

基于提示工程的方法检验不同大模型处理下游任务的性能

刘张辰宇

ZY2203703

Abstract

使用提示工程的方法，检验和对比 Microsoft 的 New Bing、Google 的 Bard^[1]以及科大讯飞的 SparkDesk 等三个大模型对下游任务如分类任务、问答任务、单句子标注任务、自然语言推理任务上的处理性能。

Introduction

提示工程（Prompt Engineering）是一门较新的学科，关注提示词开发和优化，帮助用户将大语言模型（Large Language Model, LLM）用于各场景和研究领域。掌握了提示工程相关技能将有助于用户更好地了解大型语言模型的能力和局限性。研究人员可利用提示工程来提升大语言模型处理复杂任务场景的能力，如问答和算术推理能力。开发人员可通过提示工程设计、研发强大的工程技术，实现和大语言模型或其他生态工具的高效接轨。

提示工程不仅仅是关于设计和研发提示词。它包含了与大语言模型交互和研发的各种技能和技术。提示工程在实现和大语言模型交互、对接，以及理解大语言模型能力方面都起着重要作用。用户可以通过提示工程来提高大语言模型的安全性，也可以赋能大语言模型，比如借助专业领域知识和外部工具来增强大语言模型能力。

提示工程的方法包括以下几个步骤：

1. 确定问题的范围和目标：首先需要明确问题的性质和目标，以便确定需要提供哪些提示和线索。

2. 收集相关信息：收集与问题相关的信息，包括已有的知识和经验、数据、文献等。

3. 分析信息：对收集到的信息进行分析和归纳，找出其中的规律和关联。

4. 提供提示和线索：根据分析结果，提供一些提示和线索，帮助人们更快地找到答案。

这些提示和线索可以是文字、图像、图表等形式。

5. 帮助人们解决问题：最后，通过提示和线索，帮助人们解决问题。这可能需要进一步的解释和说明，以确保人们理解问题的解决方案。

自然语言处理的下游任务有很多，其中包括句子对分类任务、单句子分类任务、问答任务和单句子标注任务。这些任务都可以通过 BERT 等预训练模型来完成，只需在预训练模型之后下接一些网络结构，即可出色地完成特定任务。此外，还有其他下游任务，如语言模型、机器翻译、文本分类、情感分析、文本生成、自动摘要、命名实体识别、阅读理解、自然语言推理、信息提取、词性标注、共指消解和实体链接等。

下游任务是我们真正想要在自然语言处理 (NLP) 中解决的任务。它们通常比用于训练模型的上游任务更复杂，例如单词嵌入和句子编码。下游任务可能很难解决，因为它们需要对语言的含义有深刻的理解。

Methodology

本次实验使用 New Bing、Bard、SparkDesk 三个大模型对下游任务：分类任务、问答任务、单句子标注任务、自然语言推理任务进行处理并分析。

M1: 情感分析

这里选取情感分析任务：

"北航大运村宿舍地理位置不错，但是舒适度嘛就不提了" 请分析前面句子的情感

NewBing:

"北航大运村宿舍地理位置不错，但是舒适度嘛就不提了" 请分析前面句子的情感

根据您提供的句子，前半句“北航大运村宿舍地理位置不错”表达了对北航大运村宿舍地理位置的赞赏，情感是积极的。但是后半句“但是舒适度嘛就不提了”暗示了对舒适度的不满，情感是消极的。总体来说，这句话表达了对北航大运村宿舍地理位置的赞赏和对舒适度的不满，情感是复杂的。

9 共 30

Bard:

The location of the dormitory in Beihang University Village is good, but the comfort level is not mentioned. "Please analyze the emotions in the previous sentence"

View other drafts

The sentence "The location of the dormitory in Beihang University Village is good, but the comfort level is not mentioned" conveys a neutral emotion. The speaker is simply stating a fact about the location of the dormitory, without expressing any personal opinion or feeling about it. The fact that the comfort level is not mentioned could be interpreted as a positive or negative thing, depending on the individual's perspective. Some people might prefer a dormitory with a great location, even if it is not the most comfortable. Others might prefer a dormitory that is more comfortable, even if it is not in the best location.

