Using Fine-Tuned LLMs to Grade Homework

Stephan Bohacek and Aishah Aseeri
Department of Electrical and Computer Engineering
bohacek@udel.edu

12th Annual CAE in Cybersecurity Community Symposium

2025 CAE Community Symposium

Outline

- Autograding scenario
- Rubrics
- Fine-tuning
- Demo – Walk Through www.guidance-rubrics.com

2025 CAE Community Symposium

Stephan Bohacek

- Associate Professor of Electrical and Computer Engineering at the University of Delaware
- Teach
  - Introduction to Cybersecurity: ~180 students
  - Introduction to Networking: ~60 students
- Continuous challenges with TAs and graders
  - Slow to grade
  - Incomplete or nonexistent comments
  - Incorrect grading
    - Experienced graders know that easy grading reduces the chances that the student will complain to the instructor

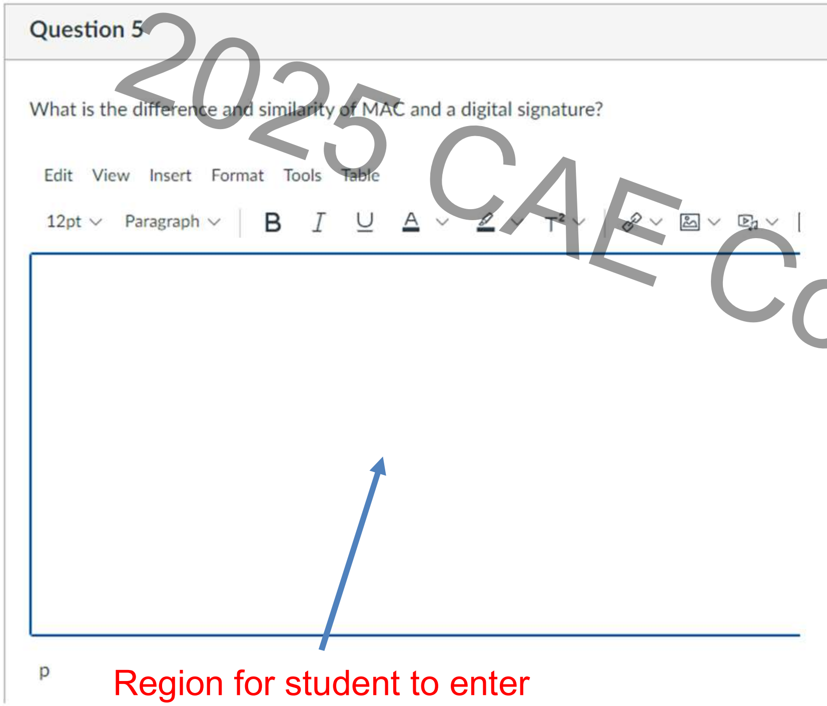
2025 CAE Community Symposium

Automatic Grading and Hint Generation Scenario

- Canvas Quizzes
  - Each student's answer is entered and assessed separately via API
- Submissions are automatically graded, and hints are automatically generated
- Students can resubmit as many times are they want, but grading takes at least 24 hours
  - Ideally, the delay reduces students' ability to "game" the system, but might cause students to lose interest
- If the student believes that the grading is wrong, they include the text "MANUAL REVIEW REQUESTED" and alerts are sent to instructors and graders

2025 CAE Community Symposium

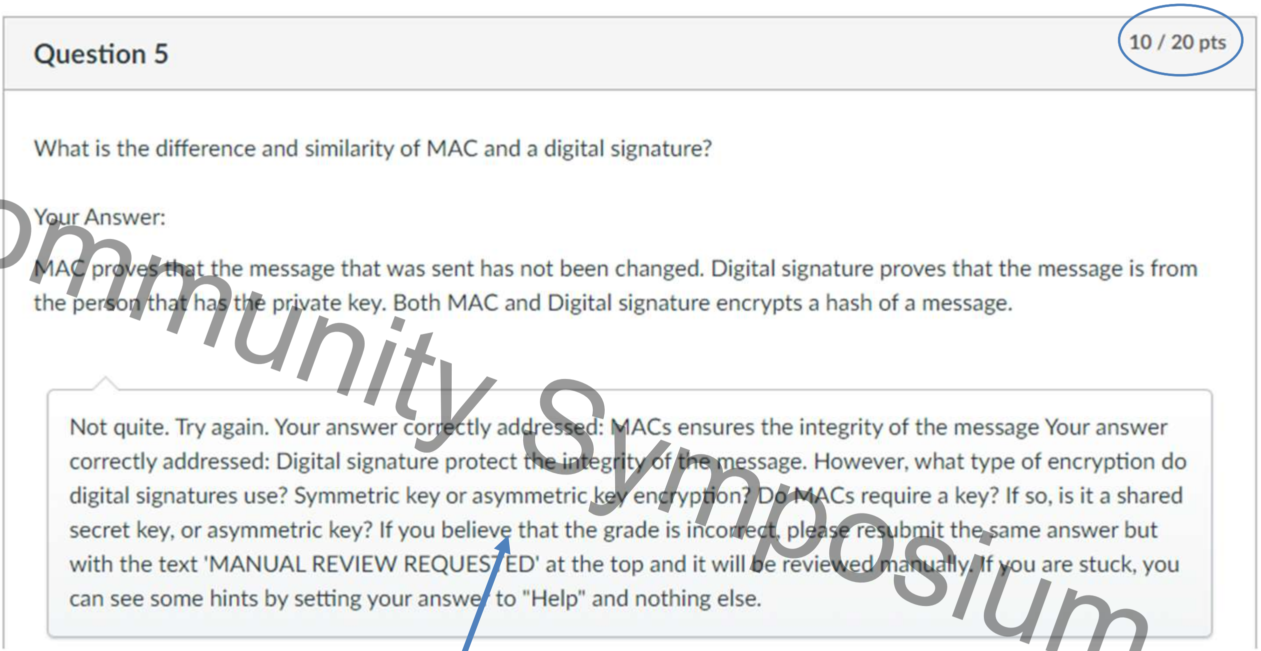
Canvas Quiz



What is the difference and similarity of MAC and a digital signature?

Region for student to enter their answer

Automatically generated grade



What is the difference and similarity of MAC and a digital signature?

Your Answer:

MAC proves that the message that was sent has not been changed. Digital signature proves that the message is from the person that has the private key. Both MAC and Digital signature encrypts a hash of a message.

Not quite. Try again. Your answer correctly addressed: MACs ensures the integrity of the message Your answer correctly addressed: Digital signature protect the integrity of the message. However, what type of encryption do digital signatures use? Symmetric key or asymmetric key encryption? Do MACs require a key? If so, is it a shared secret key, or asymmetric key? If you believe that the grade is incorrect, please resubmit the same answer but with the text 'MANUAL REVIEW REQUESTED' at the top and it will be reviewed manually. If you are stuck, you can see some hints by setting your answer to "Help" and nothing else.

