GenAI HW4
Become an AI Hypnosis Master

TA: 樊樺、呂睿超、林堅壬

ntu-gen-ai-2024-spring-ta@googlegroups.com

Deadline: 2024/04/04 23:59:59 (UTC+8)

Outline

- Overview
- Task Introduction
- Grading and Submission
- Regulations
- Colab
- Hints

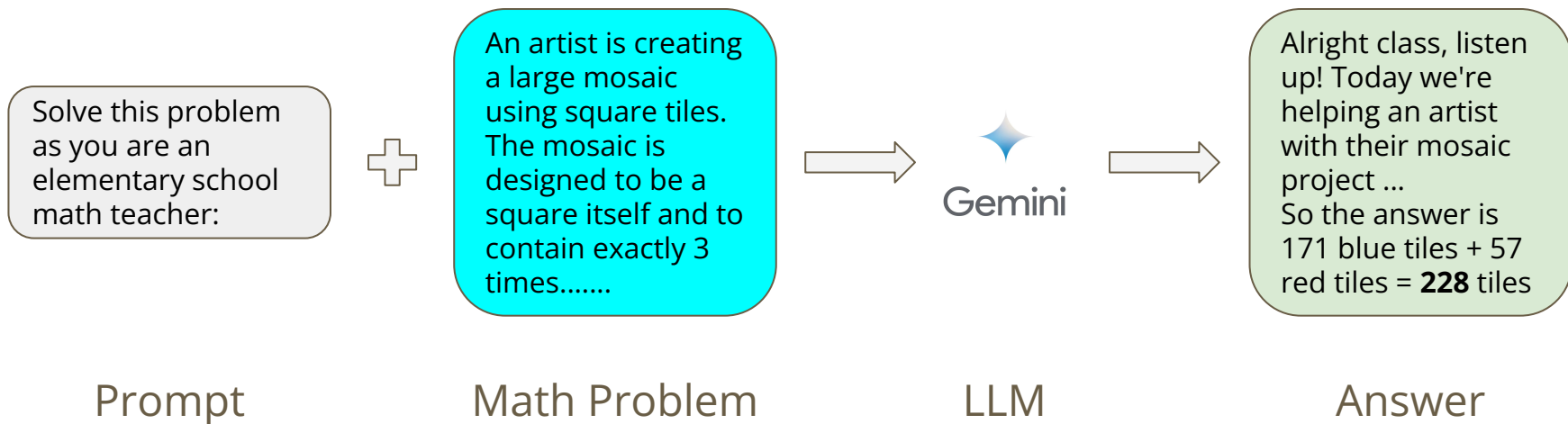# Links

- NTU Cool HW4 Forum
- HW4 Colab
- Google Gemini

Overview

Objectives

- Design prompts that improve Gemini's accuracy in solving mathematical problems.
- Understand how prompts affect the performance of a Large Language Model (LLM).

Solve this problem as you are an elementary school math teacher:

➕

An artist is creating a large mosaic using square tiles. The mosaic is designed to be a square itself and to contain exactly 3 times.......

➡️ Gemini ➡️

Alright class, listen up! Today we're helping an artist with their mosaic project ...
So the answer is 171 blue tiles + 57 red tiles = **228** tiles

Prompt   Math Problem   LLM   Answer

The importance of prompt designing

- Without prompt designing, Gemini cannot solve the problem correctly.

    ❏ **Prompt** : (Directly copying and pasting the math question)

有一張長方形壁報紙，長 36 公分、寬 40 公分，老師想要把它切成很多個一樣大的正方形，不要剩下，請問無法切出下列哪一種邊長的正方形？

(1) 1 公分

(2) 2 公分

(3) 4 公分

(4) 6 公分

The importance of prompt designing

- Without prompt designing, Gemini cannot solve the problem correctly.

  ❑ **Prompt** : (Directly copying and pasting the math question)

The importance of prompt designing

- With prompt designing, Gemini outputs correct answer with explanation

  ❏ **Prompt** : (Directly copying and pasting) + giving related information and detail

You

有一張長方形壁報紙，長 36 公分、寬 40 公分，老師想要把它切成很多個一樣大的正方形，不要剩下，請問無法切出下列哪一種邊長的正方形？

(1) 1 公分

(2) 2 公分

(3) 4 公分

(4) 6 公分

請根據公因數的概念觀察題目並一步一步小心計算答案，並注意題目要選擇的是無法切出的邊長

The importance of prompt designing

- With prompt designing, Gemini outputs correct answer with explanation

根據公因數的概念，要切出一樣大的正方形，正方形的邊長必須是長方形壁報紙的長和寬的公因數。

36 的公因數有 1、2、3、4、6、9、12、18、36。

40 的公因數有 1、2、4、5、8、10、20、40。

兩者共同的公因數有 1、2、4。

因此，可以切出邊長為 1 公分、2 公分、4 公分的正方形。

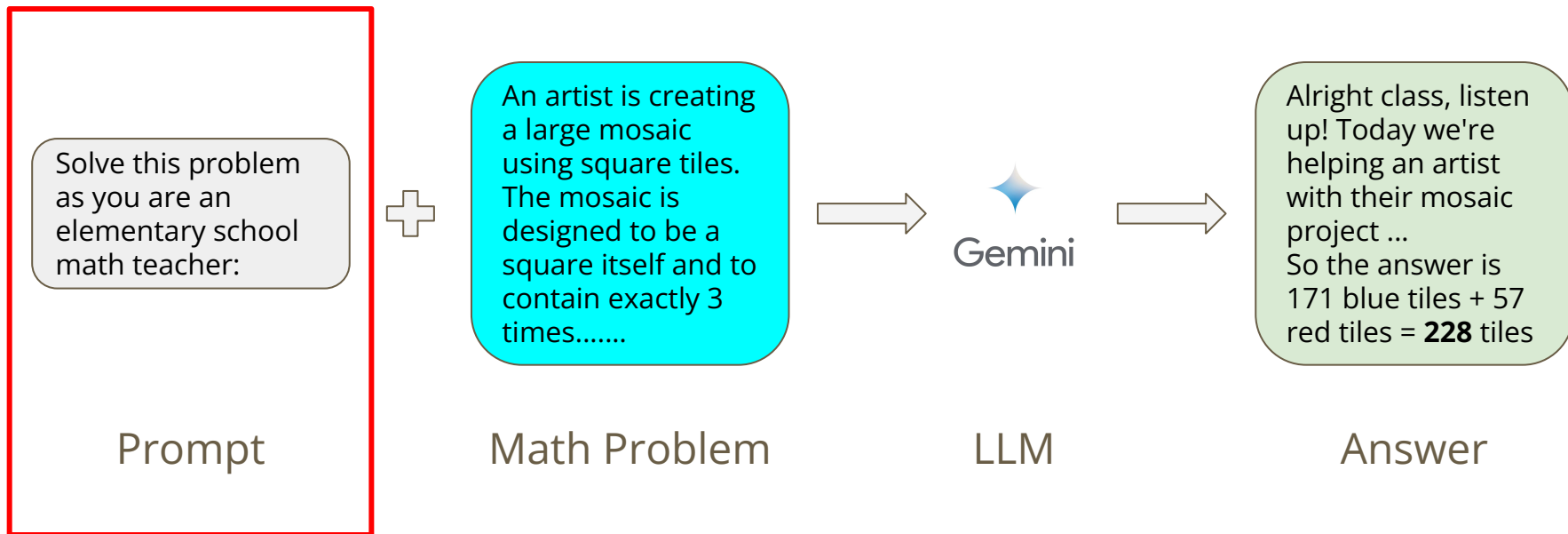而 6 公分不是 36 和 40 的公因數，因此無法切出邊長為 6 公分的正方形。

答案：(4)。

Task Introduction

In HW4:

- You need to **design a prompt** to help Gemini **solve math problems** more effectively.
- We have provided 30 math problems; you need to come up with a prompt that will enable Gemini to consistently achieve higher accuracy on those problems.

In HW4:

Solve this problem as you are an elementary school math teacher:

Prompt

An artist is creating a large mosaic using square tiles. The mosaic is designed to be a square itself and to contain exactly 3 times.......

Math Problem

Gemini

LLM

Alright class, listen up! Today we're helping an artist with their mosaic project ...
So the answer is 171 blue tiles + 57 red tiles = **228** tiles

Answer

We will focus on this part !

