语言大模型测试

潘熠暄
ZY2203112

一．摘要

本实验目的在于通过提示工程比较目前若干前沿语言大模型执行若干不同下游任务上的性能。语言大模型选用 OpenAI 开发的 GPT3.5、微软的 new Bing（GPT4）以及谷歌的 Bard。

二．介绍

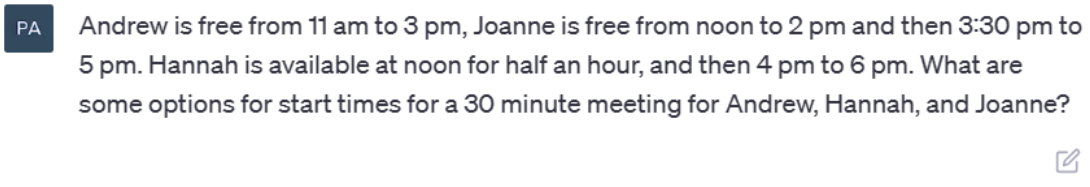
语言大模型是指基于深度学习的神经网络模型，通过大规模的文本数据进行训练，可以对自然语言进行理解、生成和处理的模型，这些模型利用了深度学习的技术，特别是 Transformer 构架，以提供对自然语言的高级语义理解和语言生成能力。

下游任务是指在自然语言处理中，基于语言大模型的预训练模型，将其应用于具体的实际任务或应用中。下游任务利用预训练模型的语言理解和生成能力，通过微调或进行特定任务的训练来解决各种自然语言处理问题。常见的下游任务包括文本分类、情感分析、命名实体识别、问答系统和文本生成等。

在使用语言模型进行特定任务时，为了引导模型产生所需的输出，我们需要进行输入提示，而提示工程就是设计和调整输入提示的过程，其目标在于帮助模型产生更准确、合理和符合预期的输出结果，合理的输入提示可以引导模型理解任务的上下文、约束输出的风格和内容，，提高模型在特定任务上的性能。

三．实验过程及结果

1．问答系统：使用 OpenAI 官方在介绍页中使用的问题：



1）GPT3.5

To find suitable options for a 30-minute meeting that accommodates Andrew, Joanne, and Hannah, we need to identify the overlapping time slots when all three individuals are available. Based on the given schedules, the following options are available:

Option 1:

- Andrew: Free from 11 am to 3 pm
- Joanne: Free from 12 pm to 2 pm and then 3:30 pm to 5 pm
- Hannah: Available at 12 pm for half an hour and then 4 pm to 6 pm

In this case, the overlapping time slots for all three individuals are from 12 pm to 12:30 pm and from 4 pm to 4:30 pm. You can schedule the 30-minute meeting during either of these time slots.

Option 2:

- Andrew: Free from 11 am to 3 pm
- Joanne: Free from 12 pm to 2 pm and then 3:30 pm to 5 pm
- Hannah: Available at 12 pm for half an hour and then 4 pm to 6 pm

In this case, the overlapping time slot for all three individuals is from 12 pm to 12:30 pm. You can schedule the 30-minute meeting during this time slot.

Please note that these are just a few options based on the given schedules. You may explore other possibilities depending on the specific requirements and preferences of the individuals involved.

