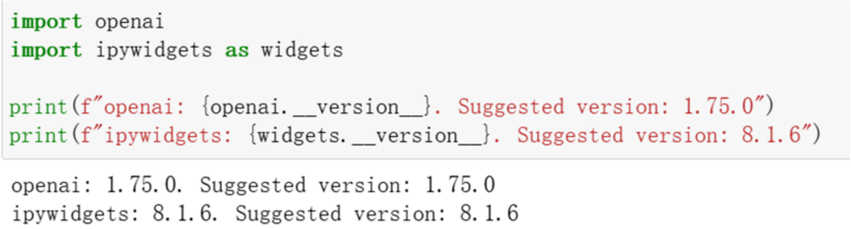
Step1: Run Project0515.ipynb with jupyter notebook

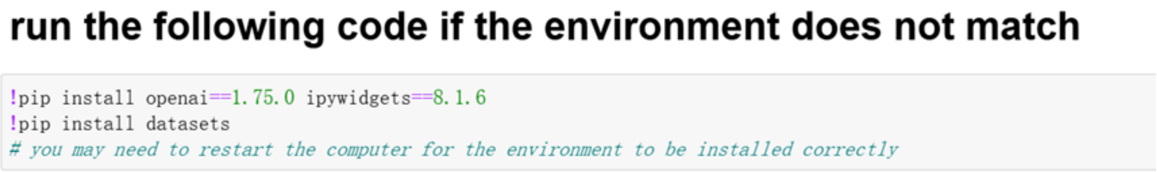
Step2: Run the first cell to make sure you have the correct version of the libraries



```
import openai
import ipywidgets as widgets

print(f"openai: {openai.__version__}. Suggested version: 1.75.0")
print(f"ipywidgets: {widgets.__version__}. Suggested version: 8.1.6")
```

```
openai: 1.75.0. Suggested version: 1.75.0
ipywidgets: 8.1.6. Suggested version: 8.1.6
```

Run the next cell if there is error. You should restart Kernel after installation, but you also may need to restart the computer sometimes for the environment to be installed correctly.

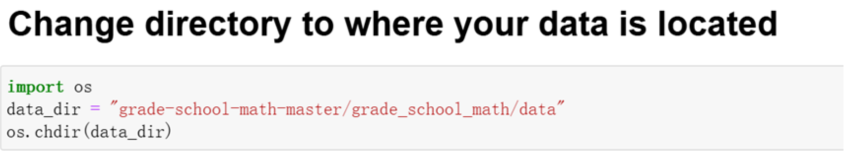


run the following code if the environment does not match

```
!pip install openai==1.75.0 ipywidgets==8.1.6
!pip install datasets
# you may need to restart the computer for the environment to be installed correctly
```

If you see error messages like ModuleNotFoundError: No module named 'random', create a new cell and run "!pip install package_name" and restart Kernel.

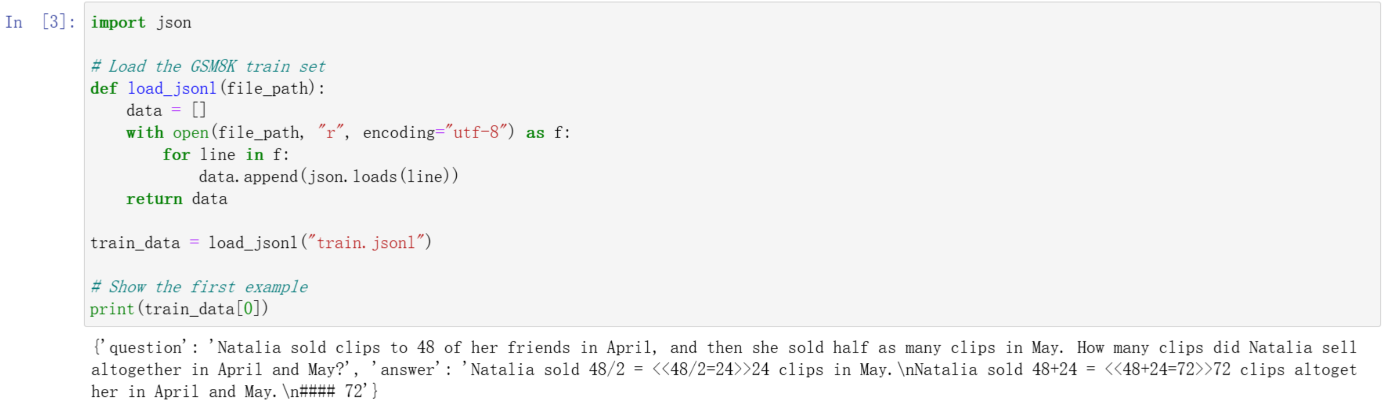
Step3: Manually edit the directory of the folder "data" under "grade-school-math-master". For example: data_dir = "/Users/29117/OneDrive/桌面 /grade-school-math-master/grade_school_math/data"