Workflow

HW4.ipynb - Colaboratory (google.com)

Step 0: Set up your Gemini API Key

Step 1: Design your prompt

Step 2: Check your prompt

Step 3: Evaluate your prompt

Step 4: Save and download your prompt

- Note: We can see that there is no training required and just like HW3, we utilize **Gemini API** for this assignment. Therefore, **No GPU resource is required.**
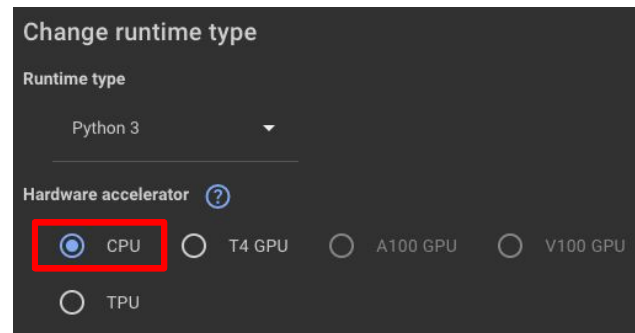
Gradio and no Gradio Version

- To avoid the inconvenience brought by gradio server crash, we prepare the codes that can be run without gradio
- You only need to either run the codes with gradio or the codes without gradios. **You DON'T need to run both of them**

Step 0

- Set up your Gemini API Key

∨ Import all necessary packages and set up your **Gemini API Key**.

> Set your Gemini API Key.
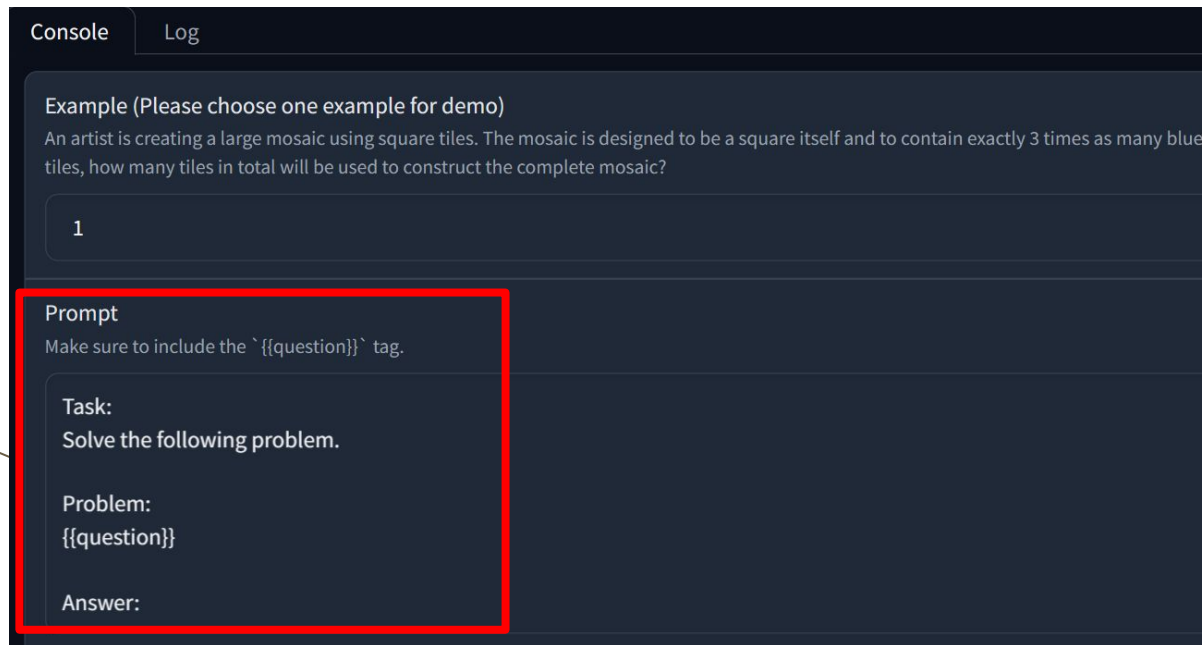
▶ **GOOGLE_API_KEY:** " 在這裡插入text

顯示程式碼

Step 1

- Design your prompt.
  - Design your prompt to improve Gemini's performance in solving math problems.
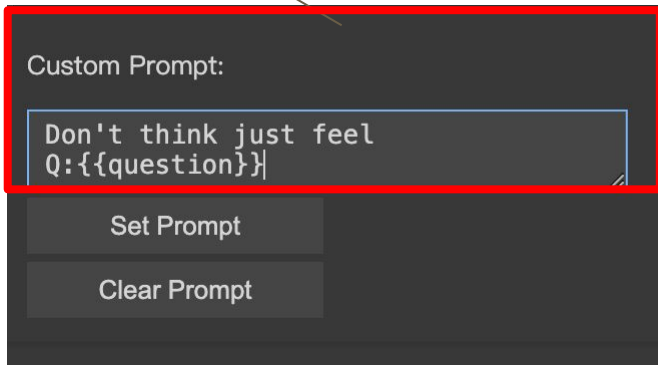
You can write your prompt in this text box.

Step 1 - No Gradio Version

- Design your prompt.
  - Design your prompt to improve Gemini's performance in solving math problems.

You can write
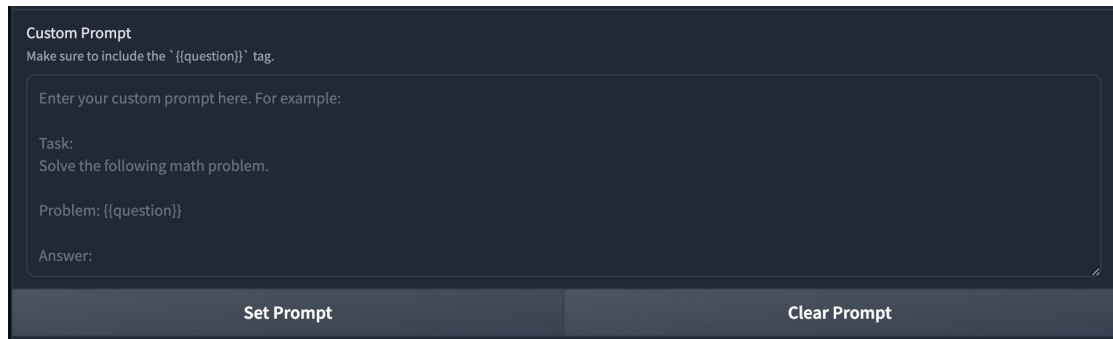your prompt
in this text
box.

Custom Prompt:

Don't think just feel
Q:{{question}}

Set Prompt

Clear Prompt

No gradio

Custom Prompt:
Make sure to include the `{{question}}` tag.

Enter your custom prompt here. For example:

Task:
Solve the following math problem.

Problem: {{question}}

Answer:

Set Prompt

Clear Prompt

gradio

Step 1

- Design your prompt.
  - Your prompt should be structured as a template with a **placeholder: {{question}}** for inserting a specific math problem.

Must include this structure with the **{{question}}** placeholder

```
Task:
Solve the following problem.

Problem:
{{question}}

Answer:
```

Step 1

- Design your prompt.
  - We will replace the **{{question}}** in your prompt with a real math question.
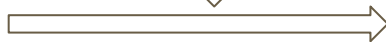 - All the math questions will use the same prompt.

A farm has chickens and cows. If there are a total of 30 heads and 88 legs, how many cows are on the farm?

Math Problem

```
Task:
Solve the following
math problem.

Problem:
{{question}}

Answer
```

```
Task:
Solve the following math problem.

Problem:
A farm has chickens and cows. If there are
a total of 30 heads and 88 legs, how many
cows are on the farm?

Answer
```

Step 1

- Valid prompt examples:

Don't think, just feel.
Q: {{question}}
A:

You are a math teacher.
Please solve this math problem.
{{question}}

Note: If you want to change line in the text box, use **Shift + Enter**.

Step 2 - Gradio

- Check your prompt
  - Choose an example
  - Click set prompt button

1.

Console | Log

Demo Example (Please choose one example for demo)

An artist is creating a large mosaic using square tiles. The mosaic is designed to be a square itself and to contain exactly 3 times as many blue tiles as red tiles. If the artist has only 57 red tiles, how many tiles in total will be used to construct the complete mosaic?

1 ▾

Custom Prompt

Make sure to include the `{{question}}` tag.

