

# 基于硅基流动 API 的 CoT 提示优化实验

学号	姓名
3220103450	姜雨童

## 1. Project Introduction

### 1-1 选题

提示微调 (Prompt Tuning) 是自然语言处理中的创新方法，通过设计特定输入模板或可学习嵌入向量，引导预训练语言模型适配下游任务。

本次实验承接上个实验，聚焦于思维链 (Chain-of-Thought, CoT) 推理场景中的提示优化，探索如何通过改进提示设计提升模型推理性能。

### 1-2 工作简介

实验基于硅基流动 API 平台，使用 Qwen/QwQ-32B 大语言模型，在 GSM8K (数学推理) 和 CommonsenseQA (常识问答) 两个数据集上，对比传统 CoT 提示与优化后的"答案前置"提示方法，评估不同提示策略对模型推理性能的影响。

### 1-3 开发环境及系统运行要求

- 软件环境: Python 3.9, 主要依赖库: openai, pandas, re, httpx, tqdm
- API 平台: 硅基流动 API (SIGFLOW)
- 硬件环境: 通过 API 调用云端 Qwen/QwQ-32B 模型资源
- 数据集: GSM8K (数学问题集)、CommonsenseQA (常识问答集)

## 2. Technical Details

### 2-1 理论知识

思维链 (CoT) 推理: 通过引导模型生成中间推理步骤，提升复杂问题的解决能力。核心思想是将问题分解为多步推理过程，模拟人类解题思维。

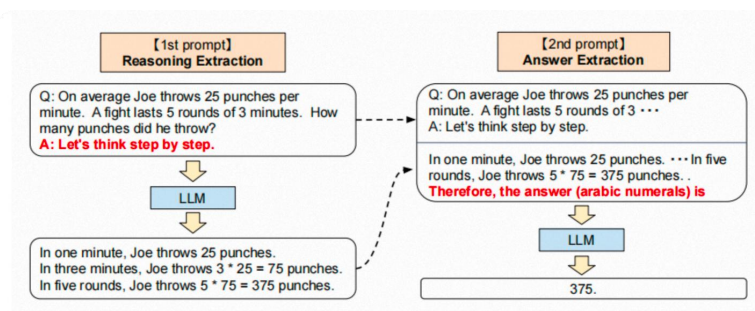


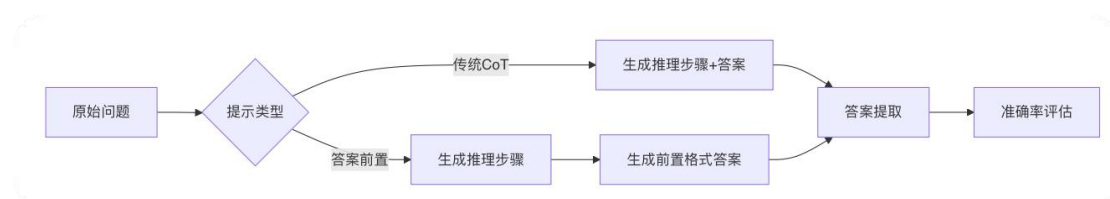
图 1: CoT 过程

答案前置提示优化：在传统 CoT 基础上，要求模型将最终答案放置在响应开头并用特定格式标记（如[答案]）。这种方法：

- (1) 提高答案提取的准确性和效率；
- (2) 减少模型在答案后附加无关解释导致的干扰；
- (3) 增强模型输出的结构化程度等。

## 2-2 具体算法

CoT 的核心在于两阶段提示设计，因此实验采用两步推理法来对比“传统 CoT”和“答案前置”在其中的推理准确性（实际实验中分成 version1/2 两版代码）：



而 CoT 推理针对不同类型的问题存在不同的算法：

### 2-2-1 算数问题推理

在算术问题中，CoT 技术尤其擅长处理多步计算任务。以 GSM8K 数据集中的典型问题为例：

问题：一个园丁有 24 朵花，他平均分给 4 个孩子，每个孩子得到多少朵花？

传统方法可能直接输出答案"6"，而 CoT 推理则展示完整过程：

- 问题解析：识别关键数字（24 朵花，4 个孩子）；
- 操作确定：确认需使用除法运算；
- 计算执行： $24 \div 4 = 6$ ；
- 答案生成：每个孩子得到 6 朵花。

