LLM Prompt Recovery

Kevin Chau [*1] **Dhawal Modi** [*1]

Abstract

Large Language Models (LLMs) are increasingly used to format and write texts in stylistic ways. The goal of this project is to recover the LLM prompt that was used to transform and rewrite a given text. This project is hosted on Kaggle and the model will be evaluated with an unseen test dataset generated by Gemma, Google's new family of open models. As a preliminary result, we submitted a pre-trained Mixtral 8x7B model to the Kaggle competition and received a Cosine Similarity score of 0.57 across all prompt predictions.

1. Introduction

The relatively recent of advent of publicly accessible Generative AI tools such as ChatGPT (OpenAI, 2024) has created an explosion in the general application of deep learning. We are now entering a new AI age where users can leverage Large Language Models to assist with everyday creative writing tasks such as generating essays, emails, articles, and stories, all with professional and authentic quality.

New LLMs with increasing capabilities are being released every quarter, but there is little literature on how to effectively prompt an LLM. This has led to the process of "Prompt Engineering", which studies how to tune prompts to get the best response from an LLM. The goal of this project is to create a model that can recover the LLM prompt that was used to rewrite a given text. By focusing on the inverse problem of **Prompt Recovery**, we hope to develop a deeper understanding of AI driven NLP and LLMs.

Our prompt recovery model will be tested on the Kaggle platform (via a public code competition) against a dataset of over 1300 original texts paired with a rewritten version from Gemma, Google's new family of open models.

[*]Equal contribution [1]Department of EECS, University of California, Merced, USA. Correspondence to: Kevin Chau <kchau15@ucmerced.edu>, Dhawal Modi <dmodi2@ucmerced.edu>.

EECS 230 Deep Learning Spring 2024, Merced, California, USA.

2. Background and Related Work

2.1. Google Gemma Open LLM

Earlier this year, Google AI released **Gemma** (Gemini Team Google, 2023)(Google AI, 2024), a family of new state-of-the-art open models based on the same technology as Google's Gemini LLM. Unlike Gemini, their largest and most capable AI application currently available, Gemma models are open-source and comparatively lightweight, as they are tailored for software developers and machine learning researchers.

Gemma model weights have been released in two sizes: Gemma 2B and Gemma 7B. JAX, Pytorch, and Tensorflow backends are supported through Keras, which enables a large variety of software toolchains for development. Gemma has readily available integrations with Colab, Kaggle, Hugging Face, TensorRT-LLM, Vertex AI, and Google Kubernetes Engine. It is highly optimized for different AI hardware platforms including NVIDIA GPUs and Google Cloud TPUs. For these reasons, we believe studying prompt recovery with Gemma is highly valuable and can have far reaching impacts in AI research.

2.2. Prompt Engineering and Recovery

The problem of how to best interact with AI models to generate the most desirable and accurate output text is relatively new and still being actively researched. **LLM Prompt Recovery** can help us gain more insights about how to accurately craft prompts to get the best outputs from an Instruction-tuned LLM. While there have been virtually no studies on prompt recovery, RewriteLM (Shu et al., 2023) talks about how an Instruction-Tuned Language Model for text rewriting tackles the problem of generating rewritten texts through a special prompt.

2.3. Dataset

The dataset for the competition consists of a tuple of the following triplets, $\langle$*Original Text*$\rangle$, $\langle$*Rewrite Prompt*$\rangle$, $\langle$*Rewritten Text*$\rangle$. We generated our own synthetic dataset using the Gemma (Google AI, 2024) model with 7 billion parameters. The model was instruction-tuned. The text corpus for $\langle$*Original Text*$\rangle$ was obtained from Wikipedia-

Movie-Plots and Kaggle forum messages (Plotts & Risdal, 2023).

3. Proposal

The goal of this project is to predict a prompt, ⟨*Rewrite Prompt*⟩, that was used to generate ⟨*Rewritten Text*⟩ by providing only ⟨*Original Text*⟩-⟨*Rewritten* Text⟩ pairs to the LLM. To tackle this challenge of prompt recovery we propose training a DeBERTa base model (He et al., 2021) that minimizes cosine similarity between the predicted embedding and the true embedding from a *sentence-t5* model (Ni et al., 2021). Then at inference, it predicts an embedding using the ⟨*Original Text*⟩ and the ⟨*Rewritten Text*⟩. This embedding is compared against a knowledge database of embeddings. The most similar embedding in this database is our model prediction. The score for each predicted / expected pair is calculated using the Sharpened Cosine Similarity, using an exponent of 3.

4. Preliminary Results

4.1. Motivating Example

To make clear what we mean by **prompt recovery** we will start with a simple motivating example for demonstration. Suppose we have some **original text** that we want an LLM to *transform* such as the following movie plot summary, which was taken from the Wikipedia Movie Plots database:

> ”Scenes are introduced using lines of the poem. Santa Claus, played by Harry Eytinge, is shown feeding real reindeer and finishes his work in the workshop. Meanwhile, the children of a city household hang their stockings and go to bed...”

We may ask any well-trained LLM to **rewrite** the text for us, using a stylistic **prompt** or specific guidance for how we want the **rewritten text** to appear. Some simple prompts could be: ”Make this rhyme”, ”Make this shorter”, ”Make this longer”, ”Rewrite this as a poem in the style of William Shakespeare”.

