Detecting LLM Authorship

May 27, 2025

1 Project Description

Our project has two main goals:

1. Given a text document, can we determine if it was human written or LLM generated?
2. Given a text document that is known to be LLM generated, can we determine which LLM wrote that document?

The dataset we will use is from the AutoTexTification dataset which can be found here.
Which contains both human and LLM generated text from 6 different models.

2 Key Performance Indicators

There is substantial business value in each of these tasks. For the first task, some possible uses include

1. Monitoring a competitors' business practices by revealing which LLM a rival company may be using to generate or edit content.
2. Help detect use of unapproved LLMs, which could pose safety or legal risks (such as if a company allows only locally run LLMs).
3. Prevent IP contamination, e.g., mixing outputs from proprietary models with open models inappropriately.
4. As an intermediate input for the second task.

For the second task, some possible uses include

1. Fraud detection, e.g. in user collected survey data
2. LLM plagiarism detection, as LLM usage can risk reputational harm (e.g. here)
3. Academic honesty

For each of these, the model should not be used as definitive proof of LLM usage, but instead as a way to flag text for further review.