Automatically generated comments

2025 CAE Community Symposium

Guidance / Hints

2025 CAE Community Symposium

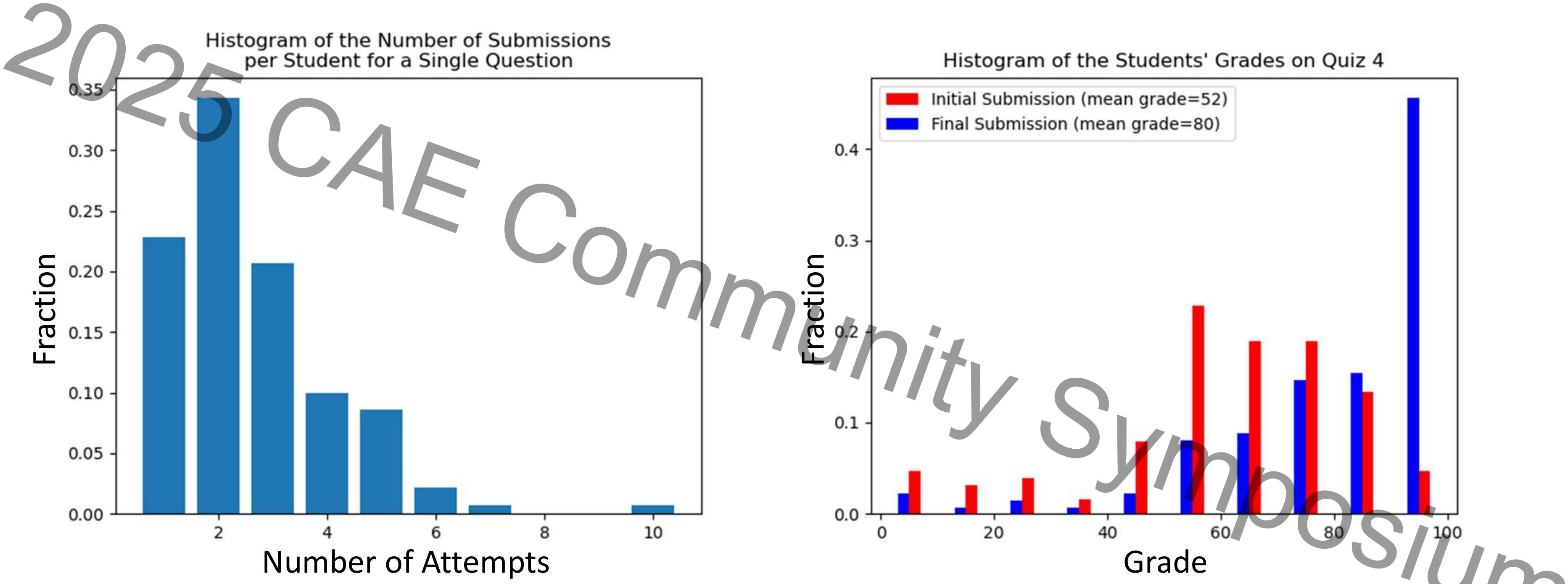
- Rubrics to determine if the submission is correct
- Rubrics to determine if a particular hint is applicable

rubric:

- prompt: “Does the statement mention that George Washington was the United States of America's first president? Statement: <submission>”
- standard hint: Please review chapter 2 of the textbook.
- common errors:
  - prompt: “Does the statement mention that Thomas Jefferson was the United States of America's first president? Statement: <submission>”
  - hint: “While Thomas Jefferson was one of the founders, he was not the first president. Please review Section 2.3 of the textbook”

Targeted hint that includes pointer to course material

Resubmissions



2025 CAE Community Symposium

Around 25,000 questions were graded in a single class with ~60 students

How to get the LLM to grade

- Prompt v0
  - Below is a student's answer to the question "Who was the first president of the United States of America?" Grade the student's answer on a scale from 0 to 10. ….
    - This approach often fails.
    - The LLM is not given enough context into
      - What the student should know given what was presented in class
      - What a student at this level should know
      - The language (words and framework, e.g., risk) that the student should use
    - The LLM might not "know" the correct answer or be able to distinguish between a lay-persons answer and an exact answer
    - The LLM might get confused by simple issues such as definitions
    - The LLM might lack any understanding of the topic

2025 CAE Community Symposium

Rubric Components

- Question: Explain the significance of the Louisiana Purchase in American history and its impact on the expansion of the United States.
- Answer components:
  - Territorial Expansion: increased the nation's land area
  - Territorial Expansion: farming, settlement, and natural resource exploration
  - Economic Growth: agricultural potential helped spur economic growth and development
  - Strategic Importance: …
  - Exploration and Manifest Destiny: …
  - Diplomatic Relations: …
- Prompts:
  - Consider the statement below. Does this statement mention that the Louisiana Purchase greatly expanded the United States' land area? Statement: <SUBMISSION>
  - Consider the statement below. Does this statement mention that the Louisiana Purchase provided more space for farming, settlement, and natural resource exploration? Statement: <SUBMISSION>
  - Consider the statement below. Does this statement mention that the Louisiana Purchase greatly expanded agricultural potential and therefore helped spur economic growth and development? Statement: <SUBMISSION>
  - …

2025 CAE Community Symposium

Rubric Component Details

- Prompts should not be too specific
  - Prompt: Consider the statement below. Does this statement mention that the Louisiana Purchase greatly expanded agricultural potential and therefore helped spur economic growth and development. Statement: <SUBMISSION>
  - Does both economic growth and development need to be discussed? The prompt indicates that they both must be discussed
    - LLMs **can** be very literal
- Use Boolean expressions
  - OR
    - Consider the statement below. Does this statement mention that the Louisiana Purchase greatly expanded agricultural potential and therefore helped spur economic growth. Statement: <SUBMISSION>
    - Consider the statement below. Does this statement mention that the Louisiana Purchase greatly expanded agricultural potential and therefore helped spur economic development. Statement: <SUBMISSION>

2025 CAE Community Symposium

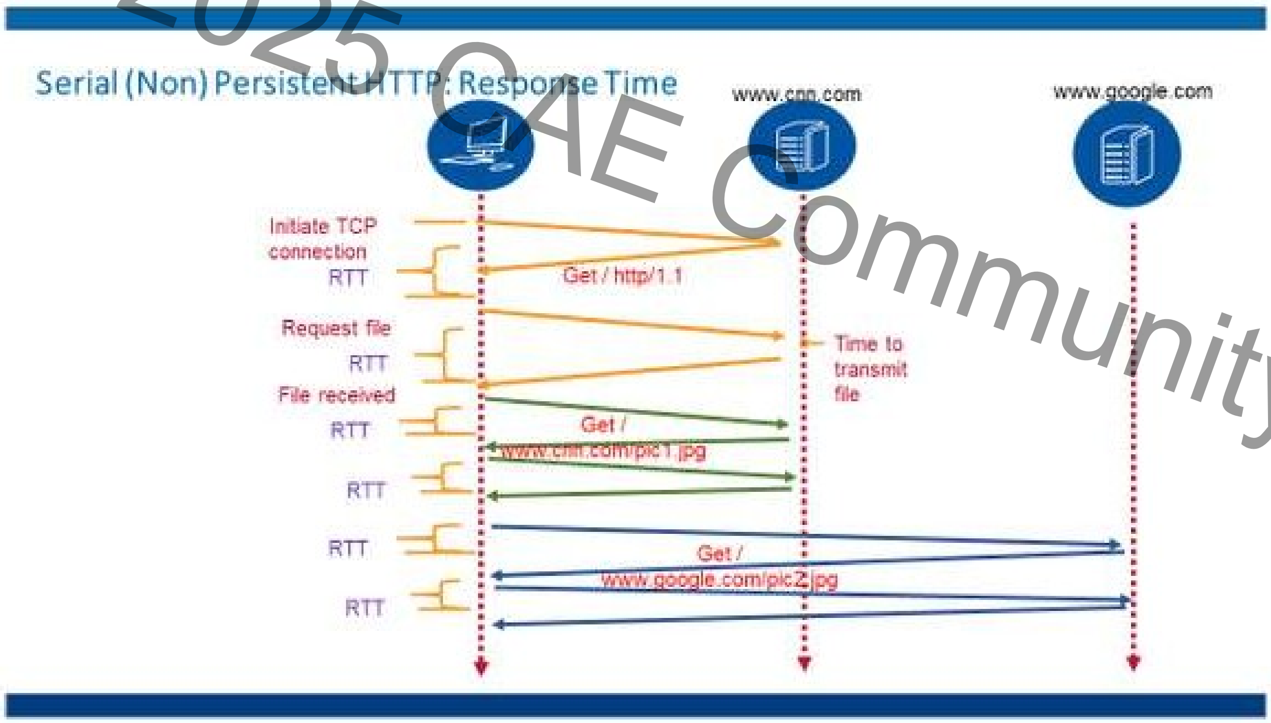
More Specific Prompts Tend to Give Better Results