Enter your custom prompt here. For example:

Task:
Solve the following math problem.

Problem: {{question}}

Answer:

2.

| Set Prompt | | Clear Prompt |

Number of prompt tokens

The token number of the assigned custom prompt.

0

Custom Prompt

Make sure to include the `{{question}}` tag.

Don't think just feel.
Q:{{question}}
A:

Set Prompt

Number of prompt tokens

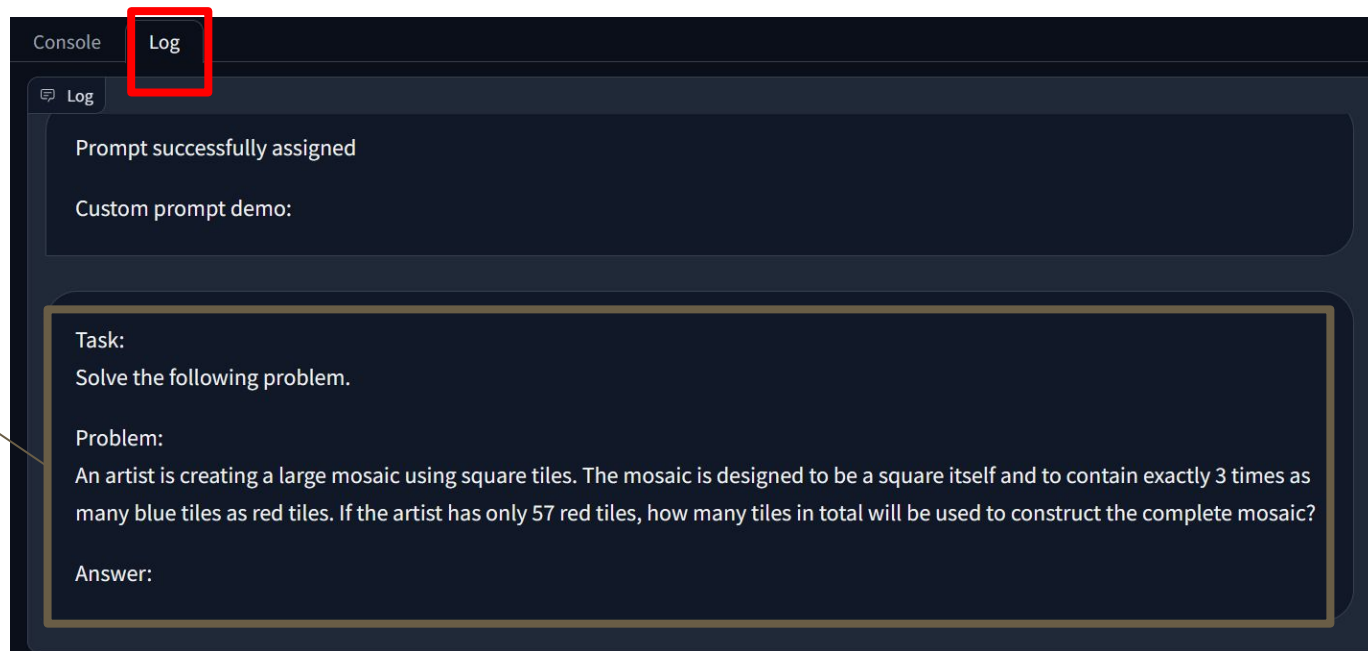The token number of the assigned custom prompt.

16

- **After setting prompt** , It will calculate the tokens of your prompt

- This button is for you to reset the prompt

20

Step 2 - Gradio Version

- Check your prompt in Log
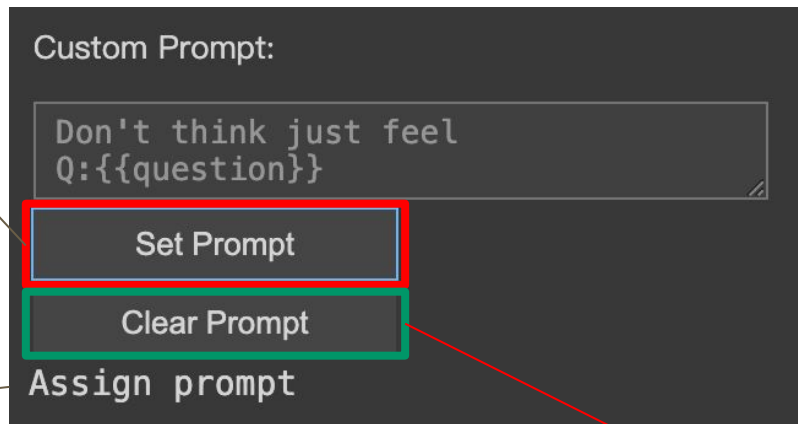 - Log is the history of interactions with Gradio.

This text will be the input of Gemini.

Console   Log

💬 Log

Prompt successfully assigned

Custom prompt demo:

Task:
Solve the following problem.

Problem:
An artist is creating a large mosaic using square tiles. The mosaic is designed to be a square itself and to contain exactly 3 times as many blue tiles as red tiles. If the artist has only 57 red tiles, how many tiles in total will be used to construct the complete mosaic?

Answer:

Step 2 - No Gradio Version

- Fill in your prompt
  - Click set prompt button

After setting prompt ,
It will output a message to tell you've successfully assigned the prompt

Custom Prompt:

Don't think just feel
Q:{{question}}

Set Prompt

Clear Prompt

Assign prompt

- This button is for you to reset the prompt

Step 2 - No Gradio Version

- Check your prompt
  - Select demo and check custom prompt demo

- Choose an example


```
custom_prompt = prompt_area.value
assert "{{question}}" in custom_prompt

#custom_prompt = "Task: {{question}}" # @param {type:"raw"}
Demo_Example = "4" # @param [1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20
demo_index = eval(Demo_Example)


environment = jinja2.Environment()
template = environment.from_string(custom_prompt)
print(f"Custom prompt demo:\n\n{template.render(question=questions[demo_index-1
```

This text will be the input of Gemini.

```
Custom prompt demo:

Don't think just feel
Q:A farmer's market sells apples in two types of mixed bags: Type A contains 4 red apples and 6 green apples, and Ty
```

Demo_Example: 4

Step 2

- We have provided some math problems: link
  - Each example contains one question and one answer

	A	B
	Question	Ground Truth
2	An artist is creating a large mosaic using square tiles. The mosaic is designed to be a square itself and to contain exactly 3 times as many blue tiles as red tiles. If th	228
3	A farmer is filling baskets with apples for a local market. He has 120 apples and wants to distribute them evenly among his baskets. If he decides to keep 15 apples	15
4	A garden has rectangular plots arranged in such a way that each plot shares sides with exactly two other plots, forming a straight line. There are 5 plots in total. Eac	132
5	A farmer's market sells apples in two types of mixed bags: Type A contains 4 red apples and 6 green apples, and Type B contains 8 red apples and 4 green apples.	0.45
6	A gardener is planting flowers in a pattern: two red flowers followed by one yellow flower. If the gardener wants to maintain this pattern and has 35 spaces in a row t	24
7	Jason is training for a marathon and runs a certain distance each day. On Monday, he runs 5 miles. Each day after, he increases his distance by 10%. If Jason cont	7.3205
8	In a triangular garden bed, there are 16 plants along each side. Each plant requires a circular space with a radius of 0.5 meters to grow properly. Assuming plants a	16
9	Dr. Wilson is planting a geometric garden with flowers arranged in concentric rings around a central fountain. Each outer ring has 6 more flowers than the inner one	720
10	A small library is reorganizing its collection. They have a total of 120 books to distribute equally among 5 shelves. The top shelf can only bear a load of books equal	30
11	A recipe for a batch of cookies requires 3 cups of flour, 2 cups of sugar, and 1 cup of chocolate chips. If Max wants to make a triple batch but only has 4 cups of sug	2
12	A pet store owner is making custom birdhouses. Each birdhouse requires 0.75 liters of wood varnish for the exterior. If the owner has a 10-liter can of wood varnish,	13
13	A farm has chickens and cows. If there are a total of 30 heads and 88 legs, how many cows are on the farm?	14
14	A local library is organizing a book fair where they sell used books to raise funds for new acquisitions. They sell 120 children's books for $2 each, 75 novels for $3 e	862.5
15	A local farmer is preparing mixed fruit baskets using apples, oranges, and bananas. Each basket must contain 3 apples, 5 oranges, and 2 bananas. The apples are	180
16	Maria has 24 apples and wants to divide them equally among her 6 friends. If each friend also gives 2 apples to their teacher, how many apples will each friend hav	2
17	Lila is planning a garden and wants to plant three types of flowers: daisies, tulips, and roses. She wants twice as many daisies as tulips and three times as many tul	6

Step 3- Gradio version

- Evaluate your prompt on the provided examples.
  - Select number of examples for assessing (from 1 to 30)
  - evaluate
- You can also evaluate your prompt at https://gemini.google.com/app

1.

| Number of examples used for evaluation | 1 |

2.

