

Discussion on the Sensitivity of Input to Large Language Models

Bochen Yuan

Department of Software Engineering, Hebei University of Science and Technology, Shijiazhuang, China

2316030130@stu.hebust.edu.cn

Abstract. The Large Language Model (LLM) was being widely used in daily life, but sometimes the LLM will provide a different reply when inputting the same content, but different format. To research this phenomenon, this paper aims to analyze the reason why this phenomenon will occur, thus this paper main to research the process of LLM dealing with the text, and research whether the prompt can influence the process of LLM to deal with the information of input, and the influence of different input modes. This paper provides some solutions to deal with this question. These solutions mainly focus on transitioning the input into a format that LLM can completely understand, and LLM provides many replies for users to choose from, and trains LLM to learn from the mapping database until it can provide replies that users need. But this research also has limitations, such as the current research lacks a details discussion on LLM, such as whether the difference in period or the size of text will cause a difference.

Keywords: Shortcut Features, input, prompt, LLM.

1. Introduction

As the development of technology advances, the large language model (LLM) has been widely used in people's daily lives. And the main function of LLM is to accept questions and deal with them by using a database. The LLM has become increasingly important for humans, as LLM can be used to predict the development of the social [1], or it can also be used in a simulation experiment. Although LLM is intelligent, it also has some problems, such as it may hallucinate, which means LLM will provide a reply, but it makes an error for the question [2], and this trouble will cause the same question, but if it has different input, the quality of the reply also has a huge difference.

The LLM gradually fit into prople daily life such as LLM often be used in medical zone such as it can help doctor to analyze medical image [3] and patients' symptom or it can also help doctor to get new medical knowledge to convenience doctors' work, and LLM also can be used in the field of robotics in that file LLM often be used to translate nature order for a code so that let robot understand the order [4], in conclusion the LLM has been used in widely but this usage scenario all need LLMs provide a precise and accurate reply but sometimes LLM often cannot give right answer this phenomenon was collectively referred to as hallucination [2], and one of the key factor to produce hallucination depend on the content of input, sometimes the quality of input often has an impact on the quality of reply, if provide LLM a same question but different prompt if the prompt can let LLM get the core idea the LLM will provide a suitable reply, if not or this idea is false the reply will mutual contradiction [5], or sometimes LLM cannot deal with complex sentence, it will rely on database to give reply directly this phenomenon is shortcut learning which is the extension of hallucination [6]. This trouble will also cause an LLM product error.

To improve the accuracy rate of LLM, this research combines the relevant references in recent years to introduce the reason why inputting different formats will cause different quality replies, and makes a summary of the solutions to deal with it, and evaluates each solution. Aim to provide the reference and basis for the development of the LLM to deal with the input message's ability

2. Study of Prompt

When using LLM, the reply often has a huge difference, such as Low-quality answers or irrelevant answers. The reason why this phenomenon occurred is that LLM produces hallucination or shortcut features, because LLM cannot deal with this question, so it will reply to the question with a sample in the question. This phenomenon will cause LLM to provide a reply, but have nothing to do with the question, or provide a reply only by the content on the context, so here are some studies of the sensitivity of LLM to deal with questions.

To research the process of LLM dealing with the information of input [7], Brown et al., train GPT-3 so that it can test its performance by text interaction, specify tasks, and provide a small sample demonstration. They make an exploration about the different levels of information LLM can get from input and divide them into some levels: FT (fine tuning), 1S (one sample), and 0S (0 sample). They use the same model as GPT-2 but with a localized band-shaped sparse attention pattern in the transformer, train 8 models, and hundreds of millions of parameters. Train the database, including downloading, filtration of Common Crawl versions, removing duplicates, and adding the high-quality Reference corpus. In the training, the bigger model often used a bigger batch size. In order to train the bigger model, they use a mix of model parallelism within matrix multiplication. And assess the experiment by randomly selecting K examples from the training set of this task, and assess each of them. After training, the ability of LLM to deal with information has a huge improvement, such as doing the cloze test, questions and answers, translating, SuperGLUE, and preventing baseline memory. They find that LLM often deals with information by context learning. If they let GPT-3 learn by 0F, it usually repeats itself on the document-level semantics and includes illogical sentences. If want to let it deal with information, by 0F extremely expanding the model scale is necessary.