- Instead of
  - OR
    - Consider the statement below. Does this statement mention that the Louisiana Purchase greatly expanded agricultural potential and therefore helped spur economic growth. Statement: <SUBMISSION>
    - Consider the statement below. Does this statement mention that the Louisiana Purchase greatly expanded agricultural potential and therefore helped spur economic development. Statement: <SUBMISSION>
- Use
  - AND
    - Consider the statement below. Does this statement mention that the Louisiana Purchase greatly expanded agricultural potential. Statement: <SUBMISSION>
    - OR
      - Consider the statement below. Does this statement mention that the Louisiana Purchase helped spur economic growth. Statement: <SUBMISSION>
      - Consider the statement below. Does this statement mention that the Louisiana Purchase helped spur economic development. Statement: <SUBMISSION>

2025 CAE Community Symposium

Prompt Engineering is not enough – e.g., Diagram Grading

2025 CAE Community Symposium

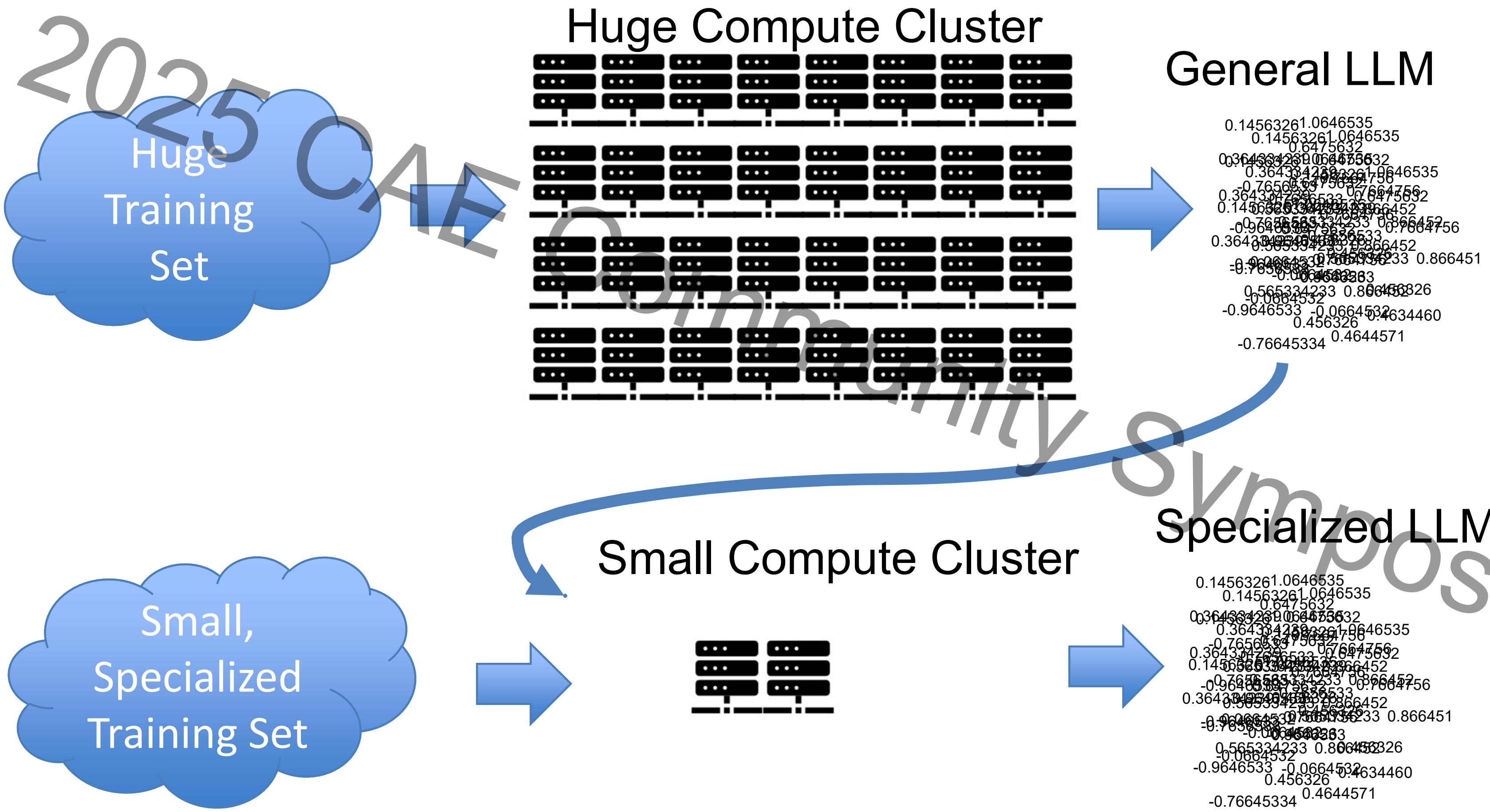


Serial (Non) Persistent HTTP: Response Time

Reply: 'In the image, there are four lines or arrows between the client and the server named www.cnn.com . There are two lines that start at the client and end at www.cnn.com , and two lines that start at www.cnn.com and end at the client.'

Wrong: There are 8 lines. 4 start at the client and 4 start at www.cnn.com
Even with many different prompts, I was unable to get consistent interpretation of the diagram.

Fine-tuning LLMs



Huge Training Set

Huge Compute Cluster

General LLM

Small, Specialized Training Set

Small Compute Cluster

Specialized LLM

2025 CAE Community Symposium

Fine-Tuning LLMs

- Leverages the huge expense spent to build general-purpose LLM
- With a high-quality fine-tuning dataset, the resulting LLM is well suited to answer questions aligned with the dataset's content.
- Most suitable for “yes” “no” types of questions
    - The training set provides inputs and desired outputs.
        - e.g., Input: “Is NYC the largest city in the world” Output: “No”
            - If the LLM answers “yes”, then the weights are adjusted
        - e.g., Input: “Explain the taste of an orange.” Output: “Oranges taste tangy”
            - If the LLM answers “Oranges have a tangy flavor”, then the weights are adjusted, even though the answer is correct.
    - Consequently, fine-tuning is poorly suited for chain-of-thought prompts.
        - Without fine-tuning, chain-of-thought often results in the best performance.

2025 CAE Community Symposium

Step to Fine-Tune for Autograding

www.guidance-rubrics.com provide a user interface to perform these steps

2025 CAE Community Symposium

- Label submissions
- Select rubric elements to include in training set
  - Elements should be similar, e.g., graphical analysis of a particular type of plot
  - Fine-tuning for a single element is acceptable
- Build training set from one or more classes/sections
- Training LLM
- Make an updated rubric that uses the new LLM
- Compare results

