Using GPT to Grade and Give Hints - Version 0

Stephan Bohacek

Department of Electrical and Computer Engineering

bohacek@udel.edu

11th Annual CAE in Cybersecurity Community Symposium

UNIVERSITY OF DELAWARE

CAE IN CYBERSECURITY COMMUNITY

Stephan Bohacek

- Associate Professor at the University of Delaware
- Teach
  - Introduction to Cybersecurity: ~175 students
  - Introduction to Networking: ~80 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
  - Inconsistent grading
  - Ambiguous questions that don't have good answers
    - This is the instructor's fault

Homework-Based Assessment Drawbacks

- When the student submits the homework, the learning stops
- Grading and comments have little impact
  - Most students don't read comments
  - When comments are read, there is little learning. Instead, grades are checked to determine whether the grader was correct rather than to learn from mistakes
  - By the time grading is complete, the students have forgotten the problem

  Partly caused by low grader quality

- Partial credit is virtually impossible
  - It is wrong, they should get 0 points.
  - Well,
    - They sort of understood part of the topics
    - This is the only chance for grade this, so 0 points will lead to the student getting a very low grade…
- Resubmission of incorrect submissions is too labor intensive for graders
- There is little opportunity for students to use homework for learning, it is mostly for assessment
  - It answers the question: Did the student learn the material?
  - Homework should be like labs:
    - Here is a problem
   - Let's work until we solve the problem

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
- If the student does not understand the comments, then they include the text "NEED BETTER COMMENTS" and alerts are sent to instructors and graders

UNIVERSITY OF DELAWARE.