For the same question, LLM usually provide different quality replies. Zhou et al., think the reply quality was greatly impacted by the quality of the prompt [8]. To verify this point, they conducted research mainly on the difference in human expression for the same question. In that experiment, they use an LLM that has been trained to provide a response to a set called U. and to optimize the question, they choose a function to standardize and measure the data set and model-generated data. To produce favorable proposals for collection, they use Iterative Monte Carlo Search to get an instruction that can achieve success more easily, and call this APE. In the last, they assess many zero samples and a small sample size to carry out context learning, and to check whether APE can be best used, they use GPT to create some instructions to verify that APE can get better performance. And then use APE on the TruthfulQA to answer questions of different styles find a question by APE can get a better answer than raising a question by a human. So, LLM deals with very sensitive questions; the tiny difference will cause it to produce a different reply.

The difference in input can have a huge influence on LLM Research by lu et al., who have shown that the order of prompts that provide LLM can affect the context learning ability of LLM [9], thus affecting the quality of answers that LLM provides. The reason is that LLM is very difficult to obtain the keywords from the question. So they designed an experiment to verify this guess. At first, they trained the sample set so that they could build the detection set directly. And based on the entropy index, to create two arrangements of Prompt Words, which are called GlobalE (main to find to avoid prompts that are imbalanced) and LocalE (main to find Insecure Prompt Words), and choose a sample ranked with 24 different arrangements, 5 different sets, and this experiment has been run 120 times. Through the standard deviations of five different sets, select four high-entropy value arrangements as prompts and compare with the oracle. The experiment result shows that GlobalE and LocalE both improved, and the prompt with high performance is more stable than the other prompts. This result shows that the order of the prompt is one of the key factors that affect the quality of the reply.

3. Discussion

The prompt that humans had input had a huge effect on LLM; the tiny differences, such as the sequence of statements or deviation of expression, can all cause a different reply; thus, the text that user providing LLM a suitable sentence so that LLM can understand is a good idea, however, this method is difficult to realize the reason is that need users have a deep understanding of the LLM procedure for processing sentences, thus to standardizing the input that LLM received it can add a function that mainly receives the sentence that people provide and converts it so that LLM can understand it completely, to reduce the variability of the text that humans provided.

LLM often relies on In-Context Learning when processing text; however, when dealing with large amounts of text, LLM will difficult to find keywords to carry out In-Context Learning. To deal with this trouble, LLM can limit the input that users provide, such as when the text that users input exceeds the limit that LLM can handle normally. LLM reminds the user should provide some keywords or the focus of the text, otherwise LLM may not provide a precise answer for users. However, this method is not the most perfect one because sometimes users may not identify the keywords in the text accurately that they had provided because of a large amount of text; thus, in order to enable the LLM can easier use In-Context Learning, it can add a processing step when dealing with text, which main aim is to receive the text and split it based on the main meaning expressed in the sentence. All these methods are key to enhancing the use of In-Context Learning by LLM.

The difference in the input may cause the LLM to provide a reply that doesn't what the user's needs. Thus, LLM can, on the basis of the question that the user provides and switch to a different format and deal with it to provide different responses. And the user can search for responses that they require.

When using an LLM, it is common to encounter the phenomenon that for the same input content, but if in a different prompt, the reply that the LLM provides will vary. To deal with this trouble, LLM can build a mapping database in advance, train the LLM to process different formats of data until it extracts a set of data results, so that let LLM learn from it so as to reduce the phenomenon that providing a reply but not meeting the requirement.

4. Limitation

This research has certain limitations. The data source of the LLM in that research is rather limited, mainly focusing on the different versions of GPT and some other LLM, thus, the result of the research may show some deviations. Show some deviations. Currently, research lacks improvement in the quality of LLM, and all current solutions need high financial support or a deep understanding of LLM, Research for the reasons why different inputs have different replies only points out that the two aspects are illusions and rely on shortcuts. Lack of some detailed discussion, such as when using LLM, the size of the input content, whether it can influence the positive of LLM to deal with questions

5. Future expectations

The future development focus of LLM should be on giving precise answers to the questions that users ask, such as increasing the accuracy when dealing with a large amount of complex text, or expanding the database of the LLM so that it reduces the dependence on In-Context Learning and the frequency of hallucinations. And pay more attention to reducing the degree of complexity to deal with the solution, such as simplification the solution to reduce the fund demand and conducting research to observe the tiny difference in each research, such as the size of input, the input time period, and the different LLM whether can provide different quality of replies when inputting the same content [10, 11]. In fact, recent survey work has emphasized taxonomy of hallucination phenomena and opportunities for mitigation via retrieval and uncertainty estimation techniques. Also, work on prompt robustness has shown that small perturbations in prompt format can greatly influence output stability, underscoring the need for prompt-adaptive design in future LLMs [12].