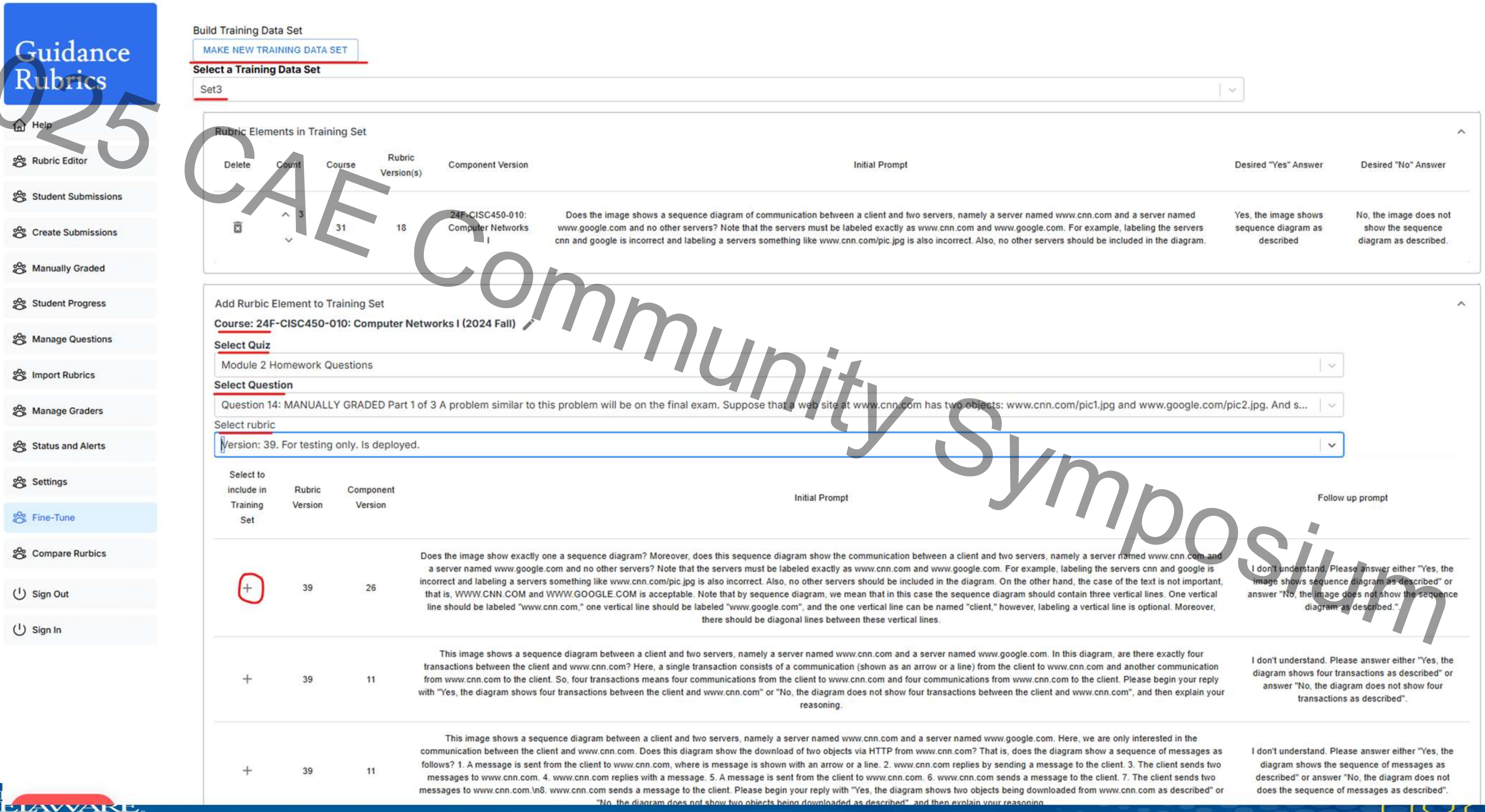
Work Flow

1. Make Rubric
2. Run grader
3. Label submissions
4. Explore errors
5. Fix errors
    1. Correct labels
    2. Correct prompts
    3. Fine-tune LLM
6. Go to 4

2025 CAE Community Symposium

Building Training Set (1)

2025 CAE Community Symposium



Build Training Data Set

Select a Training Data Set
Set3

Rubric Elements in Training Set

Delete	Count	Course	Rubric Version(s)	Component Version	Initial Prompt	Desired "Yes" Answer	Desired "No" Answer
	3	31	18	24F-CISC450-010: Computer Networks I	Does the image shows a sequence diagram of communication between a client and two servers, namely a server named www.cnn.com and a server named www.google.com and no other servers? Note that the servers must be labeled exactly as www.cnn.com and www.google.com. For example, labeling the servers cnn and google is incorrect and labeling a servers something like www.cnn.com/pic.jpg is also incorrect. Also, no other servers should be included in the diagram.	Yes, the image shows sequence diagram as described	No, the image does not show the sequence diagram as described.

Add Rurbic Element to Training Set

Course: 24F-CISC450-010: Computer Networks I (2024 Fall)

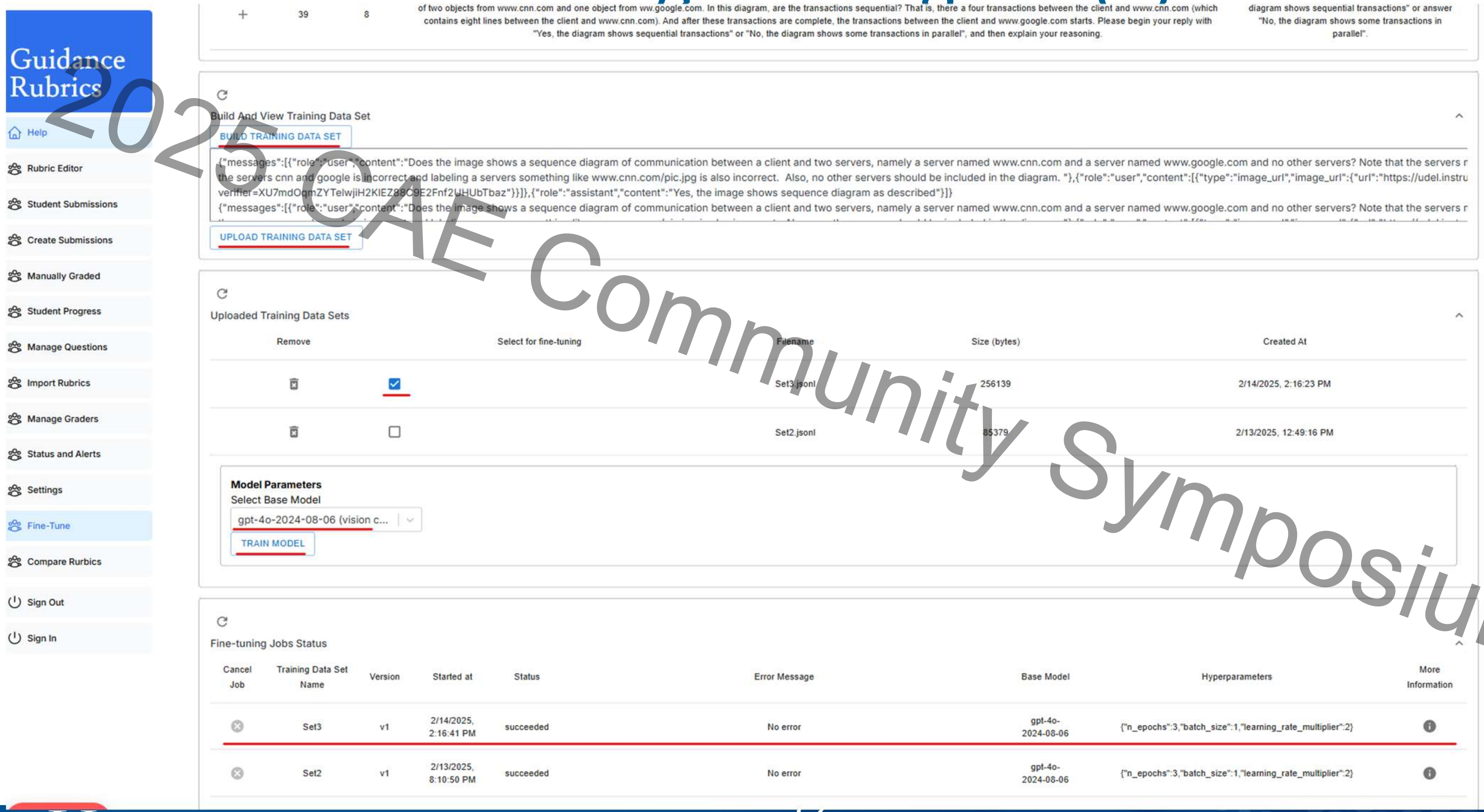
Select Quiz
Module 2 Homework Questions

Select Question
Question 14: MANUALLY GRADED Part 1 of 3 A problem similar to this problem will be on the final exam. Suppose that a web site at www.cnn.com has two objects: www.cnn.com/pic1.jpg and www.google.com/pic2.jpg. And s...

Select rubric
Version: 39. For testing only. Is deployed.

