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

Team:
The Explorer

Teammates:
Yuxuan Yang, Xufeng Zhu, Zixian Zhang, Fengqing Wu, Ming Xu

Background

GPT-3 IS NOT PERFECT !

Last June, OpenAI released the third version of a language model called GPT-3, it can not only write novels, poetry, and news reports, as well as summarise long texts, but also generate guitar tab, and create program codes.

As the training data contains the fake news, misinformation, sexual and racial discriminations, as well as the model itself could not recognize bias data, the outputs still have biases. Furthermore, training the GPT-3 was extremely financially and environmentally expensive.

Challenges

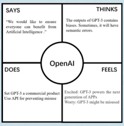
CAN WE RELY ON WAHT GPT-3 SAYS?

With the upcoming of industry 4.0, the structure of many industries would be changed due to the utilization of AI. And no matter in what kinds of industries, the biggest challenge is whether we can rely on the outputs of AI, so does the GPT-3.

Besides, many specific challenges are still waiting for overcoming, such as how the unconscious biases are embedded in the model, and how the model can be used for malicious purposes.

Problem

HOT MIGHT WE MITIGATE THE RISK OF MISUSE AND DISINFORMATION?



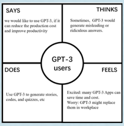
SAYS: “We would like to ensure everyone can benefit from Artificial Intelligence .”

THINKS: The outputs of GPT-3 contains biases. Sometimes, it will have semantic errors.

DOES: Set GPT-3 a commercial product
Use API for preventing misuse

FEELS: Excited: GPT-3 powers the next generation of APPs
Worry: GPT-3 might be misused

The Empathy map shows OpenAI knows the weakness of the GPT-3 and has taken some actions to protect from malicious use. It indicates that the power of GPT-3 makes the misuse become an urgent concern.



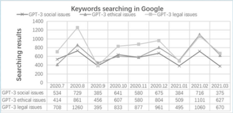
SAYS: we would like to use GPT-3, if it can reduce the production cost and improve productivity

THINKS: Sometimes, GPT-3 would generate misleading or ridiculous answers.

DOES: Use GPT-3 to generate stories, codes, and quizzes, etc

FEELS: Excited: many GPT-3 Apps can save time and cost.
Worry: GPT-3 might replace them in workplace

It points out users have high expectations from GPT-3. However, they would think about whether the outputs of GPT-3 are reliable or not, so the challenge might be to ensure the credibility of the results.



	2020.7	2020.8	2020.9	2020.10	2020.11	2020.12	2021.01	2021.02	2021.03
GPT-3 social issues	534	729	385	641	580	675	384	716	375
GPT-3 ethical issues	414	861	456	607	580	804	509	1101	627
GPT-3 legal issues	708	1260	395	833	877	961	495	1060	670

From the searching results, it shows that legal issues were the dominant concern, an However, in 2021, the ethical issues become as concerned as the legal issues, while the social issues drew the fewer attention in these 9 months.

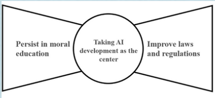
Solutions

LOTUS BLOSSOM CAN PROVIDE MANY IDEAS

Set up Regulations	Set up policies	Restriction on GPT-3 access
Improve accuracy of expression of GPT-3	How might we mitigate the risks of disinformation and misusing of GPT-3?	Building social norms
Filter outputs of GPT-3	Publicity	AI education

Recommandation

ONE CENTER TASK & TWO BASIC POINTS


Persist in moral education
Taking AI development as the center
Improve laws and regulations

One center task means to set developing artificial intelligence as the priority, because if people banned the usage of AI due to some issues, people cannot benefit from AI.

As for the two basic points, one point is improving laws and regulations to fit the future society. More and more AI issues would come up in different fields and industries as AI will become the main power in developments.

The other point is persisting moral educations. There are many ways to achieve that goal such as setting up an AI ethics course, conducting AI debate competitions, advertising, social norms campaigning, or even filming Hollywood Science Fiction movies.