Regenerate response

2）new Bing

使用 new Bing 的均衡模式生成结果如下：

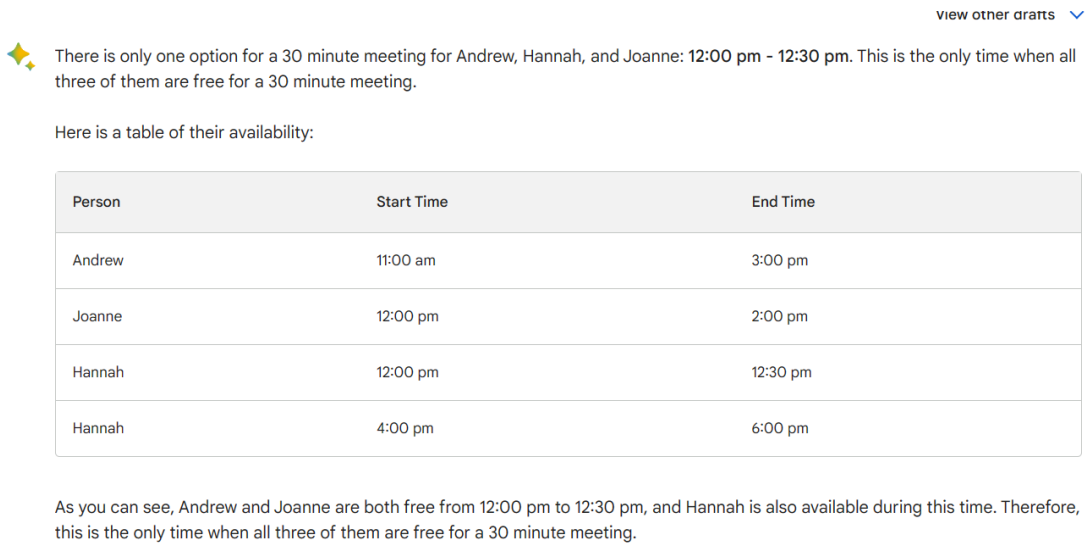
One way to find common free time slots for multiple people is to use a meeting scheduler app. These apps make it easy to find a common time slot or share your free slots for others to agree. Some of the best free meeting scheduling apps include Doodle, Calendly, Google Calendar, and Microsoft Teams [1].

However, if you want to do it manually, you can use a table like this:

Time	Andrew	Joanne	Hannah
11am	Free		
12pm		Free	Free
1pm	Free		
2pm			
3pm	Free	Free	
4pm			Free
5pm		Free	
6pm			Free

From the table above, we can see that the available timeslots for all three people are **12 pm** and **4 pm**. You can choose either one of these timeslots for your meeting [2].

3）Bard



从生成结果上看，GPT 和 new Bing 两个模型均错误的将 16 点列入可行时间，GPT 在 Option 2 中生成了正确答案。

2.情感分析

文本内容来自 Reddit 社区一篇关于暗黑破坏神 4 剧情战役的讨论帖子。选用若干楼层进行分析。素材：

Informal_Exercise_88 · 1 day ago

Not gonna lie, the campaign is an absolute slog to get through.. go here, do this or that, kill this or that.. backtrack, talk, backtrack, talk, go here... I'll be glad when I'm done with it. Even the rewards are quite pitiful as well.

The voice acting is excellent though as are the cut scenes but the music is amazing.. loved the OG Diablo music.

nanosam · 18 hr. ago · *edited 18 hr. ago*

The actual writing of the story is very poor. The plotholes are real, the ending is really bad.

I honestly hated it, the cinematics were great but whoever wrote the campaign story really dropped the ball

Best thing about the campaign is that you can skip it and forget about how bad the writing was.

Lorath was the high point - Neyrelle and the ending was poop

Why did they even have Innarius in the story? What a complete waste.

The dumb ways characters get hurt or die is just inexcusable.

Imagine having a decade between D3 and D4 and this is the best story you can come up with for D4 campaign???

Like holy crap that is not good

evillurkz · 20 hr. ago

For the first time in YEARS since diablo 3 I am taking my time reading through the whole story, going through all the side quests, and literally enjoying the content. You only play the story once, so might as well enjoy what you paid for.

t0lkien1 · 22 hr. ago · *edited 17 hr. ago*

This is not "brilliant" writing. It's not bad and has some nice moments. And let's face it, it doesn't have much competition in previous Diablo games. But there's a lot of cringe as well.

The boring, unimaginative "annoying little girl looking for her mother" trope was bad enough the first time in D3, but the moment the game flat out copied the faux death scene from *Path of Exile* I was done with trying to take the narrative seriously.