Select to include in Training Set	Rubric Version	Component Version	Initial Prompt	Follow up prompt
+	39	26	Does the image show exactly one a sequence diagram? Moreover, does this sequence diagram show the communication between a client and two servers, namely a server named www.cnn.com and a server named www.google.com and no other servers? Note that the servers must be labeled exactly as www.cnn.com and www.google.com. For example, labeling the servers cnn and google is incorrect and labeling a servers something like www.cnn.com/pic.jpg is also incorrect. Also, no other servers should be included in the diagram. On the other hand, the case of the text is not important, that is, WWW.CNN.COM and WWW.GOOGLE.COM is acceptable. Note that by sequence diagram, we mean that in this case the sequence diagram should contain three vertical lines. One vertical line should be labeled "www.cnn.com," one vertical line should be labeled "www.google.com", and the one vertical line can be named "client," however, labeling a vertical line is optional. Moreover, there should be diagonal lines between these vertical lines.	I don't understand. Please answer either "Yes, the image shows sequence diagram as described" or answer "No, the image does not show the sequence diagram as described.".
+	39	11	This image shows a sequence diagram between a client and two servers, namely a server named www.cnn.com and a server named www.google.com. In this diagram, are there exactly four transactions between the client and www.cnn.com? Here, a single transaction consists of a communication (shown as an arrow or a line) from the client to www.cnn.com and another communication from www.cnn.com to the client. So, four transactions means four communications from the client to www.cnn.com and four communications from www.cnn.com to the client. Please begin your reply with "Yes, the diagram shows four transactions between the client and www.cnn.com" or "No, the diagram does not show four transactions between the client and www.cnn.com", and then explain your reasoning.	I don't understand. Please answer either "Yes, the diagram shows four transactions as described" or answer "No, the diagram does not show four transactions as described".
+	39	11	This image shows a sequence diagram between a client and two servers, namely a server named www.cnn.com and a server named www.google.com. Here, we are only interested in the communication between the client and www.cnn.com. Does this diagram show the download of two objects via HTTP from www.cnn.com? That is, does the diagram show a sequence of messages as follows? 1. A message is sent from the client to www.cnn.com, where is message is shown with an arrow or a line. 2. www.cnn.com replies by sending a message to the client. 3. The client sends two messages to www.cnn.com. 4. www.cnn.com replies with a message. 5. A message is sent from the client to www.cnn.com. 6. www.cnn.com sends a message to the client. 7. The client sends two messages to www.cnn.com.\n8. www.cnn.com sends a message to the client. Please begin your reply with "Yes, the diagram shows two objects being downloaded from www.cnn.com as described" or "No, the diagram does not show two objects being downloaded as described", and then explain your reasoning.	I don't understand. Please answer either "Yes, the diagram shows the sequence of messages as described" or answer "No, the diagram does not does the sequence of messages as described".

Building Training Set (2)



+	39	8	of two objects from www.cnn.com and one object from ww.google.com. In this diagram, are the transactions sequential? That is, there a four transactions between the client and www.cnn.com (which contains eight lines between the client and www.cnn.com). And after these transactions are complete, the transactions between the client and www.google.com starts. Please begin your reply with "Yes, the diagram shows sequential transactions" or "No, the diagram shows some transactions in parallel", and then explain your reasoning.	diagram shows sequential transactions" or answer "No, the diagram shows some transactions in parallel".

Build And View Training Data Set

```
{"messages":[{"role":"user","content":"Does the image shows a sequence diagram of communication between a client and two servers, namely a server named www.cnn.com and a server named www.google.com and no other servers? Note that the servers r
the servers cnn and google is incorrect and labeling a servers something like www.cnn.com/pic.jpg is also incorrect.  Also, no other servers should be included in the diagram. "},{"role":"user","content":[{"type":"image_url","image_url":{"url":"https://udel.instru
verifier=XU7mdOqmZYTelwjiH2KIEZ88C9E2Fnf2UHUbTbaz"}}]},{"role":"assistant","content":"Yes, the image shows sequence diagram as described"}]}
{"messages":[{"role":"user","content":"Does the image shows a sequence diagram of communication between a client and two servers, namely a server named www.cnn.com and a server named www.google.com and no other servers? Note that the servers r
```

Uploaded Training Data Sets

Remove	Select for fine-tuning	Filename	Size (bytes)	Created At
	☑	Set3.jsonl	256139	2/14/2025, 2:16:23 PM
	☐	Set2.jsonl	85379	2/13/2025, 12:49:16 PM

Model Parameters

Select Base Model

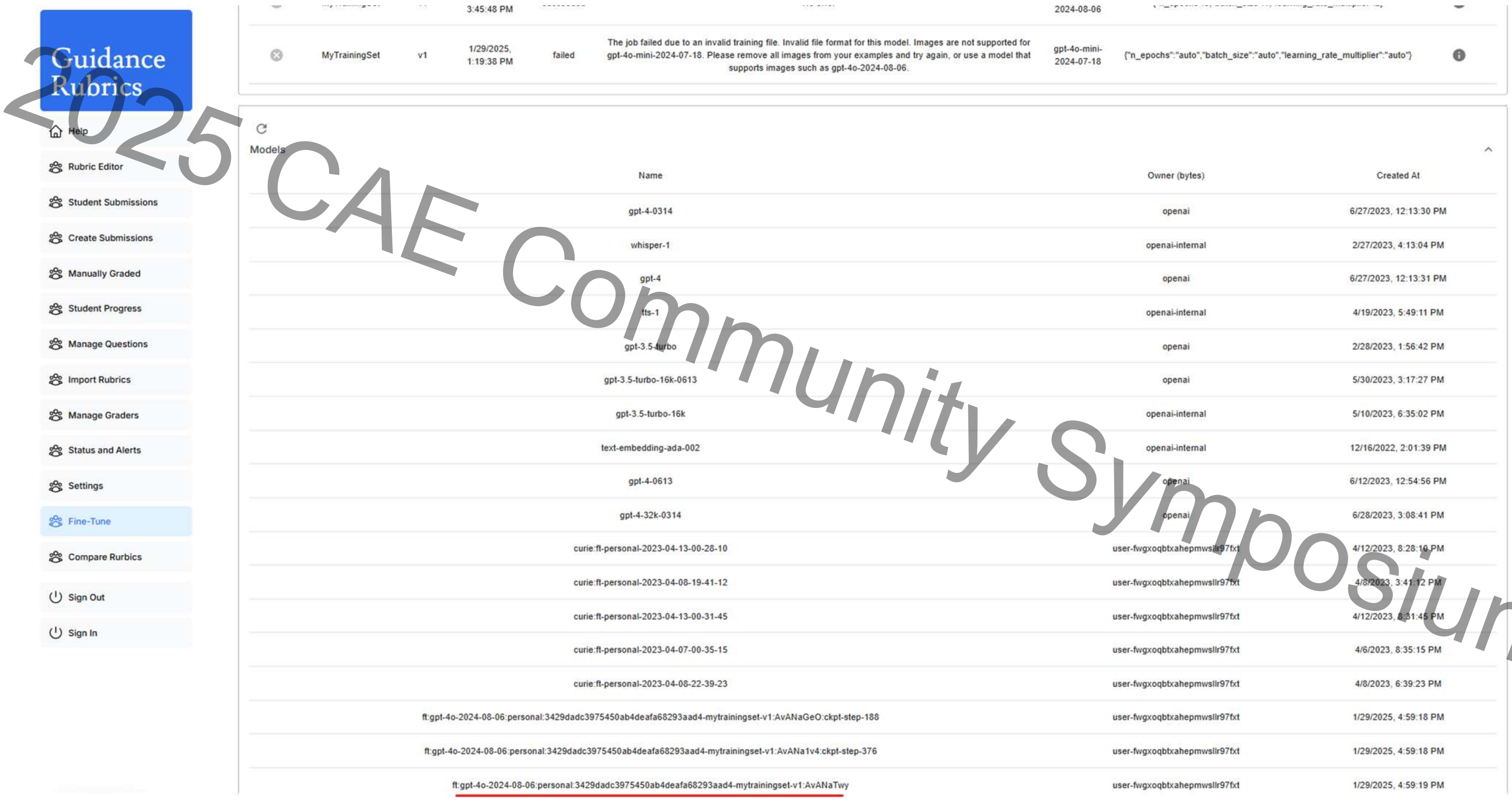
gpt-4o-2024-08-06 (vision c...