Evaluate

| Trial ID | 1 | Question ID | 1 |

Result Stats

3.

Result

No result

Save Custom Prompt

Step 3 - No Gradio Version

- Evaluate your prompt on the provided examples.
  - Select number of examples for assessing (from 1 to 30)
  - evaluate

Evaluation

1. Select a `eval_num` which determines the number of examples to evaluate.
2. Execute the cell and wait for a noticeably long time depending on `eval_num` (up to 30 mins).

2.

```
eval_num = 2 # @param {type:"slider", min:1, max:30, step:1}
assert 1 <= eval_num <= 30

# Define the template for displaying results. This includes the question, ratic
ans_template ="Prompt with Question:\n\n{{question}}\n\n--------------------\n\
```

1. eval_num: 2

Output:

```
Start trial 1
Start trial 2
('Connection aborted.', ConnectionResetError(104, 'Connection reset by peer'))

Start trial 3
('Connection aborted.', ConnectionResetError(104, 'Connection reset by peer'))

Final accuracy: Trial 1, accurate_count: 2, total_count: 2, accuracy: 100.0%
Trial 2, accurate_count: 1, total_count: 2, accuracy: 50.0%
Trial 3, accurate_count: 2, total_count: 2, accuracy: 100.0%
Final Accuracy: 100.0%
```

- It will show the accurate question counts, total questions and your accuracy of each trial.

26

Step 4

- After evaluation,check the answers generated by Gemini
  - **Trial ID** : Since your prompt will undergo three trials, you can select the answer from any specific trial.
  - **Question ID**: For the selected trial, you can select a particular question to review the answer generated by Gemini.

Step 4

- After evaluation,check the answers generated by Gemini
  - **Result** : It will show Gemini's response of the question from the specific trial.
  - **Result Stat**: It will show the accurate question counts, total questions and your accuracy of each trial.

Step 4 - No Gradio Version

- After evaluation,check the answers generated by Gemini
  - **Result** : It will show Gemini's response of the question from the specific trial.

```
trial_id = 3 # @param {type:"integer"}
question_id = 1 # @param {type:"integer"}

if trial_id not in [1,2,3]:
```

trial_id: 3

question_id: 1

Output:

```
The evaluation result of question 1 in the No.3 trial is:
Trial 3

Question 1:
----------------------

Prompt with Question:

Think and feel.
Q:An artist is creating a large mosaic using square tiles. The mosaic is designed to be a square itself and to contain exactly 3 times as many blue tiles as red tile

----------------------

Problem-solving Process:

To solve this problem, we can work backwards:
```
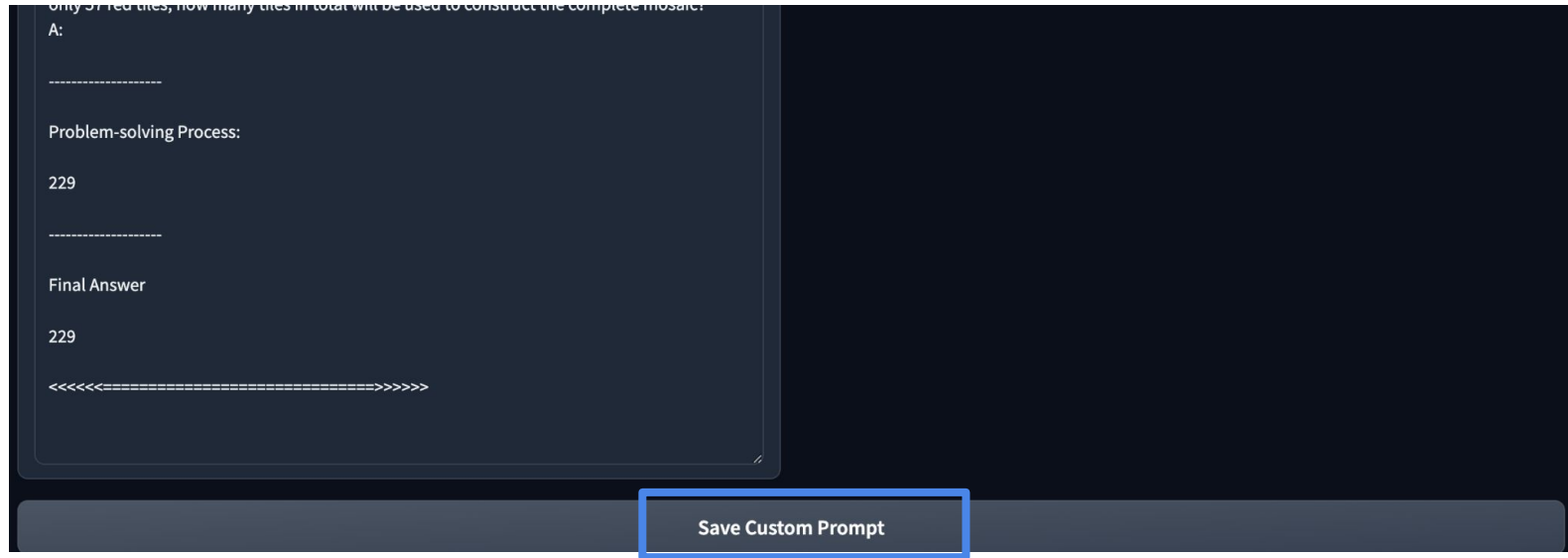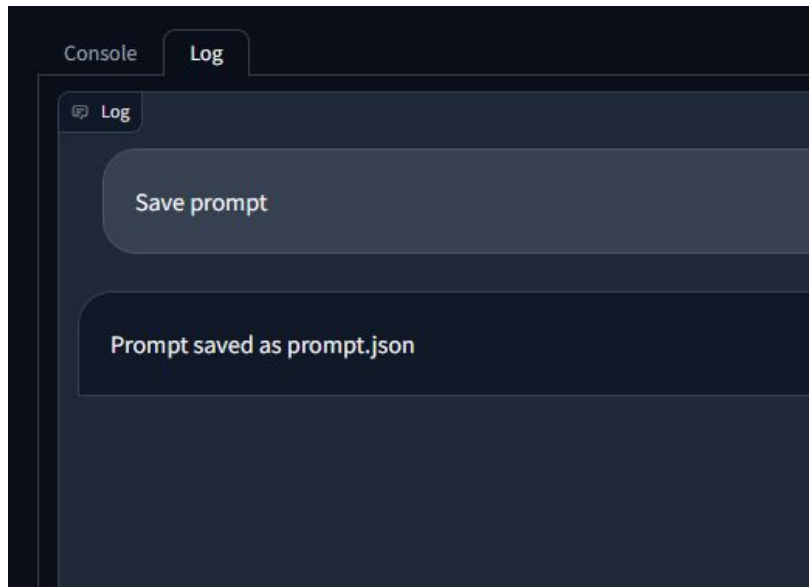
Step 5

- Save and download your prompt
  - After you set your prompt, you can save your prompt by clicking **Save Custom Prompt**.

Step 5: Gradio version

- Save and download your prompt
  - You can check if your prompt has been successfully saved in the log.

