

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

XiaoYuxiang

July 2023

1 Prompt Engineering of BlueElephant

1.1 Prompts Diversity Analysis

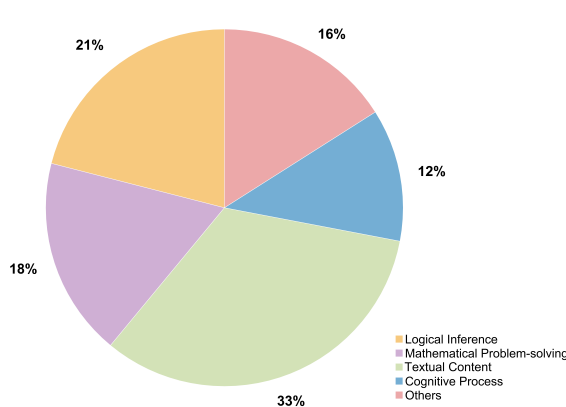


Figure 1: Distribution of Prompt Types

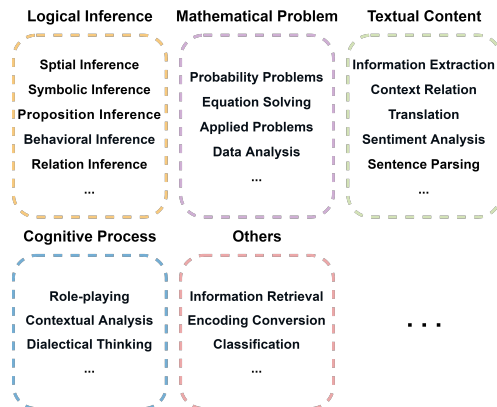


Figure 2: Subdomain of Prompt Types

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 project 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.

Prompts	Answers by BlueElephant	Corrected Answers by Human/ChatGPT	Domain
小猫、小狗、小兔、小熊排成一排做广播操：兔的左边是狗；熊在兔的右边；猫的右边是狗；熊在兔的右边。（）排在队伍的最左边	根据题目中的限制条件，可以得到以下信息：... 排列方式为小猫、小狗、小兔、小熊。 因此，小熊排在队伍的最左边。	在这个广播操的场景中，...因此，在这个广播操的场景中，小兔排在队伍的最左边。	Logical Inference/ Spatial Inference
请你尝试给出以下步骤的正确顺序：穿鞋子，剪指甲，穿袜子，系鞋带，请你一步一步思考直到给出正确答案	穿鞋子...->剪指甲...->穿袜子...->系鞋带...->穿鞋子：穿上鞋子，确保它们舒适、稳定且适合您的脚型。	好的，让我来思考一下正确的顺序：... 因此，正确的顺序应该是：剪指甲 -> 穿袜子 -> 穿鞋子 -> 系鞋带。	Logical Inference/ Behavioral Inference
请你进行数学推理，妈妈今年36岁，儿子今年12岁，几年后妈妈是儿子的4倍？	...需要用到以下公式： $(36 + 12) \div 5 = 4$... 因此，几年后妈妈的年龄将是儿子的4倍，即 $4 \times 12 = 48$ 岁。	...则根据题意可以列出以下等式： $36 + x = 4 \times (12 + x)$...因此，需要4年后妈妈的年龄是儿子的4倍。	Mathematical Problem/ Equation Solving
你是一个数据分析科学家，请对以下数据进行分析：某市五月份连续五天的日最高气温分别为23、20、20、21、26（单位： $^{\circ}$ C），这组数据的中位数和众数分别是	...位于中间位置的数值。对于这组数据，中位数是21 $^{\circ}$ C。...众数是指在一组数据中出现次数最多的数值。对于这组数据，众数是21 $^{\circ}$ C。	把这5个数据按大小顺序排列起来后，最中间的是21，所以这组数据的中位数是21。这组数据的中20出现2次是出现次数最多的，所以这组数据的众数是20。	Mathematical Problem/ Data Analysis
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	公园里春天桃花开满了迎春花，和它们一起绽放的美丽。	春天，公园里开满了迎春花和桃花	Textual Content/ Sentence Parsing
你是一个老师，请帮助我检查我的词语中是否存在错误：神机妙算 欢新鼓舞 扬眉吐气 厉排众议	神机妙算（正确）：神机妙算是指非常聪明、有智慧的人，通常用于形容高智商的人。欢新鼓舞（正确）：...扬眉吐气（正确）：...厉排众议（错误）：...	这些词语中存在一些错误，正确的写法应该是：神机妙算 欢欣鼓舞 扬眉吐气 力排众议	Textual Content/ Word Correction
作为一个儿童作家，使用有趣、生动的语言为5-8岁的孩子解释地心引力的现象，不超过100字。	地心引力是一种自然现象...想象一下，你站在一个高楼上，然后往下跳。当你跳下去时，你会感觉到一种向下的力量，这就是地心引力的作用。	地球有一个神奇的力量叫做地心引力，就像一双无形的手...就像你抓住一个小球一样，小球会被你紧紧地抓住，不会飞走...	Cognitive Process/ Role-playing
以下哪个物体与其他物体不是同类：苹果、香蕉、橙子、土豆。	苹果、香蕉、橙子、土豆都是水果，它们都属于同一类别，因此它们不是同类。	苹果、香蕉、橙子都是水果，而土豆不是水果，因此土豆与其他水果不是同类。	Others/ Classification
请帮助我将数字1234改写为16进制	将数字1234转换为16进制，可以按照以下步骤进行：将数字1234转换为二进制数：1 0 1 1 0 1 1 1 0...因此，数字1234在16进制下的表示为4D。	将数字1234转换为16进制，可以按照以下步骤进行：1...因此，数字1234的16进制表示为4D2。	Others/ Decimal Conversion

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 $3 \times 140 \div 108 = 5:4$,所以运出的为 $140 - 4 \times 5 = 90$ 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.

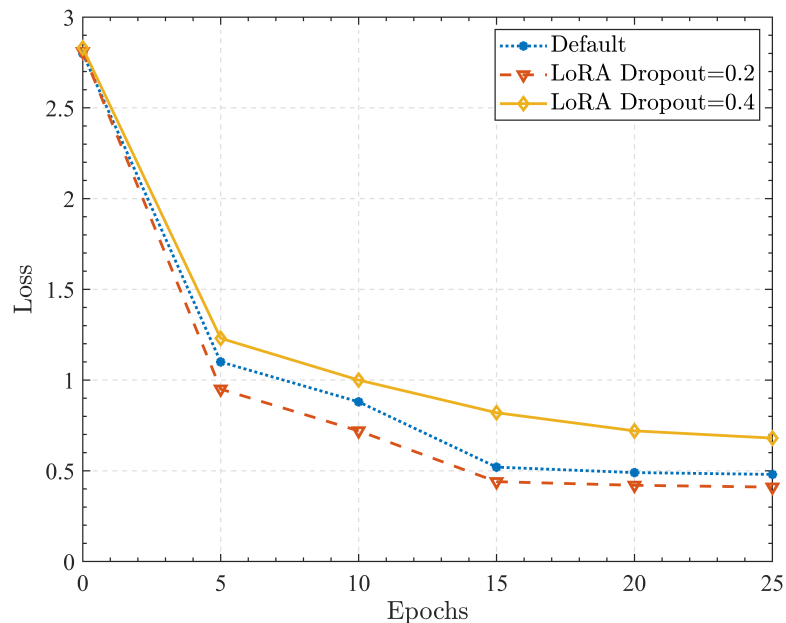


Figure 4: Loss under Different LoRA Dropouts