Fine-tuning Jobs Status

Cancel Job	Training Data Set Name	Version	Started at	Status	Error Message	Base Model	Hyperparameters	More Information
	Set3	v1	2/14/2025, 2:16:41 PM	succeeded	No error	gpt-4o-2024-08-06	{"n_epochs":3,"batch_size":1,"learning_rate_multiplier":2}	
	Set2	v1	2/13/2025, 8:10:50 PM	succeeded	No error	gpt-4o-2024-08-06	{"n_epochs":3,"batch_size":1,"learning_rate_multiplier":2}	

2025 CAE Community Symposium

Building Training Set (3)



		3:45:48 PM			2024-08-06		
MyTrainingSet	v1	1/29/2025, 1:19:38 PM	failed	The job failed due to an invalid training file. Invalid file format for this model. Images are not supported for gpt-4o-mini-2024-07-18. Please remove all images from your examples and try again, or use a model that supports images such as gpt-4o-2024-08-06.	gpt-4o-mini-2024-07-18	{"n_epochs":"auto","batch_size":"auto","learning_rate_multiplier":"auto"}	

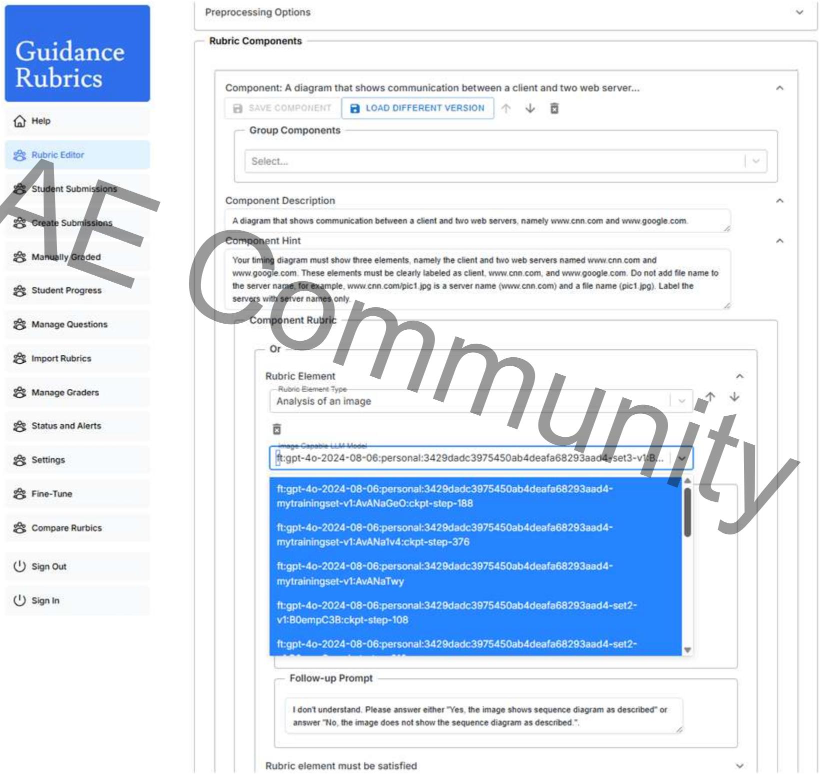
Models

Name	Owner (bytes)	Created At
gpt-4-0314	openai	6/27/2023, 12:13:30 PM
whisper-1	openai-internal	2/27/2023, 4:13:04 PM
gpt-4	openai	6/27/2023, 12:13:31 PM
tts-1	openai-internal	4/19/2023, 5:49:11 PM
gpt-3.5-turbo	openai	2/28/2023, 1:56:42 PM
gpt-3.5-turbo-16k-0613	openai	5/30/2023, 3:17:27 PM
gpt-3.5-turbo-16k	openai-internal	5/10/2023, 6:35:02 PM
text-embedding-ada-002	openai-internal	12/16/2022, 2:01:39 PM
gpt-4-0613	openai	6/12/2023, 12:54:56 PM
gpt-4-32k-0314	openai	6/28/2023, 3:08:41 PM
curie:ft-personal-2023-04-13-00-28-10	user-fwgxoqbtxahepmwsllr97fxt	4/12/2023, 8:28:10 PM
curie:ft-personal-2023-04-08-19-41-12	user-fwgxoqbtxahepmwsllr97fxt	4/8/2023, 3:41:12 PM
curie:ft-personal-2023-04-13-00-31-45	user-fwgxoqbtxahepmwsllr97fxt	4/12/2023, 8:31:45 PM
curie:ft-personal-2023-04-07-00-35-15	user-fwgxoqbtxahepmwsllr97fxt	4/6/2023, 8:35:15 PM
curie:ft-personal-2023-04-08-22-39-23	user-fwgxoqbtxahepmwsllr97fxt	4/8/2023, 6:39:23 PM
ft:gpt-4o-2024-08-06:personal:3429dadc3975450ab4deafa68293aad4-mytrainingset-v1:AvANaGeO:ckpt-step-188	user-fwgxoqbtxahepmwsllr97fxt	1/29/2025, 4:59:18 PM
ft:gpt-4o-2024-08-06:personal:3429dadc3975450ab4deafa68293aad4-mytrainingset-v1:AvANa1v4:ckpt-step-376	user-fwgxoqbtxahepmwsllr97fxt	1/29/2025, 4:59:18 PM
ft:gpt-4o-2024-08-06:personal:3429dadc3975450ab4deafa68293aad4-mytrainingset-v1:AvANaTwy	user-fwgxoqbtxahepmwsllr97fxt	1/29/2025, 4:59:19 PM

2025 CAE Community Symposium

Update Rubric to Use Fine-tuned Model

2025 CAE Community Symposium



Component: A diagram that shows communication between a client and two web server...

Component Description

A diagram that shows communication between a client and two web servers, namely www.cnn.com and www.google.com.

Component Hint

Your timing diagram must show three elements, namely the client and two web servers named www.cnn.com and www.google.com. These elements must be clearly labeled as client, www.cnn.com, and www.google.com. Do not add file name to the server name, for example, www.cnn.com/pic1.jpg is a server name (www.cnn.com) and a file name (pic1.jpg). Label the servers with server names only.

Rubric Element Type: Analysis of an image

Image Capable LLM Model: ft:gpt-4o-2024-08-06:personal:3429dadc3975450ab4deafa68293aad4-set3-v1:B...

- ft:gpt-4o-2024-08-06:personal:3429dadc3975450ab4deafa68293aad4-mytrainingset-v1:AvANaGeO:ckpt-step-188
- ft:gpt-4o-2024-08-06:personal:3429dadc3975450ab4deafa68293aad4-mytrainingset-v1:AvANa1v4:ckpt-step-376
- ft:gpt-4o-2024-08-06:personal:3429dadc3975450ab4deafa68293aad4-mytrainingset-v1:AvANaTwy
- ft:gpt-4o-2024-08-06:personal:3429dadc3975450ab4deafa68293aad4-set2-v1:B0empC3B:ckpt-step-108
- ft:gpt-4o-2024-08-06:personal:3429dadc3975450ab4deafa68293aad4-set2-

