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.