1) GPT3.5

Sentiment: Mixed

Explanation: The sentiment expressed in the text is mixed, combining both negative and positive aspects. The negative sentiments include the feeling of the campaign being a slog to get through, the repetitive nature of the tasks, and the pitiful rewards. However, there are positive sentiments expressed about the voice acting, cut scenes, and the amazing music, particularly the love for the original Diablo music. Overall, the sentiment is a mix of dissatisfaction with certain aspects and appreciation for others.

Sentiment: Negative

Explanation: The sentiment expressed in the text is predominantly negative. The text contains strong criticism of the writing of the story, highlighting poor quality, plotholes, and a bad ending. The writer expresses hatred towards the campaign and disappointment in the overall storytelling. The mention of certain characters being a high point suggests some positive aspects, but they are overshadowed by negative sentiments. The text conveys a sense of frustration and dissatisfaction with the campaign's writing and the missed opportunities for improvement.

Sentiment: Positive

Explanation: The sentiment expressed in the text is positive. The writer mentions enjoying the content and taking their time to read through the whole story and complete side quests. They emphasize the importance of enjoying what they paid for, implying a positive experience and satisfaction with the gameplay and storytelling.

Sentiment: Mixed

Explanation: The sentiment expressed in the text is mixed, combining both positive and negative aspects. The writer acknowledges that the writing is not "brilliant" but also acknowledges that it has some nice moments and doesn't have much competition in previous Diablo games. However, they express disappointment in certain elements, such as finding them cringe-worthy and mentioning the trope of an annoying little girl looking for her mother as unimaginative. The mention of copying a scene from Path of Exile indicates a negative sentiment towards the narrative. Overall, the sentiment leans towards a mix of appreciation for certain aspects and disappointment in others.

2) new Bing

Based on your message, it seems like you're expressing a negative sentiment towards the Diablo campaign.

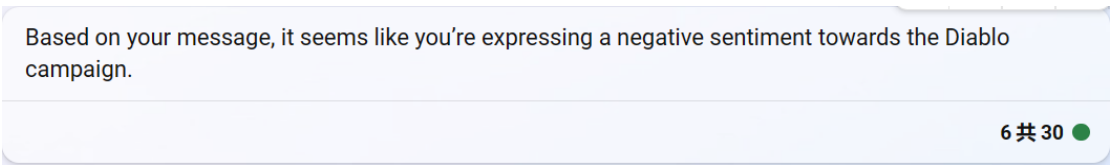
3 共 30

Based on your message, it seems like you're expressing a negative sentiment towards the Diablo campaign.