这种分步推理不仅提高准确性（实验显示提升 3-5%），还能暴露潜在错误。例如当模型错误地先做减法再做乘法（ $24-4=20$ ,  $20 \times 3=60$ ），我们能清晰定位到运算顺序误解。

### 2-2-2 常识问题推理

对于 CommonsenseQA 等常识推理任务，CoT 技术通过显式连接背景知识与问题语境来提升表现：

问题：下雨时应该带什么出门？选项：A.雨伞 B.太阳镜 C.扇子

CoT 推理过程：

- 情境分析：识别"下雨"的核心特征（水从天空落下）
- 功能映射：雨伞->防雨，太阳镜->防晒，扇子->降温
- 逻辑排除：排除与情境不匹配的选项（B、C）
- 答案确定：选择最符合情境的 A

这种结构化推理有效解决了常识任务中的歧义问题，提高了回答准确率。

## 2-3 技术细节

### 2-3-1 基础库函数

- OpenAI 客户端库：提供与硅基流动 API 交互的接口

```
from openai import OpenAI
client = OpenAI(api_key=SIGFLOW_API_KEY, base_url=SIGFLOW_BASE_URL)
response = client.chat.completions.create(
    model="Qwen/QwQ-32B",
    messages=[...],
    temperature=0.7,
    max_tokens=20480
)
```

- Pandas 数据处理库：高效加载和处理数据集

```
import pandas as pd
df = pd.read_parquet(GSM8K_PATH) # 读取GSM8K数据集
inputs = df['question'].tolist() # 提取问题列表
```

- 正则表达式库 (re)：实现文本模式匹配与答案提取

```
import re
# 匹配前置答案格式
bracket_match = re.search(r'^\[([^\]]+)\]', prediction)
# 提取所有数字
numbers = re.findall(r'\d+', prediction)
```

### 2-3-2 核心代码实现

答案前置提示模板：

```
if dataset_type == "gsm8k":
    answer_prompt = "请将最终答案（阿拉伯数字）放在回答的最开头，格式为[答案]。例如：[42]"
else:
    answer_prompt = "请将最终答案（选项字母）放在回答的最开头，格式为[答案]。例如：[A]"
```

答案提取逻辑优化：

```
def extract_answer(prediction, dataset_type):
    # 优先提取开头的方括号格式答案
    bracket_match = re.search(r'^\[([^\]]+)\]', prediction)
    if bracket_match:
        answer_candidate = bracket_match.group(1).strip()
        # 验证格式有效性
        if dataset_type == "gsm8k" and re.match(r'^\d+$', answer_candidate):
            return answer_candidate
    # ...其他验证逻辑
```

其余重要自定义函数：

- CoT 推理函数：cot\_inference()：执行两阶段思维链推理
- 答案提取函数：extract\_answer()：从模型响应中精确提取答案
- 数据集评估函数：evaluate\_dataset()：端到端评估模型在数据集上的表现

## 3. Experiment Results

### 3-1 实验过程

#### 3-1-1 传统 CoT (version1.py)

运行代码，成功连接 API 后开始推理并打印输出信息：

```
jyt555deMacBook-Pro:2025-lab4 rainy$ python3 mine.py
测试API连接...
API连接成功！

开始评估 GSM8K 数据集 (50个样本)...
处理问题: 0%|
问题 1/50:
问题: Janet's ducks lay 16 eggs per day. She eats three for breakfast every morning and bakes muffins for her friends every ainder at the farmers' market daily for $2 per fresh duck egg. How much in dollars does she make every day at the farmers'
模型输出:

Janet每天通过卖出剩余的鸭蛋在农夫市场赚取金额计算如下:

1. **每日总产蛋量**: 16个
2. **消耗的鸡蛋**:
  - 早餐使用: 3个
  - 烤松饼使用: 4个
  - **总计消耗**:  $3 + 4 = 7$ 个
3. **剩余可售鸡蛋**:  $16 - 7 = 9$ 个
4. **每日收入**:  $9 \times \$2/\text{个} = \$18$ 

因此, Janet每天在
提取的答案: 18
标签答案: 18
处理问题: 2%|
问题 2/50:
问题: A robe takes 2 bolts of blue fiber and half that much white fiber. How many bolts in total does it take?
模型输出:

The robe requires 2 bolts of blue fiber and half as much white fiber as blue.

1. **Blue fiber**: 2 bolts
2. **White fiber**:  $\frac{2}{2} = 1$  bolt

**Total bolts** = Blue + White =  $2 + 1 = 3$ .

Step-by-step explanation:
- Start with the given amount for blue (2 bolts).
- Calculate half of that for white ( $\frac{2}{2} = 1$ ).
- Add both amounts together for the total (
提取的答案: 1
标签答案: 3
处理问题: 4%|
```