Change directory to where your data is located

```
import os
data_dir = "grade-school-math-master/grade_school_math/data"
os.chdir(data_dir)
```

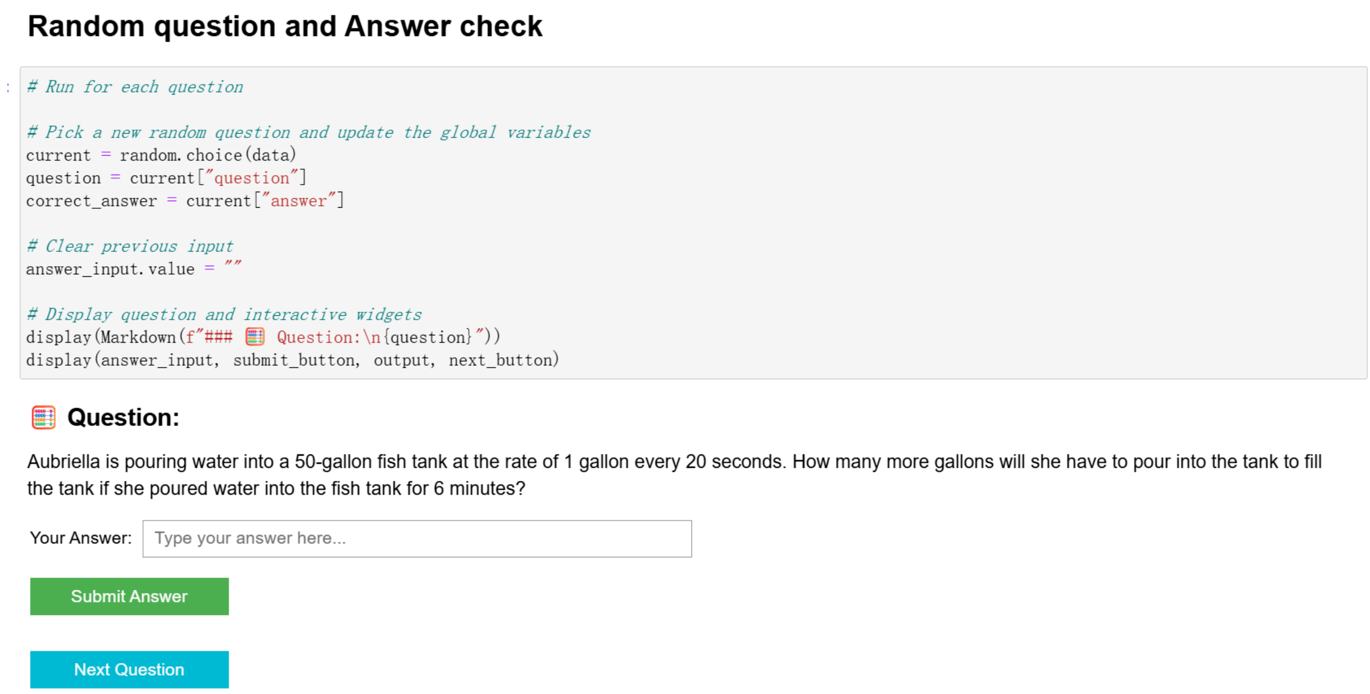
Step4: Run the next cell and see if you get the same output to make sure the dataset is loaded correctly



```
In [3]: import json

        # Load the GSM8K train set
        def load_jsonl(file_path):
            data = []
            with open(file_path, "r", encoding="utf-8") as f:
                for line in f:
                    data.append(json.loads(line))
            return data

        train_data = load_jsonl("train.jsonl")

        # Show the first example
        print(train_data[0])
```

```
{'question': 'Natalia sold clips to 48 of her friends in April, and then she sold half as many clips in May. How many clips did Natalia sell
altogether in April and May?', 'answer': 'Natalia sold 48/2 = <<48/2=24>>24 clips in May.\nNatalia sold 48+24 = <<48+24=72>>72 clips altoget
her in April and May.\n#### 72'}
```

