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 model. 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:

Our final dataset contains a collection of posts, where from each post we extracted one comment created by a human and one generated by an LLM.

We gathered the human-written comments using the Reddit API (PRAW), and the AI-generated comments were created using the Gemini API. The goal during the collection process was to create a large, reliable, and diverse dataset.

To meet those criteria, we initially collected posts with comments from several subreddits which are community feeds dedicated to various topics. The subreddits were chosen randomly from the most popular subreddits list, to insure as much diversity as possible.

As mentioned above, for each post we extracted from Reddit, we chose one human-generated comment and one AI-generated comment. To create an AI-model comment for each post, we were sending a prompt that contain a relevant text to request the model to create this comment and including the following fields: the post title, the subreddit the post was taken from, and some of the post's comments for example. In this way, we ensure that our dataset is reliable an challenging for learning, as the comments generated by the LLM is strongly based on texts created by humans.

This attempt didn't yield good results since most of the comments did not appear to match human text. For example, comments containing a URL link, such as a link to an image, "[REMOVED]" comments indicating that a user's comment was deleted, unusual characters like emojis, very short comments like 'great' or 'ok', etc. Additionally, some comments contained offensive or racist information and tone.

To achieve as authentic comments as possible from each post, we worked on cleaning the dataset. In the first stage, we removed all non-text characters from the comments, such as emojis. In the second stage, we sorted the list of comments in each post so that the top comments in each post would be the best according to the parameters mentioned above. Of course, there is no direct way to do this, so we relied on external sources and created a weighting function that gives high scores to authentic comments and low scores to comments that are not authentic or offensive.

This process proved successful, where the final dataset we created was in the format of a post with two comments: one of the human comments, and the other comment generated by the model.

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.