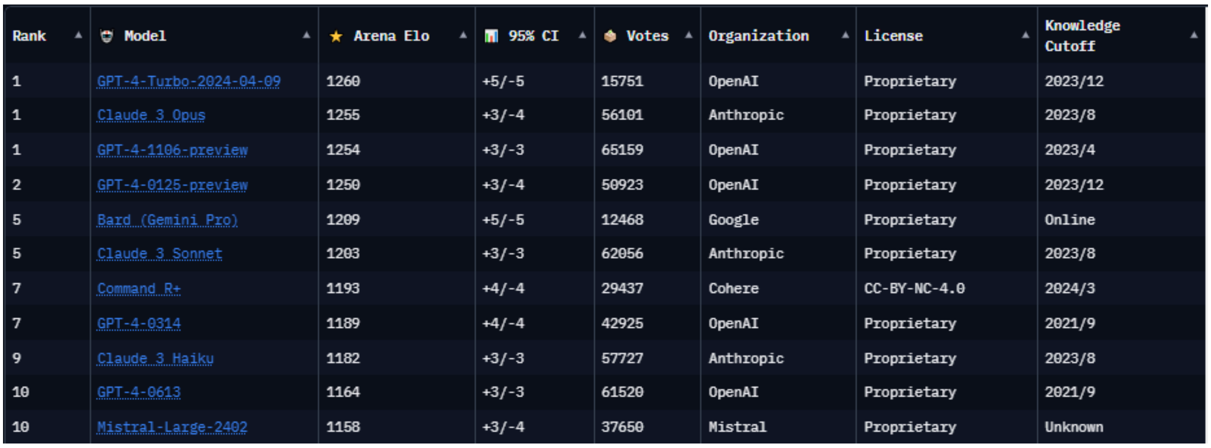
The current state of the art for LLMs is advancing rapidly. As a result, there has been a saturation of benchmarks to evaluate the best performing LLMs due to this unprecedented development pace. As a result of this, public leaderboards like the Chatbot Arena[8] and Stanford's HELM leaderboard[9] serve as the most relevant sources for understanding which LLMs have the best performance. These public leaderboards take many different benchmarks into account, with three of the main benchmarks being Massive Multitask Language Understanding (MMLU), Discrete Reasoning Over Paragraphs (DROP) and HumanEval.

MMLU is designed to assess the multitasking accuracy of an LLM across different subjects from Science Technology, Engineering and Mathematics (STEM), humanities and social sciences. It is a measure of a models' ability to perform across the board on academic questions, which underscores the model's underlying knowledge base and potential for real world use.

[8] Chatbot arena Leaderboard: https://chat.lmsys.org/?leaderboard
[9] Stanford HELM Leaderboard; https://crfm.stanford.edu/helm/lite/latest/#/leaderboard

DROP is a measure of a models' ability to perform discrete reasoning over paragraphs. It involves extracting relevant information from text to answer questions that require a series of reasoning steps that can include mathematical operations. The benchmark tests the models' comprehension and analytical abilities, giving insight into how good a model is at synthesizing information in a logical way.

HumanEval is a benchmark focused on the functional correctness of code generated by an LLM. It consists of Python programming questions, each containing a function and several unit tests. The benchmark assesses the model on language comprehension, reasoning, mathematical operations and algorithms. The performance of an LLM on this benchmark is an indication of the models' ability to understand programming tasks and generate functionally correct code.



Rank	Model	★ Arena Elo	95% CI	Votes	Organization	License	Knowledge Cutoff
1	GPT-4-Turbo-2024-04-09	1260	+5/-5	15751	OpenAI	Proprietary	2023/12
1	Claude 3 Opus	1255	+3/-4	56101	Anthropic	Proprietary	2023/8
1	GPT-4-1106-preview	1254	+3/-3	65159	OpenAI	Proprietary	2023/4
2	GPT-4-0125-preview	1250	+3/-4	50923	OpenAI	Proprietary	2023/12
5	Bard (Gemini Pro)	1209	+5/-5	12468	Google	Proprietary	Online
5	Claude 3 Sonnet	1203	+3/-3	62056	Anthropic	Proprietary	2023/8
7	Command R+	1193	+4/-4	29437	Cohere	CC-BY-NC-4.0	2024/3
7	GPT-4-0314	1189	+4/-4	42925	OpenAI	Proprietary	2021/9
9	Claude 3 Haiku	1182	+3/-3	57727	Anthropic	Proprietary	2023/8
10	GPT-4-0613	1164	+3/-3	61520	OpenAI	Proprietary	2021/9
10	Mistral-Large-2402	1158	+3/-4	37650	Mistral	Proprietary	Unknown

Figure 1: Chatbot Arena Leaderboard Scores

As of this date, OpenAI's GPT-4 Turbo, Anthropic's Claude 3 Opus and Google's Gemini Pro are performing best based on these leaderboards. For the purpose of this work, these models will be taken into consideration as the current state of the art in LLMs based on these benchmarks.

2.2.3 GPT-4

GPT-4 is Open AI's latest LLM in their series of groundbreaking Generative Pre-trained Transformer (GPT) models. GPT-4 represents a significant leap in the capabilities of LLMs mainly due to its impressive benchmark performance. The model has shown an advanced level of understanding and generation of human-like text in comparison to other LLMs, as reflected in these benchmarks. As well as this performance improvement, GPT-4 has other significant feature improvements such as a multimodal input ability, allowing both images and text to be used as input and an increased context length of over 25,000 words of text. (19)These advancements allow the model to perform tasks like generating detailed descriptions from images and summarizing content from screenshots, as well as improving document analysis skills and an ability to maintain coherence overlong passages of text. Despite these advancements, GPT-4 has some limitations, such as hallucination potential (13). In order to mitigate the risks associated with hallucinations, results produced by GPT-4 should be verified before integration. For the purposes of this work, the version used will be GPT-4 turbo, which is the latest version available to the public as of April 2024.

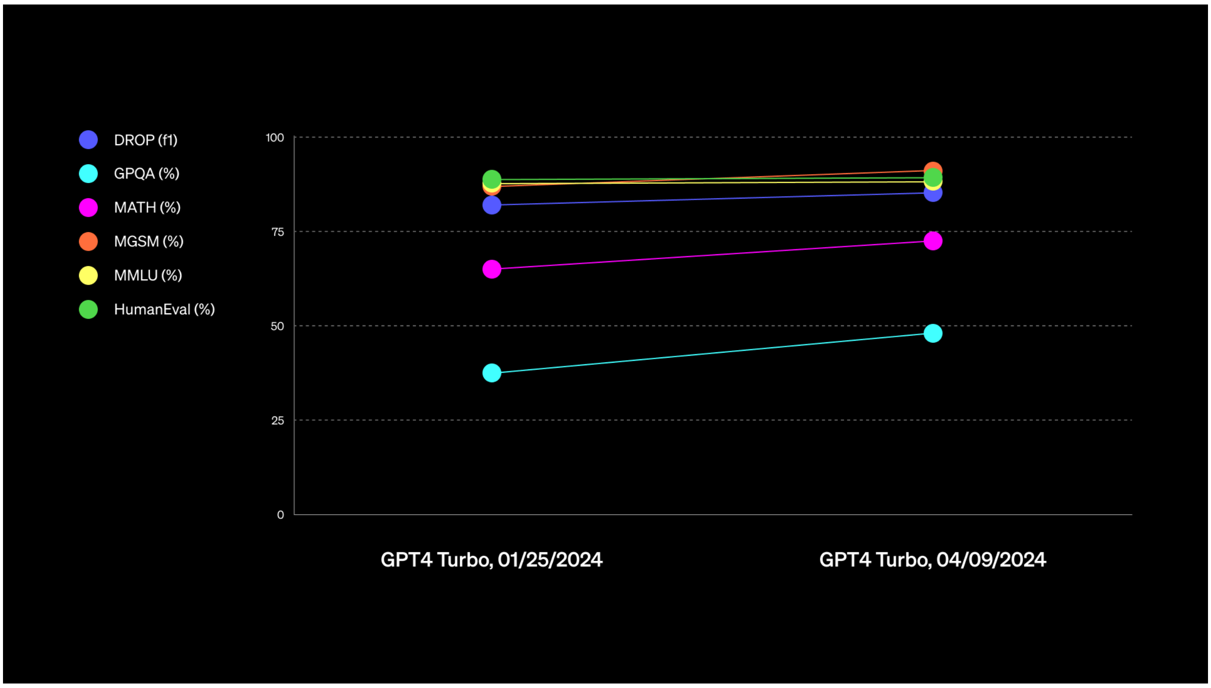
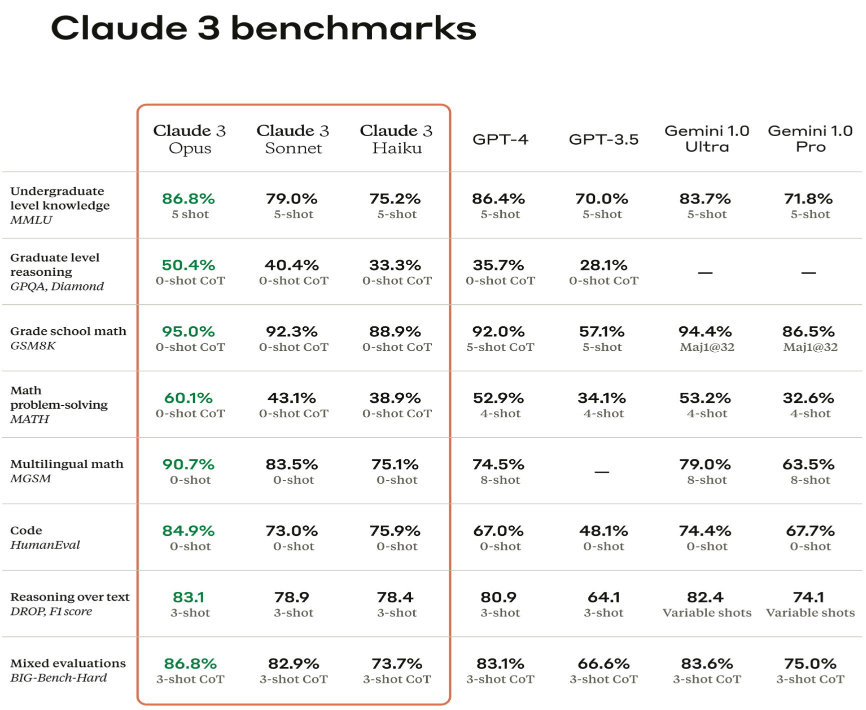


Figure 2: Latest GPT-4 Turbo improvements

2.2.4 Claudez 3 Opus

Claude 3 is a family of LLMs developed by Anthropic. This family consists of three models, Claude 3 Haiku, Claude 3 Sonnet and Claude 3 Opus. Each model offers a different balance of speed, cost and capability. For the purposes of this work, Claude 3 Opus is considered as it is designed to be the most capable of the family. Claude 3 Opus has an equally impressive benchmark performance as seen in Figure 3, outperforming GPT-4 in several benchmarks, such as MMLU[10], HumanEval[11] and DROP (4). The performance of Claude 3 Opus in these benchmarks is a suggestion of a high level of comprehension, which is important for this work. Claude 3 Opus also includes features such as multimodal inputs and a 200,000 token context window. It is important to note here that this is compared to GPT-4 and not GPT-4 Turbo, which is the latest version of GPT-4. Claude 3 Opus and GPT-4 Turbo show similar levels of performance as of April 2024.



Claude 3 benchmarks

	Claude 3 Opus	Claude 3 Sonnet	Claude 3 Haiku	GPT-4	GPT-3.5	Gemini 1.0 Ultra	Gemini 1.0 Pro
Undergraduate level knowledge *MMLU*	86.8% 5 shot	79.0% 5-shot	75.2% 5-shot	86.4% 5-shot	70.0% 5-shot	83.7% 5-shot	71.8% 5-shot
Graduate level reasoning *GPQA, Diamond*	50.4% 0-shot CoT	40.4% 0-shot CoT	33.3% 0-shot CoT	35.7% 0-shot CoT	28.1% 0-shot CoT	—	—
Grade school math *GSM8K*	95.0% 0-shot CoT	92.3% 0-shot CoT	88.9% 0-shot CoT	92.0% 5-shot CoT	57.1% 5-shot	94.4% Maj1@32	86.5% Maj1@32
Math problem-solving *MATH*	60.1% 0-shot CoT	43.1% 0-shot CoT	38.9% 0-shot CoT	52.9% 4-shot	34.1% 4-shot	53.2% 4-shot	32.6% 4-shot
Multilingual math *MGSM*	90.7% 0-shot	83.5% 0-shot	75.1% 0-shot	74.5% 8-shot	—	79.0% 8-shot	63.5% 8-shot
Code *HumanEval*	84.9% 0-shot	73.0% 0-shot	75.9% 0-shot	67.0% 0-shot	48.1% 0-shot	74.4% 0-shot	67.7% 0-shot
Reasoning over text *DROP, F1 score*	83.1 3-shot	78.9 3-shot	78.4 3-shot	80.9 3-shot	64.1 3-shot	82.4 Variable shots	74.1 Variable shots
Mixed evaluations *BIG-Bench-Hard*	86.8% 3-shot CoT	82.9% 3-shot CoT	73.7% 3-shot CoT	83.1% 3-shot CoT	66.6% 3-shot CoT	83.6% 3-shot CoT	75.0% 3-shot CoT

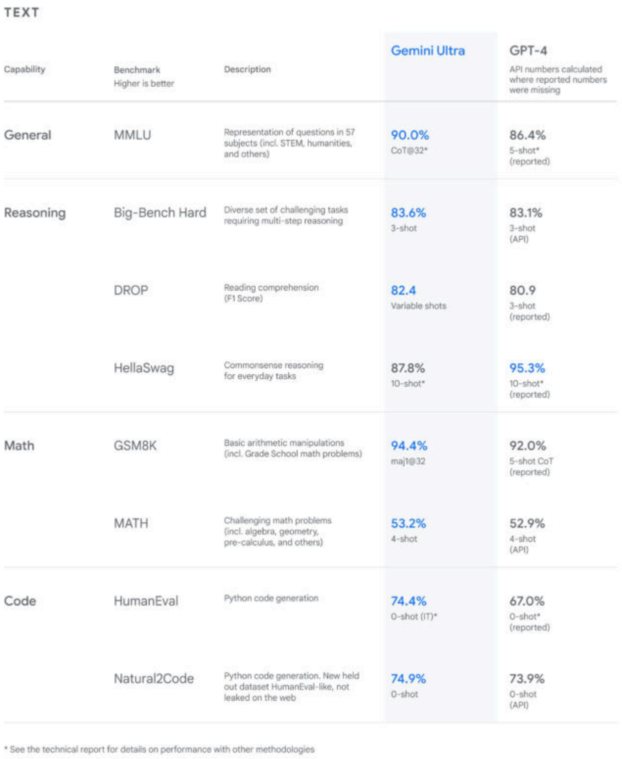
Figure 3: Claude 3 Benchmark Performance

[10] MMLU: https://paperswithcode.com/sota/multi-task-language-understanding-on-mmlu
[11] HumanEval: https://paperswithcode.com/sota/code-generation-on-humaneval

2.2.5 Gemini Ultra

Gemini is Google’s family of LLMs, developed by Google DeepMind and Google Research. It was built to be multimodal from the start, capable of understanding a multitude of different types of data, like text, audio, images and video. Gemini’s models include Ultra, Pro and nano. Gemini Ultra is the most advanced model of the lineup, It has shown impressive scores on benchmarks like MMLU, DROP and HumanEval as seen in Figure 4, surpassing previous models and in some cases even human experts[12]. Due to this impressive performance, the release of Gemini Ultra is highly anticipated, with expectations that it will offer a significant improvement on the current state of the art. However it is not currently widely available to the public, with Gemini Pro being the current available model,albeit with restrictions on use in Europe. Due to these difficulties in access, along with the consideration that Gemini Pro is outperformed by GPT-4 Turbo and Claude 3 Opus, Google’s Gemini models will not be considered for use in this work, however they stand to be a significant leap forward in the state of the art in this field once access restrictions are dealt with.