4.2. Rewriting Text with Gemma

We wrote a python notebook to programmatically interact with the Gemma 7B model, since there is currently no user-friendly GUI-based web application widely available.

Continuing with our running example, we asked Gemma to take the plot summary of *The Night Before Christmas (1905)* and **”Make it rhyme”**. The pre-trained model produced the following rewritten text:

> ”Sure, here’s the rhyme: Scenes dance with lines of rhyme, Santa’s visit, a wondrous chime. He feeds reindeer, a festive sight, And finishes his work, day and night...”

With our python notebook, we asked Gemma to process a set of original texts extracted from several databases with randomly picked prompts from a handmade list. Using this method, we are able to generate datasets of arbitrary length for testing and training (see section 2.3).

4.3. Baseline prompt prediction using existing LLMs

For a performance baseline, we asked several current LLMs to make prompt predictions given an original text and rewritten text pair. They were not given any knowledge of the original prompt. We tried several smaller LLMs such as Gemma itself and Mixtral (Jiang et al., 2024), as well as larger production models including ChatGPT, Gemini, and Claude. Specifically, we queried the LLMs in this format:

> **”Original Text: ⟨*Original Text*⟩ Rewritten Text: ⟨*Rewritten Text*⟩ What was the specific rewrite prompt for this pair of text? Limit output to one line.”**

For the *Night Before Christmas* example, they produced the following prompts:

LLM	***Predicted Prompt***
Gemini	Rewrite the passage in a rhyming and poetic style, maintaining the original story elements.
Claude Sonnet	Convert the given prose description into a rhyming poem.
ChatGPT 3.5	The specific rewrite prompt is not provided, so the details of the instructions for generating the rewritten text in one line cannot be determined.
Gemma 2B	A. to bring the poem to life B. to make the poem more interesting C. to make the poem more understandable...
Mixtral 8x7B	Transform the original essay into a rhymed poem, preserving the key events while adding poetic language and meter.

Gemini, Claude, and Mixtral inferred prompts that correctly identify the rhyming instruction, but fail to reproduce the exact prompt with respect to brevity and specific language. Interestingly, ChatGPT failed to comprehend our prompt recovery query and Gemma 2B was perhaps too small to sufficiently understand the task.

It is clear that using existing pre-trained models is not sufficient for the task of prompt recovery. We aim to develop a much better solution to this problem by training our own LLMs as outlined in the proposal section.

Software and Data Availability

All of the results and software used to develop our models are publicly available and can be found on GitHub here: https://github.com/kevin-chau/llm-prompt-recovery

References

Gemini Team Google. Gemini: A family of highly capable multimodal models. *arXiv:2312.11805 [cs.CL]*, 2023.

Google AI. Gemma: A family of lightweight, state-of-the art open models built from the same research and technology used to create the gemini models, 2024. URL https://ai.google.dev/gemma.

He, P., Liu, X., Gao, J., and Chen, W. Deberta: Decoding-enhanced bert with disentangled attention, 2021.

Jiang, A. Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D. S., de las Casas, D., Hanna, E. B., Bressand, F., Lengyel, G., Bour, G., Lample, G., Lavaud, L. R., Saulnier, L., Lachaux, M.-A., Stock, P., Subramanian, S., Yang, S., Antoniak, S., Scao, T. L., Gervet, T., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. E. Mixtral of experts, 2024.

Mistral AI. Mixtral of Experts, 2024. URL https://mistral.ai/news/mixtral-of-experts/.

Ni, J., Ábrego, G. H., Constant, N., Ma, J., Hall, K. B., Cer, D., and Yang, Y. Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models, 2021.

OpenAI. ChatGPT, 2024. URL https://chat.openai.com/chat.

Plotts, J. and Risdal, M. Meta kaggle code, 2023. URL https://www.kaggle.com/ds/3240808.

Roberts, A., Chung, H. W., Levskaya, A., Mishra, G., Bradbury, J., Andor, D., Narang, S., Lester, B., Gaffney, C., Mohiuddin, A., Hawthorne, C., Lewkowycz, A., Salcianu, A., van Zee, M., Austin, J., Goodman, S., Soares, L. B., Hu, H., Tsvyashchenko, S., Chowdhery, A., Bastings, J., Bulian, J., Garcia, X., Ni, J., Chen, A., Kenealy, K., Clark, J. H., Lee, S., Garrette, D., Lee-Thorp, J., Raffel, C., Shazeer, N., Ritter, M., Bosma, M., Passos, A., Maitin-Shepard, J., Fiedel, N., Omernick, M., Saeta, B., Sepassi, R., Spiridonov, A., Newlan, J., and Gesmundo, A. Scaling up models and data with `t5x` and `seqio`, 2022.

Shu, L., Luo, L., Hoskere, J., Zhu, Y., Liu, Y., Tong, S., Chen, J., and Meng, L. Rewritelm: An instruction-tuned large language model for text rewriting, 2023.

Will Lifferth, Paul Mooney, S. D. A. C. Llm prompt recovery, 2024. URL https://kaggle.com/competitions/llm-prompt-recovery.