6. Conclusion

This research aims to provide a review of the sensitivity of input for LLM, such as using the same content but different text input, but LLM will provide different replies. This research is mainly to discuss the process of LLM dealing with the text of input, and the main reason why LLM will cause this phenomenon. and by analyzing this question, this research provides some solutions to deal with this question; these solutions focus on changing the input to a format that LLM can completely understand, or provide many replies to users so that they can find a reply that they need. In the future. But the research on LLM still has a deficiency; the reason why this phenomenon occurs is lacking in a more detailed discussion, in the future, this situation should be dealt with.

In addition, future work should also focus on systematic evaluation frameworks that can quantify the variability of outputs under controlled input perturbations. Such frameworks would allow researchers to measure sensitivity across different tasks, languages, and domains, providing a clearer picture of when and why inconsistencies arise. Moreover, practical applications such as healthcare, education, and robotics demand stable and trustworthy outputs, making it critical to design adaptive prompting strategies and fine-tuning methods that minimize unpredictable variations. Beyond prompt engineering, integrating retrieval mechanisms, error-detection modules, or human-in-the-loop systems could significantly improve reliability. Another promising direction lies in studying the relationship between model scale and sensitivity, as larger models may exhibit both greater robustness and novel vulnerabilities. Ultimately, addressing input sensitivity will not only enhance user trust but also broaden the safe deployment of LLMs in high-stakes environments where precision and consistency are essential.

References

- [1] Yang K, Li H, Wen H, et al. Are Large Language Models (LLMs) Good Social Predictors?. arxiv preprint arxiv: 2402.12620, 2024.
- [2] Hao G, Wu J, Pan Q, et al. Quantifying the uncertainty of LLM hallucination spreading in complex adaptive social networks. *Scientific reports*, 2024, 14 (1): 16375.
- [3] Panagoulas D P, Virvou M, Tsihrintzis G A. Evaluating LLM--Generated Multimodal Diagnosis from Medical Images and Symptom Analysis. arXiv preprint arXiv: 2402.01730, 2024.
- [4] Ahn M, Brohan A, Brown N, et al. Do as i can, not as i say: Grounding language in robotic affordances. arXiv preprint arXiv: 2204.01691, 2022.
- [5] Manakul P, Liusie A, Gales M J F. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. arXiv preprint arXiv: 2303.08896, 2023.
- [6] Yuan Y, Zhao L, Zhang K, et al. Do llms overcome shortcut learning? an evaluation of shortcut challenges in large language models. arxiv preprint arxiv: 2410.13343, 2024.
- [7] Brown T, Mann B, Ryder N, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 2020, 33: 1877-1901.
- [8] Zhou Y, Muresanu A I, Han Z, et al. Large language models are human-level prompt engineers. *The eleventh international conference on learning representations*. 2022.
- [9] Lu Y, Bartolo M, Moore A, et al. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. arXiv preprint arXiv: 2104.08786, 2021.
- [10] Huang L, Yu W, Ma W, Zhong W, Feng Z, Wang H, Chen Q, Peng W, Feng X, Qin B, Liu T. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*. 2025 Jan 24; 43 (2): 1-55.
- [11] Farquhar S, Kossen J, Kuhn L, Gal Y. Detecting hallucinations in large language models using semantic entropy. *Nature*. 2024 Jun 20; 630 (8017): 625-30.
- [12] Zhu K, Wang J, Zhou J, Wang Z, Chen H, Wang Y, Yang L, Ye W, Zhang Y, Gong N, Xie X. Promptrobust: Towards evaluating the robustness of large language models on adversarial prompts. *InProceedings of the 1st ACM workshop on large AI systems and models with privacy and safety analysis 2023 Nov 19 (pp. 57-68)*.