Step5: Keep running the cells and see if you get a similar output



Random question and Answer check

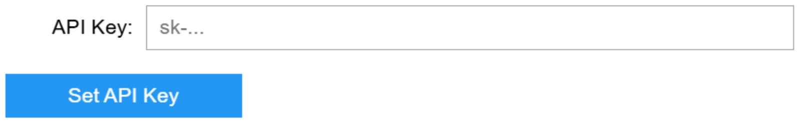
```
# Run for each question

# Pick a new random question and update the global variables
current = random.choice(data)
question = current["question"]
correct_answer = current["answer"]

# Clear previous input
answer_input.value = ""

# Display question and interactive widgets
display(Markdown(f"### 🧮 Question:\n{question}"))
display(answer_input, submit_button, output, next_button)
```

🧮 Question:

Aubriella is pouring water into a 50-gallon fish tank at the rate of 1 gallon every 20 seconds. How many more gallons will she have to pour into the tank to fill the tank if she poured water into the fish tank for 6 minutes?

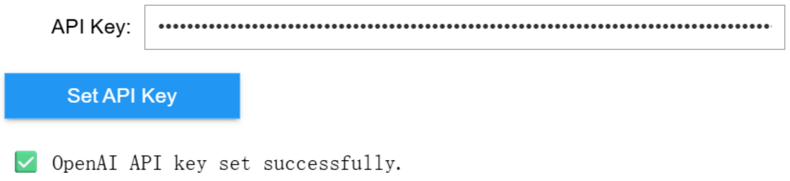
Your Answer: Type your answer here...

Step6: Run the cells and enter your own Openai API Key



API Key: sk-...