Process
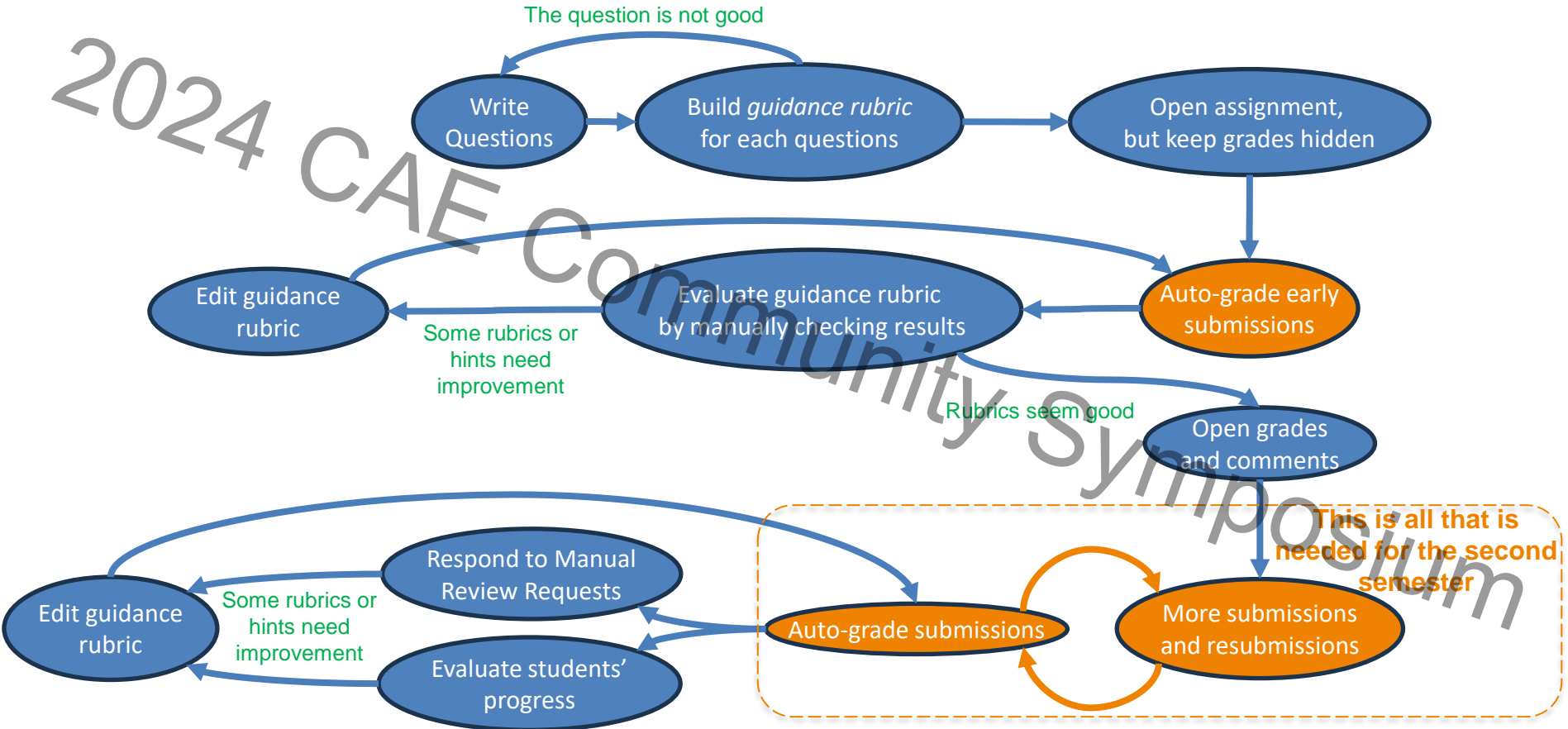
Guidance Rubric

- My Experience: AI is not useful for automatic grading. But AI can be used if it is coupled with a well written and detailed *guidance rubric*
- Guidance Rubric
  - Rubric gives rules for grading
  - Guidance rubric also gives rules for guidance (hints)
- Writing guidance rubric is labor intensive and requires practice

rule for grading →

rule for hint →

hint →

```
rubric:
o  theme: "Does the submission mention that George Washington was the United States of America's first
   president?"
o  comment error:
   o  theme: "Does the submission mention that Thomas Jefferson was the United States of America's first
      president",
   o  hint: "While Thomas Jefferson was one of the founders' he was not the first president. Please review
      Section 2.3 of the textbook"
```
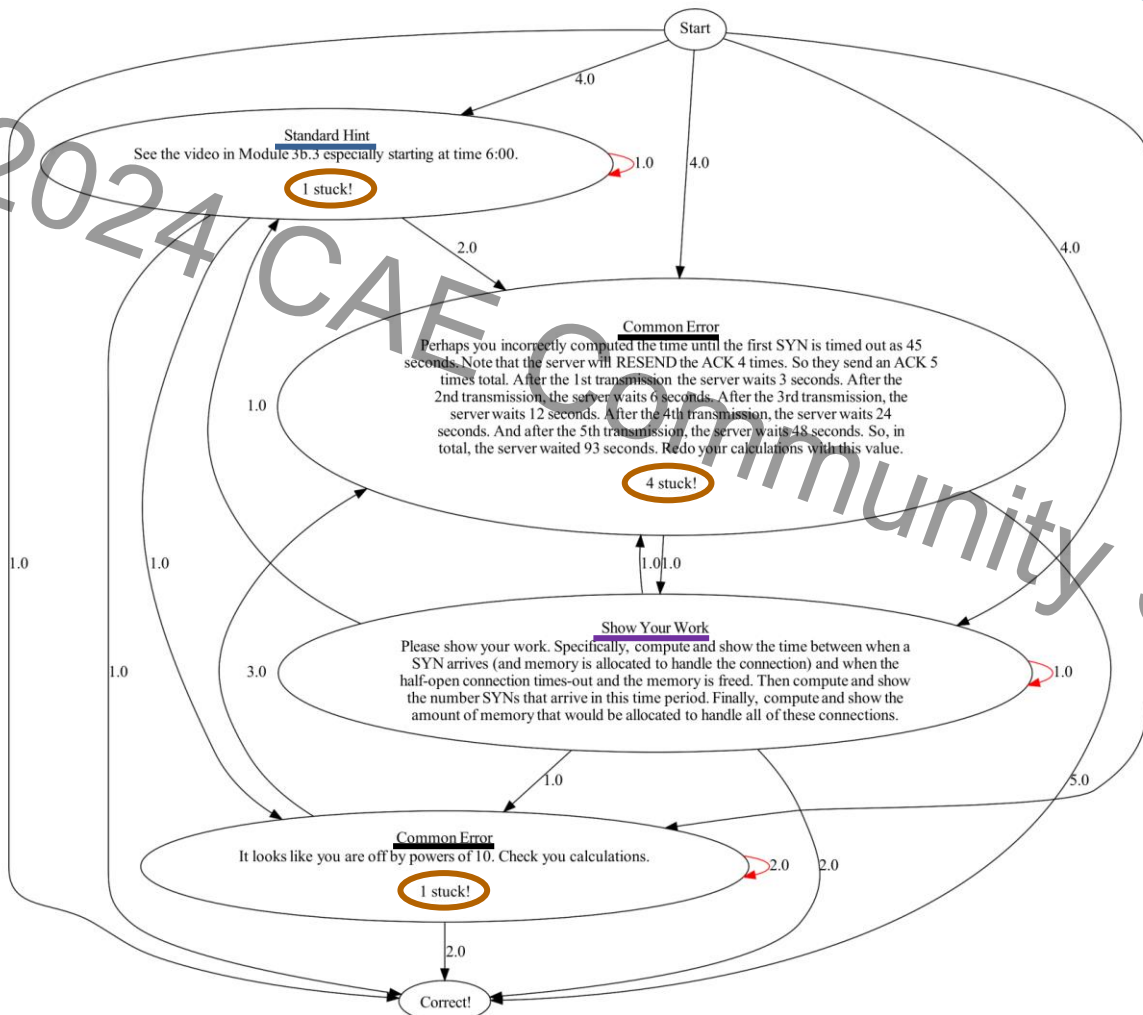
Note, this is not impressive. We need to write these highly detailed rubrics. What does the AI do?
The AI can only evaluate the guidance rubrics. AI is a text processing engine.

