ISSN 2959-6157

Prompt Recovery for Large Language Models

Ruochen Feng[1,*], Jincheng Hu[2], Yifang Chen[3]

[1]University of Toronto, Toronto, Canada, ruochen.feng@mail.utoronto.ca
[2]School of Information Science, University of Illinois, Urbana-Champaign, Champaign, United States, hu87@illinois.edu
[3]Beijing Normal University-Hong Kong Baptist University United International College, Zhuhai, China, s230026021@mail.uic.edu.cn

Abstract:

Understanding and recovering prompts in large language models (LLMs) is vital for addressing concerns related to privacy, copyright, and beyond. However, there is a lack of extensive research in this area. To fill this gap, we implemented model stacking techniques, such as utilizing mean prompts and embedding models, tailored for specific datasets. While these individual models were designed for particular datasets, our combined stacking model demonstrated improved accuracy in prompt recovery across diverse datasets. Although there was a slight decline in performance on the initial dataset, our comprehensive evaluation across multiple LLMs and prompt benchmarks indicates that our stacking model exceeds the performance of existing methods. Notably, this approach uses a single LLM without depending on external resources, making it an efficient and accessible solution for prompt recovery.

Keywords: — Large Language Models, Prompt Recovery, Pre-trained Model, Model Stacking, Predictive Entropy

1. Introduction

Large language models are complex neural network models trained by large amounts of data. In recent years, significant progress has been made in the field of natural language processing, and the performance of text generation, translation, question answering system and other tasks has been greatly improved. Models based on the Transformer architecture, such as the BERT and GPT series [1], demonstrate unprecedented language understanding and generation capabilities through self-supervised learning and extensive corpus training. In text generation tasks, LLM can generate long text that is coherent and context-relevant. In the task of machine translation, the quality of machine translation has approached or even exceeded the level of human translators. In question answering systems, LLMS can understand complex questions and provide accurate answers, which shows the strong performance of LLMS in various NLP tasks [1].

It is precisely because of the high efficiency and powerful performance of large-scale language models in handling various tasks that they are widely used in various applications. One of the most common ways to use it is to output text based on the given prompt words. However, as the size of LLM parameters continues to increase, so does the computational and storage pressure. Training a large model typically requires thousands of GPU computing hours and consumes large amounts of power and storage resources, which is a huge burden for many research institutions and enterprises. In addition, LLM training data often contains a large amount of personal information and sensitive data, which raises concerns about data privacy and security [2]. In the application process, how to effectively protect user privacy and avoid data leakage has become an urgent problem to be solved [3].

In practical application, LLM also faces a series of ethical and social issues [1]. First, models can inherit and amplify biases in training data, leading to unfair results in terms of race, gender, and so on. Secondly, the reliability of model output has also attracted much attention, especially in high-risk scenarios such as automatic driving and medical diagnosis, where wrong output may bring serious consequences. Therefore, how to reduce bias and improve output reliability in the process of model design and training has become the focus of research [2].

LLM trends will include more efficient training methods and enhanced model capabilities. Researchers are exploring new algorithms and architectures to reduce training costs and time while improving model performance. In addition, future LLMS will pay more attention to the controllability and interpretability of model outputs, making them more transparent and trustworthy in applications. Combining cutting-edge technologies such as multimodal learning and federated learning, LLM is expected to play a bigger role in a wider range of fields and advance NLP technology in a more responsible and sustainable way.

In this context, prompt word recovery technology, as one of the important applications of LLM, has attracted more and more attention. Through prompt word recovery technology, more intelligent and humanized text generation can be realized, so as to improve user experience and application effect.

As the parameter scale of large language models (LLMs) increases, there has been a discernible shift towards providing inference-only application programming interfaces (APIs) to users, exemplified by services such as ChatGPT. While this trend makes it easier for users to use the models, it also makes prompt recovery more difficult and important. For researchers, understanding which prompts lead LLMs to generate specific outputs is crucial for improving the model's understanding and interpretability. Developers and researchers may need to recover prompts from system outputs to debug models or analyze their behavior for improvement and optimization. Additionally, when LLMs generate content protected by copyright or containing sensitive information, determining the details of the original prompt is crucial for accountability and legal compliance. Furthermore, understanding and predicting how large language models (LLMs) respond to specific prompts is crucial for ensuring that their outputs are safe and reliable [3]. Prompt recovery can help identify how the model may be abused and design appropriate defenses when needed.