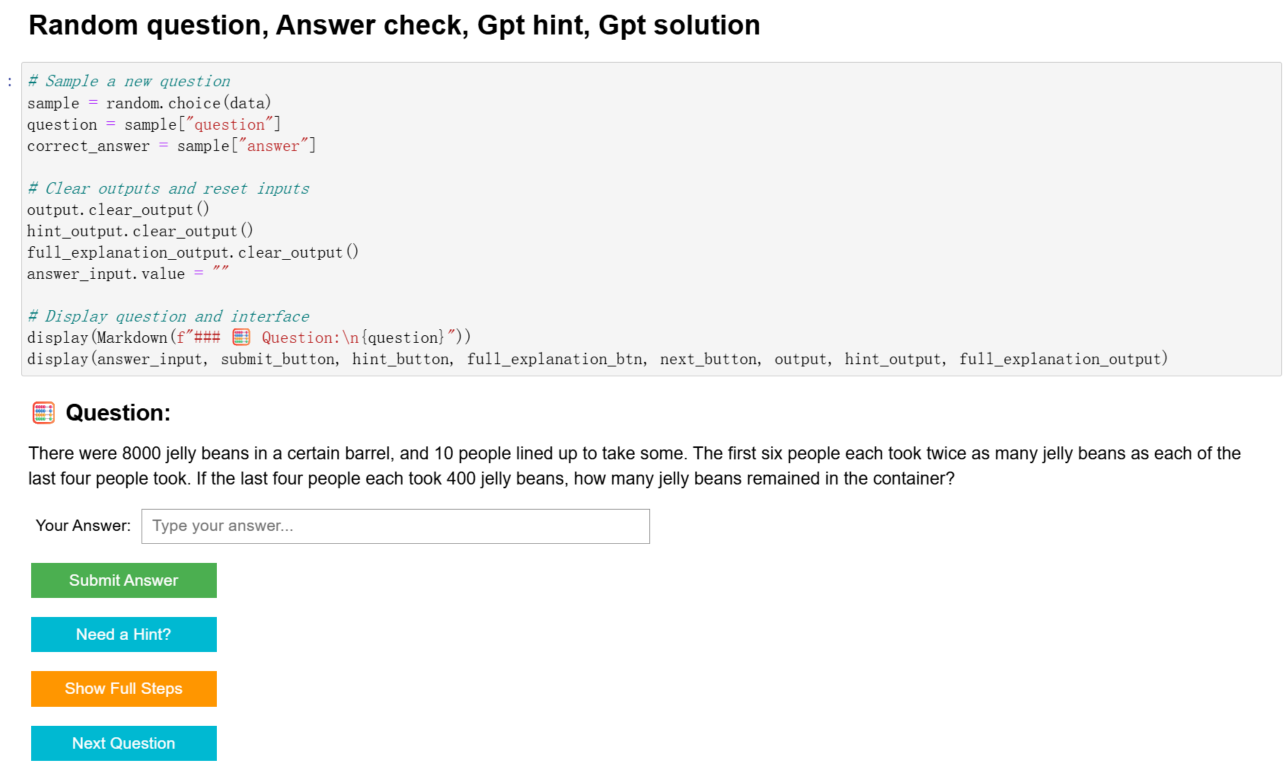
If you see "⚠️ Invalid API key format. Please double check.", double check if you are using a valid Openai API Key. It should start with "sk-". Once success, you should be able to see



API Key: ••

```
✅ OpenAI API key set successfully.
```

Step7: run the cells and you will get to a interactive window



Random question, Answer check, Gpt hint, Gpt solution

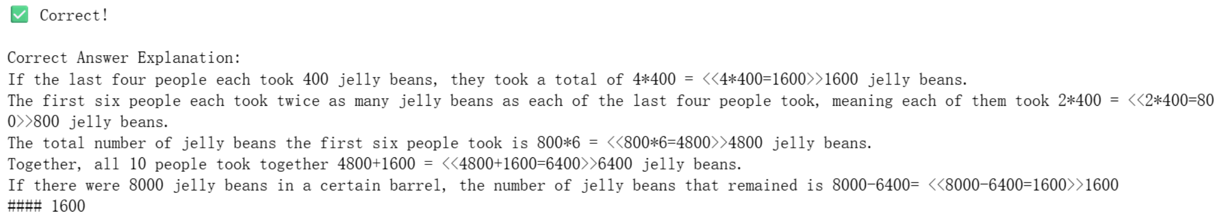
```
# Sample a new question
sample = random.choice(data)
question = sample["question"]
correct_answer = sample["answer"]

# Clear outputs and reset inputs
output.clear_output()
hint_output.clear_output()
full_explanation_output.clear_output()
answer_input.value = ""

# Display question and interface
display(Markdown(f"### 🧮 Question:\n{question}"))
display(answer_input, submit_button, hint_button, full_explanation_btn, next_button, output, hint_output, full_explanation_output)
```

🧮 Question:

There were 8000 jelly beans in a certain barrel, and 10 people lined up to take some. The first six people each took twice as many jelly beans as each of the last four people took. If the last four people each took 400 jelly beans, how many jelly beans remained in the container?

Here you will see different outputs by clicking different buttons:

If you answer the question correctly and click Submit Answer , you will see "✅ Correct!" and the correct answer explanation.