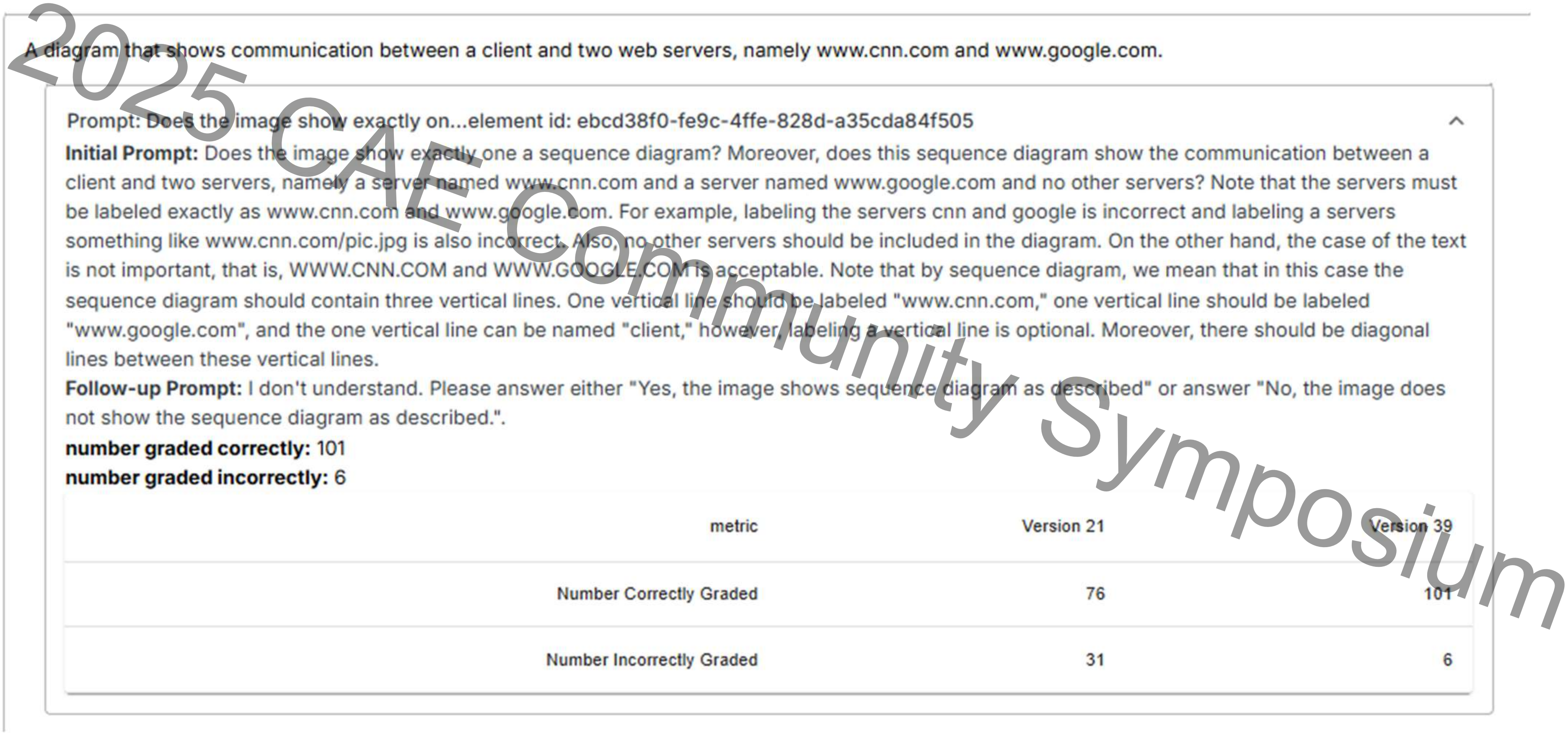
Follow-up Prompt

I don't understand. Please answer either "Yes, the image shows sequence diagram as described" or answer "No, the image does not show the sequence diagram as described.".

Next, save and deploy new rubric

Compare Rubrics

2025 CAE Community Symposium



A diagram that shows communication between a client and two web servers, namely www.cnn.com and www.google.com.

Prompt: Does the image show exactly on...element id: ebcd38f0-fe9c-4ffe-828d-a35cda84f505

Initial Prompt: Does the image show exactly one a sequence diagram? Moreover, does this sequence diagram show the communication between a client and two servers, namely a server named www.cnn.com and a server named www.google.com and no other servers? Note that the servers must be labeled exactly as www.cnn.com and www.google.com. For example, labeling the servers cnn and google is incorrect and labeling a servers something like www.cnn.com/pic.jpg is also incorrect. Also, no other servers should be included in the diagram. On the other hand, the case of the text is not important, that is, WWW.CNN.COM and WWW.GOOGLE.COM is acceptable. Note that by sequence diagram, we mean that in this case the sequence diagram should contain three vertical lines. One vertical line should be labeled "www.cnn.com," one vertical line should be labeled "www.google.com", and the one vertical line can be named "client," however, labeling a vertical line is optional. Moreover, there should be diagonal lines between these vertical lines.

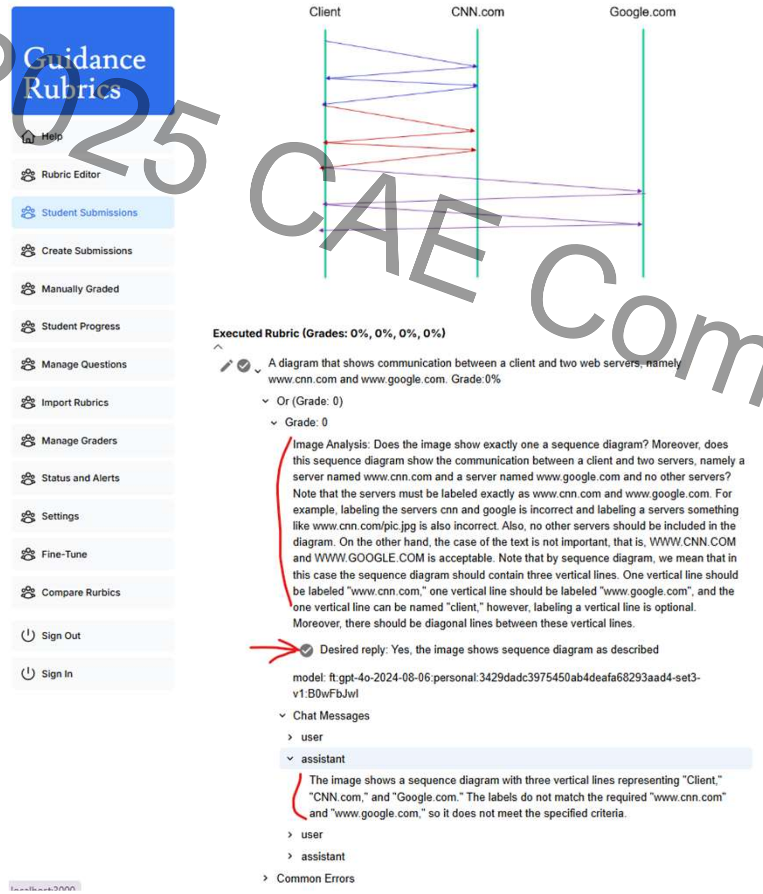
Follow-up Prompt: I don't understand. Please answer either "Yes, the image shows sequence diagram as described" or answer "No, the image does not show the sequence diagram as described.".

number graded correctly: 101

number graded incorrectly: 6

metric	Version 21	Version 39
Number Correctly Graded	76	101
Number Incorrectly Graded	31	6

Examine Incorrectly Graded Submissions



Executed Rubric (Grades: 0%, 0%, 0%, 0%)

A diagram that shows communication between a client and two web servers, namely www.cnn.com and www.google.com. Grade:0%

- Or (Grade: 0)
  - Grade: 0

Image Analysis: Does the image show exactly one a sequence diagram? Moreover, does this sequence diagram show the communication between a client and two servers, namely a server named www.cnn.com and a server named www.google.com and no other servers? Note that the servers must be labeled exactly as www.cnn.com and www.google.com. For example, labeling the servers cnn and google is incorrect and labeling a servers something like www.cnn.com/pic.jpg is also incorrect. Also, no other servers should be included in the diagram. On the other hand, the case of the text is not important, that is, WWW.CNN.COM and WWW.GOOGLE.COM is acceptable. Note that by sequence diagram, we mean that in this case the sequence diagram should contain three vertical lines. One vertical line should be labeled "www.cnn.com," one vertical line should be labeled "www.google.com", and the one vertical line can be named "client," however, labeling a vertical line is optional. Moreover, there should be diagonal lines between these vertical lines.