In our methodology, we adopted a hybrid approach that combines "average prompt" and model prediction. This approach consists of the following five main components:

1) Average prompt template: We use the formatted string of prompts predicted by the model.

2) Fine-tuned MistralForCausalLM model for predicting complete prompts.

3) Trained MistralForSequenceClassification model for filtering incorrect prompt predictions.

4) MistralForCausalLM model for predicting the labels of samples.

5) Two clustering models for clustering test samples and selecting the best average prompt template.

2. Related Works

1.1 Prompt Recovery

Large language models (LLMs) have transformed natural language processing by excelling in tasks such as text generation, translation, and question-answering [4]. These models, trained on vast datasets, can generate human-like text from given prompts. However, as these models grow in scale, understanding and interpreting their behavior becomes increasingly complex. Prompt recovery—the process of identifying the initial input that led to a specific output—has become essential for various purposes, including debugging, optimization, transparency, and ensuring the safety and reliability of the models.

Effective, prompt recovery is based on understanding the relationship between an LLM's input prompts and its generated outputs. By analyzing the prompts that produce specific outputs, researchers can gain insights into how LLMs process and generate text. This knowledge can inform improvements and optimizations in the model [5]. Developers can use prompt recovery to identify and address issues in model behavior, optimize prompts for better performance, and enhance the overall efficiency of the model. Additionally, when LLMs generate content with potential legal or ethical implications, prompt recovery can trace the output back to the original input, ensuring transparency and accountability.

1.2 Evaluation Metrics

Existing literature includes various evaluation metrics for assessing the accuracy of prompt recovery. In this study, we primarily employ two metrics: sharpened cosine similarity (SCS) and predictive entropy (PE) [4, 6]. These metrics provide a robust framework for evaluating the effectiveness of our prompt recovery approach.

To compute these metrics, we first use the sentence-t5-base model to generate embedding vectors for each row in the recovered prompt and its corresponding ground truth [6]. The sharpened cosine similarity is then calculated for each predicted/expected pair, utilizing an exponent of 3. This exponentiation serves to attenuate the inflated scores that embedding vectors may assign to incorrect answers, thereby providing a more precise measure of similarity.

Furthermore, we incorporate predictive entropy to assess the uncertainty associated with the entire output sentence [4]. Predictive entropy is calculated as the sum of the entropies of each token in the sentence, offering a comprehensive measure of uncertainty. This metric is particularly valuable for understanding the confidence level of the model's predictions and identifying areas where the model might require further refinement.

3. Method

The hybrid method utilizes the advantages of multiple models and techniques to achieve effective and rapid recovery. This workflow consists of five main components that are integrated together to provide robust and accurate real-time predictions.

1) Average prompt template

The average prompt template serves as the foundational structure for various prediction formats and is developed through several key steps. First, data generation involves using large language models (LLMs) such as Gemini and GPT-3.5 Turbo to create datasets with potential rewrite hints [7]. By generating variations of the original prompt, diversity is increased, and raw text that aligns with the prompt features is produced.

Next, subsamples are selected to match the distribution of the public ranking dataset. Beam search is employed to optimize word combinations, thereby improving the scores of these subsamples. Finally, the average prompt is optimized by using a beam search to identify the best combination of words, and the correlation between subsample scores and rankings is verified.

This structured approach ensures that the average prompt template is robust and adaptable, providing a solid foundation for accurate and diverse prediction formats.

2) Mistral For Cause LLM model for real-time prediction

The model has been fine tuned to predict complete prompts. The training includes the following steps. First is data generation, which uses LLM to generate prompt candidates. Create variations of the original prompt and generate corresponding rewritten text. Also use the Mistral-7B-Instruct-v0.2 model fine tuned with LoRA [7]. Integrate public datasets and generate variants to enhance the diversity of training data.