[12] Gemini vs Humans: https://medium.com/sofa-success-stories/google-says-its-gemini-ai-outperforms-both-gpt-4-and-expert-humans-27e743000815



TEXT

Capability	Benchmark Higher is better	Description	Gemini Ultra	GPT-4 API numbers calculated where reported numbers were missing
General	MMLU	Representation of questions in 57 subjects (incl. STEM, humanities, and others)	90.0% CoT@32*	86.4% 5-shot* (reported)
Reasoning	Big-Bench Hard	Diverse set of challenging tasks requiring multi-step reasoning	83.6% 3-shot	83.1% 3-shot (API)
	DROP	Reading comprehension (F1 Score)	82.4 Variable shots	80.9 3-shot (reported)
	HellaSwag	Commonsense reasoning for everyday tasks	87.8% 10-shot*	95.3% 10-shot* (reported)
Math	GSM8K	Basic arithmetic manipulations (incl. Grade School math problems)	94.4% maj1@32	92.0% 5-shot CoT (reported)
	MATH	Challenging math problems (incl. algebra, geometry, pre-calculus, and others)	53.2% 4-shot	52.9% 4-shot (API)
Code	HumanEval	Python code generation	74.4% 0-shot (IT)*	67.0% 0-shot* (reported)
	Natural2Code	Python code generation. New held out dataset HumanEval-like, not leaked on the web	74.9% 0-shot	73.9% 0-shot (API)

* See the technical report for details on performance with other methodologies

Figure 4: Gemini vs GPT-4

2.2.6 Perplexity AI

Perplexity AI[13] is a significant service for Language Model platforms, offering users access to powerful LLMs. As the field of artificial intelligence grows, access to these LLMs has become more important for researchers and developers. Perplexity AI provides access to multiple state of the art LLMs in one place. For the purposes of this research, access to GPT-4 was available through a premium subscription to chatGPT, while Gemini Ultra and Claude 3 Opus were not available in Europe at all. However by using Perplexity, access to GPT-4 Turbo and Claude 3 Opus was possible. Perplexity also allows for the creation of threads, where a model can be given a prompt for the entire thread, giving more context to each query. This is advantageous when building a system that prompts an LLM for the same output multiple times, like the one in this work.

[13] Perplexity AI: https://www.perplexity.ai/

2.2.7 Prompt Engineering

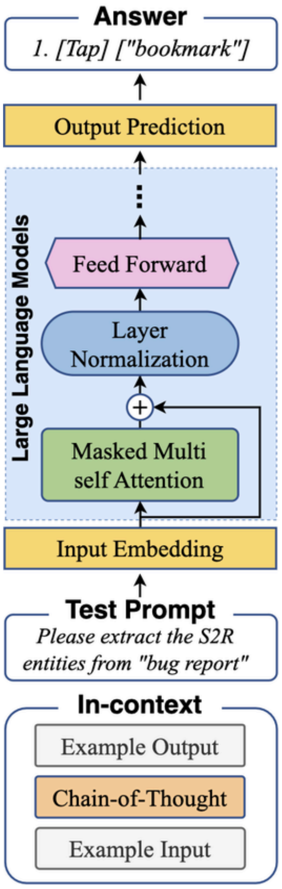


Figure 5: Prompt Engineering Process

Prompt engineering is the process of optimizing text input given to a LLM in order to get the model to perform a desired task by producing a high-quality and relevant output. This technique is vitally important in order to take full advantage of the capabilities of LLMs. The process of prompt engineering is essential for several reasons, namely to improve the model output, to adapt the model to the task and to mitigate bias.

As LLMs are trained on vast datasets and have the ability to perform a wide range of tasks, prompt engineering can help to improve the relevance and accuracy of the model's outputs by providing a clear and effective instruction to the model in order to align the task with the abilities of the model. (28) This can lead to an overall improvement in model output relative to the task given to the model.

Adapting the model to a given task is important due to the nature of different tasks. Tasks can require different formats, tones or styles of output. Prompt engineering allows for the customization of the response of the model to fit the use case of a given task. This can span from creative writing tasks, where prompt engineering is used to focus the model on a specific style of writing, to technical analysis where prompt engineering ensures that the models' analysis gives output in a particular format or language.

Due to the inherent nature of training LLMs on vast training data, biases can be unintentionally built into them. (18) By understanding how to use prompt engineering effectively, models can be steered away from such built in biases. This is also useful when attempting to avoid undesirable or unethical content from appearing in model output. (9)

In order to effectively conduct prompt engineering, various techniques can be implemented. These include in-context learning, few-shot learning and Chain-of-Thought (CoT) prompting.

In context learning is a process in which a model uses examples provided in the prompt itself in order to perform a task. This technique leverages the model's pre-training to draw on a wide array of concepts and apply them to the task without additional training. (3) By using examples in the prompt, the model can infer the tasks requirements and generate a response based on the pattern of the examples that are similar to those in its training data.

Few-Shot learning refers to the model's ability to learn new tasks or concepts from a small number of examples. In this case, the LLM is provided with a few examples that demonstrate the task at hand. (20) The examples serve as a support that helps the model to generalize new instances of the task from just a handful of examples. This reduces the need for large amounts of labeled data and improves the speed at which a model adapts to a new task.

Chain-of-Thought (CoT) prompting is a technique that encourages the LLM to articulate their reasoning process step by step when solving a problem. (27) By providing the model with examples that also explain how to arrive at the answer, CoT helps the model to generate more logically coherent responses. This technique has been shown to significantly improve model performance on arithmetic problems and commonsense reasoning. (12)

2.3 Ontologies

It is important to have a reusable way to categorize and exchange the information extracted by LLMs specifically in the context of the methodologies developed in this work. An ontology is a structured framework that captures elements in order to represent knowledge within a given domain. (15) This allows for formal logic to reason about the information and relationships defined by the ontologies. In this work, ontologies provide a standardized vocabulary for describing AI systems, specifically their risks, relationships and impacts.

2.3.1 Common Impact Data Standard CIDS

The Common Impact Data Standard (CIDS) is an ontology that represents impact as a change in outcomes, specifically using five dimensions to represent it. These dimensions are what, who, how much, contribution and risk (8). These dimensions are implemented in the ontology through the use of classes. Figure 6 shows the class diagram of CIDS.

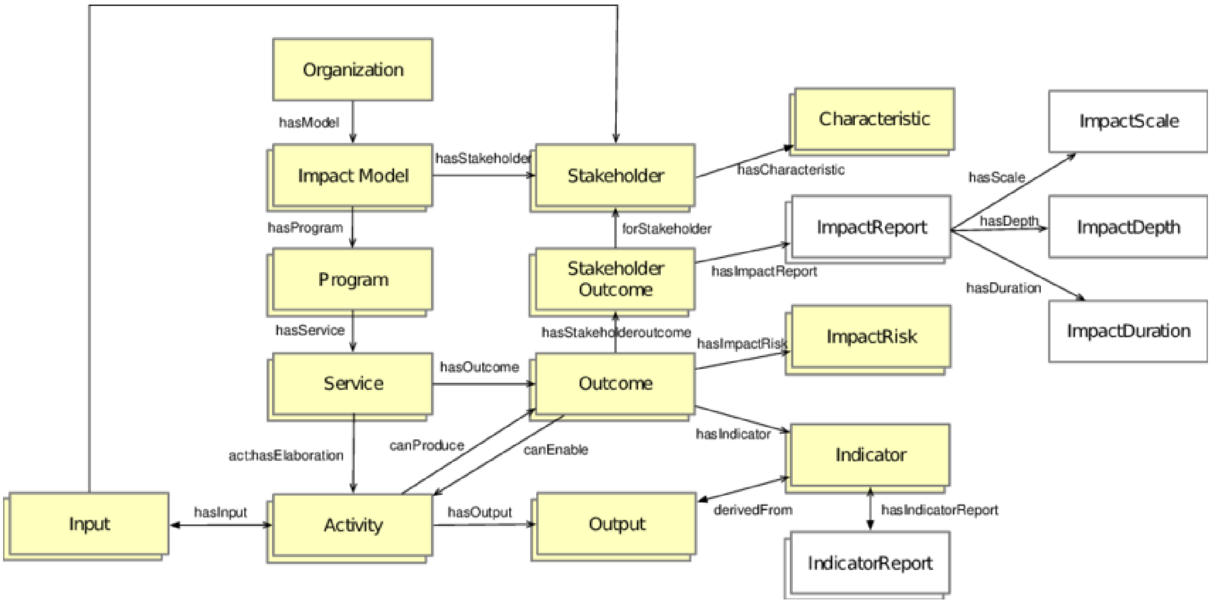


Figure 6: CIDS Core Classes

CIDS is designed to represent and communicate the impact of social purpose organizations (SPOs)[14] on society and the environment. These SPOs are entities like organizations and businesses that are focused on creating social and environmental value. This ontology is designed to represent the impact that SPOs have, and as a result can be used to model the impact that AI incidents have on society and the environment. This mapping is a useful way to use CIDS as an ontology for AI incidents as it is designed to represent impact on the same type of stakeholder, namely society and the environment, as AI incidents tend to affect.

2.3.2 Vocabulary of AI Risks (VAIR)

The Vocabulary of AI Risks (VAIR) is an open vocabulary designed to assist with the documentation of AI risks by providing a common set of concepts. (10) It gives semantic specifications for cataloging AI risks. The framework is designed to facilitate the sharing of knowledge among stakeholders in the AI value chain, such as developers, regulators and users. VAIR enables a more structured and

[14] SPO: https://www.impacteurope.net/impact-glossary

comprehensive approach to AI risk assessment, while supporting both automation and integration.

2.3.3 AI Risk Ontology (AIRO)

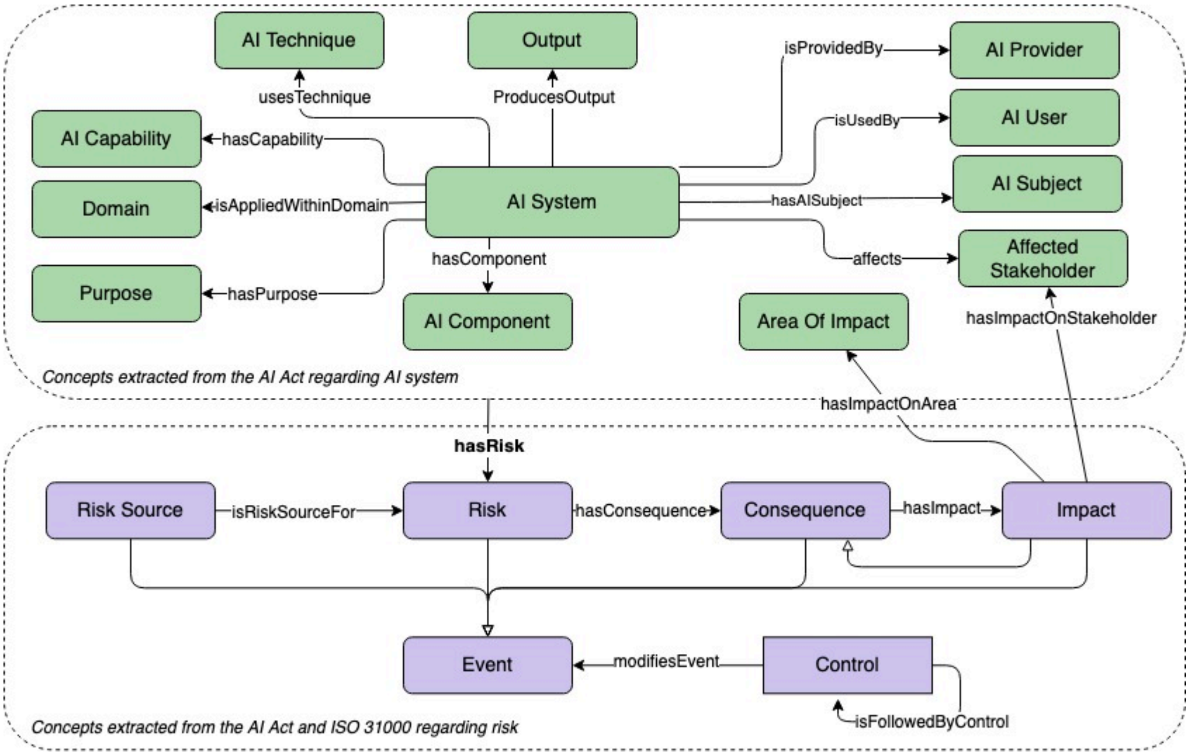


Figure 7: AIRO core concepts and relations

A subset of VAIR is the AI Risk Ontology (AIRO), designed to categorize AI systems in order to determine EU AI Act compliance. (2) It provides a formal representation of AI systems and their associated risks, based on the requirements of the EU AI Act. It is designed with the goal of assisting the relevant stakeholders in identifying high risk AI systems in order to conform with AI regulations. This ontology uses five dimensions to categorize risks, namely domain, purpose, capability, user and AI subject[15]. These dimensions are used to determine whether an AI system is to be deemed high risk. Figure 8 is used by AIRO in order to conduct this determination. If a system meets any of these conditions, it is considered high-risk unless the provider can demonstrate that “the output of the system (is) purely accessory in respect of the

[15] AIRO dimensions: https://delaramglp.github.io/airo/

relevant action or decision to be taken and is not therefore likely to lead to a significant risk to the health, safety or fundamental rights.” (6)

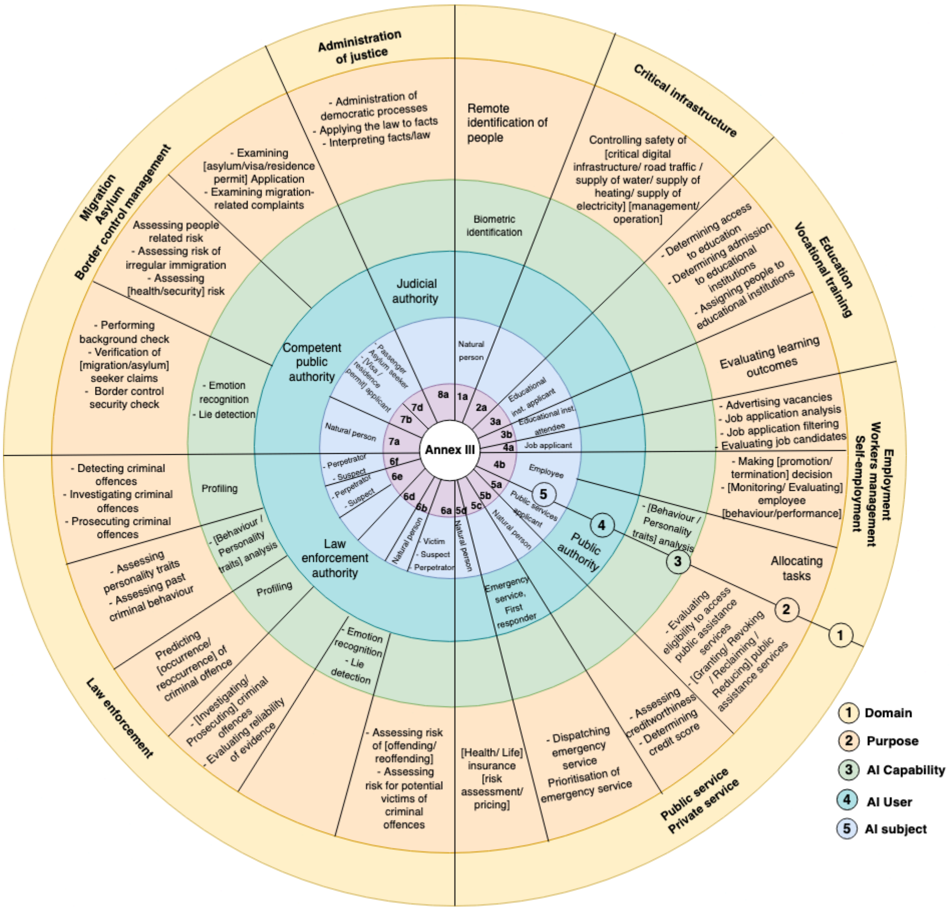

Annex III
Administration of justice
- Administration of democratic processes
- Applying the law to facts
- Interpreting facts/law
Judicial authority
Remote identification of people
Biometric identification
Natural person
Critical infrastructure
Controlling safety of [critical digital infrastructure/ road traffic / supply of water/ supply of heating/ supply of electricity] [management/ operation]
Education Vocational training
- Determining access to education
- Determining admission to educational institutions
- Assigning people to educational institutions
Evaluating learning outcomes
Educational inst. applicant
Educational inst. attendee
Employment Workers management Self-employment
- Advertising vacancies
- Job application analysis
- Job application filtering
- Evaluating job candidates
Job applicant
- Making [promotion/ termination] decision
- [Monitoring/ Evaluating] employee [behaviour/performance]
Employee
- [Behaviour / Personality traits] analysis
Allocating tasks
Public service Private service
- Evaluating eligibility to access public assistance services
- [Granting/ Revoking / Reclaiming / Reducing] public assistance services
Public services applicant
Public authority
- Assessing creditworthiness
- Determining credit score
Natural person
- Dispatching emergency service
Prioritisation of emergency service
Emergency service, First responder
[Health/ Life] insurance [risk assessment/ pricing]
Natural person
Law enforcement
- Assessing risk of [offending/ reoffending]
- Assessing risk for potential victims of criminal offences
- Victim
- Suspect
- Perpetrator
- Emotion recognition
- Lie detection
Natural person
- [Investigating/ Prosecuting] criminal offences
- Evaluating reliability of evidence
Predicting [occurrence/ reoccurrence] of criminal offence
Profiling
- Assessing personality traits
- Assessing past criminal behaviour
- [Behaviour / Personality traits] analysis
- Perpetrator
- Suspect
- Detecting criminal offences
- Investigating criminal offences
- Prosecuting criminal offences
Profiling
- Perpetrator
- Suspect
Law enforcement authority
Migration Asylum Border control management
- Performing background check
- Verification of [migration/asylum] seeker claims
- Border control security check
- Emotion recognition
- Lie detection
Natural person
Assessing people related risk
- Assessing risk of irregular immigration
- Assessing [health/security] risk
- Examining [asylum/visa/residence permit] Application
- Examining migration-related complaints
Competent public authority
- Passenger
Asylum seeker
- [Visa / residence permit] applicant
8a
1a
2a
3a
3b
4a
4b
5a
5b
5c
5d
6a
6b
6d
6e
6f
7a
7b
7d
5
4
3
2
1
1 Domain
2 Purpose
3 AI Capability
4 AI User
5 AI subject

Figure 8: Description of Annex III from the EU AI Act high-risk conditions using the dimensions of AIRO

2.4 Current Gaps in Research

Due to the unprecedented speed that AI systems and technologies are developing at, there are significant gaps in the AI incident reporting field. It can be seen that there is a lack of standardized AI incident reporting. Although repositories like AIAAIC and AIID have made progress in cataloging incidents, their primary data source is news articles. (23) Organizations like OECD are working to address this by creating a standard framework for reporting incidents, but until projects like the AI

Incidents Monitor become publicly available, research must rely on these news article based sources.

