Methods:

Objective:

The main objective of the research is to explore the ability of different language models to distinguish between texts generated by humans and those generated by LLM (Large Language Models).

Data collection:

As mentioned above, the main task was to collect and generate the appropriate data for training the network. At this stage, we needed to collect authentic texts created by humans, texts with the same context generated by LLM, and process the generated data.

During the research, we focused on three main sources of text: Reddit, Wikipedia and Newspaper websites.

Reddit:

The dataset we collected contains a collection of posts, with each post having a response written by a human and a response written by an AI model. We gathered the human-written responses using the Reddit API (PRAW), and the AI-generated responses were created using the Gemini API. The goal during the collection process was to create a large, reliable, and diverse dataset. To achieve these objectives, we conducted the collection in several stages.

Initially, we focused on diversity and volume to gather as many posts and human responses as possible, ensuring data reliability. Reddit's format is organized such that there are objects called subreddits, which represent 'groups' within the social network. In order to create a quantity and variety of data, we began by sending API requests to Reddit to retrieve posts from a few posts from each subreddit, but collected data from many subreddits.

The responses from the AI model were generated by sending prompts to each post containing relevant text to the request and including the following fields: title, subreddit, and some of the comments, for example.

לפני יצירת התגובות באמצעות מודל ה-AI שמנו לב שחלק מהתגובות שאספנו נראות לא אותטנטיות, תמונה, RMOVED, מלא שטויות.

Wikipedia:

Wikipedia is a free online encyclopedia, created and edited by volunteers around the world and hosted by the Wikimedia Foundation.

The website has a free API, which was used to extract articles from Wikipedia. Only articles which has an attribute of "summary" were extracted - a total of 4000 articles. Half of the articles comprised the human generated part of the dataset. The other half titles were inserted into an hand-crafted prompt, which in turn was fed to the Gemini model, to generate 2000 AI generated wikipedia style summaries. Many expirments were made in order to craft the most suitable prompt, and also to determine the best approach. At first the prompt didn't include a specific title, and the request was to generate a random article summary in a Wikipedia style. That attempt didn't work because the model repeated itself quite a lot. The second attempt was to give the model the title and body of the article (without the summary), and ask it to generate the summary. That attempt worked too well, generating a summary which is very similiar to the real summary - probably because the model was trained on that data. The final attempt was to give the model the title of the article, and it worked quite well. The code can be easily scaled up to generate more summaries, both human and AI generated - limited by the number of articles in the English Wikipedia, which is around 6 million.

Conclusion:

In our comprehensive study, we focused on the ability to answer the question of the origin of a given text. Ultimately, a deep learning model will answer this question by receiving a text as input and providing a classification response as output. Notably, that while the classification question is difficult and perhaps impossible for humans due to the development of LLMs today, based on the results of our research, it can be observed that this distinguishing ability is feasible. By creating a diverse, large, and balanced dataset, high performance and very high accuracy rates can be achieved.

Over time, of course, LLMs will improve significantly and rapidly, making the classification question of whether a text was created by a human or an LLM even more challenging. However, our research provides hope and indicates a close connection between the construction of the dataset intended for training the model and the ability of a deep learning model to notice differences, even those not visible to the eye, between different types of texts.