Project Overview

Reinforcing responsibility into language models: The case of OpenAI’s language generator GPT-3

Workshop01-Group03

# Project Summary

# Project Purpose

This project aims to figure out the current biggest concern of GPT-3 and comp up with useful and effective suggestions and recommendations for GPT-3 key stakeholders to help them to have better responsible developments on artificial intelligence.

# Project Challenge

Reviewing the social, ethical and legal concerns is challenging, as understanding how known discrimination and unconscious biases embedded in the language model is quite hard. Also, the efficient ways to mitigate social and ethical issues is still clouded.

**Point of View**

The developers of the project of GPT-3 need to improve the quality of the model output because the outputs of GPT-3 are still not semantically coherent and logical, as it may generate misleading information which would cause....

The clients who work on such as news institutions, education institutions, or other fields that are highly relevant to the public, need to pay more attention to potential risk of sexual and racial discriminations, and leaking sensitive information when using GPT-3, because most of datasets trained on GPT-3 come from all raw sources of the Internet, which contains many biases, discrimination information, and even identified sensitive information.

The policy makers and governments need to build new moral rules and public policy to prevent the influx of machine-generated disinformation and propaganda because a lack of safeguards and toxicity filters, a low cost of API entry, and no specialized technical required make GPT-3 successfully and effectively be weaponized by extremists to spread large-scale terrorism and radical violence.

# Project Client