## Using Fine-Tuned LLMs to Grade

Stephan Bohacek and Aishah Aseeri Department of Electrical and Computer Engineering

12th Annual CAE in Cybersecurity Community Symposium





#### Outline

ullet

Autograding scenario

Fine-tuning Demo – Walk Through www.guidance-rubrics.com ullet





#### **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 \_
- iuous challenges with TAs and graders
   ow to grade
   icomplete or nonexistent comments
   icorrect grading
   Experienced graders know that easy grading reduces the chances that the student will complain to the instructor





#### 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





#### Canvas Quiz

Question 5	20 pts	Automatically generated grade
What is the difference and similarity of MAC and a digital signature? Edit View Insert Format Tools Table	Question 5	10 / 20 pts
$12pt \lor Paragraph \lor B I \cup A \lor Z \lor T \lor Q \lor B \lor Q \lor I$	What is the difference and similarity of Your Answer: MAC proves that the message that was the person that has the private key. Bo	MAC and a digital signature? s sent has not been changed. Digital signature proves that the message is from th MAC and Digital signature encrypts a bash of a message
1	Not quite. Try again. Your answer co correctly addressed: Digital signatu	prrectly addressed: MACs ensures the integrity of the message Your answer re protect the integrity of the message. However, what type of encryption do
Region for student to enter their answer	secret key, or asymmetric key? If yo with the text 'MANUAL REVIEW R can see some hints by setting your a	by or asymmetric key encryption? Do MACS require a key? If so, is it a shared bu believe that the grade is incorrect, please resubmit the same answer but EQUESTED' at the top and it will be reviewed manually. If you are stuck, you answer to "Help" and nothing else.
	Automatic cor	ally generated mments





#### **Guidance / Hints**

Rubrics to determine if the submission is correct

Rubrics to determine if a particular hint is applicable

rubric:

- o 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.
- o common errors:
- mmon errors:
  prompt: "Does the statement mention that monueses America's first president? Statement: <submission>"
  hint: "While Thomas Jefferson was one of the founders, he was not the first president.
  Place review Section 2.3 of the textbook"

pointer to course material





#### Resubmissions



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





#### **Rubric Components**

- Question: Explain the significance of the Louisiana Purchase in American history and its impact on the expansion of the **Jnited States**.
- nswer 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:
- mmunit 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 agricultura potential and therefore helped spur economic growth and development? Statement: <SUBMISSION>





#### **Rubric Component Details**

- Prompts should not be too specific
  - Prompt: Consider the statement below. Does this statement mention that the couisiana Purchase greatly expanded agricultural potential and therefore helped spure (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 ommunit
    - LLMs **can** be very literal
- Use Boolean expressions ۰
  - OR
    - Consider the statement below. Does this statement mention that the Louisiana Purchase greatly ٠ expanded agricultural potential and therefore helped spur economic growth. Statement: <SUBMISSION>
    - Consider the statement below. Does this statement mention that the Louisiana Purchase greatly ٠ expanded agricultural potential and therefore helped spur economic development. Statement: <SUBMISSION>





#### More Specific Prompts Tend to Give Better Results

- 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

Instead of

- 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>





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



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**



#### Fine-Tuning LLMs

everages 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 Dosium though the answer is correct.
  - Consequently, fine-tuning is poorly suited for chain-of-thought prompts. ٠
    - Without fine-tuning, chain-of-thought often results in the best performance. •





#### Step to Fine-Tune for Autograding

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

- Make Rubric
- 2.
- Label submissions E Community Symposium 3.
- 4.
- 5.
- 6. Go to 4

