NLP Project: Instruction Finetuning of ChatGLM-6B with LORA

XiaoYuxiang

July 2023

1 Prompt Engineering of BlueElephant

1.1 Prompts Diversity Analysis

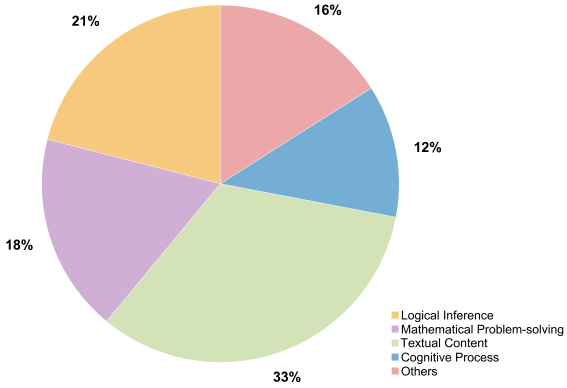

Figure 1: Distribution of Prompt Types

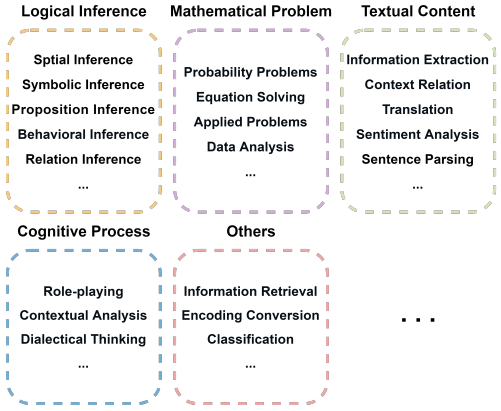

Figure 2: Subdomain of Prompt Types

In this project, we categorize prompt types into 5 main categories, shown in Figure 1, including logical inference, mathematical problems, etc,. In Figure 2, We can further subdivide these categories into more than 20 subfields. For example, in the textual content category, we explore multiple subfields related to context comprehension, sentiment analysis, and information extraction.

1.2 Prompts Content Analysis

In Figure 3, we provide example prompts from various domains and demonstrate the discrepancies between the BlueElephant model and the corrected answer (which is a combination of human and ChatGPT answers).

For instance, in the first example related to spatial reasoning, BlueElephant performed a logical reasoning process but arrived at an incorrect result. In the eighth example, BlueElephant made a judgment that went against common sense by incorrectly classifying a potato as a fruit, which led to its error. These examples highlight that BlueElephant is still a nascent llm with many shortcomings in numerous downstream NLP tasks.

Overall, We summarize here the typical reasons and prompts that lead to errors in BlueElephant across different domains:

- Logical Inference: The BlueElephant model lacks the ability to infer using logical information. When presented with a series of related conditions, it often cannot reason a correct result from them due to the lack of information utilization capability. A simplest example is as follows: "A is the father of B. A is the brother of C. B is the younger brother of D. What is the relationship between C and D?".
- Mathematical Problem: Similar to the issues encountered during logical inference, BlueElephant also struggles with understanding information when solving mathematical problems. Compared to ChatGPT, BlueElephant may provide incorrect analysis and calculations for simple math problems from the outset. This is particularly evident in many prompts, such as a problem requiring calculation of mode and median, where the model may incorrectly interpret it as a question about the mean. In simple math application problems, the model often generates an incorrect equation as the first step.
- Textual Content: In prompts related to contextual content, BlueElephant often makes typical errors in many downstream NLP tasks. For example, it lacks the ability to understand contextual information, cannot comprehend the inherent relationships between words (both semantic and syntactic), and cannot perform simple information extraction as required. As demonstrated in the examples, when asked to combine several discrete words into a sentence, the model not only fails to produce semantically fluent sentences but also adds other words that do not fit. Moreover, during the process of correcting misspelled words, BlueElephant may identify incorrect words as correct, which indicates its inability to understand the intrinsic relationships between words. In contrast, ChatGPT performs significantly better in these areas.
- Cognitive Process: In this category, we can see BlueElephant's general weaknesses. We attempted to have the model play different roles to introduce different contexts, but under these conditions, even very simple questions can lead to incorrect answers from BlueElephant. For example, when asked to play the role of a children's author, the model cannot generate vivid and lively texts. Similarly, when asked to play the role of a nutritionist guiding healthy eating habits, it cannot correctly classify suitable food ingredients.