中途输出针对 GSM8K 的推理准确度（由于 api 调用频繁导致的限制等，这里推理准确度非常低）：

```
问题 50/50:
问题: Richard lives in an apartment building with 15 floors. Each floor contains 8 units, and 3/4 of the building is occupied. What's the total number of un occupied units In the building?
模型输出:

The total number of units in the building is calculated by multiplying the number of floors by the units per floor:
 $15 \text{ floors} \times 8 \text{ units/floor} = 120 \text{ total units}$ .

Since  $\frac{3}{4}$  are occupied, the fraction unoccupied is  $1 - \frac{3}{4} = \frac{1}{4}$ .

Calculating unoccupied units:
 $\frac{1}{4} \times 120 = 30 \text{ unoccupied units}$ 
提取的答案: 30
标签答案: 30
处理问题: 100%| 50/50 [1:39:51<00:00, 119.84s/it]

GSM8K 数据集评估结果:
样本数量: 50
正确数量: 4
准确率: 8.00%
```

开始推理常识性问题（这里只采样了 5 个样本作为程序正确性的测试）：

```
28 question = f"{row['question']} 选项如下: {row['choices']}"
29 inputs.append(question)
30 labels = [row['answerKey'] for _, row in df.iterrows()]
31

问题 53 输出 调试控制台 终端 SPELL CHECKER 38 POLYGLOT NOTEBOOK
开始评估 COMMONSENSEQA 数据集 (5个样本)... | 0/5 [00:00<
处理问题: 0%
问题 1/5:
问题: The sanctions against the school were a punishing blow, and they seemed to what the efforts the school had made to change? 选项如下: {'label': 'A', 'B', 'C', 'D', 'E'}, dtype=object), 'text': array(['ignore', 'enforce', 'authoritarian', 'yell at', 'avoid'], dtype=object)}
模型输出:

The correct answer is **A (ignore)**.

### Step-by-Step Explanation:
1. **Understand the Sentence Structure**:
   - The sentence states that sanctions (punitive measures) against a school were a "punishing blow," and they "seemed to [verb] the efforts the school had made to change."
   - The key is to find a verb that logically connects the sanctions to how they treated the school's efforts.

2. **Analyze Options**:
   - **C (authoritarian)** is an adjective, not a verb, so it's grammatically incorrect.
   - **D (yell at)** is a verb, but it doesn't fit the context of the sentence.

提取的答案: A
标签答案: A
处理问题: 20% | 1/5 [00:30<02:01, 5
```

最后输出实验准确性总结（同理，这里只采样了一个数学计算问题，因此这里给出的准确性不正确）：

```
COMMONSENSEQA 数据集评估结果：
样本数量：5
正确数量：2
准确率：40.00%
```

```
实验总结：
GSM8K准确率：100.00%
CommonsenseQA准确率：40.00%
jyt555deMacBook-Pro:2025-lab4 rainy$
```

