Applying Machine Learning to Chatbot Content Detection

The large language model (LLM) ChatGPT has recently brought AI generated content into the forefront of many peoples’ minds. For educators, this can be a large problem: can an instructor verify that a students’ work is indeed their own and not the product of a statistical model? Several tools have been made available in the past few months claiming to be able to detect such forgery, though most have been easily tricked by clever prompts given to ChatGPT. A reliable tool to detect AI generated text has uses beyond the classroom though, from online forums and social media verifying users are human, to national security interests.

The purpose of this study is to investigate the construction of a statistical model that can distinguish between AI generated text and human generated text. We hope to find out if there are significant characteristics of generated text that can be distinguished by a model, and accurately predict whether a text was indeed written by a human. This abstract written by a human, but was the title?