Step 5 - No Gradio Version

- Save and download your prompt
  - You can check if your prompt has been successfully saved in the log.

```python
prompt_dict = {
    'prompt': custom_prompt
}

# Dump the prompt into a JSON file.
with open('prompt.json', 'w') as f:
    json.dump(prompt_dict, f)
    print("Prompt saved as prompt.json")
```

```
Prompt saved as prompt.json
```

Step 5

- Save and download your prompt
  - After executing the block, your **prompt.json** file will be download to your computer automatically.
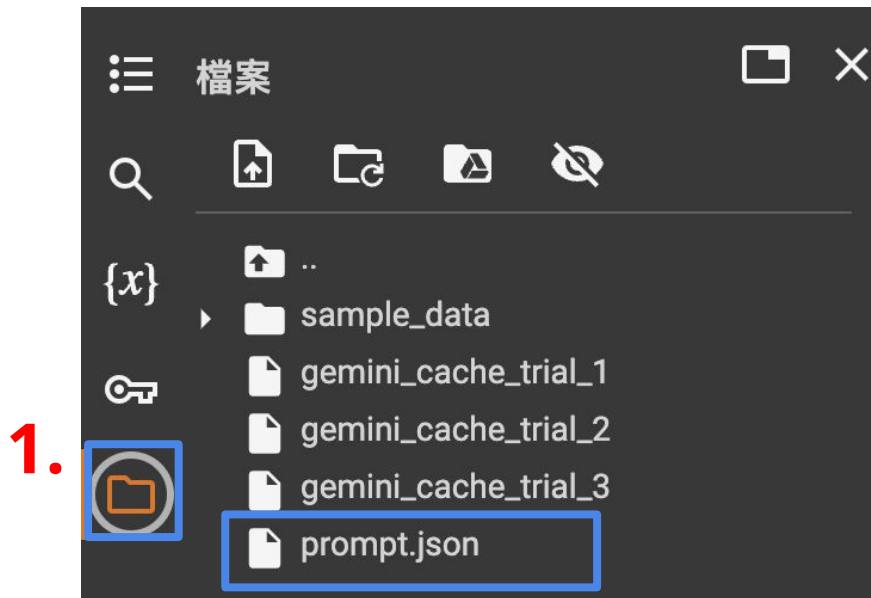 - If you want to use this block to download your prompt, make sure you have stopped the Gradio.
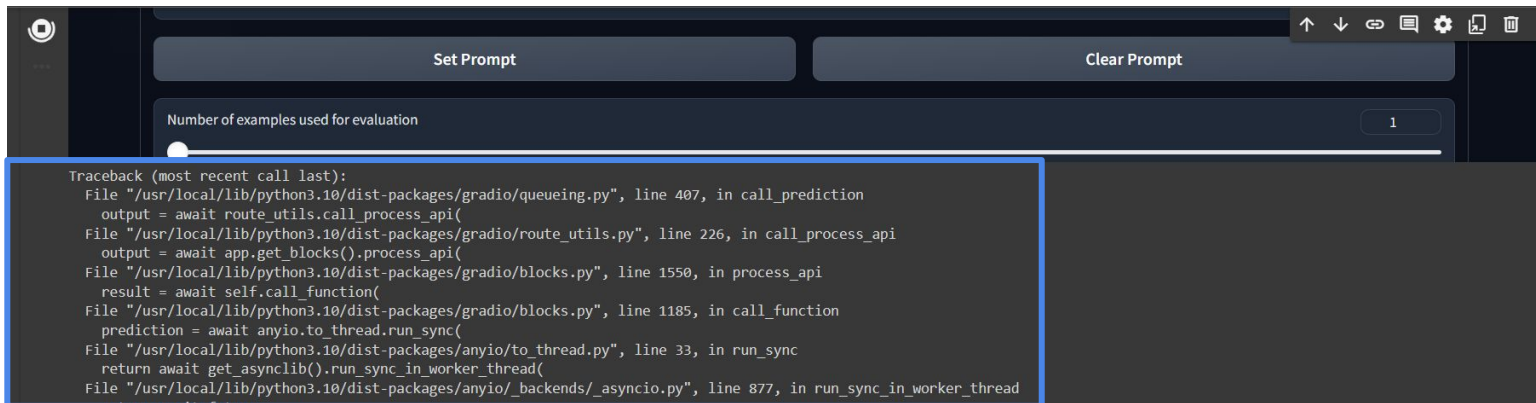
Step 5

- Save and download your prompt
  - If the code block shown in p22 results in error, double check if there exists **prompt.json**.

Issues You Might Encounter
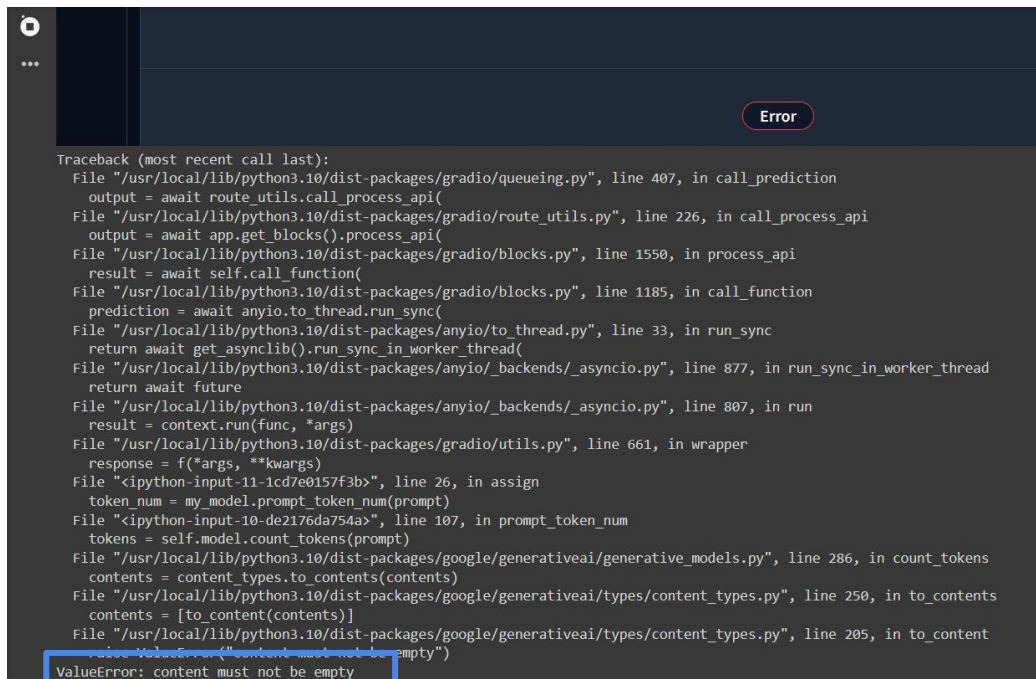
Error from code block (for debugging)

- Just like HW3, you might encounter errors in this code block.
  - The output of the code block would be below the code block, you can try to debug from the traceback
  - Or when you need assistance from TA whether via NTU COOL or mail, please also provide screenshot of this part

ValueError : content must not be empty
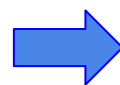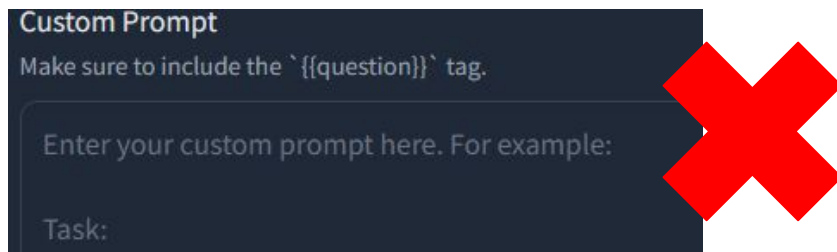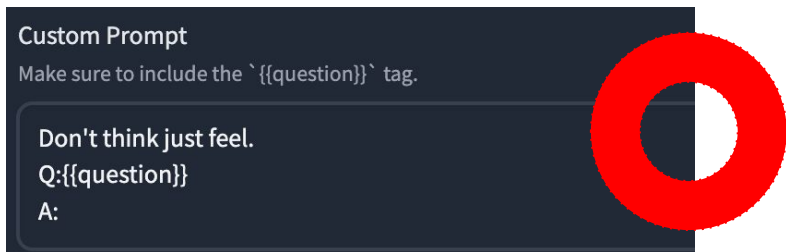
- If you encounter this error


```
Traceback (most recent call last):
  File "/usr/local/lib/python3.10/dist-packages/gradio/queueing.py", line 407, in call_prediction
    output = await route_utils.call_process_api(
  File "/usr/local/lib/python3.10/dist-packages/gradio/route_utils.py", line 226, in call_process_api
    output = await app.get_blocks().process_api(
  File "/usr/local/lib/python3.10/dist-packages/gradio/blocks.py", line 1550, in process_api
    result = await self.call_function(
  File "/usr/local/lib/python3.10/dist-packages/gradio/blocks.py", line 1185, in call_function
    prediction = await anyio.to_thread.run_sync(
  File "/usr/local/lib/python3.10/dist-packages/anyio/to_thread.py", line 33, in run_sync
    return await get_asynclib().run_sync_in_worker_thread(
  File "/usr/local/lib/python3.10/dist-packages/anyio/_backends/_asyncio.py", line 877, in run_sync_in_worker_thread
    return await future
  File "/usr/local/lib/python3.10/dist-packages/anyio/_backends/_asyncio.py", line 807, in run
    result = context.run(func, *args)
  File "/usr/local/lib/python3.10/dist-packages/gradio/utils.py", line 661, in wrapper
    response = f(*args, **kwargs)
  File "<ipython-input-11-1cd7e0157f3b>", line 26, in assign
    token_num = my_model.prompt_token_num(prompt)
  File "<ipython-input-10-de2176da754a>", line 107, in prompt_token_num
    tokens = self.model.count_tokens(prompt)
  File "/usr/local/lib/python3.10/dist-packages/google/generativeai/generative_models.py", line 286, in count_tokens
    contents = content_types.to_contents(contents)
  File "/usr/local/lib/python3.10/dist-packages/google/generativeai/types/content_types.py", line 250, in to_contents
    contents = [to_content(contents)]
  File "/usr/local/lib/python3.10/dist-packages/google/generativeai/types/content_types.py", line 205, in to_content
    raise ValueError("content must not be empty")
ValueError: content must not be empty
```