### 3-1-2 答案前置 (version2.py)

由于使用的 qwen 是中文大模型，这一版代码将提示词全部改为中文，并做出了答案以特定格式前置的要求。运行代码：

```
jyt555deMacBook-Pro:2025-lab4 rainy$ python3 mine.py
测试API连接...
API连接成功！

开始评估 GSM8K 数据集 (5个样本)... | 0/5 [00:00<?, ?it/s]
处理问题: 0%
问题 1/5:
问题: Janet's ducks lay 16 eggs per day. She eats three for breakfast every morning and bakes muffins for her friends every day with four. She sells the remainder at the farmers' market daily for $2 per fresh duck egg. How much in dollars does she make every day at the farmers' market?
模型输出:

[18]
Janet每天获得16枚鸡蛋。她每天消耗3枚用于早餐和4枚用于烘焙，总共消耗3 + 4 = 7枚。剩余可出售的鸡蛋数量为16 - 7 = 9枚。每枚售价$2，因此每日收入为9 × $2 = $18。
提取的答案: 18
标签答案: 18
处理问题: 20% | 1/5 [00:28<01:55, 28.86s/it]
问题 2/5:
问题: A robe takes 2 bolts of blue fiber and half that much white fiber. How many bolts in total does it take?
模型输出:

[3]
首先确定蓝色纤维用量为2卷。白色纤维用量是蓝色的一半，即(1/2) × 2 = 1卷。总用量为(2 + 1 = 3)卷。
提取的答案: 3
标签答案: 3
处理问题: 40% | 2/5 [00:50<01:13, 24.52s/it]
```

```
[70,000]
Josh的总投资成本为购房款$80,000加上维修费$50,000，总计$130,000。房屋价值因维修增加了原价（$80,000）的15%，即：
\[\text{增值额} = \$8,000 \times 15\% = \$1,200\]
提取的答案: 8
标签答案: 70000
处理问题: 60% | 3/5 [02:45<02:11, 65.84s/it]
问题 4/5:
问题: James decides to run 3 sprints 3 times a week. He runs 60 meters each sprint. How many total meters does he run a week?
模型输出:

[540]
James每周进行3次训练，每次训练包含3个60米的冲刺。单次训练距离为：3次冲刺 × 60米/次 = 180米。周总距离为：180米/次 × 3次 = **540米**。
提取的答案: 540
标签答案: 540
处理问题: 80% | 4/5 [03:08<00:48, 48.83s/it]
问题 5/5:
问题: Every day, Wendi feeds each of her chickens three cups of mixed chicken feed, containing seeds, mealworms and vegetables to help keep them healthy. She gives the chickens their feed in three separate meals. In the morning, she gives her flock of chickens 15 cups of feed. In the afternoon, she gives her chickens another 25 cups of feed. How many cups of feed does she need to give her chickens in the final meal of the day if the size of Wendi's flock is 20 chickens?
模型输出:

[20]
根据题目描述：
1. Wendi共有20只鸡。每只每天需要3杯饲料，因此总需求为 (20 × 3 = 60) 杯。
2. 已分配的饲料：早晨15杯 + 下午25杯 = 40杯。
3. 剩余需分配的饲料为 (60 - 40 = 20) 杯。

最终答案为20杯。
提取的答案: 20
标签答案: 20
处理问题: 100% | 5/5 [03:51<00:00, 46.27s/it]

GSM8K 数据集评估结果：
样本数量：5
正确数量：4
准确率：80.00%
```



可以直观感受到模型推理的准确性提高了很多，针对常识问题的推理也是一样：

```
开始评估 COMMONSENSEQA 数据集 (5个样本)...
处理问题: 0% | 0/5 [00:00<?, 7it/s]
问题: 1/5:
问题: The sanctions against the school were a punishing blow, and they seemed to what the efforts the school had made to change? 选项如下: {'label': array(['A', 'B', 'C', 'D', 'E'], dtype=object), 'text': array(['ignore', 'enforce', 'authoritarian', 'yell at', 'avoid'], dtype=object)}
模型输出:
[A]
提取的答案: A
标签答案: A
处理问题: 20% | 1/5 [00:29<01:56, 29.01s/it]
问题: 2/5:
问题: Sammy wanted to go to where the people were. Where might he go? 选项如下: {'label': array(['A', 'B', 'C', 'D', 'E'], dtype=object), 'text': array(['ace track', 'populated areas', 'the desert', 'apartment', 'roadblock'], dtype=object)}
模型输出:
[答案]
[**B**]
**推理步骤:**
1. **题干分析:** Sammy想要去“people were”（人群所在）的地方，需选择人群密集或聚集的地点。
2. **排除法:**
   - **沙漠 (C):** 自然环境人烟稀少。
   - **路障 (E):** 临时交通管制设施，非人群主动聚集地。
   - **公寓 (D):** 住宅区以居住为主，公共区域人流有限。
3. **剩余选项对比:**
提取的答案: B
标签答案: B
处理问题: 40% | 2/5 [00:59<01:30, 30.07s/it]
```

