Abstract:

Natural Language Processing (NLP) has made remarkable progress in recent years across various language-related tasks, and the use of large language models (LLMs) for text generation, summarization, and writing is becoming increasingly widespread.

However, these advancements have also introduced a new challenge, as distinction between human-generated and machine-generated text becomes less clear.

With LLMs becoming more sophisticated, they can produce text that closely resembles human-written content in terms of style, syntax, and coherence.

This blurring of lines poses significant implications for various applications where textual authenticity is crucial, such as in journalism, social media, and online encyclopedia (Wikipedia) where they all serve as primary sources of knowledge for people worldwide.

Our research project is designed to help answering this fundamental question: Given a text, was it generated by a human or by a large language model (LLM)?

To address this question, we first design a pipeline to gather diverse human-generated texts in specific domains (Wikipedia, Reddit posts and comments, newspaper articles).

By selecting a wide array of sources, we ensure that our dataset captures a broad spectrum of writing styles, topics, and formats. This diversity is crucial for creating a robust challenge.

Based on these human-texts, we then design a pipeline to collect LLM-based texts mimicking the same diversity and domains as the natural ones. This involves using the large language model Gemini to generate texts that replicate the styles and contexts of the human texts, and can be easily modified to use other LLM's.

Our goal is to create a dataset where the machine-generated texts are as indistinguishable as possible from the human-generated ones at first glance.

Subsequently, the dataset will form the foundational training infrastructure for machine learning models. These models will serve as classifiers for addressing our research question.

Introduction:

To address the classification question, we aim to build a robust infrastructure for creating a Neural Network model that, given a text input, can determine whether it was generated by a human or by an LLM. The model's performance will largely depend on the quality of the dataset we build, which will be used for the model's learning process.

The main challenge of our project was to create two text corpora: one of texts generated by humans and the other of texts generated by LLMs. Both corpora need to be representative, diverse, and balanced in size.

We started by building the corpus of human-generated texts. To meet the requirements mentioned above, we chose to extract texts from various sources such as newspapers, posts and comments from Reddit (a social network), and Wikipedia.

To create this text corpus, we wrote python programs that use open APIs such as Wikipedia API (Wikipedia-api), Reddit API (praw) and Newspaper3k API. From these, we extracted summaries of Wikipedia pages, posts and comments from Reddit, and published newspaper articles from various newspaper websites (such as BBC and NYT), which together form the corpus of human-generated texts. The topics in each domain were selected using a random mechanism to ensure as much diversity as possible.

Simultaneously, to create the dataset of texts generated by LLMs, we used the Gemini free API (gemini-1.5-flash). We hand-crafted and expirimented several prompts for each domain to achieve the most appropriate one to use when generating texts.

In order to evaluate the dataset a python program was written for analyzing basic features of texts such as the number of words in a given text, the number of lines in a given text, histogram of the words frequency and more.

Eventually a BERT classifier was created and trained on the dataset to demonsrate how to use the dataset and show a basic ability to classify correctly a given text.

Given more resources this pipleline can be easily scaled up. For example, using more domains (such as books and poems), for every domain use more sources (such as post-comments text from Twitter and Facebook in addition to Reddit) to gather human generated text, generating more text from each source and using varoius LLMs to generate texts in order to ensure diversity. All of the above can enrich the dataset which in turn will lead to a better generalized calssifier.

Related work

Our project focuses on assessing the ability to distinguish between texts generated by humans and those generated by machines. Specifically, we aim to create a pipeline to gather dataset that serves as a robust infrastructure for training models to excel in this task.

In order to perform this task to the best of our ability, we first researched academic papers that were published, addressing the same topic and gained insights regarding the most suitable working method for our project.

In a recent study published on May 2024 by Qazi, Shiao and Papalexakis, they introduced GRiD (GPT Reddit Dataset), a new collection of texts generated by the Generative Pretrained Transformer (GPT). This dataset is designed to evaluate detection model's performance in identifying responses generated by ChatGPT.

Despite the valuable insights gained from this study, there remain gaps and limitations in our understanding of the cognitive skills of NLP models and their alignment with human cognition. In the following sections, we present our methodology, experimental setup, and results

Methodology:  
To address the classification question, we aim to build a robust infrastructure for creating a Neural Network model that, given a text input, can determine whether it was generated by a human or by an LLM. The model's performance will largely depend on the quality of the dataset we build, which will be used for the model's learning process.

The main challenge of our project was to create two text corpora: one of texts generated by humans and the other of texts generated by LLMs. Both corpora need to be representative, diverse, and balanced in size.

We started by building the corpus of human-generated texts. To meet the requirements mentioned above, we chose to extract texts from various sources such as newspapers, posts and comments from Reddit (a social network), and Wikipedia, which is an online encyclopedia containing information on a wide range of topics written by a diverse group of people around the world.

To create this text corpus, we wrote python scripts that use open APIs such as Wikipedia API, Subreddit API, and Newspaper3k API. From these, we extracted summaries of Wiki pages, posts and comments from Reddit, and published newspaper articles from the BBC, which together form the corpus of human-generated texts. The topics in each domain were selected using a random mechanism to ensure as much diversity as possible.

Simultaneously, to create the dataset of texts generated by LLMs, we used the Gemini free API. We hand-crafted several prompts for each domain to achieve the most appropriate one to use when generating texts.

For example, for every human-generated comment we extracted from Reddit, we provided an appropriate prompt to Gemini, which included 5 additional comments on the same post along with a request to generate a suitable AI-generated comment that aligns with the rest of the comments. In this way, we ensure that our dataset is diverse, balanced, and challenging for learning, as the response generated by the LLM is strongly based on texts created by humans.

Similarly, we expanded our dataset with 2 additional databases as described above.