Discussion about misleading chatGPT and corresponding solutions

Jingtong.Cai21

April 5, 2023

1 Introduction

The widely spread and usage of chatGPT demonstrated that Large Language Models (LLMs) will have a far-reaching influence on varieties of fields [1]. ChatGPT developed and maintained by OpenAI is a successful case of highly capable LLM. It is an interactive question-and-answer based chatting system, which is likely to assist users with professional or daily kind of tasks. Its great success mainly owes to the powerful ability to generate reasonable answer and interactive dialog interface which allows users to get timely response. Large language models are kind of like neural networks based on deep and machine learning and are pre-trained on massive collected data. LLM such as chatGPT has shown good performance in natural language processing (NLP) and natural language generation (NLG) work.

However, despite its strengths, LLM also shows some weaknesses in several particular fields. Since the power of chatGPT has raised concerns about plagiarism and cheating in academic area, the use of LLMs is forbidden in many educational institutions. Some may hold the question that whether it can be misguided or confused when generating corresponding essay. This article mainly discuss To mislead chatGPT to generate off topic or contradictory text, these particular fields mentioned above can be useful breakthroughs.

2 chatGPT failure

2.1 multilingual failure

It is commonly held that the amount of training data affects the final performance of LLMs [2][3]. Similarly, language understanding and generation ability of chatGPT could be influenced by the corresponding task language to some extent. There are mainly two significant differences for Multilingualism problems. The first is the difference between high resource, medium resource and low resource languages, and the second is the difference between Latin scripts and non- Latin scripts.

Evidence has shown that chatGPT reacted worse when dealing with questions in low resource language such as Buginese [2]. According to Bang et al [2], chatGPT achieves its accuracy of 84

The second point focuses on the performance difference between Latin and non-Latin scripts. It seems that chatGPT can better understand than generate non-Latin scripts which also demonstrated that answer in Chinese could cause more errors [4]. Not only that, but chatGPT also conduct an inferior translation from Latin to non-Latin scripts. Transformation between that is a useful trick to induce chatGPT to make mistakes.

2.2 reasoning failure

ChatGPT, as a large language model which is basically trained on textual data, is often reported to be lacking in ability of non-textual semantic reasoning such as temporal[5], spatial and mathematical reasoning[6]. There is an opinion states that AI model like chatGPT suffers the shortage of a "world model", which means it lacks a full understanding of the physical world, only being able to generate answer according to the training patterns[1]. Because of that, chatGPT could not correctly answer the question "how many birds are left on tree when shooting one of the five." This is very simple for human but difficult and crucial to test the intelligence for LLMs[7].

ChatGPT tends to be a lazy reasoner which can sometimes be easily mislead. There is an example: when asked about what gender will the first female president be, chatGPT said it is unpredictable. It also answered the same way for race and height. This shows that chatGPT sometimes fail to recognize the textual entailment, resulting in confusion in the context.

Since chatGPT possesses low capacity of connecting the entities and concepts together, misleading chatGPT is possible when topic is related with physical reasoning.

3 solutions and future work

To solve the inadequacy mentioned above, there are mainly two aspects to think through. The first is from the user's point of view. Interaction is an important evaluation of LLMs and chatGPT have made a great progress in that. Improving the interaction can improve the performance of chatGPT as well. Research shows that when the rounds of conversation increase, chatGPT tends to give more accurate answer in in turn. Consequently, user can make good use of the multi-turn character of chatGPT. If it makes mistakes, make it aware of the incorrectness step by step, correcting it by asking more in details. In addition to that, there are several specific phrases that can improve the quality of the receiving answer[8].

Secondly, from the developers' point of view, two most relative point is dataset and algorithm respectively. Dataset build up the fundament of LLMs as input model, allowing chatGPT construct a pattern of concept and entity and algorithm matters with matching the input with that pattern. The data input is not big enough or it contains some basis will all result in the deficiencies of chatGPT. The more comprehensive the dataset is, the better will the learning model be built. Algorithm concerns with the underlying kernel of AI's behavior, since chatGPT can often run on the separate way of our custom to solve the problem, we should truly understand the logic behind its every choice[9].

Apart from that, there is a little tip for developers to consider: enable chatGPT to show the level of confident about the answer, avoiding representing incorrect reply with certain tone[10].

4 conclusion

In conclusion, this article summarized two typical failures of chatGPT and provide some general suggestions for reference. In addition to the points mentioned above, there are many other aspects to think through, the more specific details such as translation accuracy and illusion towards things that do not exist. To make a complete understanding of chatGPT, recognize the its shortcoming is a key point.

References

[1] A. Borji, "A categorical archive of chatgpt failures," *ArXiv*, vol. abs/2302.03494, 2023.

[2] Y. Bang, S. Cahyawijaya, N. Lee, W. Dai, D. Su, B. Wilie, H. Lovenia, Z. Ji, T. Yu, W. Chung, Q. V. Do, Y. Xu, and P. Fung, "A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity," *ArXiv*, vol. abs/2302.04023, 2023.

[3] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, P. Schuh, K. Shi, S. Tsvyashchenko, J. Maynez, A. Rao, P. Barnes, Y. Tay, N. M. Shazeer, V. Prabhakaran, E. Reif, N. Du, B. C. Hutchinson, R. Pope, J. Bradbury, J. Austin, M. Isard, G. Gur-Ari, P. Yin, T. Duke, A. Levskaya, S. Ghemawat, S. Dev, H. Michalewski, X. García, V. Misra, K. Robinson, L. Fedus, D. Zhou, D. Ippolito, D. Luan, H. Lim, B. Zoph, A. Spiridonov, R. Sepassi, D. Dohan, S. Agrawal, M. Omernick, A. M. Dai, T. S. Pillai, M. Pellat, A. Lewkowycz, E. Moreira, R. Child, O. Polozov, K. Lee, Z. Zhou, X. Wang, B. Saeta, M. Díaz, O. Firat, M. Catasta, J. Wei, K. S. Meier-Hellstern, D. Eck, J. Dean, S. Petrov, and N. Fiedel, "Palm: Scaling language modeling with pathways," *ArXiv*, vol. abs/2204.02311, 2022.

[4] A. Wan, "Fairness in representation for multilingual nlp: Insights from controlled experiments on conditional language modeling," in *International Conference on Learning Representations*, 2022.

[5] L. Qin, A. Gupta, S. Upadhyay, L. He, Y. Choi, and M. Faruqui, “Timedial: Temporal commonsense reasoning in dialog,” in *Annual Meeting of the Association for Computational Linguistics*, 2021.

[6] M. Kosinski, “Theory of mind may have spontaneously emerged in large language models,” *ArXiv*, vol. abs/2302.02083, 2023.

[7] G. Kortemeyer, “Could an artificial-intelligence agent pass an introductory physics course?,” 2023.

[8] S. Kambhampati, “Changing the nature of ai research,” *Communications of the ACM*, vol. 65, pp. 8 – 9, 2022.

[9] A. Saparov and H. He, “Language models are greedy reasoners: A systematic formal analysis of chain-of-thought,” *ArXiv*, vol. abs/2210.01240, 2022.

[10] E. Davis, “Benchmarks for automated commonsense reasoning: A survey,” *ArXiv*, vol. abs/2302.04752, 2023.