As well as this, due to the reliance on benchmarking, the focus on the performance of AI technologies is primarily on task performance rather than considering the potential negative impacts of AI technologies. While benchmarks measure how well an AI system performs for a specific task when compared to a predefined test, they do not account for the risks or adverse effects that the integration of these systems could have on society, individuals or the environment. As a result, data regarding the real-world impacts of these AI systems remain unanalyzed, which makes mitigating the potential harms of these systems difficult.

2.5 Related Work

As the use of LLMs in AI incident annotation is relatively new, this section consists of what is available at the time of writing. The related work for this dissertation is focused around the use of ontologies to improve LLM performance, using LLMs to annotate textual data and to use LLMs to align ontologies.

(20) presents a methodology for utilizing an ontology-driven structured prompt system with ChatGPT. The resulting system, named OntoChatGPT enhances the performance of the chatbot by extracting entities from contexts, classifying them and generating the relevant responses.

(25) presents a system for applying LLMs to perform semantic annotation of text from a large corpus of legal documents of various types. This system is designed to integrate LLMs in the semantic annotation workflow of legal texts, which can be related to the work of this dissertation on annotating AI incident reports.

(22) discusses using LLMs to optimize the ontology matching process. It discusses using LLMs to understand the semantic interoperability between different ontologies by aligning their related entities. This work is related to the ontology alignment task seen in this dissertation.

2.6 Summary

This chapter aimed to give an overview of the current state of the art in artificial intelligence with a focus on incident reporting, the regulatory frameworks involved in Europe, generative AI and the semantic models used for documenting AI risks.

The chapter began by discussing the rapid development of AI technologies, specifically LLMs like GPT-4 , Claude 3 and Gemini Ultra. These new technologies have led to a need for regulation. In Europe, the AI Act has recently come into legislation, being the first of it’s kind and covering the entire European Union. It categorizes AI systems based on their risk level and imposes obligations on developers of high risk AI systems. The chapter then moved on to look at how these risks were determined, understood and mitigated. The tool used for this is the incident report. Their aim is to document failures and biases in AI systems. However there is a lack of standardized AI incident reporting, with repositories like AIAAIC and AIID, which catalog AI incidents primarily from news articles being the current main source of standardized reports. The chapter then looked at how LLMs are assessed, namely through public leaderboards that provide benchmarks for performance. GPT-4 Turbo and Claude 3 Opus were determined to be the highest performing models that are currently available to the public. Next, prompt engineering is examined as the primary method for optimizing the output of these LLMs in order to adapt them to the task, and mitigates biases. The task that these models would be focusing on comes in the form of ontology classification, which was the next section that this chapter examined. Semantic models used for documenting AI risks like the CIDS and AIRO ontologies were presented as tools to categorize and communicate AI impacts and risks. The chapter concludes with a discussion of the gaps in the current research, namely the need for a more comprehensive approach to evaluate AI systems that considers the potential negative impacts of AI technologies.

Chapter 3

Design

3.1 Design Methodology

The focus of this research is to investigate how semantic models can be leveraged to optimize the process of annotating AI incident reports, as well as to analyze the relationships and correlations between ontologies used for AI incident report classification. In order to implement this focus, the design methodology used is designed around the objectives set, as informed by the state of the art analysis. The primary data source for this research will be the AIAAIC repository. GPT-4 and Claude 3 will be the LLMs used in this work, as selected from public LLM leaderboards discussed in 2.2.2, with the models accessed through Perplexity AI. The CIDS and VAIR ontologies will be the primary frameworks used for cataloging the AI risks. The aim of this design process is to use LLMs in optimizing the process of annotating AI incident reports through ontologies, as it is an important step in improving on the slow process of human annotation. The design is also intended to facilitate the understanding of the relationships and correlations between different ontologies in order to provide valuable insights into developing AI systems safely, complying with policies and regulations like the EU AI Act and the ethics and safety concerns when integrating AI into critical domains such as healthcare (21) and law enforcement. (1)

3.2 Design Overview

This section will look at the workflow of the system, describe the use of LLMs for ontology annotation and RDF extraction, look at the ontology choice, explain the correlation evaluation process and give an overview of the main design decisions taken.

3.2.1 LLMs for Ontology Annotation

The design of the ontology annotation system using LLMs is centered around the goal of converting unstructured AI incident reports into a structured, machine readable Resource Description Framework (RDF) format. The system intends to automate the extraction of information and its encapsulation into RDFs which are aligned with the chosen ontology. The proposed system for this work intends to implement LLMs for ontology annotation by implementing a system which takes unstructured incident reports as input,, applies a LLM to annotate the reports and outputs a structured RDF. This design is further outlined below in 3.3.1 Workflow of the system.

3.2.2 Workflow of the System

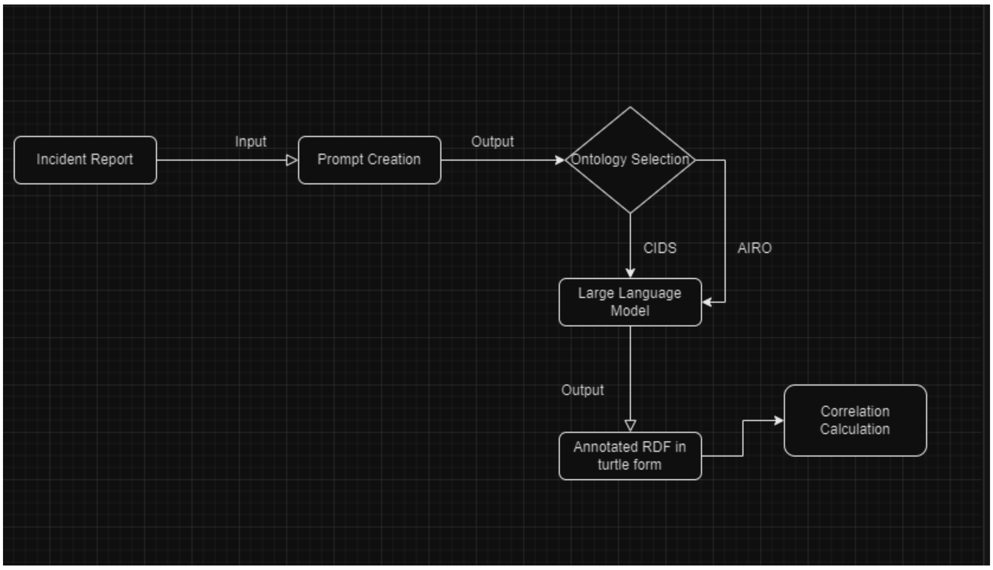


Figure 9: System workflow diagram

“You are designed to convert Text to RDF format when given an ontology to use as a guide”
Listing 1: Thread Prompt

The workflow of the system can be seen in Fig 9 First, an incident report is taken from the AIAAIC repository. This incident report is processed into a promptable form that can be inputted into an LLM. For the purposes of this work, Perplexity.AI's threads feature was applied to allow for ease of model choice. This feature allows for the creation of a thread using a selected model, in this case GPT-4 Turbo or Claude 3 Opus. Once the thread has been selected, a prompt can be applied to it. Listing X shows the prompt that was applied. This gave the LLM the relative context needed. The next prompt contained the ontology selected by the user, along with instructions on how to interpret it in order to annotate the incident reports efficiently. Next the user can attach a file containing the incident reports that are to be annotated, along with a prompt asking the LLM to perform the annotations. The output is an annotated Resource Description Framework (RDF) containing the annotated incidents that can be used for the correlation calculations.

3.3.1 Ontology Design for CIDS

CIDS is an ontology that represents impact, and provides a structured framework for capturing elements to represent knowledge within the domain of AI impact. The ontology consists of multiple classes relating to the five dimensions discussed in 2.3.1. The CIDS ontology is a large model with many classes and properties. In order to effectively implement CIDS for this work, a subset of CIDS was considered for use in annotation. This was partially due to the relevancy of the classes as CIDS was not originally designed for AI incident reporting as discussed in 2.3.1, The second reason for this subset was due to the context limits of the LLMs in question. Currently, it is not feasible to use the entirety of CIDS for the annotation task when using the publicly available versions of GPT-4 Turbo and Claude 3 due to their token limits. A possible workaround would be to implement the APIs of these models and send the ontology piece by piece to the LLM, however this would be very expensive and unlikely to add enough value to the work to justify it. As a result, the subset of CIDS considered can be seen in listing 2. These listed properties encapsulate the dimensions of the ontology, as well as being relevant to AI incidents.

- hasName (title): A title for the stakeholder as a string.
- hasDescription: A description of this risk.
- fromPerspectiveOf: Identifies the Stakeholder who is determining the importance of the Impact.
- forStakeholder: Identifies the Stakeholder affected.
- forOutcome: Identifies the general outcome of the incident.
- hasImportance: Specifies the nature of the importance. One of {“high importance”, “moderate importance”, “neutral”, “unimportant”}. assess importance level based on your perceived level of relative seriousness of the incident.
- intendedImpact: Identifies the intended direction of the change – note that ImpactReport captures the actual direction.
- hasCatchmentArea: Specifies the regional span of the stakeholders.
- hasStakeholderCharacteristic: Specifies characteristics of the stakeholder
- hasLikelihood: Identifies the likelihood that the incident will happen again.
- hasConsequence: Identifies the degree of impact the risk could have.
- hasMitigation: A string that specifies a mitigation plan or references a document.

Listing 2: relevant CIDS properties

3.3.2 Ontology Design for AIRO

The VAIR ontology is designed to categorize and analyze the risks associated with AI systems. This ontology provides a comprehensive framework for detailing the various aspects of AI risks. Similarly to CIDS, the implementation of VAIR has been tailored to suit the needs of this work. Conveniently VAIR has a defined subclass, the AI Risk Ontology(AIRO) which is useful for this work. AIRO is designed to categorize high risk AI systems, as explained in 2.3.2. Figure 10 shows the dimensions of AIRO, which serve as the foundation for defining the properties of the ontology. In order to implement AIRO, the properties of AIRO[16] were inputted to GPT-4 Turbo along with a prompt to output label:definition pairings similar to those seen in 3.3.1. These pairings were used as the definition of AIRO for this work.

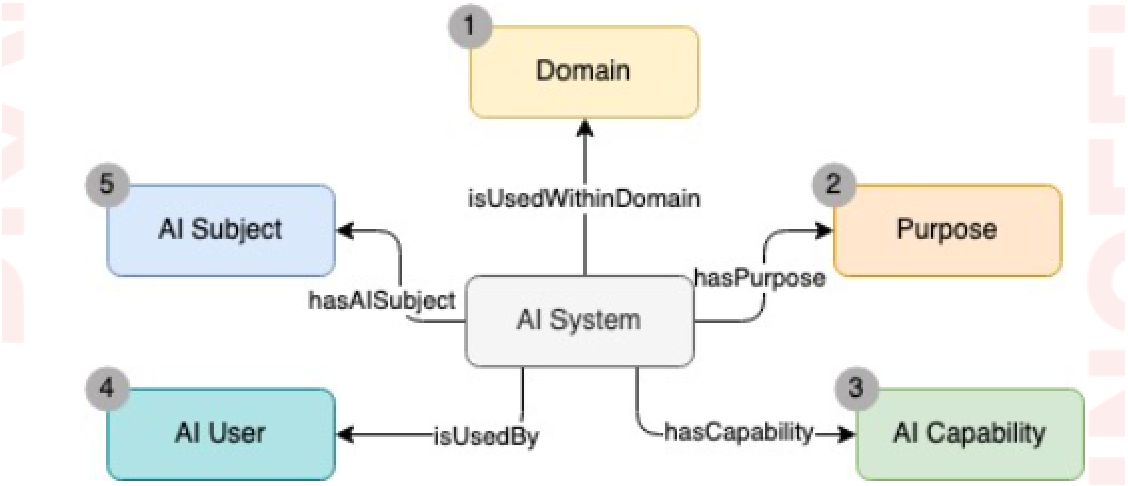


Figure 10: What is a high risk system[17]

- is applied within domain: Specifies the domain an AI system is used within.
- has purpose: Indicates the intended purpose of an AI system.
- has capability: Specifies capabilities implemented within an AI system to materialize its purposes.
- uses technique: Indicates the AI techniques used in an AI system.
- produces output: Specifies an output generated by an AI system.
- has component: Indicates components of an AI system.
- has risk: Indicates risks associated with an AI system, an AI component, etc.One of {“unacceptable,high risk, “moderate risk, low risk}. assess risk level based on your percieved level of relative seriousness of the incident.
- is risk source for: Specifies risks caused by materialization of a risk source
- has consequence: Specifies consequences caused by an incident
- has impact: Specifies impacts caused by materialization of an incident
- has stakeholder: Specifies stakeholders that are affected by an incident
- mitigation: specifies a mitigation plan

[16] AIRO dimensions: https://delaramglp.github.io/airo/
[17] AIRO high risk diagram: https://raw.githubusercontent.com/DelaramGlp/airo/main/figures/airo-annexIII-concepts.jpeg

- has documentation: Indicates documents related to an entity, e.g. AI system
- is provided by: Indicates provider of an AI system
- is used by: Indicates user of an AI system
- has AI subject: Indicates subject of an AI system
- has version: Indicates the version of an AI system
- has severity: Indicates severity of an incident
- has likelihood: Indicates the probability of reocurrence of an incident
- has lifecycle phase: Indicates the AI system's lifecycle phase

Listing 3: AIRO properties

The lists of properties seen in 3.3.1 and 3.3.2. represent the definitions and designs of CIDS and AIRO used for this work.

3.3.3 Evaluation System

For each ontology, the list of relevant properties is used to represent each incident report for the respective ontology. For example, when a given incident is fed into the LLM along with the ontology, the LLM’s job is to extract relevant information that relates to the definition of each property from the incident and assign the property label to the extracted information in RDF form. This gives an output of RDFs for each incident, for each ontology. These lists of RDFs can be examined both relevant to each other, as well as in pairs where each pair represents one incident categorized into each ontology. This gives two outputs to analyze, the first being the lists of RDFs for each ontology, per incident, and the second being pairs of RDFs, where each pair contains two RDFs, one for each ontology, per incident. Once these datasets are outputted, an evaluation process will occur.