2 Instruction Finetuning of ChatGLM-6B with LoRA

2.1 Dataset

The data utilized in this projecy was obtained from a shared document up to June 30th, 2023. Answers were mixed from both human and ChatGPT sources, and subsequently screened for quality control. Specifically, Answers that solely originated from ChatGPT were excluded, as their efficacy for the particular prompt, BlueElephant, remained unverified. Additionally, answers lacking sufficient complexity, including an absence of reasoning or proof, were also removed. Finally, A total of **2,287** valid data were retained after the screening process.

2.2 Inference Results Analysis

Table 1 displays a selection of fundamental questions across different fields and shows the varying answers provided by ChatGLM-6B before and after finetuning.

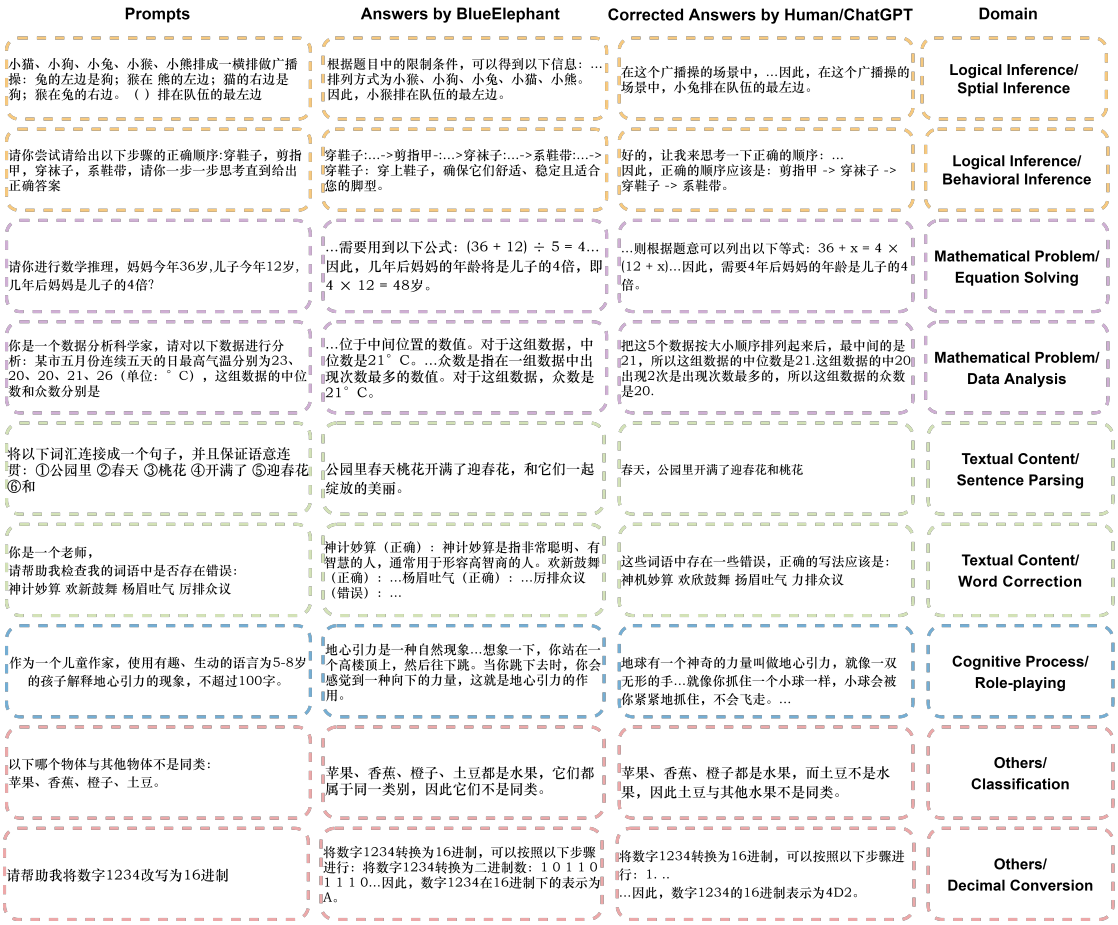

Figure 3: Examples of Prompts

By comparing the differences between the inference and fine-tuned inference, we can observe a significant improvement in the model's ability to complete certain tasks. For example, in the fourth instruction, ChatGLM-6B is already capable of performing some logical reasoning. However, it is worth noting that for Instruction 2-type mathematical problems, the model can provide deeper analysis but may still make basic arithmetic errors.