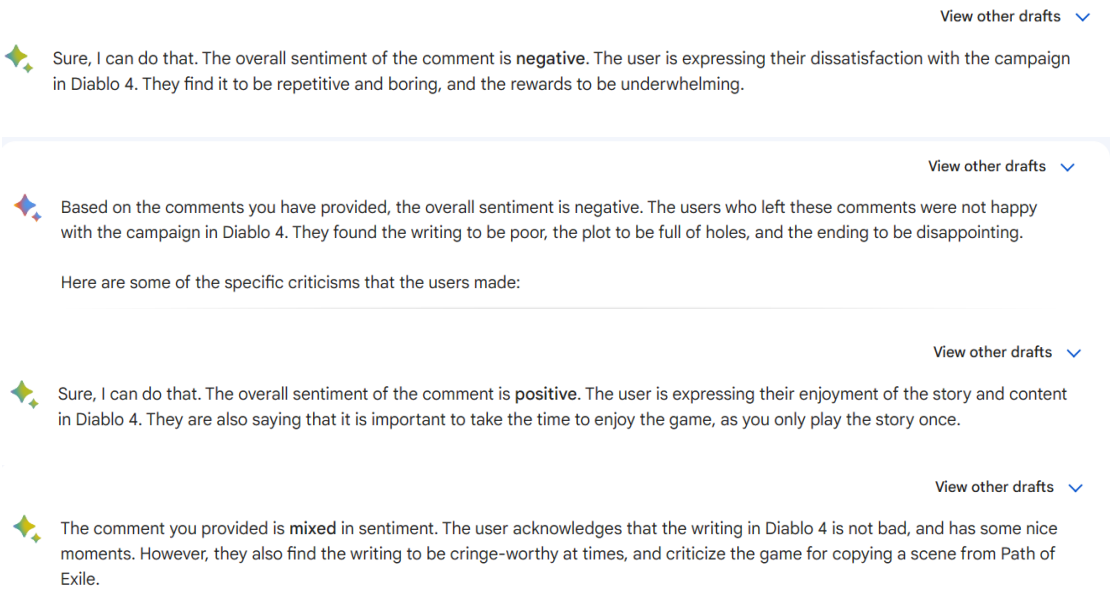
4 共 30

It's great to hear that you're enjoying the Diablo campaign this time around. It's always good to take your time and enjoy the content you paid for.

5 共 30



3)Bard



在只要求语言模型"do sentiment analysis"的情况下，GPT3.5 在分类的同时也生成了原因，而 new Bing 的分类更为简单，只有 Positive 和 Negative 两类。在准确度上，new Bing 在对第一和第四个评论的分类上与另外两个模型的结果出现了分歧，只有第一个可以称得上是 GPT 给出的 mixed 归类，第四个则完全可以说是负面评价。当要求 GPT 在只有 Positive 和 Negative 情形下重新分析时：

Positive Sentiments:

1. "For the first time in YEARS since Diablo 3, I am taking my time reading through the whole story, going through all the side quests, and literally enjoying the content. You only play the story once, so might as well enjoy what you paid for."

Mixed Sentiments:

1. "Not gonna lie, the campaign is an absolute slog to get through.. go here, do this or that, kill this or that.. backtrack, talk, backtrack, talk, go here... I'll be glad when I'm done with it. Even the rewards are quite pitiful as well. The voice acting is excellent though as are the cut scenes but the music is amazing.. loved the OG Diablo music."

Negative Sentiments:

1. "The actual writing of the story is very poor. The plotholes are real, the ending is really bad. I honestly hated it, the cinematics were great but whoever wrote the campaign story really dropped the ball. Best thing about the campaign is that you can skip it and forget about how bad the writing was. Lorath was the high point - Neyrelle and the ending was poop. Why did they even have Innarius in the story? What a complete waste. The dumb ways characters get hurt or die is just inexcusable. Imagine having a decade between D3 and D4 and this is the best story you can come up with for D4 campaign??? Like holy crap that is not good."
2. "This is not 'brilliant' writing. It's not bad and has some nice moments. And let's face it, it doesn't have much competition in previous Diablo games. But there's a lot of cringe as well. The boring, unimaginative 'annoying little girl looking for her mother' trope was bad enough the first time in D3, but the moment the game flat out copied the faux death scene from Path of Exile, I was d[illegible] narrative seriously."

GPT 依旧没有放弃 mixed 这一类别，但正确地将第四个评论归入 Negative 类别。
而对 Bard，当提问第四个评论是“more positive or more negative”时，它给出了正确的答案：

I would say that the comment is **more negative than positive**. The user's criticisms of the writing are more detailed and specific than their praise. They also seem to be more bothered by the negative aspects of the writing than they are impressed by the positive aspects.

同时，在情感分析任务中，Bard 模型似乎不擅长将对话作为一个整体进行处理，即便明确要求它对接下来“分开”发送的四个评论进行情感分析，它依旧需要我在在后发送的三个评论前加上进行情感分析请求，才能正确执行任务。

3.代码编写
选择使用 leetcode 的每日一题（2023.06.05），题干如下：
You are given a 0-indexed array nums of size n consisting of non-negative integers.
You need to apply n - 1 operations to this array where, in the ith operation (0-indexed), you will apply the following on the ith element of nums:
If nums[i] == nums[i + 1], then multiply nums[i] by 2 and set nums[i + 1] to 0. Otherwise, you skip this operation.
After performing all the operations, shift all the 0's to the end of the array.
For example, the array [1,0,2,0,0,1] after shifting all its 0's to the end, is [1,2,1,0,0,0].
Return the resulting array.
Note that the operations are applied sequentially, not all at once.