### 3-1-3 结果对比

调整输出提示信息格式后，分别对两版代码采样 100 个问题（数学问题 50 个，常识问题 50 个），结果如下：

**GSM8K 数据集评估结果：**

样本数量：50  
正确数量：19  
准确率：38.00%

**COMMONSENSEQA 数据集评估结果：**

样本数量：50  
正确数量：16  
准确率：32.00%

**实验总结：**  
GSM8K准确率：38.00%  
CommonsenseQA准确率：32.00%

**GSM8K 数据集评估结果：**

样本数量：50  
正确数量：40  
准确率：80.00%

**COMMONSENSEQA 数据集评估结果：**

样本数量：50  
正确数量：37  
准确率：74.00%

**实验总结：**  
GSM8K准确率：80.00%  
CommonsenseQA准确率：74.00%

（左图 version1-传统 CoT；右图 version2-答案前置）

### 3-2 结果分析

根据实验结果，不添加任何修改的 CoT 推理准确率较低；而通过答案前置等方式能够有效提高推理准确率。

通过过程输出的提示信息，推理准确度较低的原因可能有以下几点：

1. API 调用频繁、网络/API 连接不稳定等原因导致的 API 调用失败，在经过多次尝试且超时后，直接略过该题目，导致该题推理出错：

```
问题 42/50:
问题: The great dragon, Perg, sat high atop mount Farbo, breathing fire upon anything within a distance of 1000 feet. Polly could throw the gold javelin only as far as the dragon's flames. But when Polly held the gemstone, she could throw the javelin three times farther than when not holding the gemstone. If holding the gemstone, how far outside of the reach of the dragon's flames could Polly stand and still hit the dragon with the gold javelin?
模型输出: API请求失败: Error code: 429 - {'message': 'Request was rejected due to rate limiting. Details: TPM limit reached.', 'data': None}
提取的答案: 429
标签答案: 200
处理问题: 84% | 42/50 [56:40<01:18, 9.0it/s]
API错误 (尝试 1/3): Error code: 429 - {'message': 'Request was rejected due to rate limiting. Details: TPM limit reached.', 'data': None}
API错误 (尝试 2/3): Error code: 429 - {'message': 'Request was rejected due to rate limiting. Details: TPM limit reached.', 'data': None}
API错误 (尝试 3/3): Error code: 429 - {'message': 'Request was rejected due to rate limiting. Details: TPM limit reached.', 'data': None}
问题 43/50:
问题: Grandma Jones baked 5 apple pies for the fireman's luncheon. She cut each pie into 8 pieces and set the five pies out on the buffet table for
```

```

答案: 18
提取的答案: 18
标签答案: 18
处理问题: 2%|
问题 2/50:
问题: A robe takes 2 bolts of blue fiber and half that much white fiber. How many bolts in total does it take?
模型输出:
3
提取的答案: 3
标签答案: 3
处理问题: 4%|
API错误 (尝试 1/3): Request timed out.
API错误 (尝试 2/3): Request timed out.
API错误 (尝试 3/3): Request timed out.
问题 3/50:
问题: Josh decides to try flipping a house. He buys a house for $80,000 and then puts in $50,000 in repairs. This increased the value of the ho
How much profit did he make?
模型输出: API请求失败: Request timed out.
提取的答案: None
标签答案: 70000
处理问题: 6%|
问题 4/50:
问题: James decides to run 3 sprints 3 times a week. He runs 60 meters each sprint. How many total meters does he run a week?
模型输出:
James每周进行的总跑步距离计算如下:
1. **每次训练的冲刺次数**:: 3次
2. **每周训练次数**:: 3次
   - 每周总冲刺次数 =  $(3 \times 3 = 9)$  次
3. **每次冲刺的距离**:: 60米
   - 总距离 =  $(9 \times 60 = 540)$  米
因此, James每周通过冲刺训练总共跑了**540米**.*
答案: 540
提取的答案: 540
标签答案: 540
处理问题: 8%|

```

2. 读取 token 的限制: token 数过小时, 程序不能获得大语言模型完整的输出, 而传统 CoT 在推理数学问题时选择提取 token 中最后一个数字作为答案, 因此非完整的输出直接导致了提取到的答案错误, 即使大语言模型给出了正确的答案:

```

问题 9/50:
问题: John drives for 3 hours at a speed of 60 mph and then turns around because he realizes he forgot something very important at home. He spends the next half-hour driving at a speed of 30mph, before beginning his remaining time of the 4 hours going at 80 mph. How far is he from home at the end of those 4 hours?
模型输出: API请求失败: Error code: 400 - {'code': '20015', 'message': 'max_tokens: Must be less than or equal to 32768', 'data': None}
提取的答案: 32768
标签答案: 45