You pressed set prompt button before filling it

ValueError : content must not be empty

- Solution: Restart the block and fill in the prompt correctly

Custom Prompt
Make sure to include the `{{question}}` tag.

Don't think just feel.
Q:{{question}}
A:

Custom Prompt
Make sure to include the `{{question}}` tag.

Enter your custom prompt here. For example:

Task:

Read timed out

- If you encounter these error:
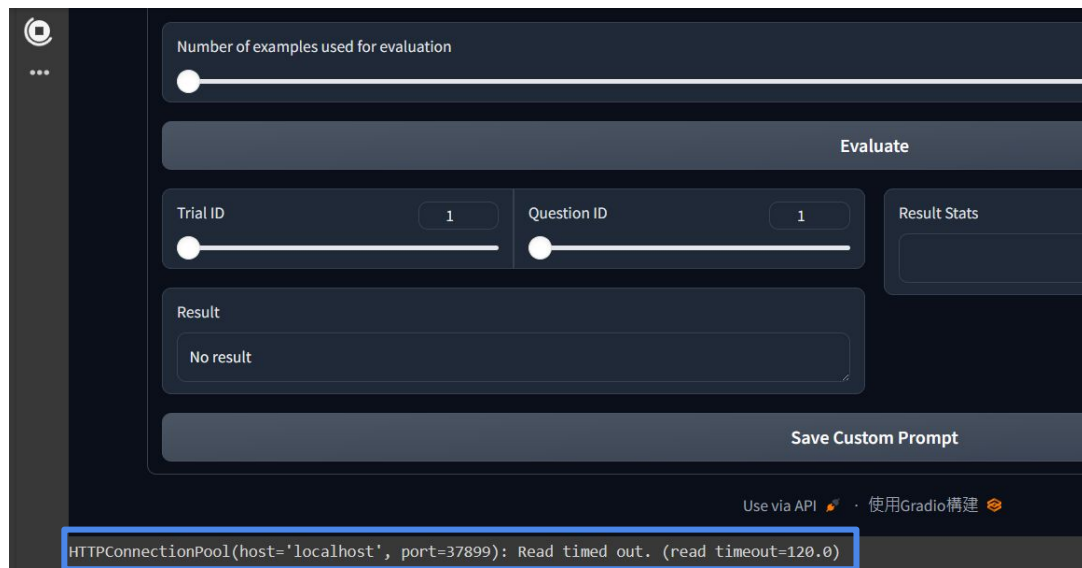
This means that Gemini doesn't provide any response.

Read timed out

- Solution
  - This error usually occurs when Gemini is not properly terminated, such as when pressing the stop key of the Colab block while executing Evaluate. After encountering it once, any subsequent calls to Gemini will trigger a "Read timed out," so it is recommended to directly restart Colab.
    1. Restart the colab.
    2. If it still doesn't work, restart the colab then change the prompt.

None was returned

- If you encounter these error:
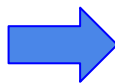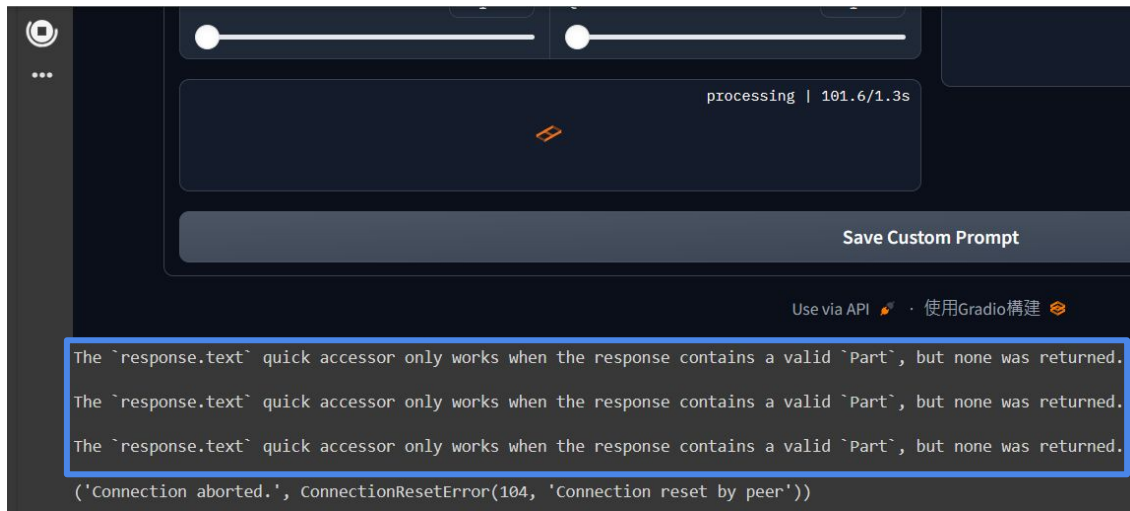
This means that Gemini doesn't provide any response.

None was returned

- Solution
    1. Restart the colab.
    2. If it still doesn't work, change the prompt.

Connection aborted.

- If you encounter this problem, **just ignore it !!**
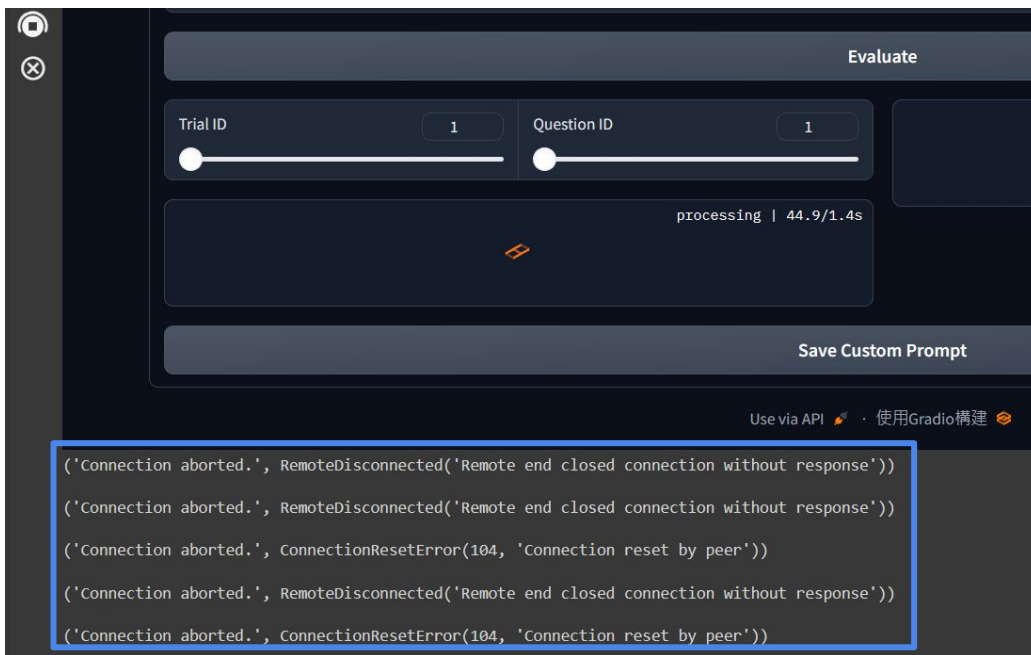
Grading and Submission