For the first output, a CSV file is created which represents each list and the presence or absence of a description for each property. By evaluating these lists, conclusions can be drawn regarding the research objective, of how LLMs can be leveraged to optimize the annotation process for AI incident reports, and how effective GPT-4 Turbo and Claude 3 are respectively at the annotation task.

For the second output, a mapping will occur between the properties of each ontology. Once the properties have been mapped, an analysis will be carried out to understand how similar each ontology annotates the pair. This CSV file will contain the similarity scores for each pair of properties. By evaluating these scores, conclusions can be drawn regarding the research objective of analyzing the relationships and correlations between ontologies for AI incident report classification by looking at similarities and differences between the annotations of incidents by the same LLM for each pair of ontologies.

3.4 Main Design Decisions

The main design decisions outlined in this chapter revolve around the implementation of LLMs and ontologies to enhance the process of annotating AI incident reports. These decisions aim to create a system that achieves all of the research objectives outlined in 1.3 in order to answer the research question of this work. These main decisions can be seen below:

Ontology Subset Implementation

The decision to use subsets of CIDS and VAIR was an important decision in this work. By implementing these subsets, practical limitations of context limits of the LLMs and relevancy of classes to the task were mitigated while still maintaining the usefulness of the ontologies.

Workflow Design

The design of the workflow was planned to allow flexibility in the ontology and LLM used. By implementing Perplexity's thread feature over API calls, the system is able to seamlessly switch between the LLM used for annotation, while also being cost efficient, removing the need for API calls to multiple different LLM services.

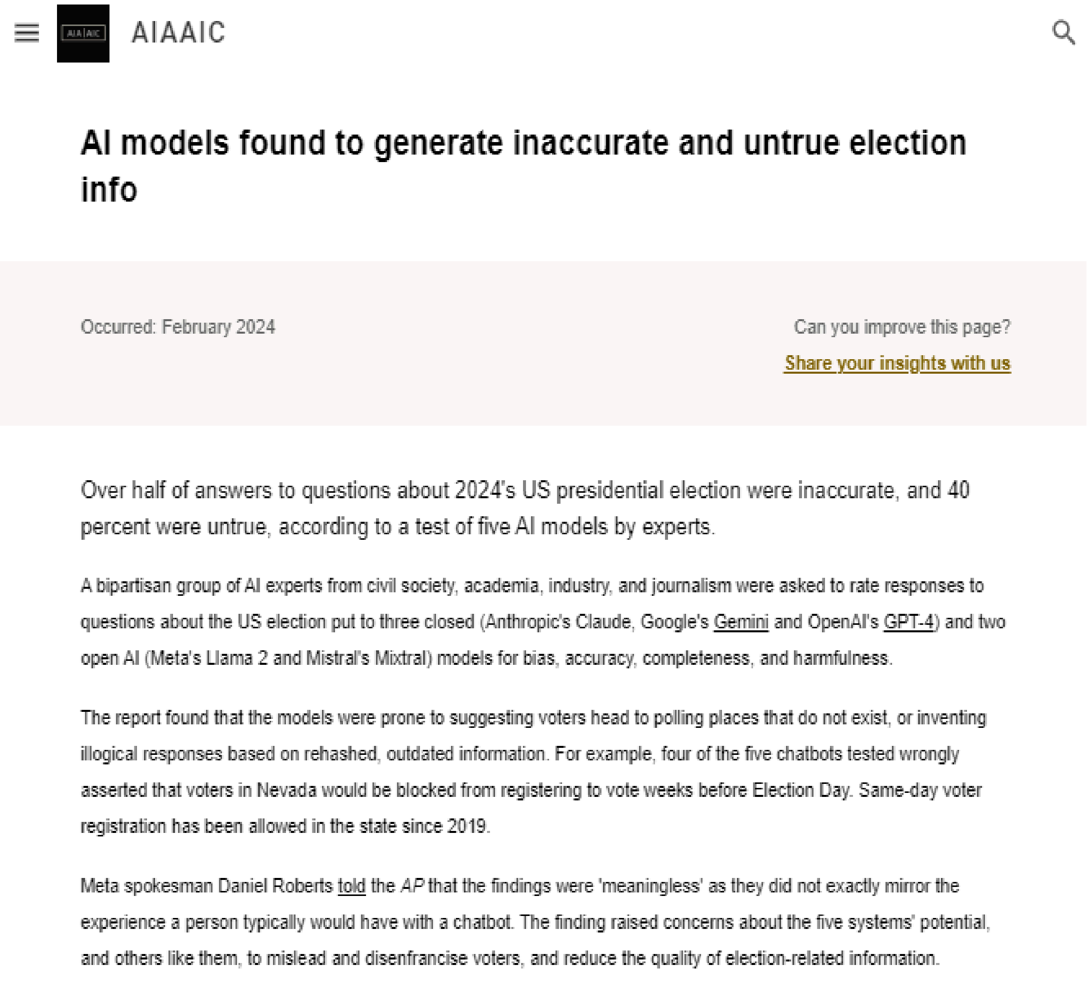
Choice of Evaluation System

The choice of evaluation system to analyze the annotated RDFs places an emphasis on dataset analysis. As the data outputted by the system is mostly qualitative and test-based, it is difficult to effectively evaluate it. The proposed system intends to implement a system to analyze the data according to the main research objectives of this work.

Chapter 4

Implementation

4.1 Data pre processing



AI models found to generate inaccurate and untrue election info

Occurred: February 2024

Can you improve this page?
Share your insights with us

Over half of answers to questions about 2024's US presidential election were inaccurate, and 40 percent were untrue, according to a test of five AI models by experts.

A bipartisan group of AI experts from civil society, academia, industry, and journalism were asked to rate responses to questions about the US election put to three closed (Anthropic's Claude, Google's Gemini and OpenAI's GPT-4) and two open AI (Meta's Llama 2 and Mistral's Mixtral) models for bias, accuracy, completeness, and harmfulness.

The report found that the models were prone to suggesting voters head to polling places that do not exist, or inventing illogical responses based on rehashed, outdated information. For example, four of the five chatbots tested wrongly asserted that voters in Nevada would be blocked from registering to vote weeks before Election Day. Same-day voter registration has been allowed in the state since 2019.

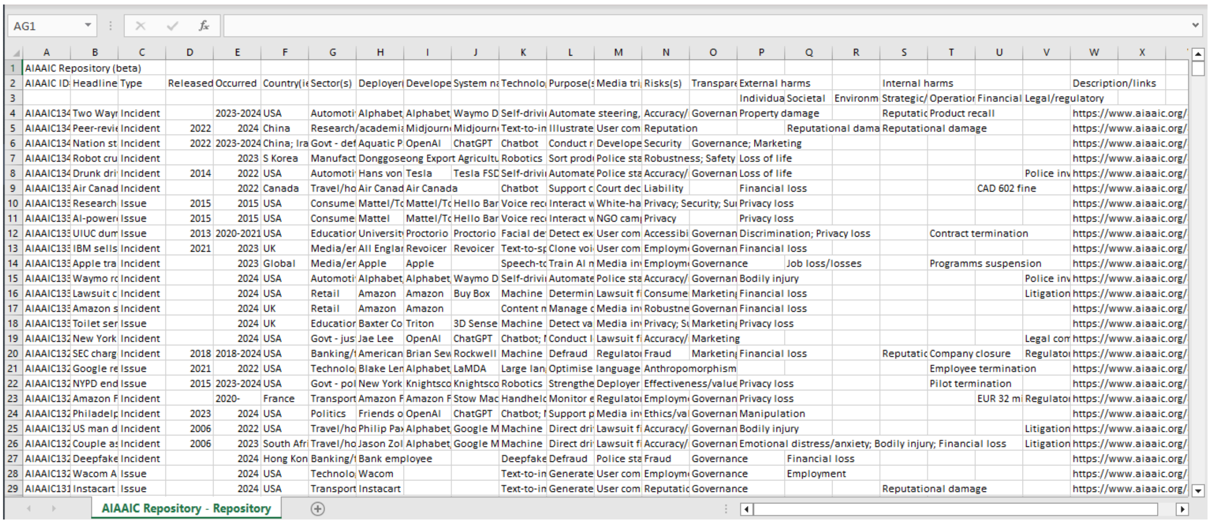
Meta spokesman Daniel Roberts told the *AP* that the findings were 'meaningless' as they did not exactly mirror the experience a person typically would have with a chatbot. The finding raised concerns about the five systems' potential, and others like them, to mislead and disenfrancise voters, and reduce the quality of election-related information.

Figure 11: Incident report example from AIAAIC repository [18]

[18] AIAAIC Inccident link:https://www.aiaaic.org/aiaaic-repository/ai-algorithmic-and-automation-incidents/ai-models-found-to-generate-inaccurate-and-untrue-election-info

Figure 11 is an example of an AI incident from the AIAAIC repository. In order to use reports like the one shown in Figure 11, data preprocessing is required. The following steps were taken in order to perform this preprocessing:

1. Incident Report Database Processing



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
1	AIAAIC Repository (beta)																							
2	AIAAIC ID	Headline	Type	Released	Occurred	Country(i	Sector(s)	Deployer	Develope	System na	Technolo	Purpose(s	Media tri	Risks(s)	Transpare	External harms			Internal harms				Description/links	
3																Individua	Societal	Environm	Strategic/	Operation	Financial	Legal/regulatory		
4	AIAAIC134	Two Wayr	Incident		2023-2024	USA	Automoti	Alphabet,	Alphabet,	Waymo D	Self-drivi	Automate	steering,	Accuracy/	Governan	Property damage			Reputatic	Product recall			https://www.aiaaic.org/	
5	AIAAIC134	Peer-revie	Incident	2022	2024	China	Research/academi	Midjourn	Midjourn	Text-to-in	Illustrate	User com	Reputation				Reputational dama		Reputational damage				https://www.aiaaic.org/	
6	AIAAIC134	Nation st	Incident	2022	2023-2024	China; Ira	Govt - def	Aquatic P	OpenAI	ChatGPT	Chatbot	Conduct r	Develope	Security	Governance; Marketing								https://www.aiaaic.org/	
7	AIAAIC134	Robot cru	Incident		2023	S Korea	Manufact	Donggoseong Export		Agricultu	Robotics	Sort prod	Police sta	Robustness; Safety		Loss of life							https://www.aiaaic.org/	
8	AIAAIC134	Drunk dri	Incident	2014	2022	USA	Automoti	Hans von	Tesla	Tesla FSD	Self-drivi	Automate	Police sta	Accuracy/	Governan	Loss of life						Police inv	https://www.aiaaic.org/	
9	AIAAIC133	Air Canad	Incident		2022	Canada	Travel/ho	Air Canad	Air Canada		Chatbot	Support c	Court dec	Liability		Financial loss					CAD 602 fine		https://www.aiaaic.org/	
10	AIAAIC133	Research	Issue	2015	2015	USA	Consume	Mattel/To	Mattel/To	Hello Bar	Voice rec	Interact w	White-ha	Privacy; Security; Su		Privacy loss							https://www.aiaaic.org/	
11	AIAAIC133	AI-powere	Issue	2015	2015	USA	Consume	Mattel	Mattel/To	Hello Bar	Voice rec	Interact w	NGO cam	Privacy		Privacy loss							https://www.aiaaic.org/	
12	AIAAIC133	UIUC dum	Issue	2013	2020-2021	USA	Educatior	Universit	Proctorio	Proctorio	Facial de	Detect ex	User com	Accessibi	Governan	Discrimination; Privacy loss				Contract termination			https://www.aiaaic.org/	
13	AIAAIC133	IBM sells	Incident	2021	2023	UK	Media/er	All Englar	Revoicer	Revoicer	Text-to-sp	Clone voi	User com	Employm	Governan	Financial loss							https://www.aiaaic.org/	
14	AIAAIC133	Apple tra	Incident		2023	Global	Media/er	Apple	Apple		Speech-tc	Train AI n	Media in	Employm	Governance		Job loss/losses			Programms suspension			https://www.aiaaic.org/	
15	AIAAIC133	Waymo rc	Incident		2024	USA	Automoti	Alphabet,	Alphabet,	Waymo D	Self-drivi	Automate	Police sta	Accuracy/	Governan	Bodily injury						Police inv	https://www.aiaaic.org/	
16	AIAAIC133	Lawsuit c	Incident		2024	USA	Retail	Amazon	Amazon	Buy Box	Machine	Determin	Lawsuit fi	Consume	Marketing	Financial loss						Litigation	https://www.aiaaic.org/	
17	AIAAIC133	Amazon s	Incident		2024	UK	Retail	Amazon	Amazon		Content n	Manage c	Media in	Robustne	Governan	Financial loss							https://www.aiaaic.org/	
18	AIAAIC133	Toilet ser	Issue		2024	UK	Educatior	Baxter Co	Triton	3D Sense	Machine	Detect va	Media in	Privacy; Su	Marketing	Privacy loss							https://www.aiaaic.org/	
19	AIAAIC132	New York	Incident		2024	USA	Govt - jus	Jae Lee	OpenAI	ChatGPT	Chatbot; I	Conduct l	Lawsuit fi	Accuracy/	Marketing							Legal com	https://www.aiaaic.org/	
20	AIAAIC132	SEC charg	Incident	2018	2018-2024	USA	Banking/f	American	Brian Sew	Rockwell	Machine	Defraud	Regulator	Fraud	Marketing	Financial loss			Reputatic	Company closure		Regulator	https://www.aiaaic.org/	
21	AIAAIC132	Google re	Issue	2021	2022	USA	Technolo	Blake Len	Alphabet,	LaMDA	Large lan	Optimise	language	Anthropomorphism						Employee termination			https://www.aiaaic.org/	
22	AIAAIC132	NYPD end	Issue	2015	2023-2024	USA	Govt - pol	New York	Knightsco	Knightsco	Robotics	Strengthe	Deployer	Effectiveness/value		Privacy loss				Pilot termination			https://www.aiaaic.org/	
23	AIAAIC132	Amazon F	Incident		2020-	France	Transport	Amazon F	Amazon F	Stow Mac	Handhelc	Monitor e	Regulator	Employm	Governan	Privacy loss					EUR 32 mi	Regulator	https://www.aiaaic.org/	
24	AIAAIC132	Philadelp	Incident	2023	2024	USA	Politics	Friends o	OpenAI	ChatGPT	Chatbot; I	Support p	Media in	Ethics/val	Governan	Manipulation							https://www.aiaaic.org/	
25	AIAAIC132	US man d	Incident	2006	2022	USA	Travel/ho	Philip Pa	Alphabet,	Google M	Machine	Direct dri	Lawsuit fi	Accuracy/	Governan	Bodily injury						Litigation	https://www.aiaaic.org/	
26	AIAAIC132	Couple as	Incident	2006	2023	South Afri	Travel/ho	Jason Zol	Alphabet,	Google M	Machine	Direct dri	Lawsuit fi	Accuracy/	Governan	Emotional distress/anxiety; Bodily injury; Financial loss						Litigation	https://www.aiaaic.org/	
27	AIAAIC132	Deepfake	Incident		2024	Hong Kon	Banking/f	Bank employee			Deepfake	Defraud	Police sta	Fraud	Governance		Financial loss						https://www.aiaaic.org/	
28	AIAAIC132	Wacom A	Issue		2024	USA	Technolo	Wacom			Text-to-in	Generate	User com	Employm	Governance		Employment						https://www.aiaaic.org/	
29	AIAAIC131	Instacart	Issue		2024	USA	Transport	Instacart			Text-to-in	Generate	User com	Reputatic	Governance				Reputational damage				https://www.aiaaic.org/	