NIVERSITY OF ELAWARE



15

#### Building Training Set (1)

0	Guidance	Build Training Dat	ta Set	SET				
	Rubrics	Set3	Data Set				~	
	ter Help	Rubric Eleme	ents in Train	ing Set				^
	Rubric Editor	Delete	Count C	ourse Versio	ric Component Version	Initial Prompt	Desired "Yes" Answer	Desired "No" Answer
	Student Submissions				245-0100460.040	Deep the justice above a sequence distance of computicities between a steel and two secures a supply a secure stand interview secure and a secure stand	Ves the image shows	No the image does not
	Create Submissions			31 18	Computer Networks	Does the image shows a sequence olagiam or communication between a client and two servers, named view.chn.com and a server inamed www.google.com and no other servers? Note that the servers must be labeled exactly as www.chn.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.	sequence diagram as described	show the sequence diagram as described.
	Manually Graded				<b>UN</b>	b.		
	Student Progress	Add Rurbic E	element to T	raining Set	Notworks I (2024 Eall)	nn		^
	Hanage Questions	Select Quiz	0130430	ono. Computer	Networks (2024 Fail)			
	S Import Rubrics	Module 2 H	lomework Q	uestions		'Unit.	v	
	8 Manage Graders	Question 14	: MANUALL	Y GRADED Par	t 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	
	Status and Alerts	Version: 39	. For testing	only. Is deploy	ed.		×	
	A Settings	Select to include in	Rubric	Component			Fallow	un cromot
	😤 Fine-Tune	Training Set	Version	Version				ap prompt
	Compare Rurbics				Does the image show exactly	y one a sequence diagram? Moreover, does this sequence diagram show the communication between a client and two servers, namely a server named www.cn.com and le com and no other servers? Note that the servers must be labeled exactly as www.cnn.com and www.concode.com. For example, labeling the servers cnn ad google is	I don't understand / Pie	ase answer either "Yes, the
	() Sign Out	+	39	26	incorrect and labeling a serve that is, WWW.CNN.COM an line should be labeled "www	ers 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 of WWW.GOOGLE.COM is acceptable. Note that by sequence diagram, we mean that in this case the sequence diagram should contain three vertical lines should be labeled "www.google.com", and the one vertical line should be labeled "www.google.com", and the one vertical line should be labeled "www.google.com", and the one vertical line should be labeled "www.google.com", and the one vertical line should be labeled "www.google.com", and the one vertical line can be named "client," however, thabeing a vertical line is optional. Moreover,	image shows sequence answer "No, the image diagram	e diagram as described" or does not show the sequence as described.".
	() Sign In					there should be diagonal lines between these vertical lines.		
		+	39	11	This image shows a seq transactions between the cli from www.cnn.com to the cl with "Yes, the diagram shows	uence 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 ent 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 ent. So, four transactions means to inter a the to www.cnn.com and four communications from www.cnn.com to the client. Please begins 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 you reasoning.	I don't understand. Ple diagram shows four tra answer "No, the dia transaction	ase answer either "Yes, the ansactions as described" or gram does not show four s as described".
UTS I		+	39	11	This image shows a se communication between the follows? 1. A message is sen messages to www.cnn.com messages to www.cnn.com.tr	equence diagram between a client and two servers, namely a server named www.cnn.com and a server named www.gogle.com. Here, we are only interested in 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 if 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 is down with the message is a client. The client to www.cnn.com 6. www.cnn.com sends a message to the client. 3. The client sends two 6. 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 6. where explain which we downloaded as described" on "No. the diagram down and how two wholers being downloaded as described" on "No. the diagram down and how two work on being two workloaded as described" on "No.	I don't understand. Ple diagram shows the s described" or answer does the sequence of	ase answer either "Yes, the equence of messages as "No, the diagram does not messages as described".

#### Building Training Set (2)



#### Building Training Set (3)

				3:45:48 PM			2024-08-06			-
Guidance Rubrics	8	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.	gpl-4o-mini- 2024-07-18	{"n_epochs","auto","batch_size","auto","le	arning_rate_multiplier":"auto"}	0
P Heip	C Models					News		Ourse (hotes)	Created M	,
Student Submissions						rvaine		Owner (bytes)	Citaled Al	
						gpt-4-0314		openai	6/27/2023, 12:13:30 PM	
S Create Submissions			4			whisper-1		openai-internal	2/27/2023, 4:13:04 PM	
Manually Graded				• (		gpt-4		openai	6/27/2023, 12:13:31 PM	
Student Progress						15-1		openal-internal	4/19/2023, 5:49:11 PM	
Manage Questions						gpt-3.5-turbo		openai	2/28/2023, 1:56:42 PM	
g Import Rubrics						gpl-3.5-turbo-16k-0613		openai	5/30/2023, 3:17:27 PM	
Manage Graders						gpl-3.5-turbo-16k		openai-internal	5/10/2023, 6:35:02 PM	
Status and Alerts						text-embedding-ada-002	Q.	openai-internal	12/16/2022, 2:01:39 PM	
g Settings						gpt-4-0613	$\mathbf{O}$	ođenaj	6/12/2023, 12:54:56 PM	
§ Fine-Tune						gpl-4-32k-0314		openai	6/28/2023, 3:08:41 PM	
Compare Rurbics					curie	:fl-personal-2023-04-13-00-28-10	us	er-fwgxoqbtxahepmwsilr97fxt	4/12/2023, 8:28:10.PM	
) Sign Out					curie	:fl-personal-2023-04-08-19-41-12	us	er-fwgxoqbtxahepmwslir97fxt	4/8/2023, 3:41 12 PM	
					curie	:fl-personal-2023-04-13-00-31-45	us	er-fwgxoqbtxahepmwsllr97fxt	4/12/2023, 8,31:45 PM	
) sign in					curie	:ft-personal-2023-04-07-00-35-15	u	er-fwgxoqbtxahepmwsllr97fxt	4/6/2023, 8:35:15 PM	
					curie	:ft-personal-2023-04-08-22-39-23	u	er-fwgxoqbtxahepmwsllr97fxt	4/8/2023, 6:39:23 PM	
			ft:gpt	I-4o-2024-08-06:perso	nal:3429dadc39	975450ab4deafa68293aad4-mytrainingset-v1:AvANaGeO:ckpt-step-188	u	er-fwgxoqbtxahepmwsiir97fxt	1/29/2025, 4:59:18 PM	
			ft:gp	/t-4o-2024-08-06:persc	nal:3429dadc3	975450ab4deafa68293aad4-mytrainingset-v1:AvANa1v4:ckpi-step-376	us	er-fwgxoqbtxahepmwsllr97fxt	1/29/2025, 4:59:18 PM	
				ft:gpt-4o-2024-08-0	6:personal:342	9dadc3975450ab4deafa68293aad4-mytrainingset-v1:AvANaTwy	us	er-fwgxoqbtxahepmwsllr97fxt	1/29/2025, 4:59:19 PM	