1）GPT3.5

```cpp
std::vector<int> shiftZerosToEnd(std::vector<int>& nums) {
    int n = nums.size();
    int countZeros = 0;

    // Perform the operations
    for (int i = 0; i < n - 1; i++) {
        if (nums[i] == nums[i + 1]) {
            nums[i] *= 2;
            nums[i + 1] = 0;
            i++;  // Skip the next element since it has been set to 0
        }
    }

    // Shift the zeros to the end of the array
    for (int i = 0; i < n; i++) {
        if (nums[i] != 0) {
            nums[countZeros++] = nums[i];
        }
    }

    // Fill the remaining elements with zeros
    while (countZeros < n) {
        nums[countZeros++] = 0;
    }

    return nums;
}
```

运行结果：

2）new Bing

```cpp
class Solution {
public:
    vector<int> applyOperations(vector<int>& nums) {
        int n = nums.size();
        while (true) {
```

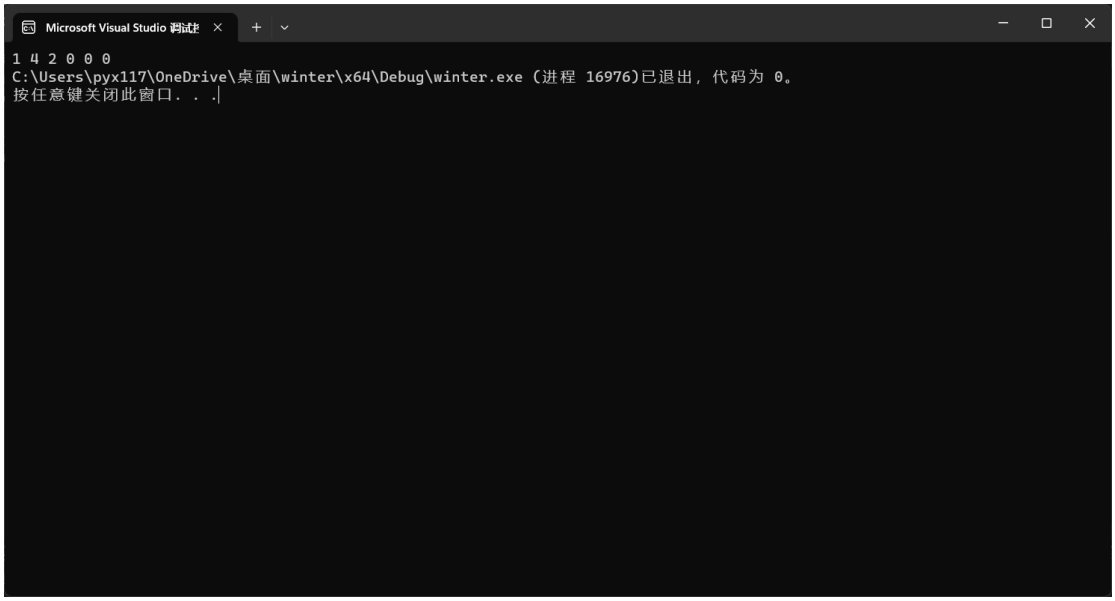
```
            bool flag = false;
            vector<int> temp(n);
            temp[0] = nums[0];
            temp[n - 1] = nums[n - 1];
            for (int i = 1; i < n - 1; ++i) {
                if (nums[i] == nums[i - 1] && nums[i] != 0) {
                    temp[i] = nums[i] * 2;
                    nums[i] = 0;
                    flag = true;
                } else if (nums[i] == nums[i + 1] && nums[i] != 0) {
                    temp[i] = nums[i];
                    nums[i] = 0;
                    flag = true;
                } else {
                    temp[i] = nums[i];
                }
            }
            if (!flag) {
                return temp;
            }
            for (int i = 0; i < n; ++i) {
                if (temp[i] == 0) {
                    for (int j = i + 1; j < n; ++j) {
                        if (temp[j] != 0) {
                            swap(temp[i], temp[j]);
                            break;
                        }
                    }
                }
            }
            nums = temp;
        }
    }
};
```