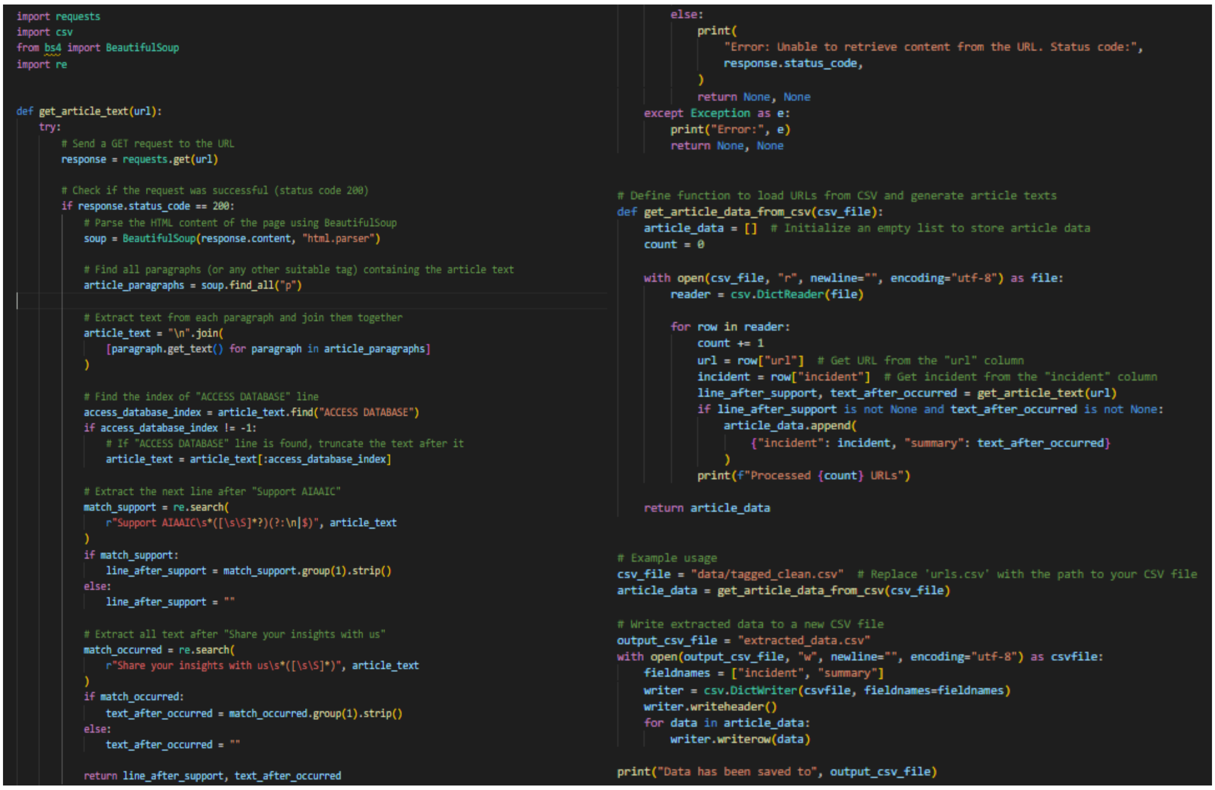
Figure 12: AIAAIC repository dataset[19]

The AIAAIC database was downloaded as an excel spreadsheet from the AIAAIC repository website[20]. Figure 12 shows the layout of the spreadsheet. For the purposes of this work, the dataset was reduced to two columns, namely the headline and description/links columns. The purpose of this is to have each line in the spreadsheet represent an incident and the link to the incident report. This spreadsheet was then exported as a csv to be used for URL Scraping.

[19] Github: https://github.com/paddyocallaghan/dissertation

[20] AIAAIC repository:https://www.aiaaic.org/aiaaic-repository

2. URL Scraping



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


def get_article_text(url):
    try:
        # Send a GET request to the URL
        response = requests.get(url)

        # Check if the request was successful (status code 200)
        if response.status_code == 200:
            # Parse the HTML content of the page using BeautifulSoup
            soup = BeautifulSoup(response.content, "html.parser")

            # Find all paragraphs (or any other suitable tag) containing the article text
            article_paragraphs = soup.find_all("p")

            # Extract text from each paragraph and join them together
            article_text = "\n".join(
                [paragraph.get_text() for paragraph in article_paragraphs]
            )

            # Find the index of "ACCESS DATABASE" line
            access_database_index = article_text.find("ACCESS DATABASE")
            if access_database_index != -1:
                # If "ACCESS DATABASE" line is found, truncate the text after it
                article_text = article_text[:access_database_index]

            # Extract the next line after "Support AIAAIC"
            match_support = re.search(
                r"Support AIAAIC\s*([\s\S]*?)(?:\n|$)", article_text
            )
            if match_support:
                line_after_support = match_support.group(1).strip()
            else:
                line_after_support = ""

            # Extract all text after "Share your insights with us"
            match_occurred = re.search(
                r"Share your insights with us\s*([\s\S]*)", article_text
            )
            if match_occurred:
                text_after_occurred = match_occurred.group(1).strip()
            else:
                text_after_occurred = ""

            return line_after_support, text_after_occurred
        else:
            print(
                "Error: Unable to retrieve content from the URL. Status code:",
                response.status_code,
            )
            return None, None
    except Exception as e:
        print("Error:", e)
        return None, None


# Define function to load URLs from CSV and generate article texts
def get_article_data_from_csv(csv_file):
    article_data = []  # Initialize an empty list to store article data
    count = 0

    with open(csv_file, "r", newline="", encoding="utf-8") as file:
        reader = csv.DictReader(file)

        for row in reader:
            count += 1
            url = row["url"]  # Get URL from the "url" column
            incident = row["incident"]  # Get incident from the "incident" column
            line_after_support, text_after_occurred = get_article_text(url)
            if line_after_support is not None and text_after_occurred is not None:
                article_data.append(
                    {"incident": incident, "summary": text_after_occurred}
                )
            print(f"Processed {count} URLs")

    return article_data


# Example usage
csv_file = "data/tagged_clean.csv"  # Replace 'urls.csv' with the path to your CSV file
article_data = get_article_data_from_csv(csv_file)

# Write extracted data to a new CSV file
output_csv_file = "extracted_data.csv"
with open(output_csv_file, "w", newline="", encoding="utf-8") as csvfile:
    fieldnames = ["incident", "summary"]
    writer = csv.DictWriter(csvfile, fieldnames=fieldnames)
    writer.writeheader()
    for data in article_data:
        writer.writerow(data)