Evaluate Students' Progress

Standard hint: The hint received when full credit is not received for this component (question). Standard hints are not specific. If many students are stuck in standard hint, then perhaps a more specific hint is needed.

Stuck: The number of students where this is the last stage they are at.

Looping: The number of students that received the same hint twice.

Common errors: There can be many common errors, each with its own specific hint.

Show Your Work: A special type of common error triggered when not enough text was submitted.
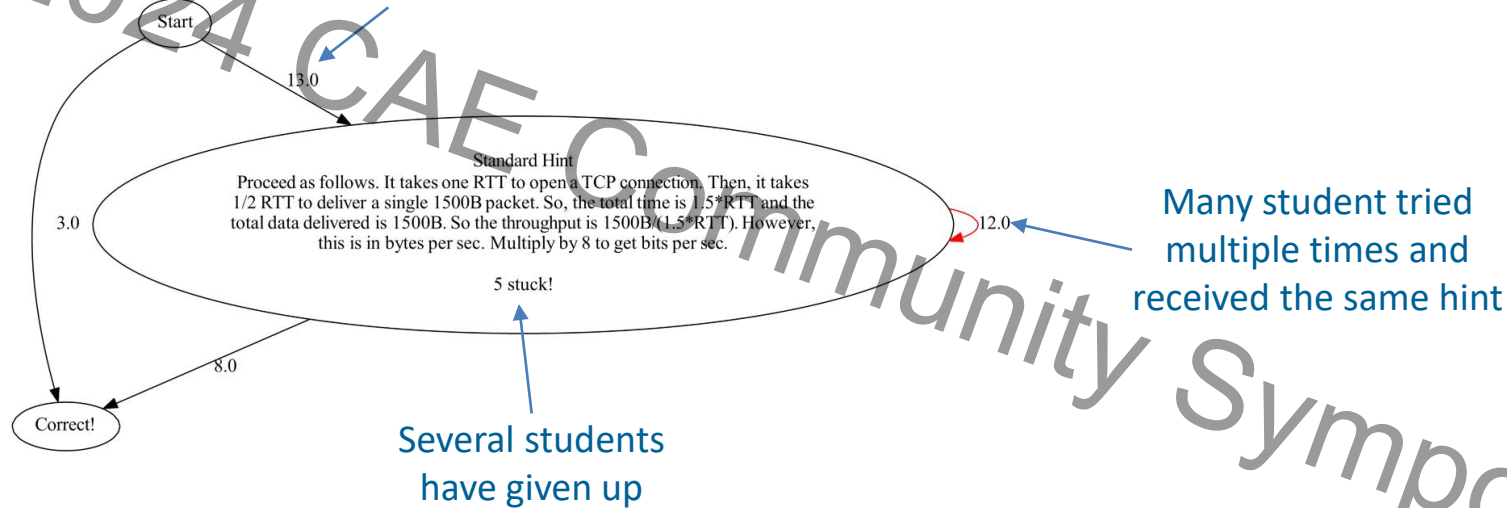
An Example With Many Common Errors