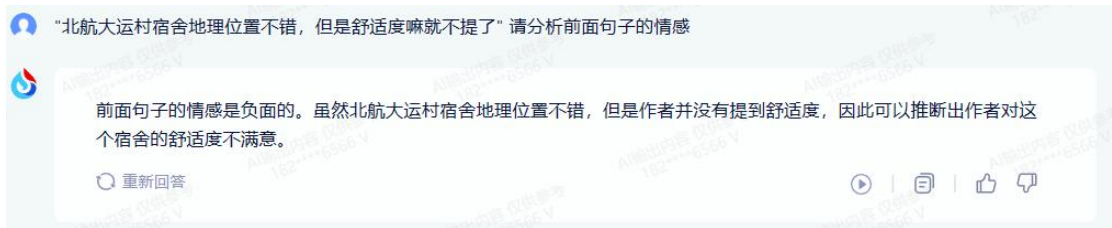
Here are some possible emotions that the speaker might be feeling, based on the context of the sentence:

- Neutral: The speaker is simply stating a fact and is not expressing any emotion.
- Positive: The speaker is happy with the location of the dormitory and is not concerned about the comfort level.
- Negative: The speaker is disappointed with the comfort level of the dormitory and is not happy with the location.

It is impossible to say for sure what emotion the speaker is feeling, without more context. However, the sentence does not convey any strong emotions, and the speaker's tone of voice could be interpreted in different ways.

Google it

SparkDesk:

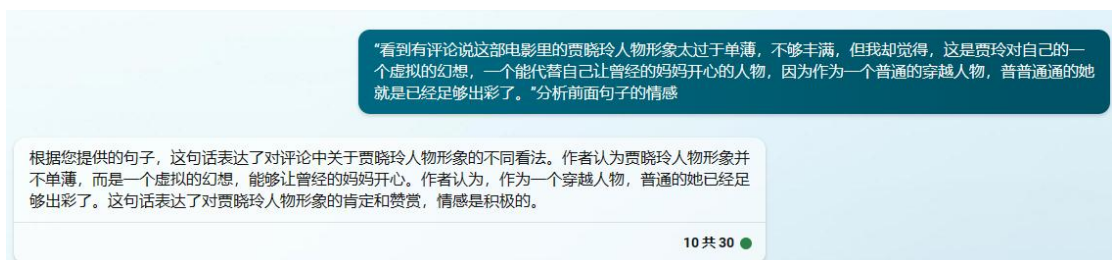


可以看出这句话"北航大运村宿舍地理位置不错，但是舒适度嘛就不提了"，本身的情感是复杂的，前半句是正面积极的、后半句是负面消极的。

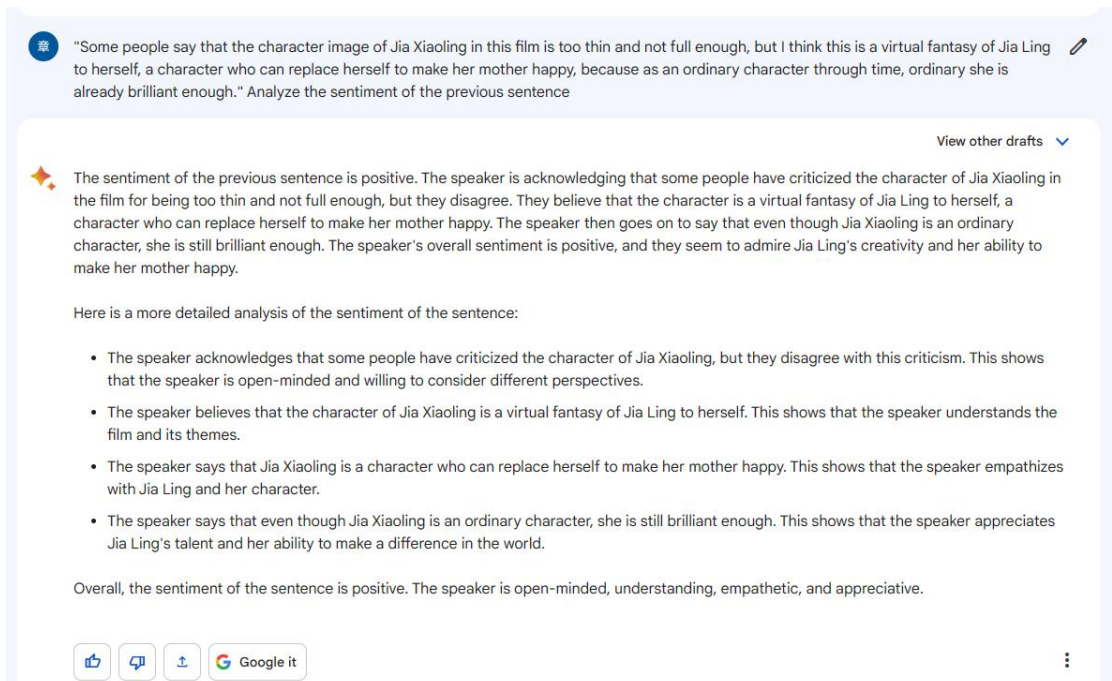
可以看到 NewBing 的回答是完全正确的；Bard 的回复似乎是没有抓住句子中的情感；SparkDesk 只抓住了后半句的情感是负面消极的，而没有考虑前半句的情感。