print("Data has been saved to", output_csv_file)
```

Listing 4: URL Scraping Code[21]

URL Scraping: Listing 4 shows the code used to scrape the URLs from the csv exported from the AIAAIC repository. The get_article_text() function extracts the body of the AI incident report. This text is then saved in a csv file where the first column is the headline, and the second is the description. This file is then cleaned using the clean_data() function, which removes any row with empty data in either the headline or description column. At this point, the dataset is ready to be annotated using an LLM.

[21] Github for code: https://github.com/paddyocallaghan/dissertation

4.2 Ontology preparation

In order to use both the CDS and AIRO ontologies as defined in 3.3.1 and 3.3.2, their definitions needed to be converted to a prompt. For AIRO, the properties and descriptions from 3.3.2 were fed to GPT-4 Turbo and along with the following prompt:

“take this list of properties and definitions and refine it into a readable format, where first you give the name of the property, and secondly the definition”

The list of properties and definitions were then used to annotate the first 10 incident reports. Using the outputs of this annotation, the list of properties and definitions were refined to the listing 3 in 3.3.2. Once this list was obtained, a thread was created in Perplexity for AIRO using the prompt from listing 5. The list from 3.3.1 that defines the CIDS ontology was then used with the same initial prompt to set up a separate Perplexity thread for it.

The final task in the ontology preparation process was to create a mapping of properties from the two ontologies that were to be used. For this process GPT-4 turbo was used. by inputting the two lists of properties and the following prompt:

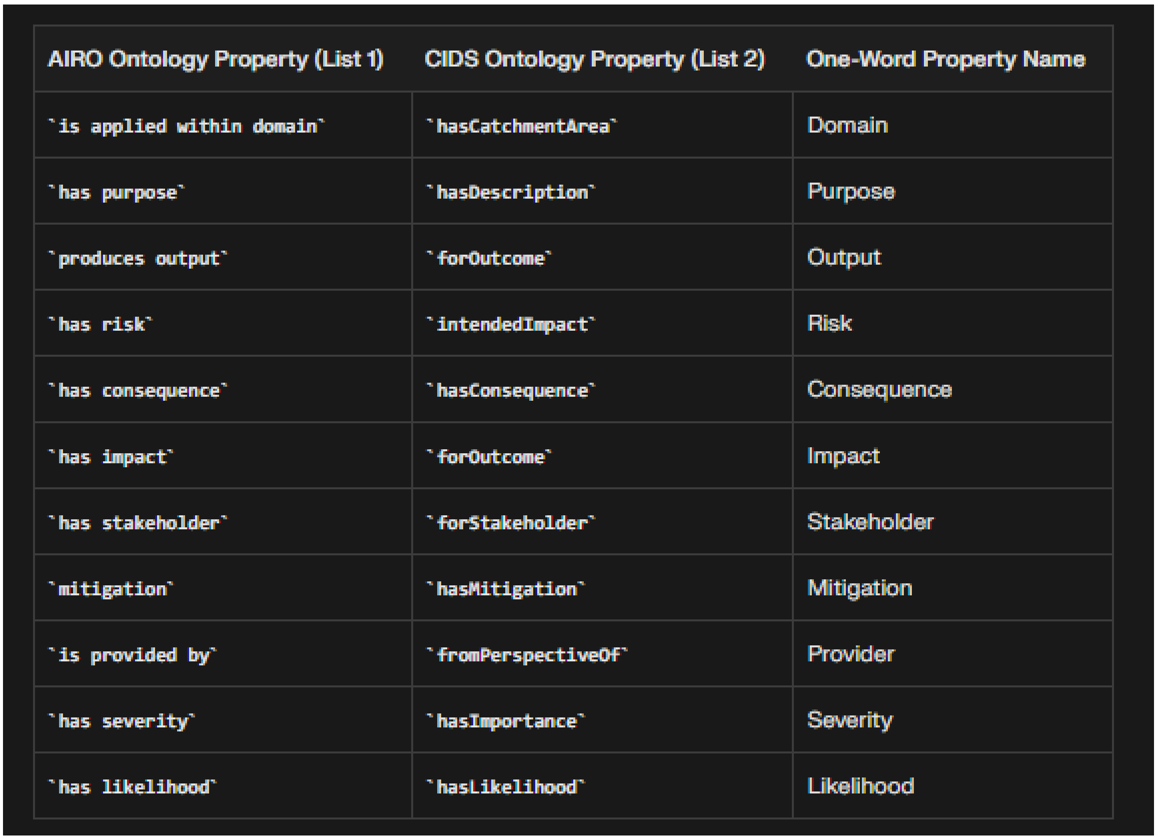
create a one to one mapping between these two lists of properties:

…
Lists of CIDS and AIRO properties
…

Output the mapping as a table, and add any properties that do not map one to one at the end of the table. Add a third column called property name to the mapped property table and assign an appropriate one word name to each pair

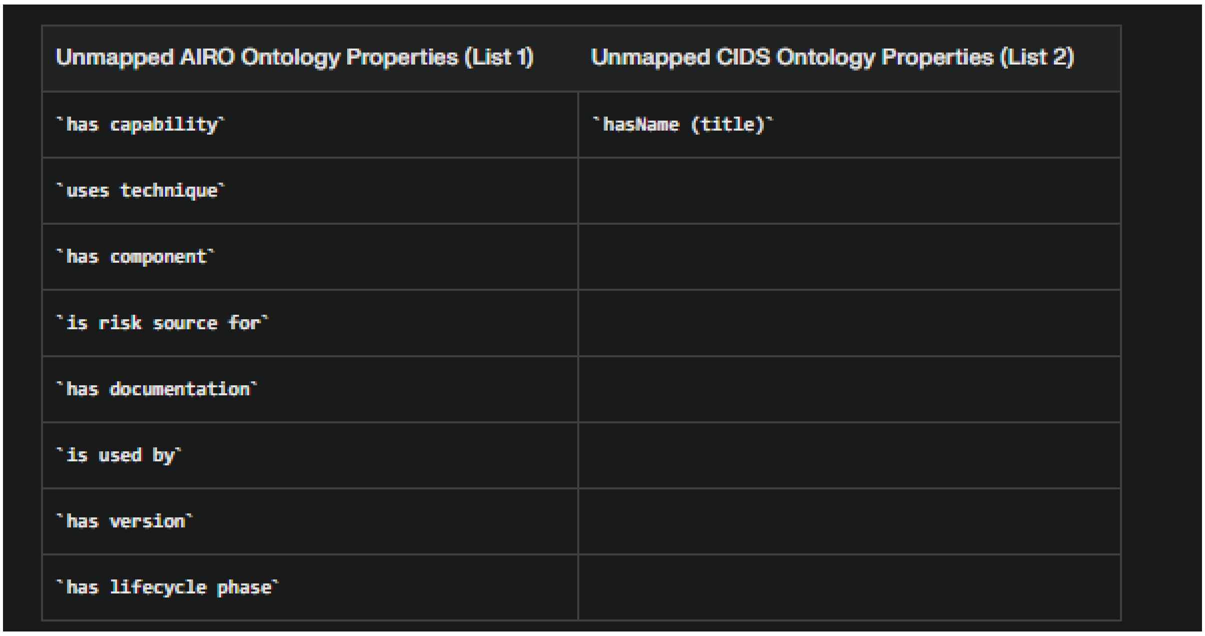
Listing 5: Ontology Prompt

The LLM was able to produce a mapping of similar properties in the two ontologies that could be used to evaluate their correlations and relationships. It is important to note here that not all properties mapped 1 to 1. This would be taken into account and examined in depth in the evaluation chapter. The mapping produced can be seen in listing 6.



AIRO Ontology Property (List 1)	CIDS Ontology Property (List 2)	One-Word Property Name
`is applied within domain`	`hasCatchmentArea`	Domain
`has purpose`	`hasDescription`	Purpose
`produces output`	`forOutcome`	Output
`has risk`	`intendedImpact`	Risk
`has consequence`	`hasConsequence`	Consequence
`has impact`	`forOutcome`	Impact
`has stakeholder`	`forStakeholder`	Stakeholder
`mitigation`	`hasMitigation`	Mitigation
`is provided by`	`fromPerspectiveOf`	Provider
`has severity`	`hasImportance`	Severity
`has likelihood`	`hasLikelihood`	Likelihood

Listing 6: AIRO and CIDS mapping



Unmapped AIRO Ontology Properties (List 1)	Unmapped CIDS Ontology Properties (List 2)
`has capability`	`hasName (title)`
`uses technique`	
`has component`	
`is risk source for`	
`has documentation`	
`is used by`	
`has version`	
`has lifecycle phase`	

Listing 7: Unmapped terms

4.3 Prompt engineering

Before beginning the annotation process, the prompts used to annotate the data must be designed. The prompt engineering process involves the following steps:
These lists from 3.3.1 and 3.3.2 were used as the definition of the ontologies for the purposes of this work. By taking this list of properties, prompt engineering was used to refine the LLM output. The finalized prompt used can be seen below:

Here is the __ ontology:

…

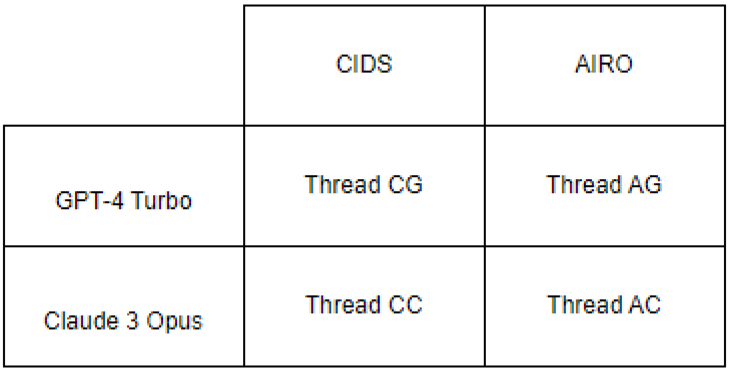
List of properties from either CIDS or AIRO depending on the thread

…

Now, when given an incident report, you find information for each part of the definition and fill in an RDF in turtle format using the info found. Always include all of the properties in your return. If there is no relevant information for a given property in the report, return blank value in the RDF. Do you understand? if so reply yes with no further follow up and await an incident report

Listing 8: RDF creation prompt

When the LLM returned “Yes.” the prompt engineering was complete. At this stage, each incident report could be fed into the LLM. By taking the preprocessed incidents obtained from the process defined in 4.1, the fist 50 incidents were annotated using GPT-4 Turbo and Claude 3 Opus, for both CIDS and AIRO. This took place across four different Perplexity threads, one for each combination as explained in Figure X.



	CIDS	AIRO
GPT-4 Turbo	Thread CG	Thread AG
Claude 3 Opus	Thread CC	Thread AC

Figure 14: Perplexity threads

These four threads, CG,AG,CC and AC represent the possible combinations of LLMs and ontologies. Once these first 50 incidents were annotated with each combination, the evaluation process could begin.

4.4 Evaluation Analysis Process

The process involved for analyzing the output of this work involved the creation of csv files that could be used for evaluation.

The first section of CSV files included a file for each thread, where the columns of the file were incidents 1 to 50, and the rows were the properties of the ontology used. The values in the CSV were binary, with 1 representing a description being successfully extracted from the incident report for that property, and a 0 representing no description extracted. If a description was not extracted, this could mean that the property is not relevant to the incident or that the model did not extract it successfully. The four CSV files processed here were used for the first part of the evaluation process, aimed to evaluate firstly if LLMs are effective for extracting relevant data from AI incident reports, secondly if LLMs are effective at annotating impact categorisation ontologies using the extracted data, and thirdly which LLM performs most optimal at the extraction and annotation tasks.

The second section of analysis involved creating a CSV file with the columns representing the mapped pairs of properties, and the rows representing incidents 1 to 50. The values in the CSV contain a score from 0-2, where 0 indicates the descriptions for the pairing have no similarity for that property in the incident, 1 indicating some sort of similarity between the descriptions and 2 representing the same or synonymous descriptions for the pair of properties for both ontologies. This CSV file was then used for the similarity analysis in the evaluation chapter in order to evaluate whether there are correlations or relationships between the two ontologies.

Chapter 5
Evaluation

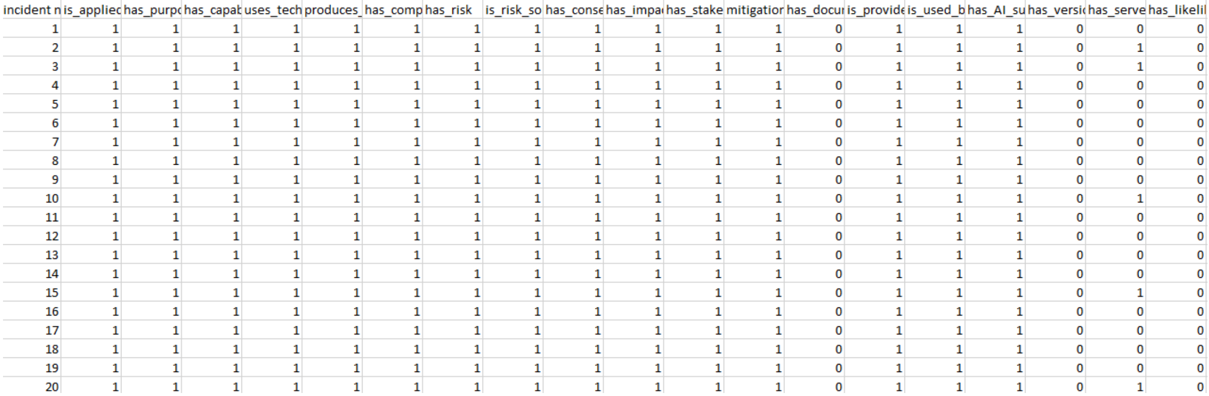
This chapter will look at the evaluation techniques used in this work, and their advantages and disadvantages. The chapter will then present the results of the evaluation, and a discussion of these results will be presented afterwards.

5.1 Evaluation Techniques

In order to evaluate the performance of the LLMs at the annotation task, and the correlations or relationships between ontologies, some evaluation techniques were decided on. These techniques were previously described in 4.4. Namely, for the annotation task performance evaluation, frequency analysis was employed on the CSV files created to represent the LLM performance. For the correlation evaluation between ontologies, ontology alignment and similarity scoring were undertaken. These techniques have advantages and disadvantages, which will be discussed below.

5.1.1 Frequency Analysis

The frequency analysis was conducted as the initial step in evaluating the performance of LLMs in extracting and annotating AI incident reports into impact categorisation ontologies. By analyzing the CSV files obtained from the process in 4.4, a straightforward means to quantify the effectiveness of each LLM in recognizing and extracting relevant data from the incident reports was obtainable. This method had several advantages and disadvantages, which are discussed further here.



incident n	is_appliec	has_purpc	has_capab	uses_tech	produces_	has_comp	has_risk	is_risk_so	has_conse	has_impa	has_stake	mitigatior	has_docur	is_provide	is_used_b	has_AI_su	has_versi	has_serve	has_likelil
1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	0	0	0
2	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	0	1	0
3	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	0	1	0
4	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	0	0	0
5	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	0	0	0
6	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	0	0	0
7	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	0	0	0
8	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	0	0	0
9	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	0	0	0
10	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	0	1	0
11	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	0	0	0
12	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	0	0	0
13	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	0	0	0
14	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	0	0	0
15	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	0	1	0
16	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	0	0	0
17	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	0	0	0
18	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	0	0	0
19	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	0	0	0
20	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	0	1	0

Figure 15: example of AIRO_GPT frequency analysis data[22]

Advantages of Frequency Analysis

Simplicity and Clarity: This analysis offers a clear and straightforward approach to evaluate the presence or absence of property descriptions within incident reports. The simplicity of a binary value makes it easy to understand and interpret the results. Quantitative evaluation: This analysis provides a quantitative measure of each LLMs' performance, allowing for a direct comparison between different models based on the number of properties successfully extracted.

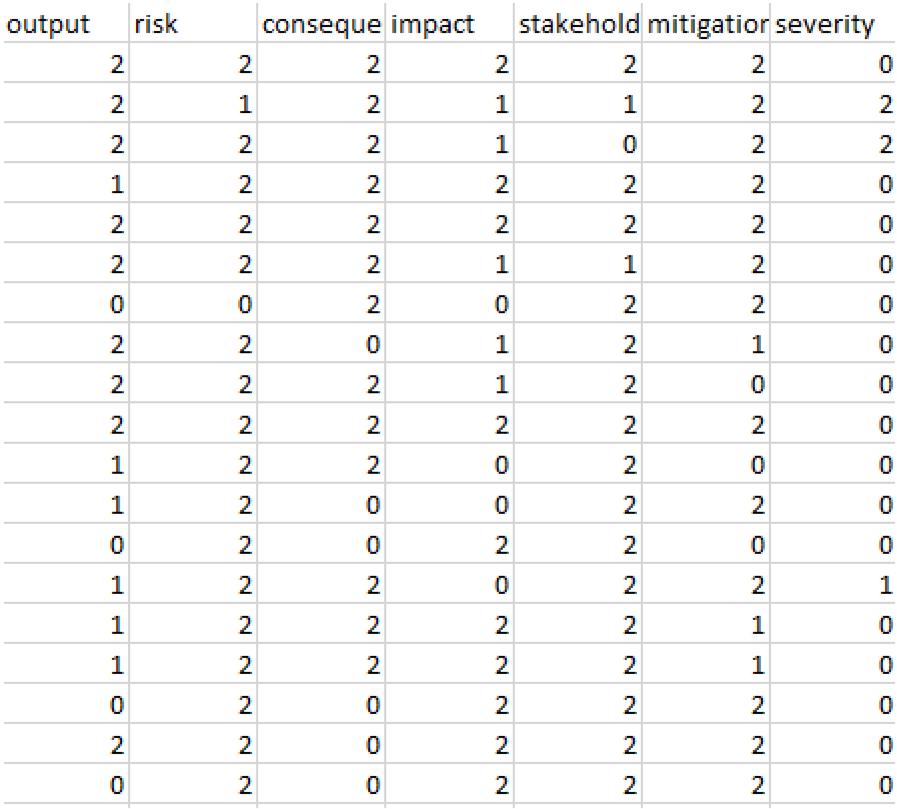
Disadvantages of Frequency Analysis

Lack of depth: while this analysis quantifies how often each property is extracted, it does not assess the quality or accuracy of the extracted descriptions. However, due to the nature of the annotation task, there is no simple ground truth to compare them to. As a result, the accuracy of these descriptions is not taken into account for the frequency analysis, rather the omission of annotations being the deciding factor.

[22] Github: https://github.com/paddyocallaghan/dissertation

5.1.2 Ontology Alignment

As discussed in 4.2, the ontology mapping alignment was handled by GPT-4 Turbo, by mapping similar properties across the two ontologies. The output gave a total of 11 pairs of properties to be used for similarity evaluations. As both ontologies are structured to capture and describe mainly text based qualitative information, this alignment of ontologies was required to effectively evaluate their similarities. By assigning a value to each property pair for each incident, the dataset required to conduct a similarity evaluation can be created. Figure 16 gives a sample of what this data looked like:



output	risk	conseque	impact	stakehold	mitigatior	severity
2	2	2	2	2	2	0
2	1	2	1	1	2	2
2	2	2	1	0	2	2
1	2	2	2	2	2	0
2	2	2	2	2	2	0
2	2	2	1	1	2	0
0	0	2	0	2	2	0
2	2	0	1	2	1	0
2	2	2	1	2	0	0
2	2	2	2	2	2	0
1	2	2	0	2	0	0
1	2	0	0	2	2	0
0	2	0	2	2	0	0
1	2	2	0	2	2	1
1	2	2	2	2	1	0
1	2	2	2	2	1	0
0	2	0	2	2	2	0
2	2	0	2	2	2	0
0	2	0	2	2	2	0

Figure 16: CSV of property pairs for CIDS and AIRO[23]

Ontology alignment advantages and disadvantages are seen below:

[23] Github: https://github.com/paddyocallaghan/dissertation

Advantages of Ontology Alignment

Facilitates comparative analysis: by aligning these two ontologies, as seen in the 11 pairs of properties between AIRO and CIDS, a comparative analysis can be conducted. This is crucial for evaluating the correlations and relationships between the two ontologies which would not be possible without this ontology alignment due to the nature of the data, namely text-based descriptive, qualitative data.

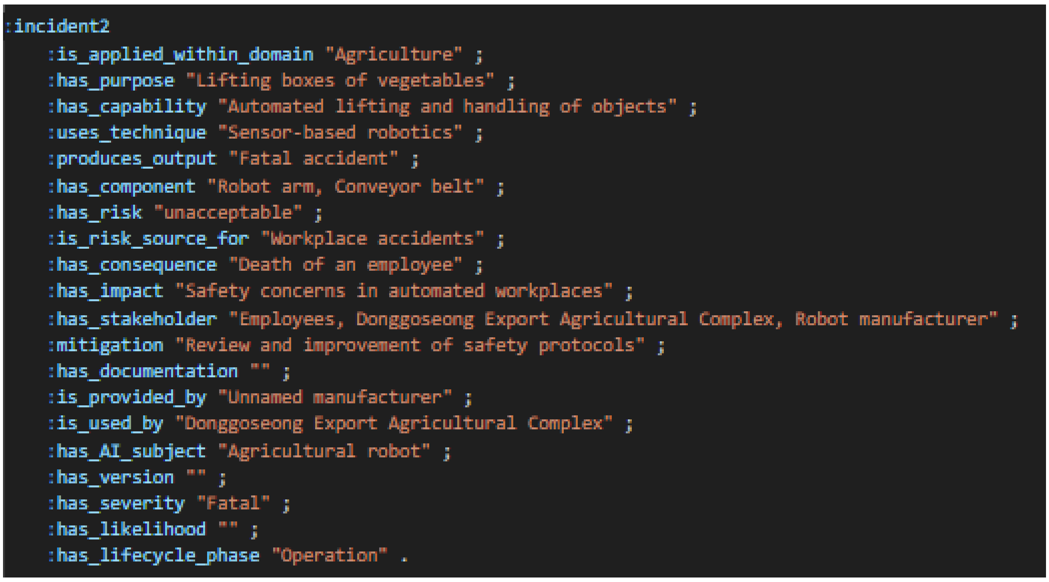
Disadvantages of Ontology Alignment

Imperfect mappings: As not all properties map perfectly to each other, there are some outlier properties that have no mapping. For the purpose of this analysis, these properties cannot be considered for similarity scoring. This was mainly due to AIRO having more properties than CIDS in this work. However all of the properties are used for the frequency analysis section, so they are not irrelevant.

5.1.3 Similarity Scoring

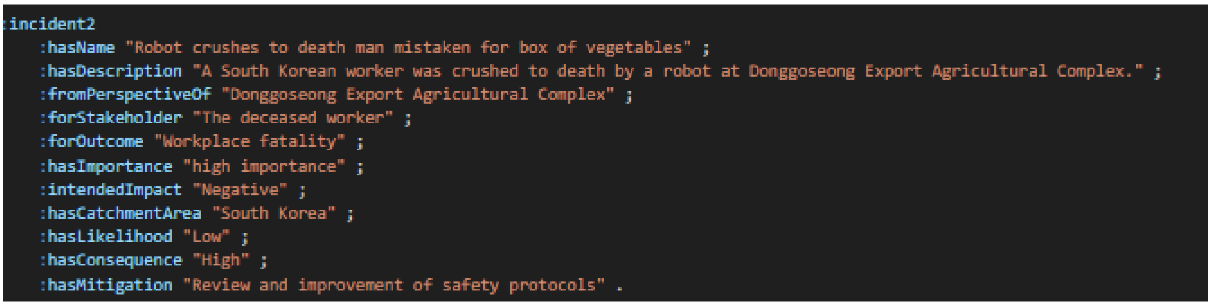
The similarity scoring process involved manually applying an encoding of 0, 1 or 2 to the pairs of properties discussed in 4.2 based on their similarity. This process was done manually, by the author of this work. Due to lack of resources, an outside annotator was not available, so there will be some inherent biases to the encodings. Figure X shows examples of a scoring of 0, 1 and 2 to give the reader a better understanding as to what was classified as having 0: no similarity/no description, 1: some similarity and 2: identical or synonymously.

Scoring Example[24]



```
:incident2
    :is_applied_within_domain "Agriculture" ;
    :has_purpose "Lifting boxes of vegetables" ;
    :has_capability "Automated lifting and handling of objects" ;
    :uses_technique "Sensor-based robotics" ;
    :produces_output "Fatal accident" ;
    :has_component "Robot arm, Conveyor belt" ;
    :has_risk "unacceptable" ;
    :is_risk_source_for "Workplace accidents" ;
    :has_consequence "Death of an employee" ;
    :has_impact "Safety concerns in automated workplaces" ;
    :has_stakeholder "Employees, Donggoseong Export Agricultural Complex, Robot manufacturer" ;
    :mitigation "Review and improvement of safety protocols" ;
    :has_documentation "" ;
    :is_provided_by "Unnamed manufacturer" ;
    :is_used_by "Donggoseong Export Agricultural Complex" ;
    :has_AI_subject "Agricultural robot" ;
    :has_version "" ;
    :has_severity "Fatal" ;
    :has_likelihood "" ;
    :has_lifecycle_phase "Operation" .
```

Figure 17: Incident 2 AIRO



