Here I described how I searched and analyzed models and how I found the best model for this task. Initially, I embarked on an exploration of three distinct approaches to tackle the detoxification challenge: utilizing the top-performing model, employing a freely available pre-trained model, or crafting a custom model. To evaluate the effectiveness of each approach, I relied on the BLEU metric, which provides a clear and informative score.

Approach 1 - Leveraging the Best Model

To identify the most capable model for detoxification, my choice naturally gravitated towards OpenAI's GPT3.5-turbo model. I did not use GPT4 and I did not fine-tune the model, because it costs some money and I was too lazy to pay for that:) For a detailed understanding of its usage, you can refer to the code in the "str/models/gpt3.5-turbo" Python file (ensure that you've included the API key in the .env file before implementation). This approach boasts several advantages:

- No text preprocessing is required; we can directly input the text and request detoxification.
- Furthermore, no additional training is necessary as this model comes pre-trained, saving valuable time.
- The model's performance is highly impressive (although I didn't provide nice metric measurements, it's evident from the results).

However, the model's score of 7.2 in 100 samples may seem relatively modest, possibly due to some divergence in translations compared to the original dataset. Nonetheless, it excels in detoxification. An important caveat, though, is that it's not available for free, which prompted me to explore alternative solutions.

Approach 2 - Utilizing Free Pretrained Models

I delved into available models outside of OpenAI, with the T5 model being one of the prominent choices. Initially, I attempted to install and evaluate the "t5-11b" model, but its memory requirement of 45 GB proved to be a significant obstacle. Subsequently, I opted for the "t5-base" model and put it to the test. Surprisingly, it achieved a score of 8.8, surpassing the performance of GPT3.

My quest led me to the discovery of the "bart-base-detox" model, specifically designed for detoxification purposes. Leveraging this pretrained model for the entire dataset yielded a remarkable score of 19.6, surpassing the performance of

both the T5 and GPT models. Consequently, I made the decision to adopt this model as my ultimate choice for training and detoxification tasks.

Approach 3 - Creation of own models for this task

At first, the concept of developing a text generation model that produces non-toxic content appeared quite promising. Regrettably, certain challenges came to my attention, ultimately influencing my decision to abandon this endeavor:

- **Limited NLP Proficiency**: I encountered difficulties in grasping the intricacies of Natural Language Processing. Consequently, I realized my inability to construct models that could outperform specialists dedicated to excelling in the field of detoxification.
- **Time Constraints**: My schedule proved too restrictive, leaving me insufficient time to acquire the extensive knowledge required for crafting a robust model. While I remain hopeful that I might revisit this project in the coming months, for the present, I am unable to commit the entirety of my time to its development.

In light of these considerations, I opted to discontinue this particular approach.

Result of the search

My quest to find the most suitable model for text detoxification led me to explore various options, including OpenAI's GPT3.5-turbo and free pre-trained models like T5. Ultimately, the "bart-base-detox" model proved to be the most effective solution.