3) Mistral For Sequences Classification model for error prompts filtering

The gate model filtered out incorrect prompt predictions. The training process includes tip construction and training strategy. Create a prompt that combines the original text, candidate prompts, and rewritten text. Using positive and negative samples (randomly select incorrect prompts and similar prompts in the T5 embedding space) [7]. Train the model to classify whether the rewritten text matches the candidate prompts.

4) Mistral For Cause LLM model for label prediction

This model predicts labels related to the sample, such as tone and style. Labels provide additional context for prompt prediction. Label prediction: Train the model to predict labels based on samples. Ensure to include tags that have not yet appeared in the complete prompt prediction [7].

5) The Clustering model for optimal prompt selection

Two clustering models, Mistral clustering and KMeans clustering, are used to group test samples and select the best mean prompt template. Clustering strategy: Based on T5 embedding, fit the KMeans model with 12 clusters [7]. Optimize the average prompt for each cluster using LBFGS. Cluster protocol: If two cluster models are consistent, the average prompt template in the agreed cluster is used. Otherwise, please use the global mean prompt template.

6) Output Structure

The final output is constructed by combining the selected average prompt template with unique words from complete prompt predictions and labels. First are unique words. It will remove overlapping words to ensure uniqueness. Furthermore, also inserting a label and complete prompt after the third word in the middle prompt, as this placement displays the best results in validation.

The hybrid method combining mean prompt and model prediction provides a robust solution for fast recovery in large language models. By utilizing the advantages of multiple models and optimizing the fast selection process, this method improves the interpretability, reliability, and performance of LLM [7]. This method not only helps debug and optimize models but also ensures accountability and compliance in scenarios involving sensitive or legally protected content. The detailed workflow and component integration demonstrates the practical application of theoretical principles in achieving effective and rapid recovery.

4. Limitation

Although DORY has achieved remarkable results on multiple mainstream large language models (LLMs), the method has not been validated on more advanced models such as GPT-4 due to cost constraints. The rapid development of LLMs has led to continuous improvement in their understanding and reasoning capabilities [4]. Advanced LLMs such as GPT-4 have further expanded the capabilities of natural language processing with their enhanced features and refined architectures. This evolution has brought some uncertainty to the effectiveness of the method.

Evaluating the method on these cutting-edge models requires a lot of resources, including considerable computing power and access to the models themselves [4]. These principles are inherently applicable to the architectures of various LLMs, indicating that the method may have potential even for state-of-the-art models.

5. Conclusion

In conclusion, this paper presents a comprehensive exploration of prompt recovery in large language models (LLMs), an important component in enhancing model interpretability, optimizing performance, and ensuring accountability in complex NLP applications. Utilizing a hybrid methodology, we combined an average prompt template with fine-tuned models, alongside clustering techniques, to develop a robust framework for accurate and efficient prompt recovery.

The hybrid method for prompt recovery offered a scalable solution for improving the transparency and control of LLM outputs. This approach not only aids in debugging and optimizing models but also supports legal and ethical considerations in sensitive content generation. As LLMs continue to evolve, prompt recovery will become increasingly important for ensuring the responsible deployment of these powerful models across various domains.

References

[1] J. Schulman et al., “Introducing ChatGPT,” OpenAI, 2022. [Online]. Available: https://openai.com/blog/chatgpt.

[2] F. Zheng, “Input reconstruction attack against vertical federated large language models,” arXiv preprint arXiv:2311.07585, 2023.

[3] J. X. Morris, W. Zhao, J. T. Chiu, V. Shmatikov, and A. M. Rush, “Language model inversion,” 2023.

[4] L. Gao, R. Peng, Y. Zhang, J. Zhao, “DORY: Deliberative Prompt Recovery for LLM,” arXiv:2405.20657, 2024

[5] W. Zhao et al., “A Survey of Large Language Models,” arXiv:2303.18223, 2023

[6] S. Bubeck et al., “Sparks of Artificial General Intelligence: Early experiments with GPT-4,” arXiv:2303.12712, 2023

[7] LLM prompt recovery. Kaggle. (n.d.-b). https://www.kaggle.com/competitions/llm-prompt-recovery/discussion/494621