Desired reply: Yes, the image shows sequence diagram as described

model: ft:gpt-4o-2024-08-06:personal:3429dadc3975450ab4deafa68293aad4-set3-v1:B0wFbJwl

- Chat Messages
  - user
  - assistant

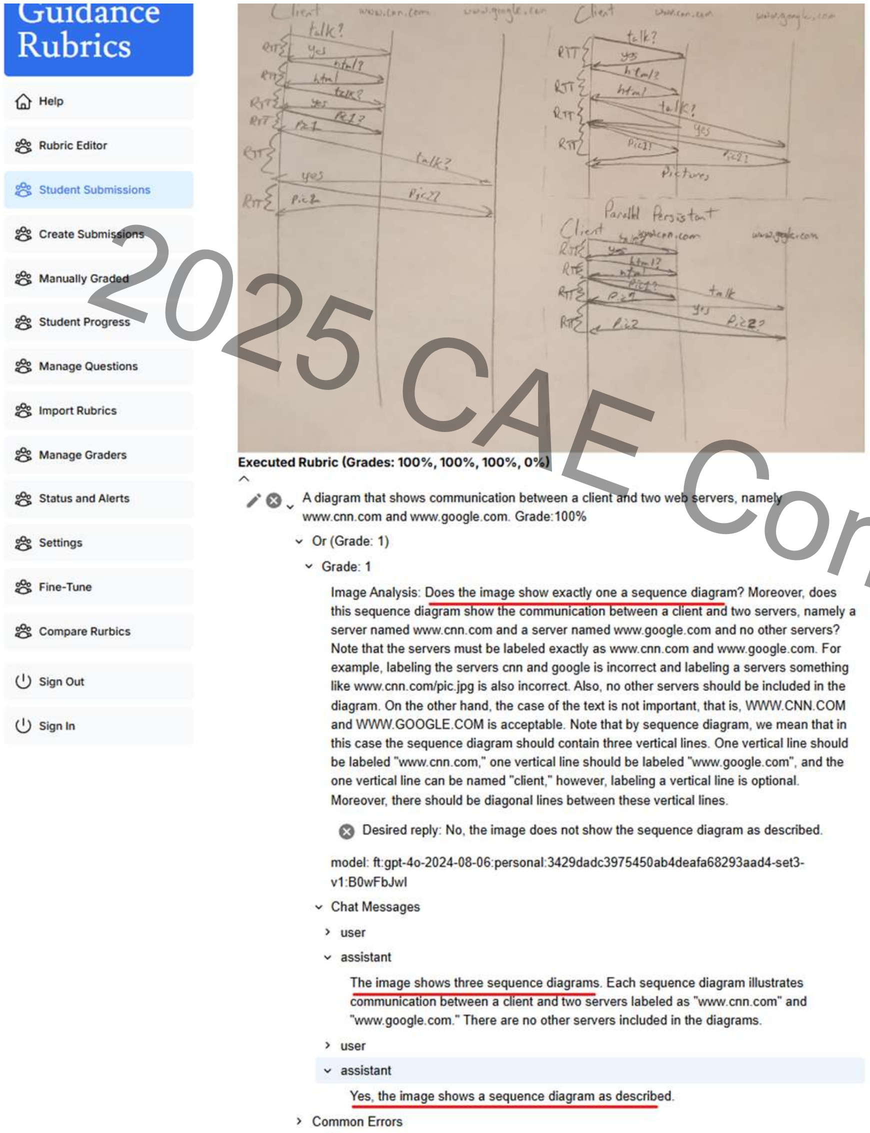
The image shows a sequence diagram with three vertical lines representing "Client," "CNN.com," and "Google.com." The labels do not match the required "www.cnn.com" and "www.google.com," so it does not meet the specified criteria.

  - user
  - assistant
- Common Errors

This was mislabeled.
After fixing labeling, repeating the fine-tuning process might give better results.

2025 CAE Community Symposium

Examine Incorrectly Graded Submissions



Executed Rubric (Grades: 100%, 100%, 100%, 0%)

A diagram that shows communication between a client and two web servers, namely www.cnn.com and www.google.com. Grade:100%

Or (Grade: 1)

Grade: 1

Image Analysis: Does the image show exactly one a sequence diagram? Moreover, does this sequence diagram show the communication between a client and two servers, namely a server named www.cnn.com and a server named www.google.com and no other servers? Note that the servers must be labeled exactly as www.cnn.com and www.google.com. For example, labeling the servers cnn and google is incorrect and labeling a servers something like www.cnn.com/pic.jpg is also incorrect. Also, no other servers should be included in the diagram. On the other hand, the case of the text is not important, that is, WWW.CNN.COM and WWW.GOOGLE.COM is acceptable. Note that by sequence diagram, we mean that in this case the sequence diagram should contain three vertical lines. One vertical line should be labeled "www.cnn.com," one vertical line should be labeled "www.google.com", and the one vertical line can be named "client," however, labeling a vertical line is optional. Moreover, there should be diagonal lines between these vertical lines.

☒ Desired reply: No, the image does not show the sequence diagram as described.

model: ft:gpt-4o-2024-08-06:personal:3429dadc3975450ab4deafa68293aad4-set3-v1:B0wFbJwl

Chat Messages

user

assistant

The image shows three sequence diagrams. Each sequence diagram illustrates communication between a client and two servers labeled as "www.cnn.com" and "www.google.com." There are no other servers included in the diagrams.

user

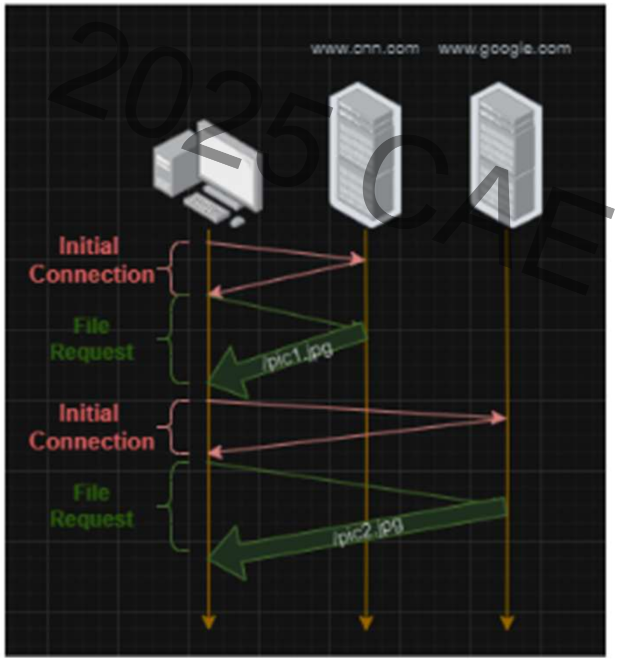
assistant

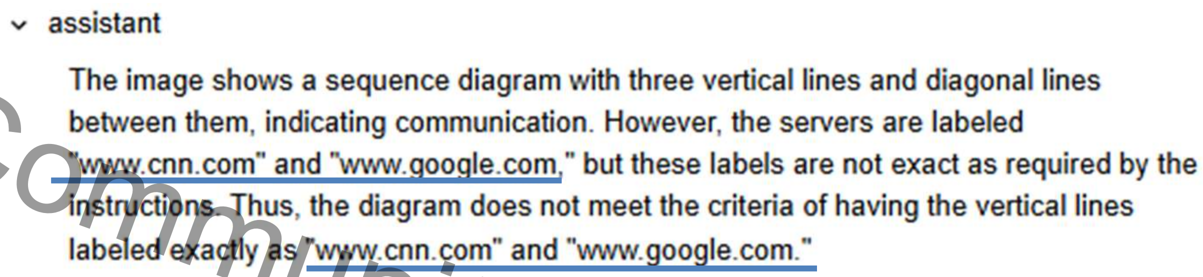
Yes, the image shows a sequence diagram as described.

Common Errors

2025 CAE Community Symposium

Examine Incorrectly Graded Submissions




⌄ assistant

The image shows a sequence diagram with three vertical lines and diagonal lines between them, indicating communication. However, the servers are labeled "www.cnn.com" and "www.google.com," but these labels are not exact as required by the instructions. Thus, the diagram does not meet the criteria of having the vertical lines labeled exactly as "www.cnn.com" and "www.google.com."

Might be solvable with a different prompt

2025 CAE Community Symposium

Conclusions

- LLM enable autograding of open-ended questions
- Detailed rubrics are needed
- In some case, iteration is required to improve rubrics
- Iteration can include fine-tuning, prompt engineering, and rubric design
- Once rubrics are perfected, autograding is extremely useful

2025 CAE Community Symposium