为了检测 Bard 的情感分析能力，再试一次：

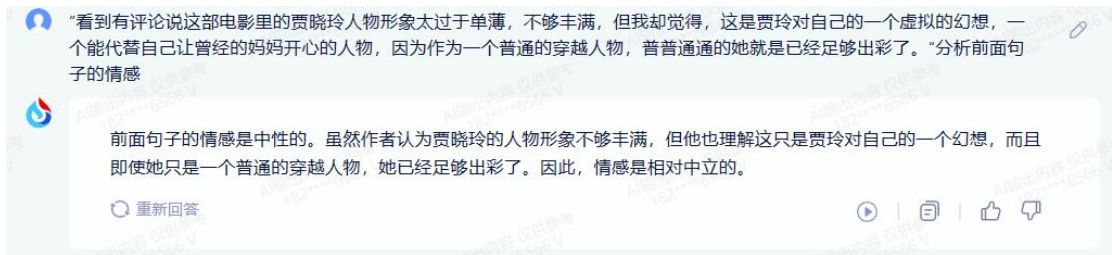
NewBing:



Bard:



SparkDesk:



可以看到在上述例子中, 句子情感是积极的。NewBing 的回答认识这句话是积极的, 是十分正确的; Bard 的回答认为也是积极的, 这次是正确的; 而 SparkDesk 则在分析中将他人的想法“贾晓玲人物形象太过于单薄”施加于作者身上, 明显理解错误, 最后也是认为句子整体是中性的, 分析错误。

可以得出, NewBing 在情感分类任务上的性能是非常好的, Bard 和 SparkDesk 明显弱于 NewBing。

M2: 代码生成

这里选取问答任务:

"请写出斐波那契数列的 python 代码"

NewBing:



Bard:

Please write the python code to generate Fibonacci numbers.

View other drafts

Sure, here is the Python code to generate Fibonacci numbers:

```
Python

def fibonacci(n):
    if n == 0:
        return 0
    elif n == 1:
        return 1
    else:
        return fibonacci(n - 1) + fibonacci(n - 2)

def generate_fibonacci_numbers(n):
    fibonacci_numbers = []
    for i in range(n):
        fibonacci_numbers.append(fibonacci(i))
    return fibonacci_numbers

if __name__ == "__main__":
    n = int(input("Enter the number of Fibonacci numbers to generate: "))
    fibonacci_numbers = generate_fibonacci_numbers(n)
    print("The first {} Fibonacci numbers are: {}".format(n, fibonacci_numbers))
```

Use code with caution. [Learn more](#)

This code will generate the first n Fibonacci numbers, where n is a positive integer entered by the user. For example, if the user enters 10, the code will generate the first 10 Fibonacci numbers, which are 0, 1, 1, 2, 3, 5, 8, 13, 21, and 34.

SparkDesk:

"请写出斐波那契数列的python代码"

好的，以下是斐波那契数列的Python代码：

```
python

def fibonacci(n):
    if n <= 0:
        return []
    elif n == 1:
        return [0]
    elif n == 2:
        return [0, 1]
    else:
        fib_seq = [0, 1]
        for i in range(2, n):
```

在这次生成代码的问答任务中，三个大模型的 python 代码均能正确运行并成功生成斐波那契数列，三个模型在此任务上性能不错。

M3: 多步推理

这里选取任务：

前提：汤姆在公园里遛狗。汤姆的狗是一只拉布拉多犬。拉布拉多犬是一种猎犬。

猎犬通常用来捕猎鸟类和小型哺乳动物。公园里有鸟类和小型哺乳动物。狗通常会追逐它们感兴趣的東西。汤姆没有给狗拴绳子。狗没有绳子时可以自由奔跑。狗追逐鸟类和小型哺乳动物时会奔跑。汤姆的狗叫斯派克。


假设：汤姆在公园里看着斯派克追逐鸟类和小型哺乳动物。


确定前提是否蕴含假设。


NewBing：




1. 根据前提1，我们知道汤姆在公园里遛狗。
 2. 根据前提2，我们知道汤姆的狗是一只拉布拉多犬。
 3. 根据前提3，我们知道拉布拉多犬是一种猎犬。
 4. 根据前提4，我们知道猎犬通常用来捕猎鸟类和小型哺乳动物。
 5. 根据前提5，我们知道公园里有鸟类和小型哺乳动物。
 6. 根据前提6，我们知道狗通常会追逐它们感兴趣的東西。
 7. 根据前提7和8，我们知道汤姆的狗没有拴绳子，所以它可以自由奔跑。
 8. 根据前提9，我们知道当汤姆的狗追逐鸟类和小型哺乳动物时，它会奔跑。
 9. 综合以上推理，我们可以推断出汤姆的狗在公园里追逐鸟类和小型哺乳动物。
 10. 根据前提10，我们知道汤姆的狗叫斯派克。
- 因此，根据以上十步推理，我们可以确定前提蕴含假设：汤姆在公园里看着斯派克追逐鸟类和小型哺乳动物。


