# Methodology

## Logic of choices

Upon reviewing the instructions given, it was clear that the problem presented was a multi-label classification problem. This was distinctly different from its multi-class classification problem with the user submissions being scored between 0-1 for match between the LLMs. Therefore, upon review of all existing open-source LLMs, I determined that Distillbert (uncased) stood the best chance with its light-weight, good performance as well as its ability to understand code as given in its documentation on hugging face. Since the problem essentially boiled down to textual similarity both meaning and pattern, distillbert’s native ability (with its unique architecture of looking for meaning as well as pattern), made it the obvious choice.

Additionally, Pytorch as well as Pytorch Lightning were used for tensor-based operations due to their proven speed and efficiency especially in training LLMs for down-stream tasks.

## Preprocessing

Here, the process was to take the code submissions (Both AI generated and User), and remove all punctuation, extra spaces and new line characters. This meant that only the textual data will be fed into the model for training. However, a run was performed with non-processed data to compare the accuracy of results. The processed data gave slightly better True-positive vs False-positive results so it maybe required in order make sure the model performs at its best.

The link to the Colab notebook is as follows:

<https://colab.research.google.com/drive/1PK0FCfAQiidD0saWUtSRc9RWLbBcykFz?usp=sharing>

Unzip the datafile in the main directory for proper operation