- Complicated problems (especially problems that include calculations) can have many common errors
- Students can pass through several common errors and hints before arriving at the correct solution
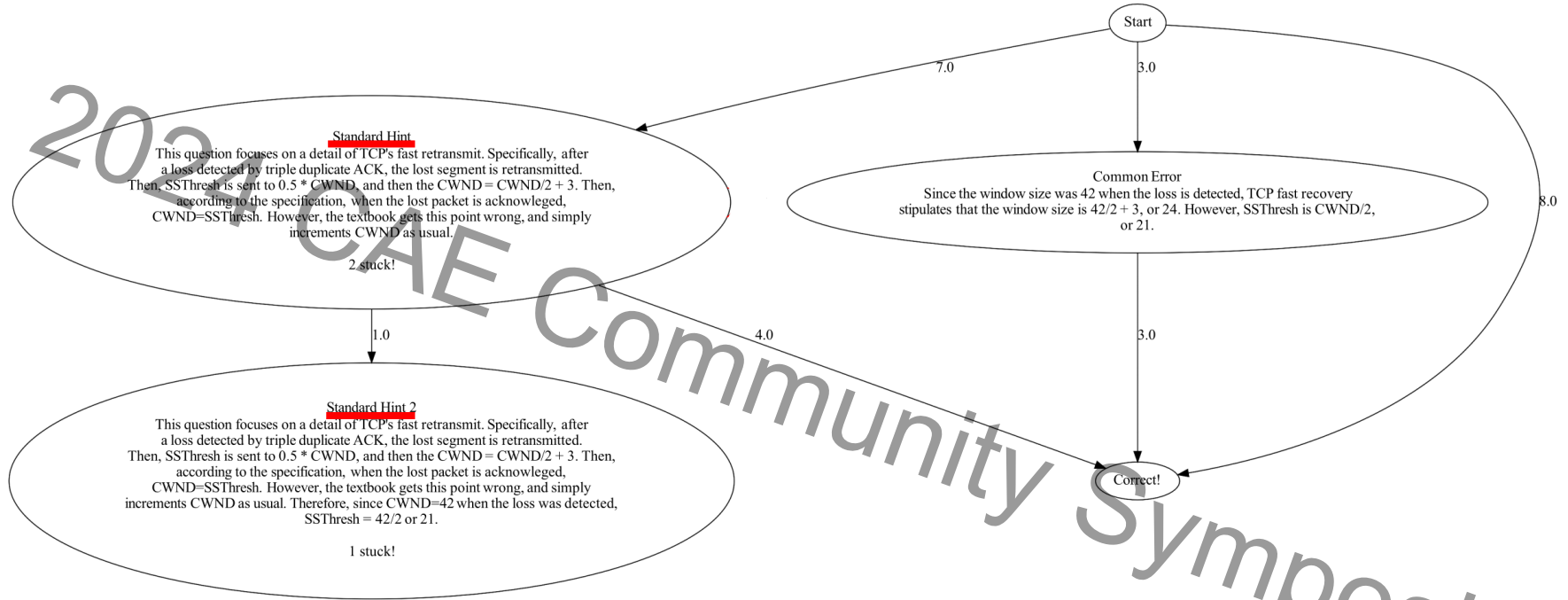- I feel like I'm having a conversation with the students

Process for Developing Common Errors

Most students received a (non-specific) Standard Hint

Start

13.0

Standard Hint
Proceed as follows. It takes one RTT to open a TCP connection. Then, it takes 1/2 RTT to deliver a single 1500B packet. So, the total time is 1.5*RTT and the total data delivered is 1500B. So the throughput is 1500B/(1.5*RTT). However, this is in bytes per sec. Multiply by 8 to get bits per sec.

3.0

12.0

Many student tried multiple times and received the same hint

5 stuck!

8.0

Correct!

Several students have given up

In such cases, manually examine student submissions and then enhance the guidance rubric and/or provide extra learning material.