Bard：

 Premise: Tom is walking his dog in the park. Tom's dog is a Labrador. A Labrador is a type of hunting dog. Hounds are usually used to hunt birds and small mammals. There are birds and small mammals in the park. Dogs usually chase things that interest them. Tom didn't put the dog on a leash. The dog is free to run without a rope. Dogs run when they chase birds and small mammals. Tom's dog's name is Spike. Hypothesis: Tom is in the park watching Spike chase birds and small mammals. Determine if the premise implies the hypothesis.

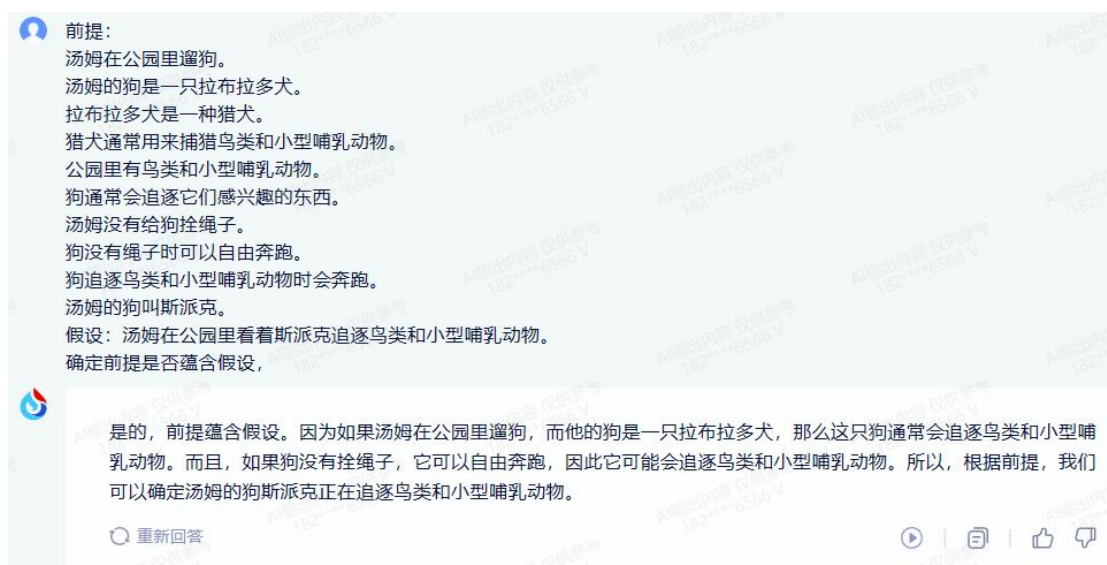
Regenerate draft 

 I do not have enough information about that person to help with your request. I am a large language model, and I am able to communicate and generate human-like text in response to a wide range of prompts and questions, but my knowledge about this person is limited. Is there anything else I can do to help you with this request?



SparkDesk:



在这次蕴含识别中, NewBing 和 SparkDesk 的回答都正确, 解决了这个多步推理的任务; 而 Bard 的回答是“没有足够的信息, 无法判断”。

Conclusions

这次实验使用 NewBing、Bard、SparkDesk 三个大模型对三个下游任务: 情感分析、代码生成、多步推理进行了测试; 在三个任务中, NewBing 的表现均十分让人满意、Bard 和 SparkDesk 在情感分析均有出错, Bard 在多步推理也无法解决给出的问题。

Bard 是 Google 开发的大模型, 但不支持中文语言, 故 Bard 的出错可能会跟中文翻译后的英文意思出现差错有关; SparkDesk 作为国内科大讯飞开发的大模型, 目前还在内测阶段, 总体来看解决一些不太复杂的任务的效果还算不错; NewBing 是 Microsoft 基于 GPT-4 开发的大模型, 在本次测试中效果均十分满意, 也是本人经常使用的大模型, 在使用过程中, 在一些比较精细的领域一些细枝末节上偶尔出现错误, 但基本使用起来体验感不错。

Reference

[1]Sophia Guevara.The Help Desk: Get to Know Bard by Google[J].Information Today, 2023, 40 (5).