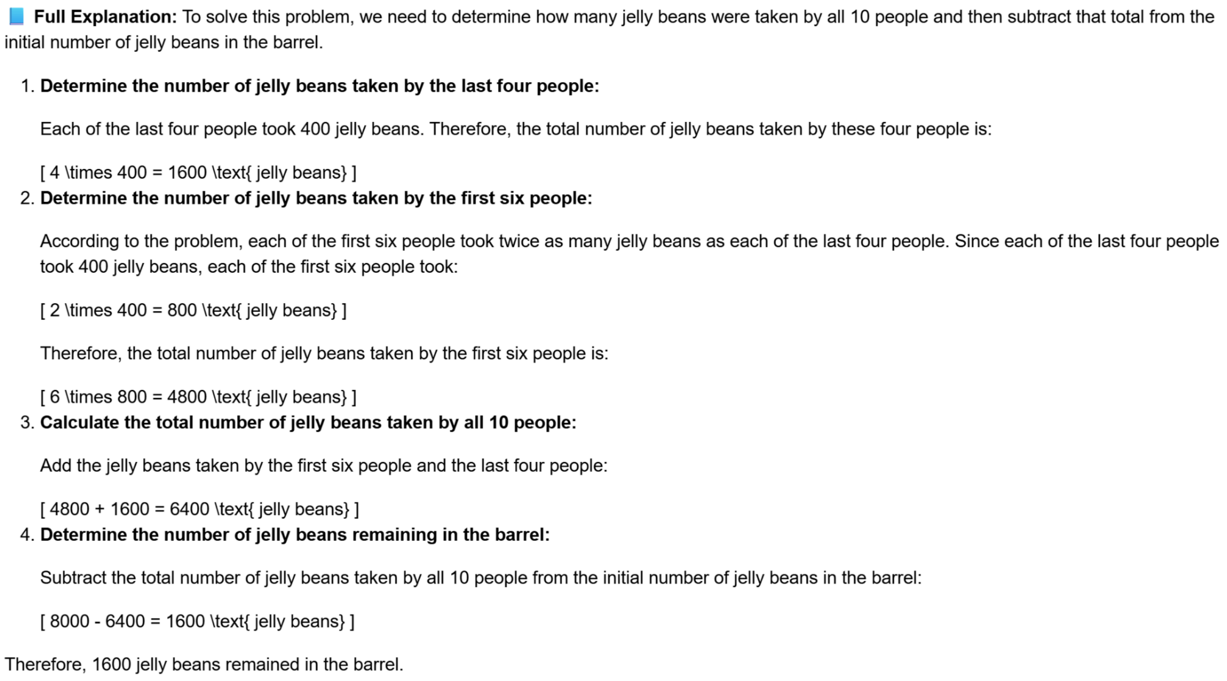


```
✅ Correct!

Correct Answer Explanation:
If the last four people each took 400 jelly beans, they took a total of 4*400 = <<4*400=1600>>1600 jelly beans.
The first six people each took twice as many jelly beans as each of the last four people took, meaning each of them took 2*400 = <<2*400=80
0>>800 jelly beans.
The total number of jelly beans the first six people took is 800*6 = <<800*6=4800>>4800 jelly beans.
Together, all 10 people took together 4800+1600 = <<4800+1600=6400>>6400 jelly beans.
If there were 8000 jelly beans in a certain barrel, the number of jelly beans that remained is 8000-6400= <<8000-6400=1600>>1600
#### 1600
```

If you answer the question incorrectly and click Submit Answer , you will see something like "❌ Incorrect. The correct answer is: 1600" and the correct answer explanation.
If you leave the space blank and click Submit Answer , you will see"⚠️ Error in processing the answer: could not convert string to float: ' ' " and the correct answer explanation.

If you click Need a Hint? , you will see ⏳ Generating hint... shortly and then get some output like: 💡 **Hint:** Try breaking down the problem by identifying the total number of jelly beans taken by the last four people and then using that information to find the total number of jelly beans taken by the first six people.

If you click Show Full Steps , you will see ⏳ Generating full solution... shortly and get output like this:



📘 **Full Explanation:** To solve this problem, we need to determine how many jelly beans were taken by all 10 people and then subtract that total from the initial number of jelly beans in the barrel.