Looping and Hint Sequences

Start

7.0   3.0

Standard Hint
This question focuses on a detail of TCP's fast retransmit. Specifically, after a loss detected by triple duplicate ACK, the lost segment is retransmitted. Then, SSThresh is sent to 0.5 * CWND, and then the CWND = CWND/2 + 3. Then, according to the specification, when the lost packet is acknowleged, CWND=SSThresh. However, the textbook gets this point wrong, and simply increments CWND as usual.

2 stuck!

Common Error
Since the window size was 42 when the loss is detected, TCP fast recovery stipulates that the window size is 42/2 + 3, or 24. However, SSThresh is CWND/2, or 21.

8.0

1.0   4.0   3.0

Standard Hint 2
This question focuses on a detail of TCP's fast retransmit. Specifically, after a loss detected by triple duplicate ACK, the lost segment is retransmitted. Then, SSThresh is sent to 0.5 * CWND, and then the CWND = CWND/2 + 3. Then, according to the specification, when the lost packet is acknowleged, CWND=SSThresh. However, the textbook gets this point wrong, and simply increments CWND as usual. Therefore, since CWND=42 when the loss was detected, SSThresh = 42/2 or 21.

1 stuck!

Correct!

- Looping is when a student gets the same hint more than once.
- Providing a sequence of hints can help students in this situation
- That is, for a specific common error (or the standard hint), if that hint has been given to the student before, the next hint in the sequence is given.
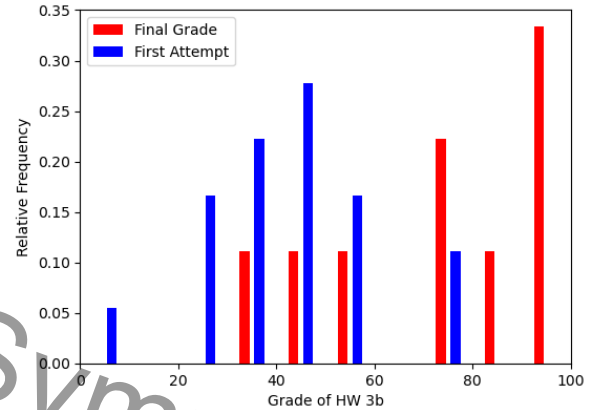
Impact on Grades
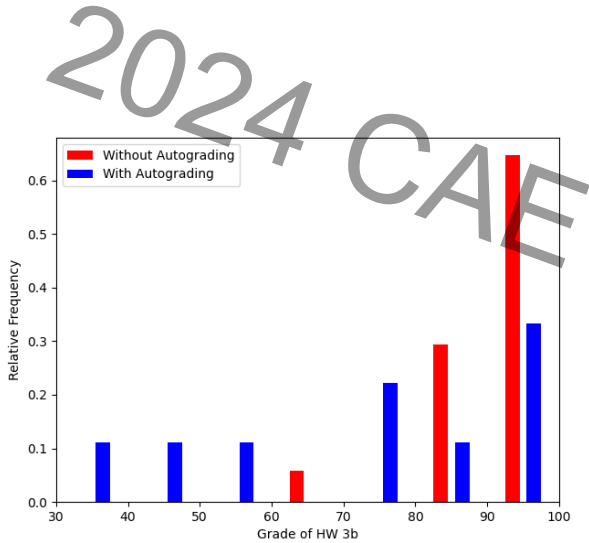
Grades are significantly lower!

- Previous grading quality was low
- Auto-grading is very strict, there is limited partial credit
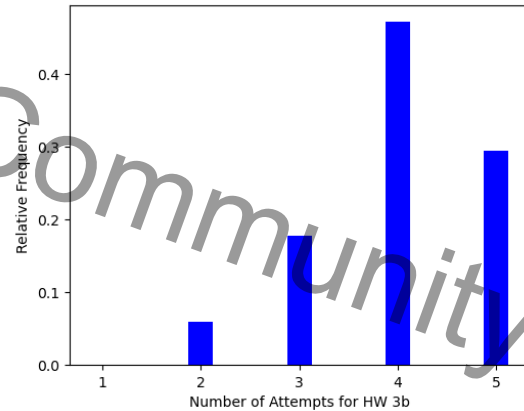
Students take advantage of resubmission

