EVALUATION OF ARTIFICIAL INTELLIGENCE TEXT DETECTION MODELS

Member:
Dana Goh Siew Yuen
(Nanyang Girls' High School)

Mentor:
Tan Wei Peng, Andrew
(Defence Science and Technology Agency)

Aim

The project aims to evaluate the open-source landscape and pick the best generative text detection model which can be compared to paid software to aid with the development of an LLM detector in Singapore.

Background

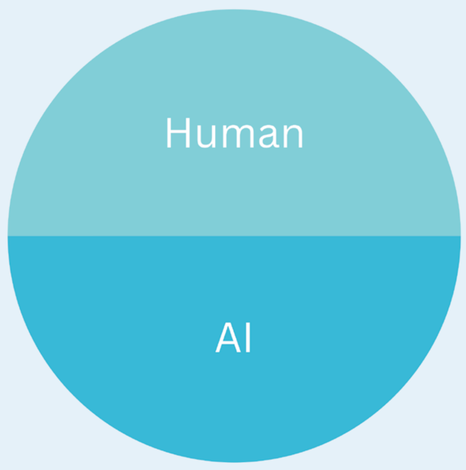
Currently, a key focus in technology is the development of AI and Large Language Models (LLMs). Due to the rise of LLMs like ChatGPT and Google Gemini, these products have become accessible to the general public. However, very few existing models have high accuracy rates in detecting Malay and Chinese text.

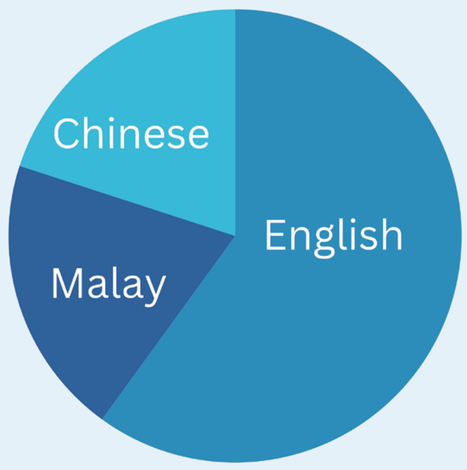
Overview

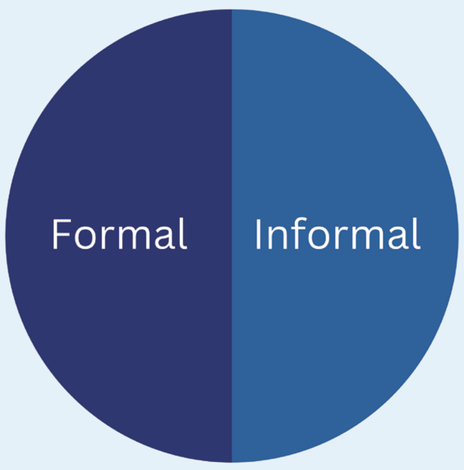
Ten existing open-source offline Artificial Intelligence (AI) text detector models were run through a self-curated dataset with English text. The four models that performed the best were then run through an expanded dataset including Chinese and Malay text to determine the accuracy of the detectors.

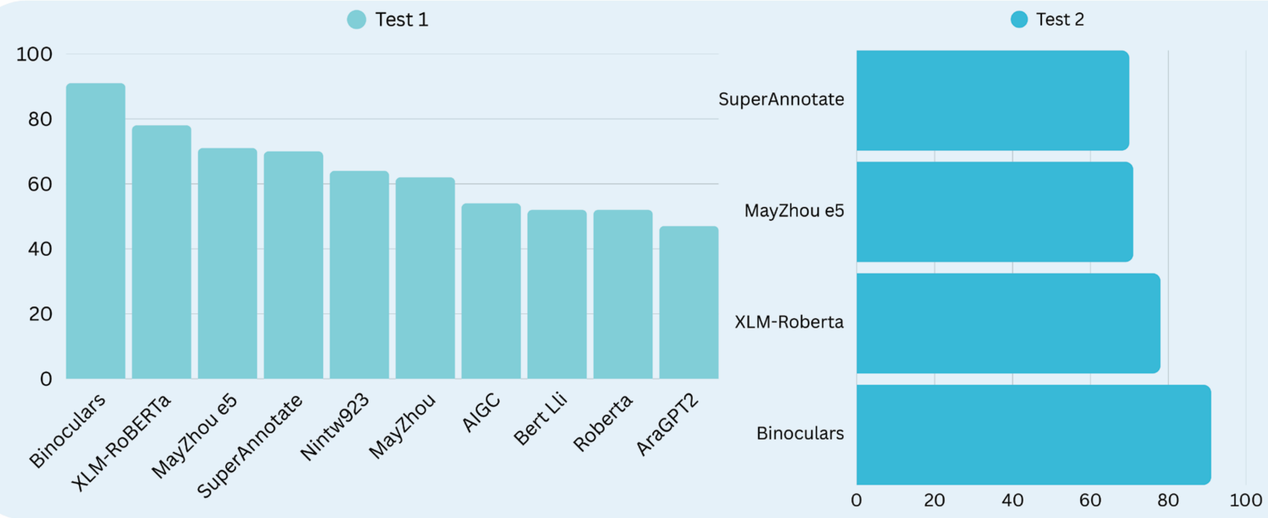
Dataset

There were 200 text samples in total. AI-generated text samples were generated through ChatGPT, while human-generated text samples were taken from Kaggle and Huggingface.









Analysis of Results

Binoculars was the most proficient at distinguishing English and Chinese text, while all models had a low accuracy score for Malay text. Overall, the Binoculars model had the highest accuracy across all languages.

Conclusion

The majority of open-source offline models are inadequate at detecting GPT 4.0 text as they were trained on older versions of GPT. For English and Chinese text, Binoculars has a high accuracy rate, thus the code used for Binoculars can be examined to train an AI text detector in Singapore.