```
:incident2
    :hasName "Robot crushes to death man mistaken for box of vegetables" ;
    :hasDescription "A South Korean worker was crushed to death by a robot at Donggoseong Export Agricultural Complex." ;
    :fromPerspectiveOf "Donggoseong Export Agricultural Complex" ;
    :forStakeholder "The deceased worker" ;
    :forOutcome "Workplace fatality" ;
    :hasImportance "high importance" ;
    :intendedImpact "Negative" ;
    :hasCatchmentArea "South Korea" ;
    :hasLikelihood "Low" ;
    :hasConsequence "High" ;
    :hasMitigation "Review and improvement of safety protocols" .
```

Figure 18: Incident 2 CIDS

For example, incident 2 had a score of 2 for mitigation as the description for mitigation in figure X and has_mitigation in Figure 17 are identical. Incident 2 had a score of 1 for risk as has_Risk had a description of unacceptable which is close to the negative intended impact description in CIDS. A score of 0 was given to likelihood as the value is blank in Figure 17

.

Advantages of Similarity Scoring

[24] Github: https://github.com/paddyocallaghan/dissertation

In-depth analysis: This process is richer in value detail when compared to a simple binary scoring. By having three distinct categories, a more nuanced understanding of the data can be used for analysis.

Disadvantages of Similarity Scoring

Subjectivity: The process is inherently subjective, relying on the author's judgment to determine the degree of similarity between the property descriptions. This can introduce biases and inconsistencies to the scoring.

Time Consuming: The process of manually reviewing and scoring each pair is a time consuming process especially when working with large datasets.

Lack of Reproducibility: Without an external or standardized scoring system, the results may lack reproducibility. The migration of this is that there does not exist a simple standardized scoring system for this type of work, so in this case manual annotation is the best solution.

5.2 LLM Annotation Performance

The results of the frequency analysis discussed in 5.1.1 can be seen in the figures below. The section is broken down into two parts, AIRO and CIDS analysis. Section 5.2.1 contains the results of the AIRO frequency analysis, and section 5.2.2 contains the results of the frequency analysis using CIDS.

5.2.1 AIRO Frequency Analysis

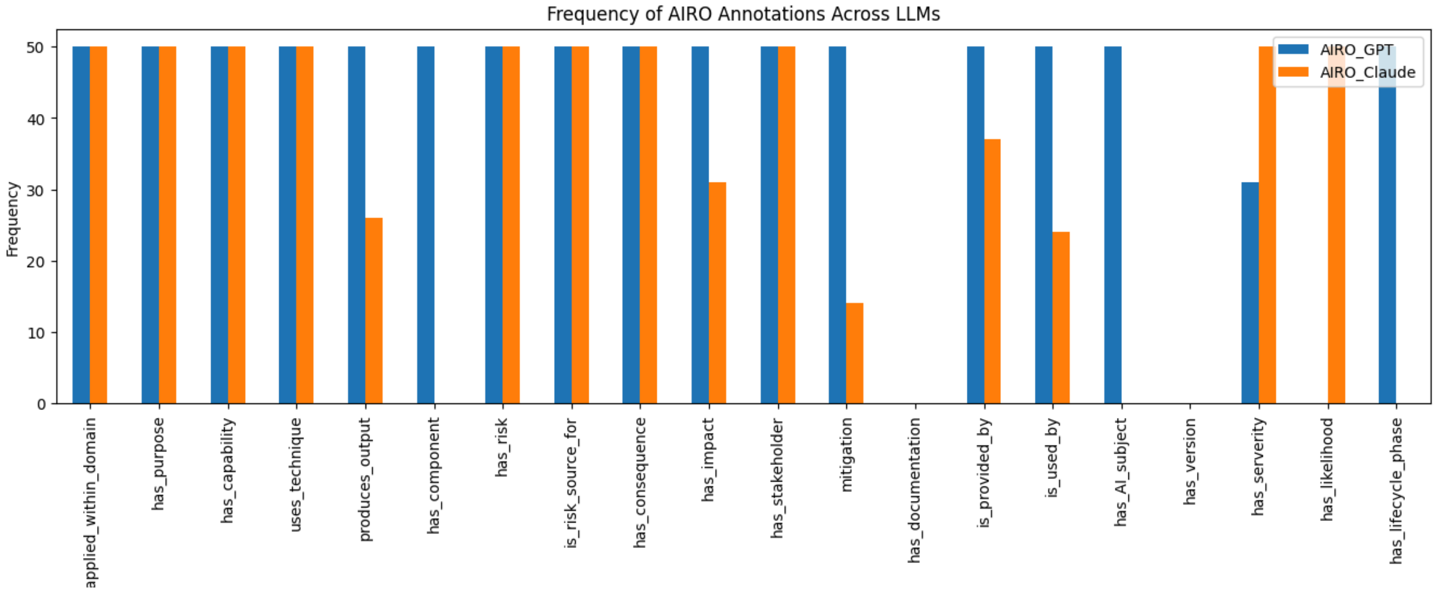


Figure 19: AIRO annotation frequency

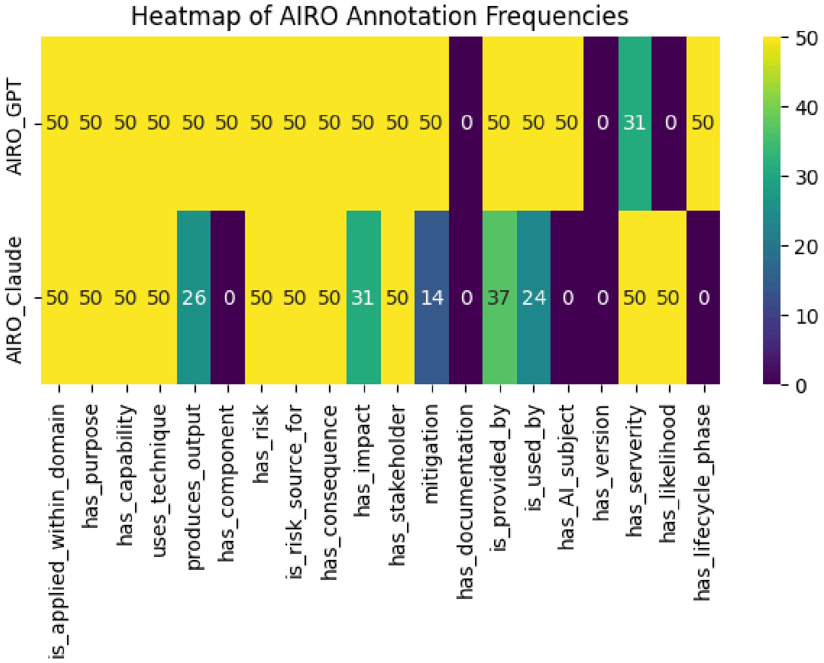


Figure 20: Heatmap of AIRO Annotation Frequencies

Figure 19 is a frequency graph of all properties of AIRO annotated using GPT-4 Turbo and Claude 3 Opus. The length of each bar represents the frequency of each respective property over the 50 incident reports. By analyzing this graph, it is clear that both models successfully annotate the majority of properties. When looking at the heatmap in figure 20, it is clear that some properties have a 0 value for annotations. Has_documentation and has_version are the only properties which are 0 for both LLMs, so it is clear that these two properties are not considered relevant by the LLMs to these incidents. Other properties have missing values, like has_technique or has_impact in Claude 3 Opus. However, these properties were successfully annotated by GPT-4 Turbo. In this case, it seems that GPT-4 Turbo produced a better set of annotations due to its higher annotation rate.

5.2.2 CIDS Frequency Analysis

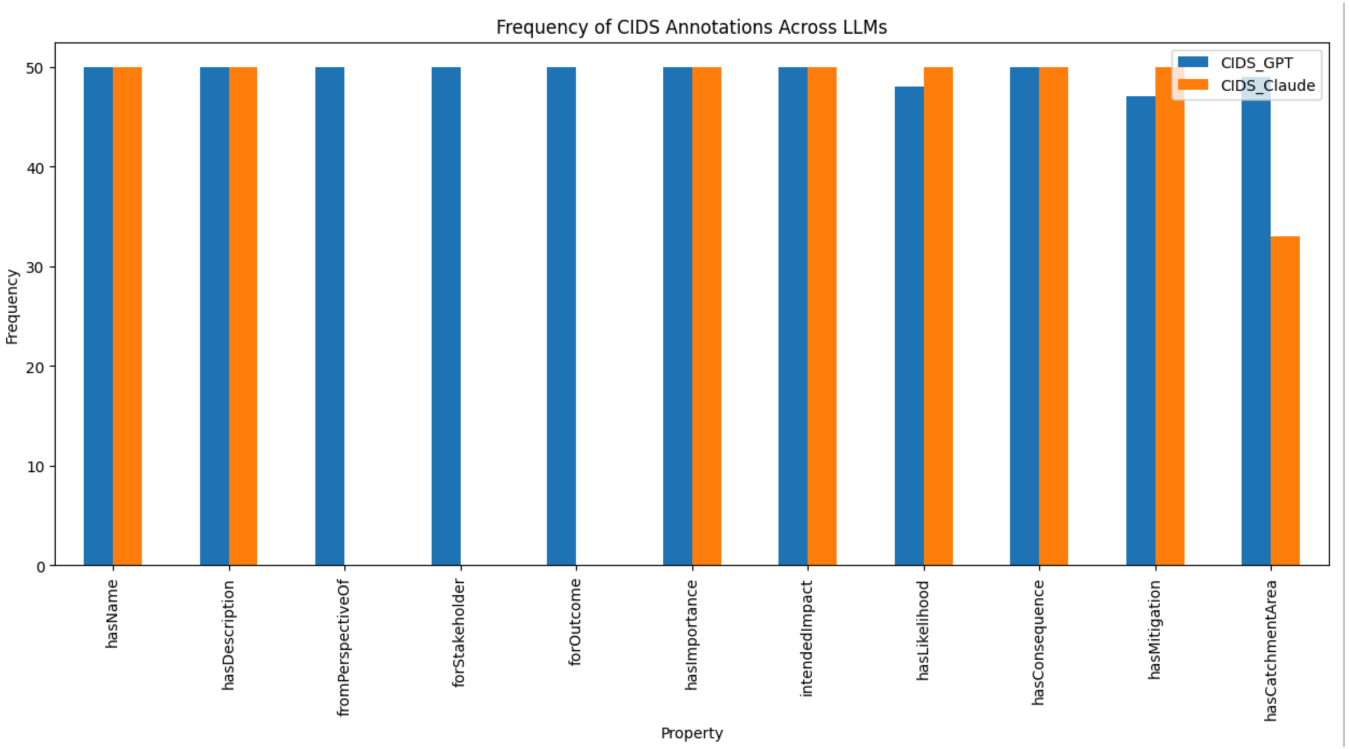


Figure 21: CIDS annotation frequency

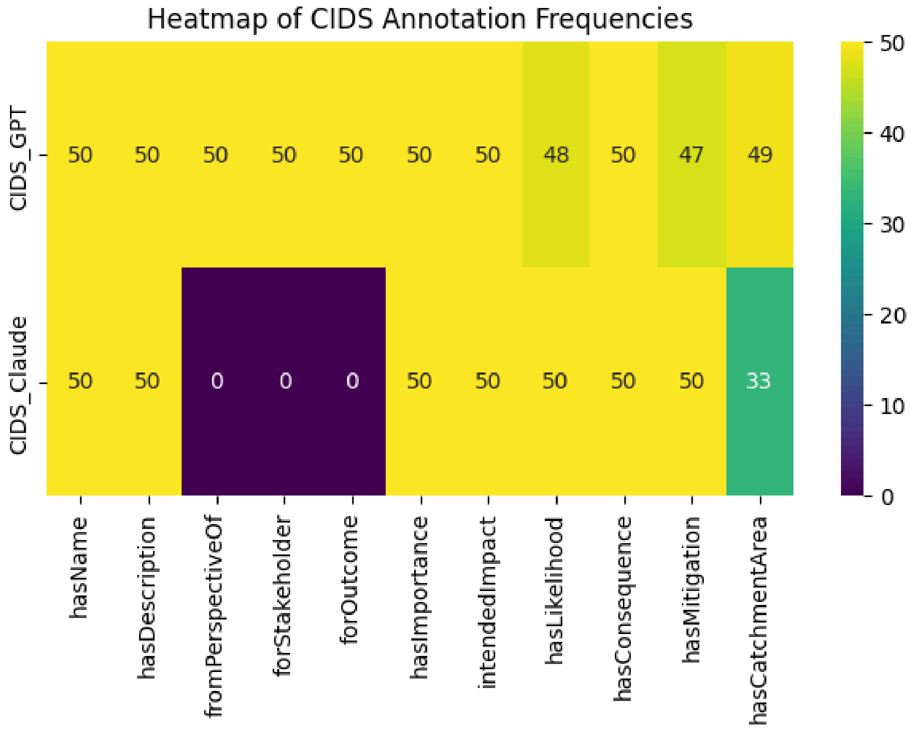


Figure 22: Heatmap of CIDS annotation frequencies

Figure 21 shows a frequency graph of all properties of CIDS annotated using GPT-4 Turbo and Claude 3 Opus.Similarly to 5.2.1, the length of each bar represents the frequency of each respective property over the 50 incident reports. By analyzing this graph, it is clear that both models successfully annotate the majority of properties, with GPT-4 Turbo having no 0 annotations. When looking at the heatmap in figure 22, it is clear that some properties have a 0 value for annotations for Claude 3 Opus. fromPerspectiveOf, forStakeholder and forOutcome are 0 for both Claude 3 Opus, while 8/11 properties are 50 for GPt-4 Turbo, with the remaining 3/11 all 47 or over.In this case, it seems that GPT-4 Turbo produced a better set of annotations due to its higher annotation rate across the board.

5.2.3 Frequency Analysis Discussion

From analyzing the frequency of properties in annotations for both GPT-4 Turbo and Claude 3 Opus, applied to both CIDS and AIRO, it can be concluded that GPT-4 Turbo is better suited to the annotation task due to it’s higher annotation frequency.

As an aside here, a subjective analysis of the accuracy of these annotations was conducted by the author due to the lack of accuracy analysis. This was performed as a sanity check, to ensure that the annotations were not completely incorrect, as an annotation frequency of 50 did not necessarily mean that the 50 annotations were correct. Over the course of assigning frequency scores to each property for each ontology/LLM pairing, it was seen by the author that there were no apparent hallucinations by either model, and that all of the annotations could be reasonably considered to be correct.

One flag during this process was the 0 frequency for some properties that did not seem to be caused by irrelevance to the task. Examples of this were fromPerspectiveOf, forStakeholder and forOutcome in CIDS that had a frequency of 0 using Claude 3 Opus and 50 using GPT. This was seen to be caused by a lapse in understanding by Claude, as both models received the same definitions, this could be attributed to GPT being superior in understanding and extracting text data in this work. Further examples of this could be seen in AIRO also, with Claude having a 0 frequency for Has_component has_AI_Subject has_lifecycle_phase while GPT had 50. GPT was seen to have a 0 for has_likelihood with Claude having 50 in AIRO also. Due to this, it can be said that GPT had the overall better performance for this task, as it only scored lower than Claude on one property across the two ontologies.

5.3 Correlation and Relationship Analysis

For the Correlation and relationship analysis, the ontology alignment and similarity scoring methods are implemented. First, the list of mapped properties from listing X was manually annotated with scores from the range [0,1,2] according to the scoring definitions outlined in 5.1.3. The data from AIRO_GPT and CIDS_GPT was used to do this, as 5.2 revealed that GPT-4 Turbo was more optimal than Claude 3 Opus for this task The resulting dataset was used for correlation and relationship analysis by visualizing the average scores by property and incident. These results can be seen below.

5.3.1 Average Score by Property

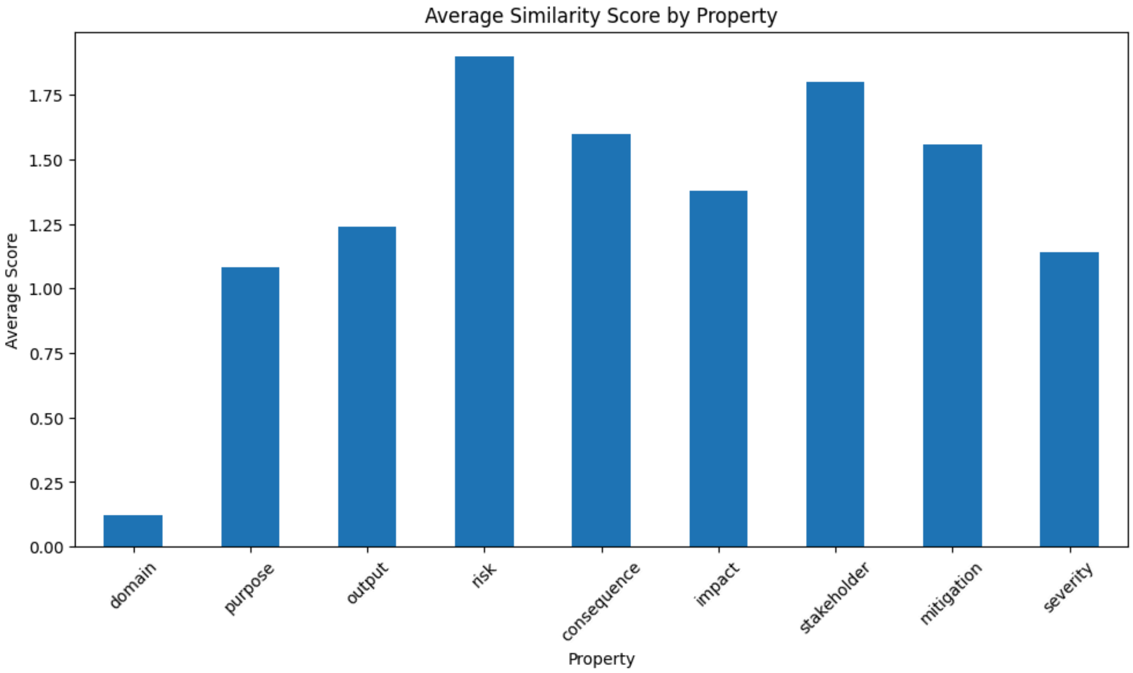


Figure 23: Average similarity score by property

The results of the correlation and relationship analysis process produced from the similarity scoring method discussed in 5.1.2 can be seen in figure 23. The bar chart shows which of the properties from 4.2 had the highest average similarity score. For this analysis, it was found that the Provider and Likelihood properties were irrelevant as both scored a 0 for every property. As a result, they were omitted from the analysis. It can be seen that risk and stakeholder have the highest average score, indicating that CIDS and AIRO correlate strongly for these properties. Consequence, mitigation and impact score relatively high at 1.5 +- .15, indicating good correlations for these properties. Output, severity and purpose all score above 1 meaning there is some correlation. Domain scored poorly, indicating little to no correlation. The Heatmap in Figure 24 visualizes the correlations more clearly, as the extent of the range is evident from the color coding.

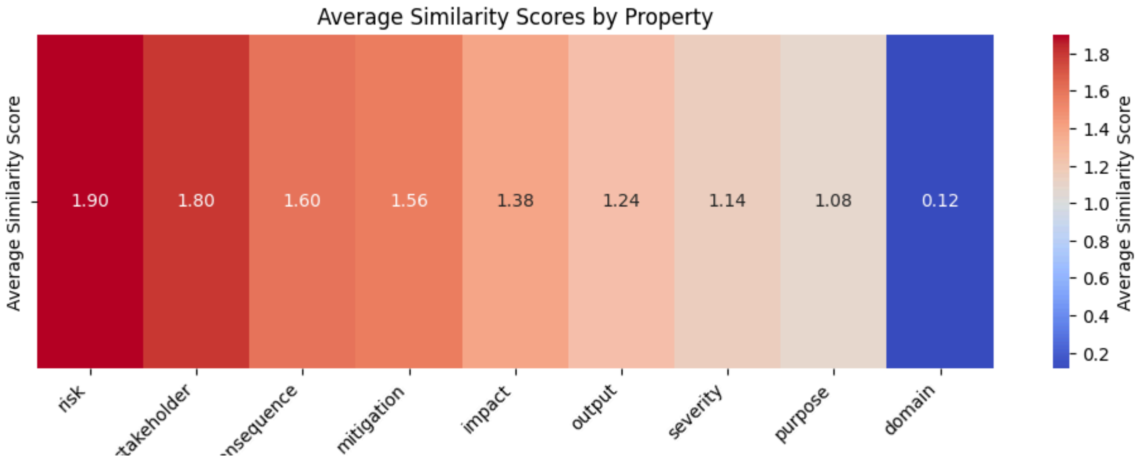


Figure 24: Heatmap of average score by property

5.3.2 Average Score by Incident

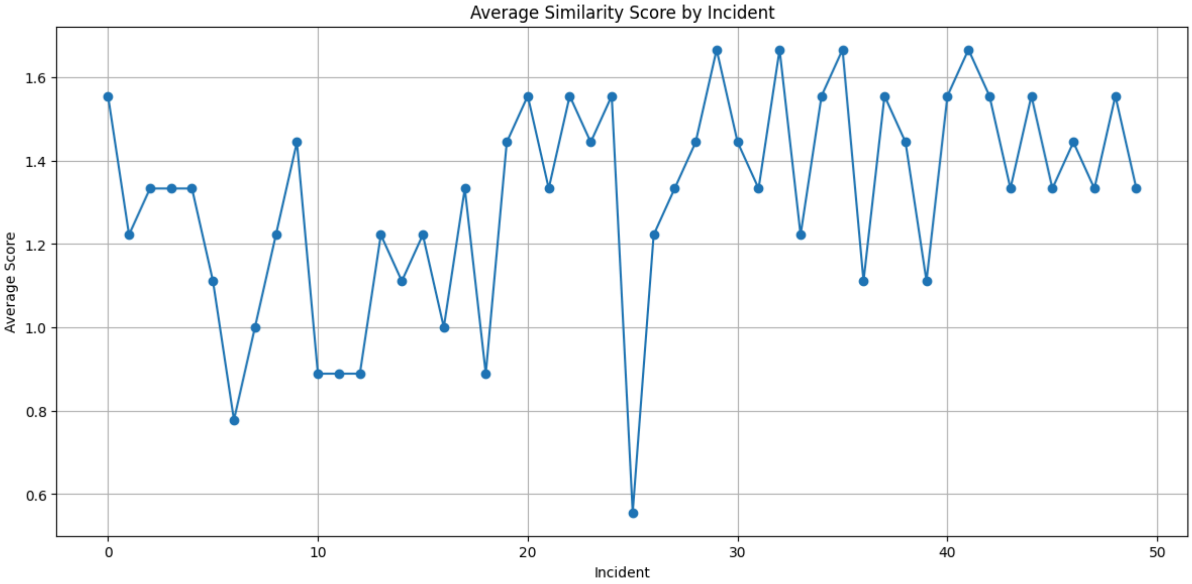


Figure 25: Average similarity scores by incident

Similarity scores by incident can be seen in Figure X. These scores show that all incidents had a score of above 0.6, indicating that no incident had 0 for all of its similarity scores. The average score per incident was calculated at 1.31333. This indicates that there is a general level of similarity across the properties of the incidents, which indicates that the ontologies used to categorize them have some level of correlation.

5.3.3 Correlation & Relationship Analysis Discussion

The correlation and relationship analysis using similarity scoring provides an insight into how well the AIRO and CIDS ontologies relate across their properties. The analysis shows a general trend of moderate to strong correlation for most properties with some exceptions. This suggests that not all properties of these ontologies correlate, but that the ontologies do relate for the stronger correlated properties.

Strong Correlations

It is seen that the risk and stakeholder properties show a strong correlation.This is an indication that the ontologies categorize risk and stakeholder information similarly, thus showing a correlation between them. This alignment is significant as it shows that there is a shared understanding for the critical elements of risk and stakeholder across the ontologies.

Moderate Correlation

Properties like consequence, mitigation and impact exhibit a moderate correlation. This suggests that while there may be differences in the specifics of how each ontology details consequences, or suggests mitigation strategies, there is an overall similar theme. This indicates that the ontologies have similar explanations for these categories, however the semantics might differ in how they describe them. This is expected due to the nature of LLM output and the text based data.

Weak or No Correlation