#### Update Rubric to Use Fine-tuned Model

		Preprocessing Options V
20-	Guidance	Rubric Components
$\leq 112$	Rubrics	Component: A diagram that shows communication between a client and two web server
-0/5-	G Help	
	😤 Rubric Editor	Select v
	Student Submissions	Component Description
	Create Submissions	A diagram that shows communication between a client and two web servers, namely www.cnn.com and www.google.com.
	8 Manually Graded	Your timing diagram must show three elements, namely the client and two web servers named www.cnn.com and
	Student Progress	www.google.com. Insee exements must be exempt able as client, www.clin.com, and www.google.com. Do no a do tien hame to the server example for example, www.cin.com/parts www.clin.com/ and a file name (pict [gg], Label the servers with server namea only
	答 Manage Questions	Component Rubit:
	😤 Import Rubrics	Or Rubric Element
	S Manage Graders	Rubic Banes Type Analysis of an image
	Status and Alerts	ā var salatīta s
	& Settings	ft:gpt-4o-2024-08-06:personal:3429dadc3975450ab4deafa68293aadd-set3-vHB
	龄 Fine-Tune	ft:gpt-40-2024-08-06:personal:3428dadc3975450ab4deafa68283aad4- mytrainingset-v1:AvANaGeO:ckpt-step-188
	😤 Compare Rurbics	ftsgn+4o-2024-08-06:personal:3429dadc3975450ab4deafa68293aad4- mytralningset-v1:AvANatV4:ckpt-step-376
	() Sign Out	ft:gpt-40-2024-08-06:personal:3429dadc3975450ab4deafa68293aad4- myftrainingset-yf:AvANaTwy
	() Sign In	ft:gpt-40-2024-08-06:personal:342b4dadc3975450ab4dosfe68293aad4-set2-vt:B0empC38-ckpt-step-108
		ft:gpt-4o-2024-08-08:pterisonal:3429dadc3975450ab4deafa68283aad4-set2-
		- Follow-up Prompt
		I don't understand. Please answer either "l'es, the image shows sequence diagram as described" or answer "Ito, the image does not show the sequence diagram as described.".
		Rubric element must be satisfied

Next, save and deploy new rubric



#### **Compare Rubrics**

			-0///
	metric	Version 21	Version 39
omething like www.cnn.com/pic.jpg is also incorrect not important, that is, WWW.CNN.COM and WWW.0 equence diagram should contain three vertical lines. www.google.com", and the one vertical line can be not nes between these vertical lines. ollow-up Prompt: I don't understand. Please answer ot show the sequence diagram as described.". umber graded correctly: 101 umber graded incorrectly: 6	t. Also, no other servers should be in GOOGLE.COM is acceptable. Note th . One vertical line should be labeled named "client," however, labeling a ver- er either "Yes, the image shows seque	cluded in the diagram. On the other ha nat by sequence diagram, we mean tha "www.cnn.com," one vertical line shou ertical line is optional. Moreover, there ence diagram as described" or answer	nd, the case of the text t in this case the Id be labeled should be diagonal "No, the image does
lient and two servers, namely a server named www. e labeled exactly as www.cnn.com and www.google	equence diagram? Moreover, does th cnn.com and a server named www.g e.com. For example, labeling the serv	is sequence diagram show the commu google.com and no other servers? Note vers cnn and google is incorrect and la	inication between a that the servers must beling a servers
		8041505	^

#### **Examine Incorrectly Graded Submissions**



This was mislabeled.

After fixing labeling, repeating the finetuning process might give better results.

ELAWARE

Guidance **Rubrics** A Help Rubric Editor Student Submissio 2 Create Submissio S Manually Graded P.222 Student Progres Anage Questions Se Import Rubrics Anage Graders Executed Rubric (Grades: 100%, 100%, 100%, Status and Alerts A diagram that shows communication between a clie www.cnn.com and www.google.com. Grade:100% Settings Or (Grade: 1) ~ Grade: 1 Se Fine-Tune 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 S Compare Rurbics 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 (<sup>1</sup>) Sign Out 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 () Sign In and WWW.GOOGLE.COM is acceptable. Note that by sequence diagram, we mean that in

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

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.

- model: ft:gpt-4o-2024-08-06:personal:3429dadc3975450ab4deafa68293aad4-set3v1: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
```



# Examine Incorrectly Graded Submissions

#### **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



#### **Conclusions**

Venable 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 Posium
- Once rubrics are perfected, autograding is extremely useful ۲