Grades significantly improve over multiple attempts

Minimal AI Guidance Rubric Elements

- Writing prompts is difficult.
- Ideally, don't use AI to grade. Instead use
  - Does the submission include specific text?
    - That is, the answer must have some key word(s)
  - Does the submission include a specific number (rounded to some precision)?
  - Does the submission include a number in a range?
  - Does the submission include a mathematical expression?
  - Does the submission include a mathematical expression with value?
    - 0.14s + 10Mb/15Mbps is an expression with value 81ms (within 2 significant digits)
    - 24 + sqrt(4+8) is an expression with value 27.5 (within 3 significant digits)
  - Does the submission include a number with specific units (rounded to some precision)?
    - E.g., 10Mbps (which is the same as 10,000Kbps, 10,000,000bps, and 0.01Gbps)
  - Does the submission include a number that is close to a power of 10 of a specific value?
    - The correct answer is 26,666,666.7, but the student gave the answer 26,666.6 because they confused that M is 10^3
  - Does the submission include a table with specific columns?
    - e.g., Source IP address, Source Port, Destination IP Address, Destination Port
  - Does the table include a row where (some) columns take specific values?
    - E.g., Source IP address = 192.168.2.1 and Source Port = 2345,…
- These are not the prompts. These are the objectives.
  - Some of these objectives are implemented without any AI.
  - Some used straightforward prompts.
  - Some use more complicate prompt engineering. Examples to follow.

UNIVERSITY OF DELAWARE.

Some Prompt Engineering is Still Needed (Part 1)

- Prompt Objective:
  - Does the submission include the number 277GB, within 2 significant digits?
- Submission:
  - With 4 retries and RTO multiplied by 2 for each retry the connection will last for: 3 + 3(2) + 3(4) + 3(8) + 3(16) = 3 + 6 + 12 + 24 + 48 = 93 seconds. Given 93 seconds as the window for attack, the attacker can send 290,625 SYN packets. 100KB of memory is allocated per SYN requiring at least 29,062,500 KB of memory to within the SYN attack assuming nothing else is also coming in.
- Reply:
  - Yes, the number 29,062,500 KB is approximately equal to 277GB.
- … prompts …
- Reply:
  - No

Some Prompt Engineering is Still Needed (Part 2)

- Prompt Objective:
  - Does the submission include an integer that is equal to 63
- Submission:
  - What is the maximum number of subscribers that my server can handle?
    28 Subscribers is the maximum number of subscribers the server can handle if the goal is to keep the probability of a crash less than 1%.
    Question 2. What is the maximum number of subscribers I can support if my server is twice as large and can handle 20 simultaneous players?
    67 Subscribers is the maximum number of subscribers the server can handle if the goal is to keep the probability of a crash less than 1% and the server can handle up to 20 simultaneous subscribers.
- Reply:
  - Yes
- … prompts …
- Reply:
  - No

Cleaning Submissions

- Student submission can contain significant typographical errors, spelling mistakes, etc.
- Submission with these types of errors
  - are hard for the AI to grade
  - are hard for humans to grade
- Option 1 (mimic the normal approach to grading)
  - Make a system that can infer what the student is trying to say
  - My Experience
    - I spent several months trying to get AI to grade previous years' homework.
    - I couldn't get satisfying results.
    - There were too many grammatical errors and similar errors, so the accuracy metrics were too low (even with human graders)
- Option 2
  - Force students to submit well written answers
    - Train the students
    - We all learn to write good google queries. We can all learn to work with auto-grading
  - My Experience
    - I have spent a considerable amount time writing rules to handle problemed submissions
    - But this greatly simplified auto-grading

Cleaning Submissions – Inconsistencies

- Submission:
  - I believe 2 RTT will be required to download an object in addition to the DNS lookup. When I used the google dig tool (https://toolbox.googleapps.com/apps/dig/#NS/) I received a TTL of 5 hours 5 minutes and 30 seconds when I looked in the NS tab. I think a total of 5 RTT will be needed as both DNS servers will have to be contacted as well as the website in addition to the download.
- Consider
  - Rubric: Does the submission mention that is takes 2 RTTs to download the object? Reply: Yes
  - Rubric: Does the submission mention that is takes 5 RTTs to download the object? Reply: Yes
- Instead, first check for inconsistencies. Reply from GPT:
  - The statement contradicts itself. Initially, it is stated that 2 RTT (Round Trip Time) will be required to download an object in addition to the DNS lookup. However, later in the statement, it is mentioned that a total of 5 RTT will be needed. These two parts of the statement are in contradiction.
- Reply to student (with help from AI):
  - Your submission seems to have inconsistencies. Specifically, the statement contradicts itself. Initially, it is stated that 2 RTT (Round Trip Time) will be required to download an object in addition to the DNS lookup. However, later in the statement, it is mentioned that a total of 5 RTT will be needed. These two parts of the statement are in contradiction. Please correct your answer and resubmit.