Other properties like domain, likelihood and provider had weak or no correlation, which indicates that the ontologies did not relate in the way that they capture these properties. This is indicative of a lack of similarity in approach to define these aspects of Ai incidents. These make up the minority of the properties seen, however it is still important to show that these ontologies do not align across the board.

Chapter 6
Conclusion and Future work

6.1 Conclusion

Over the course of this work, a method for applying the power of large language models for the capturing of impact assessments into impact categorisation ontologies. .This methodology shows how to use LLMs effectively to extract relevant data by building prompts using prompt engineering techniques. By applying these techniques to the CIDS and AIRO ontologies, it was shown that LLMs, specifically GPT-4 Turbo can successfully annotate AI incident reports into structured RDF format.
The evaluation of the annotation performance revealed that GPT-4 Turbo outperformed Claude 3 Opus in terms of annotation frequency across both ontologies. This suggests that GPT-4 Turbo is better suited to the task of extracting and annotating relevant information from AI incident reports.

The correlation and relationship analysis carried out using the similarity scoring method provided insights into the alignment between CIDS and AIRO. The results showed moderate to strong correlations for most properties, indicating an underlying understanding of critical elements in AI incident categorization across these models.

Overall this work demonstrates the potential for leveraging LLMs to optimize the process of capturing impact assessments from AI incident reports, as well as the potential for using LLMs to annotate these impact assessments into categorization ontologies. The methodology for aligning and analyzing different ontologies also outlines the potential for using LLMs for ontological mapping and matching. These methodologies can serve as a foundation for future research in the area, with the goal of improving the analysis process of AI incidents and of the understanding of the relationships between different ontologies. By improving this process and understanding, risks associated with AI systems can be identified and mitigated easier. This ultimately contributes to the goal of a safer and more responsible AI system development process, in line with the EU AI Act regulations.

.

6.2 Future Work

This work had laid the groundwork for several potential directions for future work, such as:

Expansion of ontologies

Future work here could explore the use of other ontologies beyond CIDS and AIRO to expand the properties considered. This work could involve adapting the methodology to handle more complex ontology structures and relationships, or use LLMs with a larger context window to handle entire ontologies, rather than subsets as used in this work.

Larger dataset of incidents

Research could be done to grow the database used in this methodology. In this case, 50 incidents were used due to the time intensive manual scoring process used. By adapting this methodology, further work could increase the number of incidents used. Another possibility here could use other sources than just the AIAAIC repository in order to get a more diverse set of incidents.

Model Fine Tuning

Work could be done to fine tune a model for the specific task of incident annotation. An investigation into the benefits of fine-tuning GPT-4 Turbo could lead to improved performance and more accurate property extraction.

Integration with AI Regulation Frameworks

By exploring how the methodology in this work could be integrated into broader AI governance frameworks, such as the EU AI Act, an advancement in standardization could be achieved. As there is no current standard AI incident reporting method, further work could be beneficial to determine how applicable these laid out methodologies are to the regulations in the AI Act.

References

[1] Alzou'bi, S. (n.d.). Artificial Intelligence In Law Enforcement, A Review. Retrieved April 15, 2024, from https://d1wqtxts1xzle7.cloudfront.net/36050961/Artificial_Intelligence_in_Law_Enforcement__A_Review-libre.pdf?1419546897=&response-content-disposition=inline%3B+filename%3DARTIFICIAL_INTELLIGENCE_IN_L_AW_ENFORCEM.pdf&Expires=1713215607&Signature=YZPPc9hAn3

[2] Dimou, A. (2022). Airo: An ontology for representing ai risks based on the proposed eu ai act and iso risk management standards. Towards a Knowledge-Aware AI: SEMANTiCS, 51.

[3] Dong, Q. (2022, December 31). [2301.00234] A Survey on In-context Learning. arXiv. Retrieved April 15, 2024, from https://arxiv.org/abs/2301.00234

[4] Dua, D., & Wang, Y. (2019, June). DROP: A Reading Comprehension Benchmark Requiring Discrete Reasoning Over Paragraphs. ACL Anthology. Retrieved April 15, 2024, from https://aclanthology.org/N19-1246/

[5] Edwards, L. (2022, April 11). The EU AI Act: a summary of its significance and scope. Ada Lovelace Institute. Retrieved April 15, 2024, from https://www.adalovelaceinstitute.org/wp-content/uploads/2022/04/Expert-explainer-The-EU-AI-Act-11-April-2022.pdf

[6] EU AI Act. (n.d.). Art. 6 Classification Rules for High-Risk AI Systems. EU AI Act. Retrieved April 15, 2024, from https://www.euaiact.com/article/6

[7] Feuerriegel, S., Hartmann, J., & Zschech, P. (2023, September 12). generative AI. Springer. Retrieved April 15, 2024, from https://link.springer.com/article/10.1007/s12599-023-00834-7

[8] Fox, M., & Ruff, K. (2021, September 18). CIDS-OntoBeSS v1. EIL. Retrieved April 15, 2024, from https://eil.mie.utoronto.ca/wp-content/uploads/2021/09/CIDS-OntoBeSS-v1.pdf

[9] Giray, L. (2023, June 7). Prompt Engineering with ChatGPT: A Guide for Academic Writers. Retrieved April 15, 2024, from https://link.springer.com/article/10.1007/s10439-023-03272-4

[10] Golpayegani, D., Pandit, H., & Lewis, D. (2023, June). ,To Be High-Risk, or Not To Be—Semantic Specifications and Implications of the AI Act’s High-Risk AI Applications and Harmonised Standards. Retrieved April 15, 2024, from https://dl.acm.org/doi/10.1145/3593013.3594050

[11] GPT-4 Technical Report. (2023, March 27). OpenAI. Retrieved April 15, 2024, from https://cdn.openai.com/papers/gpt-4.pdf

[12] Imani, S., Du, L., & Shrivastava, H. (2023, March 4). [2303.05398] MathPrompter: Mathematical Reasoning using Large Language Models. arXiv. Retrieved April 15, 2024, from https://arxiv.org/abs/2303.05398

[13] Jahanbakhsh, K., & Hajiabadi, M. (n.d.). Beyond Hallucination: Building a Reliable Question Answering & Explanation System with GPTs. gaied.org. Retrieved April 15, 2024, from https://gaied.org/neurips2023/files/4/4_paper.pdf

[14] Johnson, O., & Alyasiri, O. (2023, 12 29). Image Analysis through the lens of ChatGPT-4. Wikipedia, the free encyclopedia. Retrieved April 15, 2024, from

https://www.researchgate.net/profile/Osamah-Alyasiri/publication/376957906_Image_Analysis_through_the_lens_of_ChatGPT-4/links/6592ea9b0bb2c7472b2646fd/Image-Analysis-through-the-lens-of-ChatGPT-4.pdf

[15] Kommineni, V., König-Ries, B., & Samuel, S. (2024, March 13). [2403.08345] From human experts to machines: An LLM supported approach to ontology and knowledge graph construction. arXiv. Retrieved April 15, 2024, from https://arxiv.org/abs/2403.08345

[16] McGregor, S. (2021, May 18). Preventing Repeated Real World AI Failures by Cataloging Incidents: The AI Incident Database. Preventing Repeated Real World AI Failures by Cataloging Incidents: The AI Incident Database | Proceedings of the AAAI Conference on Artificial Intelligence. Retrieved April 15, 2024, from https://ojs.aaai.org/index.php/AAAI/article/view/17817

[17] Naveed, H. (n.d.). A Comprehensive Overview of Large Language Models. Wikipedia. Retrieved April 15, 2024, from https://arxiv.org/pdf/2307.06435.pdf

[18] Navigli, R., Conia, S., & Ross, B. (2023, June 22). Biases in Large Language Models: Origins, Inventory, and Discussion. Retrieved April 15, 2024, from https://dl.acm.org/doi/full/10.1145/3597307

[19] OpenAI. (2023, March 15). [2303.08774] GPT-4 Technical Report. arXiv. Retrieved April 15, 2024, from https://arxiv.org/abs/2303.08774

[20] Palagin, O. (2023, July 11). [2307.05082] OntoChatGPT Information System: Ontology-Driven Structured Prompts for ChatGPT Meta-Learning. arXiv. Retrieved April 15, 2024, from https://arxiv.org/abs/2307.05082

[21] Petric, D. (2021, September 25). Applications of Artificial Intelligence (AI) in healthcare: A review. ScienceOpen. Retrieved April 15, 2024, from

https://www.scienceopen.com/hosted-document?doi=10.14293/S2199-1006.1.SOR-.PPVRY8K.v1

[22] Qiang, Z. (2023, December 1). [2312.00326] Agent-OM: Leveraging Large Language Models for Ontology Matching. arXiv. Retrieved April 15, 2024, from https://arxiv.org/abs/2312.00326

[23] Rodrigues, R., Resseguier, A., & Santiago, N. (2023). When Artificial Intelligence Fails: The Emerging Role of Incident Databases. Pub. Governance, Admin. & Fin. L. Rev., 8, 17.

[24] Ruschemeier, H. (2024, March 5). AI as a challenge for legal regulation. Springer.com. Retrieved April 15, 2024, from https://link.springer.com/article/10.1007/s12027-022-00725-6

[25] Savelka, J. (2023, May 8). [2305.04417] Unlocking Practical Applications in Legal Domain: Evaluation of GPT for Zero-Shot Semantic Annotation of Legal Texts. arXiv. Retrieved April 15, 2024, from https://arxiv.org/abs/2305.04417

[26] Wang, Y., & Yao, Q. (2020, June 12). On the Dangers of Stochastic Parrots | Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. ACM Digital Library. Retrieved April 15, 2024, from https://dl.acm.org/doi/abs/10.1145/3386252?casa_token=OuAySB_1cZ0AAAAA:NbL6-BPpGOomvbpYHGGNTGU9OZ-CtkCs34W5bPzXNUH6s0HLKOi7KwCl8dqlPOoZqHALhWNNyoQn

[27] Wei, J., & Wang, X. (2020). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. NeurIPS Proceedings. Retrieved April 15, 2024, from https://proceedings.neurips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html

[28] White, J., & Fu, Q. (2023, February 21). [2302.11382] A Prompt Pattern Catalog to Enhance Prompt Engineering with Chat*GPT*. arXiv. Retrieved April 15, 2024, from https://arxiv.org/abs/2302.11382