1. **Determine the number of jelly beans taken by the last four people:**

   Each of the last four people took 400 jelly beans. Therefore, the total number of jelly beans taken by these four people is:

   [4 \times 400 = 1600 \text{ jelly beans}]
2. **Determine the number of jelly beans taken by the first six people:**

   According to the problem, each of the first six people took twice as many jelly beans as each of the last four people. Since each of the last four people took 400 jelly beans, each of the first six people took:

   [2 \times 400 = 800 \text{ jelly beans}]

   Therefore, the total number of jelly beans taken by the first six people is:

   [6 \times 800 = 4800 \text{ jelly beans}]
3. **Calculate the total number of jelly beans taken by all 10 people:**

   Add the jelly beans taken by the first six people and the last four people:

   [4800 + 1600 = 6400 \text{ jelly beans}]
4. **Determine the number of jelly beans remaining in the barrel:**

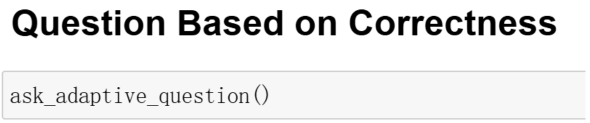
   Subtract the total number of jelly beans taken by all 10 people from the initial number of jelly beans in the barrel:

   [8000 - 6400 = 1600 \text{ jelly beans}]

Therefore, 1600 jelly beans remained in the barrel.

If you click Next Question , you will see a new question with the default setup with no output.

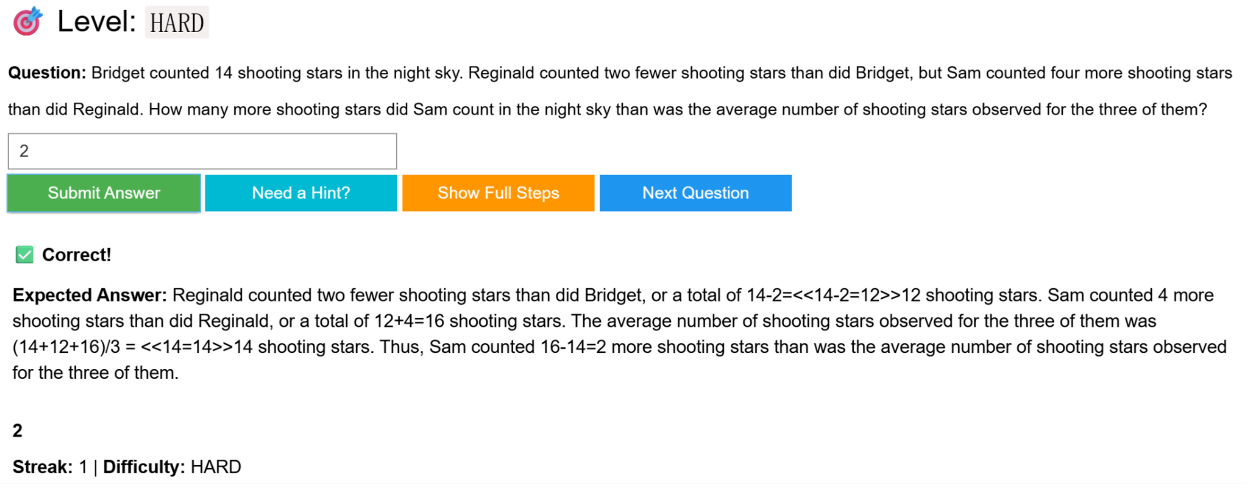
Step8: Keep running the cells until you see our final version math tutor:



Question Based on Correctness

```
ask_adaptive_question()
```

Running this cell with generate an Easy level question like:



🎯 Level: HARD

Question: Bridget counted 14 shooting stars in the night sky. Reginald counted two fewer shooting stars than did Bridget, but Sam counted four more shooting stars than did Reginald. How many more shooting stars did Sam count in the night sky than was the average number of shooting stars observed for the three of them?

2

✅ **Correct!**

Expected Answer: Reginald counted two fewer shooting stars than did Bridget, or a total of 14-2=<<14-2=12>>12 shooting stars. Sam counted 4 more shooting stars than did Reginald, or a total of 12+4=16 shooting stars. The average number of shooting stars observed for the three of them was (14+12+16)/3 = <<14=14>>14 shooting stars. Thus, Sam counted 16-14=2 more shooting stars than was the average number of shooting stars observed for the three of them.

