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.