Furthermore, for some problems, ChatGLM-6B may merely "memorize" their answers and fail to extract the implied information, which is likely due to the quantity, quality, and diversity of the instructions.

2.3 Hyper-parameter Analysis

LoRA dropout is the dropout probability used in the LoRA layer. It is a regularization technique that can prevent overfitting of the model. The number of epochs is a hyper-parameter that determines the number of times the entire training dataset is used to update the model during training. Choosing an appropriate number of epochs and dropout rate are essential for achieving optimal model performance.

The search range of LoRA dropout rates is $\{0.1, 0.2, 0.4\}$. As shown in Figure 4, 0.2 yields

Instruction 1	公园在博物馆以北的三个街区。超市在公园以南的两个街区。请问博物馆是在什么的北边?
Inference 1	博物馆在公园以北的三个街区,因此它在北边。
Finetuned Inference 1	博物馆在超市的北边，因为公园在博物馆以北，超市在公园以南，所以博物馆在超市的北边。
Instruction 2	甲仓库存粮108吨,乙仓库存粮140吨...乙仓库的3倍,必须从乙仓库运出( )吨放入甲仓库。
Inference 2	甲仓库和乙仓库的比为3×140÷108=5:4,所以运出的为140-4×5=90吨。
Finetuned Inference 2	要使甲仓库存粮数是乙仓库的3倍,甲、乙仓库的存粮数必须成比例...因此,乙仓库的存粮数为70
Instruction 3	将以下词汇连接成一个句子,并且保证语意连贯:1公园里 2春天 3桃花 4开满了 5迎春花 6和
Inference 3	春天到了,公园里桃花和迎春花都开满了,美丽极了。
Finetuned Inference 3	春天，公园里开满了迎春花和桃花
Instruction 4	请您按照正确的顺序排列以下步骤:穿上袜子,穿上鞋子,剪指甲,系鞋带。
Inference 4	正确的顺序是:1.穿上袜子2.穿上鞋子3.剪指甲4.系鞋带
Finetuned Inference 4	剪指甲,系鞋带,穿上袜子,穿上鞋子。

Table 1: Inference Results

the best performance and minimum average loss among all epochs. When the number of epochs is sufficiently large, the model parameters converge and the loss function only slightly changes, indicating that the model has learned all the relevant features and patterns in the training data.

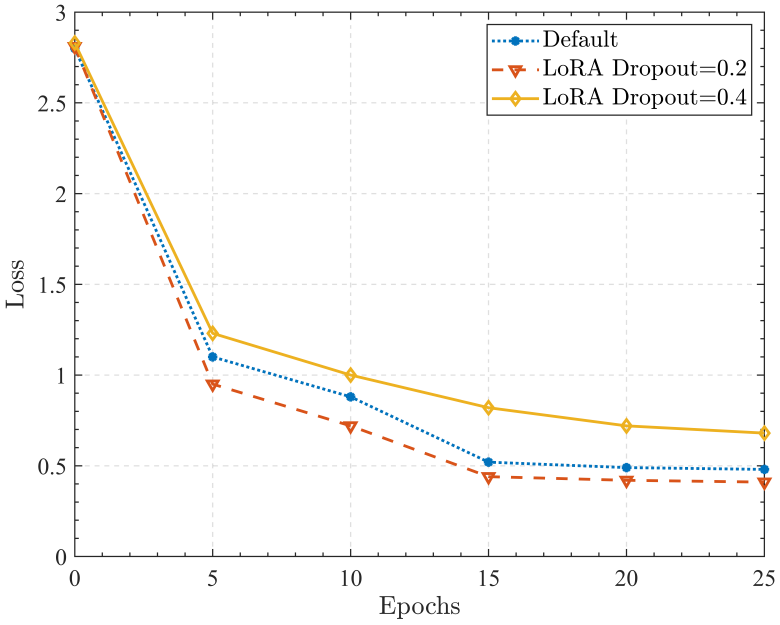

Figure 4: Loss under Different LoRA Dropouts