运行结果：超出时间限制，无法通过 leetcode 的测试。

若是在 visual studio 中调用测试，结果是正确的。

```
int main() {
    vector<int> vs = { 1,2,2,1,1,0 };
    vector<int> v = applyOperations(vs);
    for (int i : v)
        cout << i << " ";

}
```



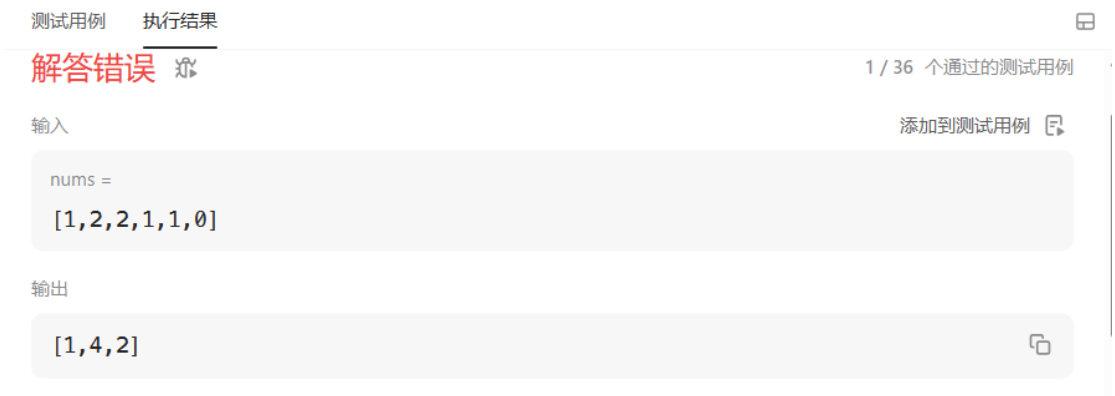
3）Bard

```
vector<int> solve(vector<int> nums) {
  for (int i = 0; i < nums.size() - 1; i++) {
    if (nums[i] == nums[i + 1]) {
      nums[i] *= 2;
      nums[i + 1] = 0;
    }
  }

  vector<int> result;
  for (int i = 0; i < nums.size(); i++) {
    if (nums[i] != 0) {
      result.push_back(nums[i]);
    }
  }

  return result;
}
```

运行结果：



当使用如下提示：

actually your code will print the following output: 1 2 1. it seems that you forget to push back all the 0 at the end of the result

后，Berd 可以改进错误：

```
vector<int> solve(vector<int> nums) {
  for (int i = 0; i < nums.size() - 1; i++) {
    if (nums[i] == nums[i + 1]) {
      nums[i] *= 2;
      nums[i + 1] = 0;
    }
  }

  vector<int> result;
  for (int i = 0; i < nums.size(); i++) {
    if (nums[i] != 0) {
      result.push_back(nums[i]);
    }
  }

  for (int i = 0; i < nums.size() - result.size(); i++) {
    result.push_back(0);
  }

  return result;
}
```

然而，Bard 忽略了 for 循环语句的 cond-expression 每次都执行一次，nums.size() – result.size()的值并非固定，依旧错误。

给出提示如下：

Y the value of "nums.size() - result.size()" will change after every loop, the number of 0 in the answer is still not right

Bard 将第三个 for 循环修正为：

```
int num_zeros = nums.size() - result.size();
for (int i = 0; i < num_zeros; i++) {
  result.push_back(0);
}
```

运行结果：

综合来看，三种模型都可以完成较简单的编程问题，GPT 的结果最佳，同时还自带了注释，new Bing 的结果可读性极差，而且时间占用大，Bard 则需要一点额外的帮助。