Grading

- You need to submit two prompts
- We will evaluate the accuracy of **gemini-pro** on the **public set math questions (with 30 questions)** and a **private set math questions (with 100 questions)** using the prompts you submitted
- Temperature: 1.0

Grading

- For each math question, we will use a two-stage method to extract the answer from Gemini's output
  - **First Stage**: Gemini will be asked to provide a problem-solving process using the **prompt** you provided.
  - **Second Stage**: Gemini will be requested to **extract the final answer** based on the problem-solving process given in the first stage. This step allows the TAs to parse the answer and calculate the accuracy automatically.
    - Use the prompt "**The answer to the original question is (a number only):**" to extract the final answer.
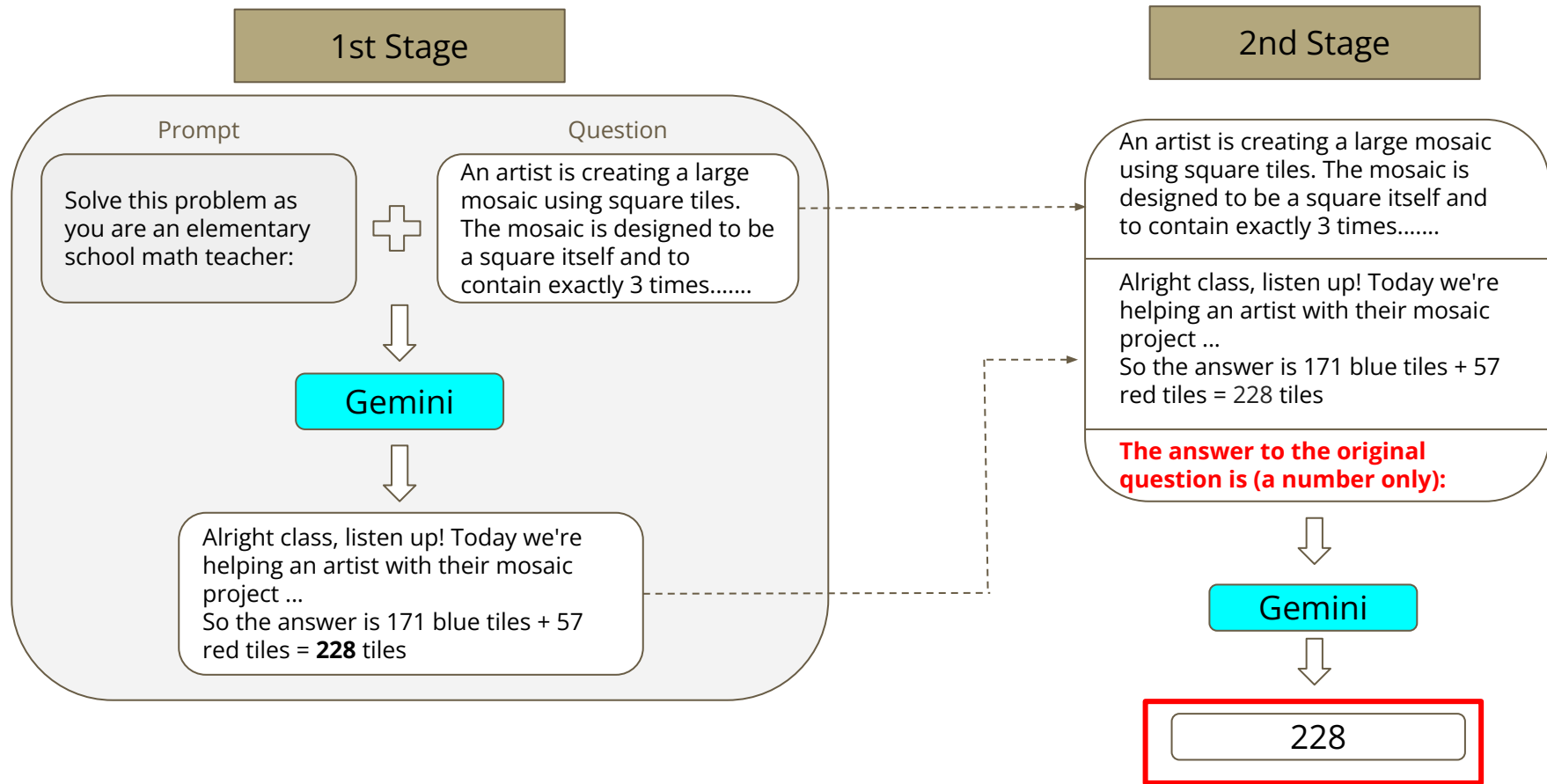
- **We will only consider the extracted answer for each problem, regardless of the accuracy of the model's problem-solving process.**

Grading

- For each prompt, we will sample the answer of each question three times. A question will be considered correct only if it is answered correctly in at least two of the trials.

- The final accuracy for each prompt will be determined by calculating the percentage of questions answered correctly.
- Among the two prompts, we will only consider the one with a higher final accuracy.

Final Score

- Average accuracy on 30 questions in public set
  - >= 0.60
    - +5pts
- Average accuracy on 100 questions in private set
  - >= 0.62
    - +1pts
  - >= 0.67
    - +3pts
  - >= 0.72
    - +5pts

Disclaimers

- The questions in public and private set have **no overlap**

Public Set

An artist is creating a large mosaic using square tiles. …

A question from public set

Gemini

Accuracy on public set

Solve this problem as you are an elementary school math teacher:

Prompt

???????????????????????????????