```

```

1. **Per Session Calculation:**
Each session consists of **3 sprints**, and each sprint is **60 meters**:

$$3 \times \text{3 sprints} \times 60 \text{ meters/sprint} = 180 \text{ meters/session}$$


2. **Weekly Total:**
He trains **3 times a week**, so multiply the per-session distance by the number of sessions:

$$180 \times 3 = 540$$

提取的答案: 180
标签答案: 540
处理问题: 8%|
问题 5/50:
问题: Every day, Wendi feeds each of her chickens three cups of mixed chicken feed, containing seeds, mealworms and vegetables to help keep them healthy. She gives the chickens their feed in three separate meals. In the morning, she gives her flock of chickens 15 cups of feed. In the afternoon, she gives her chickens another 25 cups of feed. How many cups of feed does she need to give her chickens in the final meal of the day if the size of Wendi's flock is 20 chickens?
模型输出:
To determine how many cups of feed Wendi needs to give her chickens in the final meal of the day:

1. **Total Daily Requirement:**
Each of Wendi's **20** chickens eats **3** cups/day.

$$\text{Total} = (20 \times 3) = 60 \text{ cups/day}$$


2. **Feed Already Given:**
Morning: **15** cups
Afternoon: **25** cups
Total so far = 40 cups
提取的答案: 25
标签答案: 20
处理问题: 10%|
问题 6/50:
问题: Kylar went to the store to buy glasses for his new apartment. One glass costs $5, but every second glass costs only 60% of the price. Kylar wants to buy 16 glasses. How much does he need to pay for them?
模型输出:
Kylar buys 16 glasses where every second glass is priced at 60% of the original $5. Here's the breakdown:
- **Full-price glasses:** Every odd-numbered glass (1st, 3rd, ..., 15th) costs **$5**. There are  $(\frac{16}{2}) = 8$  such glasses.
- **Discounted glasses:** Every even-numbered glass (2nd, 4th, ..., 16th) costs **$3** ( $(5 \times 0.6)$ ). There are also 8 such glasses.
提取的答案: 6
标签答案: 64
处理问题: 12%|

```

3. 题库自身题目可能存在一定问题, 例如下面这题根据题干只能排除 CDE 三个选项, 并不能真正判断答案是 A 还是 B (二者皆有道理)。根据输出 token 来看, 大语言模型也在 A 和 B 选项中抉择, 最终错误判断答案为 B:

```
问题 3/5:
问题: To locate a choker not located in a jewelry box or boutique where would you go? 选项如下: {'label': array(['A', 'B', 'C', 'D', 'E'], dtype=object), 't
ext': array(['jewelry store', 'neck', 'jewelry box', 'jewelry box', 'boutique'],
dtype=object)}
模型输出:
[B]
问题要求寻找不在首饰盒或精品店的choker位置。排除选项C（拼写错误的首饰盒）、D（首饰盒）、E（精品店）后，剩下A（珠宝店）和B（颈部）。珠宝店是销售场所而非存放
位置，而choker作为贴颈项链，最可能佩戴在颈部（neck），故选B。
[答案]
[B]
提取的答案: B
标签答案: A
处理问题: 60% | 3/5 [01:25-00:56, 28.06s/it]
```

## References:

1 API 调用接口: [创建文本对话请求 - SiliconFlow](#)

2 [一文读懂: 思维链 CoT \(Chain of Thought\)](#)

### 附录-关于报告在ddl后上交的说明:

最开始仍旧试图使用华为云的 mindspore 平台，但是代金券不够了，因此在五月底的时候向助教申请了新的代金券。由于此前下发代金券的时候平台有短信提示，因此在没收到短信的时候一直没想到登录平台查看代金券，后来时间比较久了才登录平台查看，最后发现这次发的短信被归到垃圾短信里拦截了，因此我发现代金券下发的时间比它实际下发的时间迟了很多。

而在 mindspore 运行的时候发现推理速度非常慢，而且一不小心中断了就得重新开始，耗时过久，因此转而选择了调用“硅基流动”的 api。而实验过程中又发现 api 调用过于频繁时会被锁定，因此延长了代码中 api 调用间隔时间，最终导致实验耗时较久，拖延 ddl 较久。