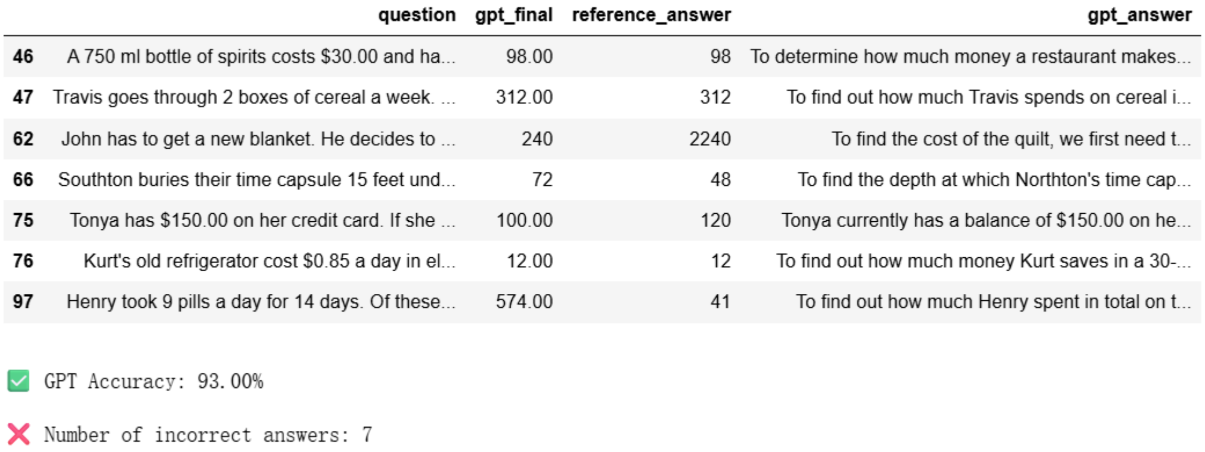
2

Streak: 1 | **Difficulty:** HARD

All the buttons have the same function, except that Next Question will show the next question without removing the output history. When you check the answer by clicking Submit Answer , you will also see a winning streak **Streak:** 1 | **Difficulty:** EASY .

This difficulty level will increase if the winning streak hits 2. While if an answer is considered incorrect, the difficulty level will decrease (i.e.) The minimum level is EASY and maximum level is HARD.

Step9: This step should **NOT** be running since it will override the result.



	question	gpt_final	reference_answer	gpt_answer
46	A 750 ml bottle of spirits costs $30.00 and ha...	98.00	98	To determine how much money a restaurant makes...
47	Travis goes through 2 boxes of cereal a week. ...	312.00	312	To find out how much Travis spends on cereal i...
62	John has to get a new blanket. He decides to ...	240	2240	To find the cost of the quilt, we first need t...
66	Southton buries their time capsule 15 feet und...	72	48	To find the depth at which Northton's time cap...
75	Tonya has $150.00 on her credit card. If she ...	100.00	120	Tonya currently has a balance of $150.00 on he...
76	Kurt's old refrigerator cost $0.85 a day in el...	12.00	12	To find out how much money Kurt saves in a 30-...
97	Henry took 9 pills a day for 14 days. Of these...	574.00	41	To find out how much Henry spent in total on t...

```
✅ GPT Accuracy: 93.00%

❌ Number of incorrect answers: 7
```

This code is a systematic evaluation of GPT-4o’s ability to solve math word problems by comparing its answers to reference answers from the GSM8K dataset. The process begins by randomly selecting 100 questions and prompting GPT-4o to solve each one, instructing the model to end its response with a clearly formatted final answer (e.g., #### 72). The code then extracts this final number and compares it to the correct answer provided in the dataset. If the model omits the expected format, a fallback method retrieves the last number in the response to maintain evaluation consistency. Each result is recorded along with whether the model’s answer matches the reference exactly, and overall accuracy is calculated. Additionally, the code highlights incorrect responses for further review.