A question from private set

Gemini

Accuracy on private set
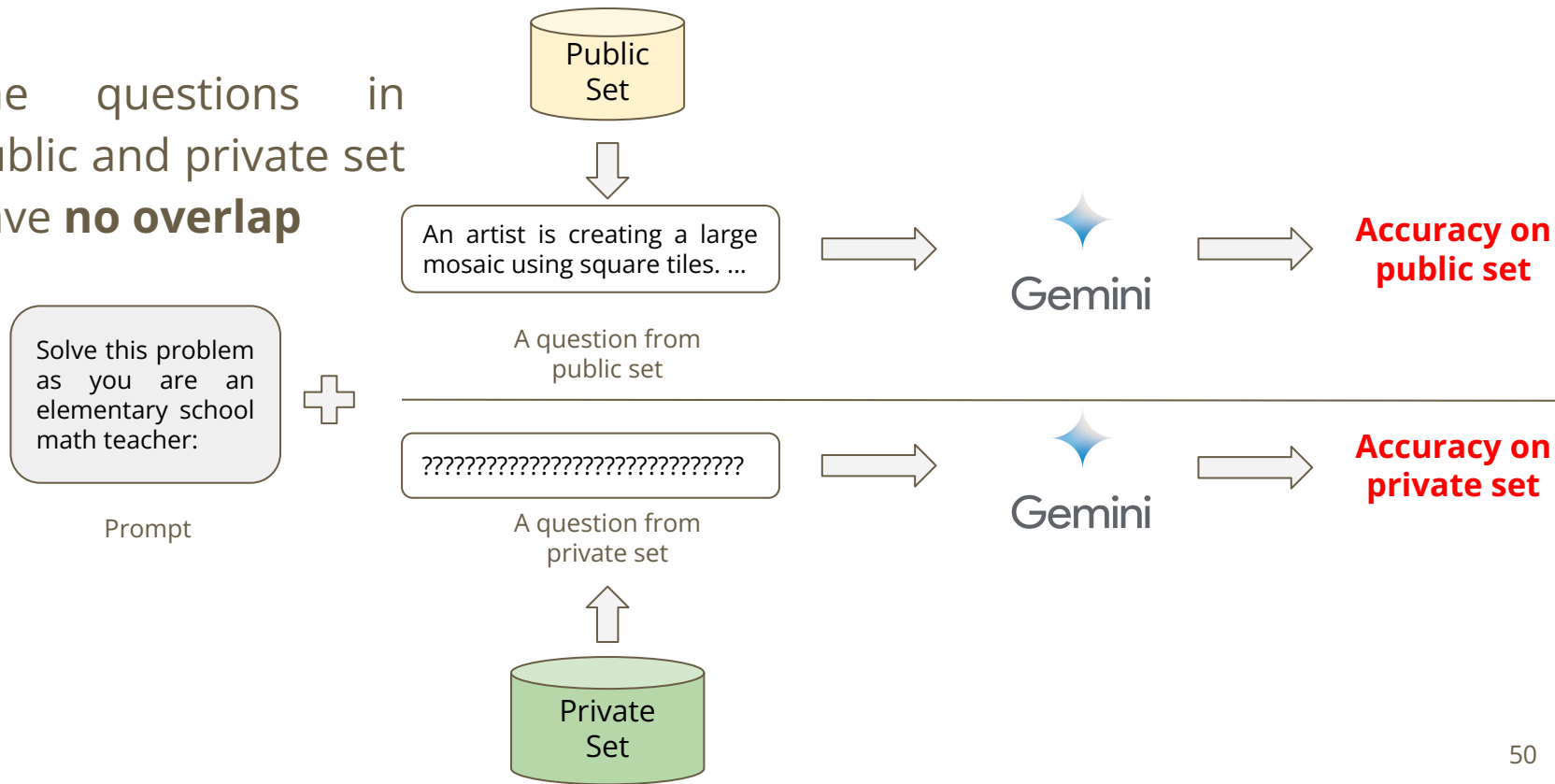
Private Set

Disclaimers

- The accuracy on the public and private set can be **different**

Public Set

An artist is creating a large mosaic using square tiles. …

A question from public set

Accuracy on public set

Solve this problem as you are an elementary school math teacher:

Prompt

?????????????????????????????????

A question from private set

Accuracy on private set
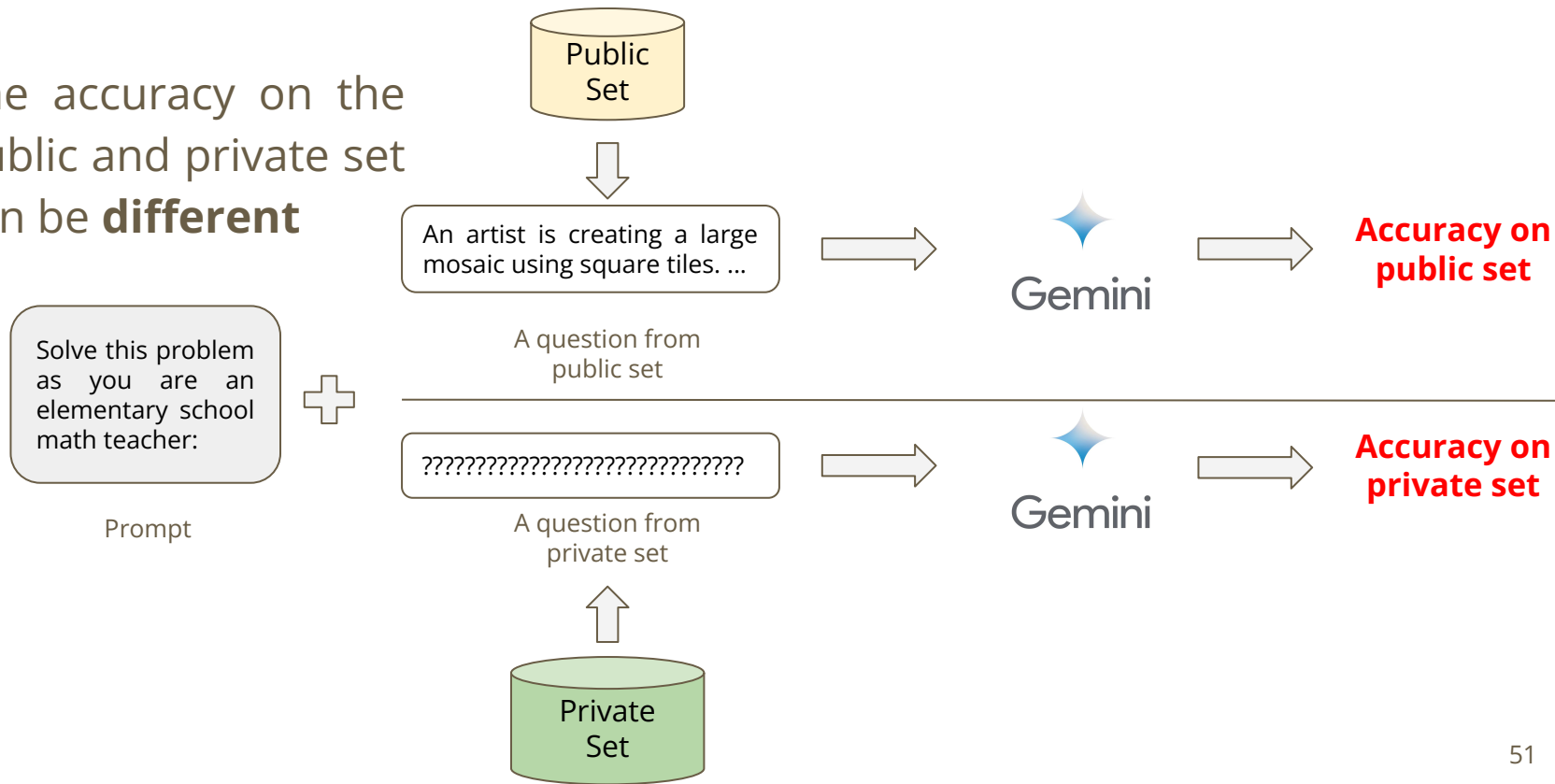
Private Set

Disclaimer (Emphasized again)

- The public set and private set has no overlap. This is a **realistic** setting in machine learning problems

Public Set ≠ Private Set
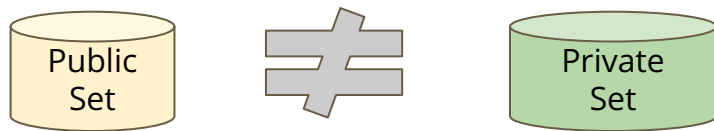
- The accuracy on the public set can be (very) different from the private set

Accuracy on public set ≠ **Accuracy on private set**

- You cannot argue with the TA about your final score if you score high on the public set but score low on the private set.

Submission

- Submit your homework to **NTU Cool**
- We will open two submission areas for HW4, and each HW4 submission area can only receive one prompt.
  - Your submission file should be a **.json** file.
- No late submission is allowed

Important Dates

- Deadline for Submission (NTU Cool)

    2024/04/04 23:59:59 (UTC+8)

- Grading Release Date

    2024/04/25 23:59:59 (UTC+8)

Regulations

Rules

- Plagiarism in any form is prohibited.
- Do NOT share your prompts with others.
- 第一次違反以上規定，**該作業0分，學期總成績再乘以0.9**
- 第二次違反以上規定，**學期成績F**
- Prof. Lee & the TAs preserve the rights to change the rules & grades.

If You Have Any Questions

- NTU Cool [HW4 Forum](#)
  - Recommended if no assignment answers or privacy issues are involved in your questions
  - Prioritize answering questions on the homework forum
- Email: ntu-gen-ai-2024-spring-ta@googlegroups.com
  - Title should start with [GenAI 2024 Spring HW4]
  - Email with the wrong title will be moved to trash automatically
- TA Hours
  - Time:
    - **3/22(16:30 - 17:20)**
    - **3/29(13:20 - 14:10, 16:30 - 17:20)**
  - Location: **綜合大講堂**

Colab

Link

Colab: HW4.ipynb - Colaboratory (google.com)

Hints

Hint

1. Zero Shot Chain of Thought

Q: {{question}}
A: Let's think step by step

[2205.11916] Large Language Models are
Zero-Shot Reasoners (arxiv.org)

Hint

2. In-Context Learning

Q: Shawn has five toys.For Christmas, he got two toys each from his mom and dad.How many toys doe she have now?
A: He has 5 toys. He got 2 from mom, so after that he has 5+2=7 toys. Then he got 2 more from dad, so in total he has 7+2=9 toys. The answer is 9.
Q: {{question}}
A:

[2201.11903] Chain-of-Thought Prompting Elicits Reasoning in Large Language Models (arxiv.org)

Hint

3. Emotional Stimuli

Q: {{question}} This is very important to my career.

[2307.11760] Large Language Models Understand and Can be Enhanced by Emotional Stimuli (arxiv.org)