Cleaning Submissions –Off Track

- Spelling
  - Multiple tests are used (some free and some not free)
  - If the errors are small, then the submission is corrected, and the grading continues
  - Need to manually allow unusual or domain specific words
- Off track
  - Question: Suppose that a NAT is between the Internet and a company's network. Now suppose that the NAT crashes, but quickly restarts, in a very short amount of time. How would a short-lived reboot impact the users and machines in the company's network?
  - Submission: NAT or network address translation protocol is used to translate or substitute private IPv4 network addresses with public IPv4 addresses to gain access to internet resources. Private IPv4 addresses are not routable on the internet and due to the exhaustion of the IPv4 addresses, the NAT maps the private network with a limited amount of public addresses (public IPv4 network addresses are routable on the internet).
  - Reply to student: Your answer seems to be off track and not relevant to the question. Please review the question and try again. Consider the following hint on making your answer more relevant: You've provided a good explanation of what a NAT is and its function, but this doesn't answer the question asked. The question is asking about the impact of a short-lived reboot of a NAT on the users and machines in a company's network. To improve your response, you should focus on explaining what happens when a NAT reboots, even briefly. Consider the effects on ongoing connections, how the reboot might disrupt network services, and how quickly the network can recover. Also, consider the potential impact on users, such as whether they might experience a temporary loss of internet access or other disruptions. Your answer should directly address the consequences of a NAT reboot, as that is the core of the question. Also, the following hints might be helpful: When a NAT reboots, the NAT table is erased, so only statically assigned entries remain. That means that all existing connections are broken. The application needs to have reliably (e.g., retry methods) to recover from this scenario. Depending on the NAT implementation, packets going from the clients behind the NAT to servers in the Internet will reconstruct the NAT table entries. But some NATs might only build NAT table entries from SYN packets. In any case, if the client is simply waiting for data from the server, the server's transmissions will be blocked by the NAT and the connection will fail.

AI generated (with careful prompts)

Cleaning Submissions - Typos and grammar

- Significant challenge
- Objective: For small typos and grammar errors, fix the submission and continue. For larger errors, request the student fix the errors.
  - Otherwise, too many submissions are rejected.
- Challenge 1: The fixed submission changes the submission making it correct!
  - Solution: only small changes are allowed
- Challenge 2: The "fixed" submission changes numbers, so the submission is changed from incorrect to correct or from correct to incorrect
  - GPT does not have a scream mode. If it did, I would scream "DO NOT CHANGE ANY NUMBERS"

UNIVERSITY OF DELAWARE.

Cleaning Submissions - Typos and grammar

Submission

This looks like the 7th transmission round.
Round 1 = 1 segments
Round 2 = 2 segments
Round 3 = 4 segments
Round 4 = 8 segments
Round 5 = 16 segments
Round 6 = 32 segments
1+2+4+8+16+32 = 63
Round 7 = 33 segments.
63-96 segments are sent in the 7th round.

Correct!

"Corrected" Text

This looks like the 7th transmission round.
Round 1 = 1 segments
Round 2 = 2 segments
Round 3 = 4 segments
Round 4 = 8 segments
Round 5 = 16 segments
Round 6 = 32 segments
1+2+4+8+16+32 = 63
Round 7 = 33 segments.
63-63 segments are sent in the 7th round.

Wrong!

Themes vs Smart Searching

- Smart Searching: Does the submission include a specific number, specific text, etc.

- Themes: Use AI to detect if the answer addresses a theme.

- E.g., Theme: Does the submission mention that TCP is a popular transmission layer protocol?

- There is considerable prompt engineering around themes. But themes are still difficult to write.

Distribution of Rubric Element Types
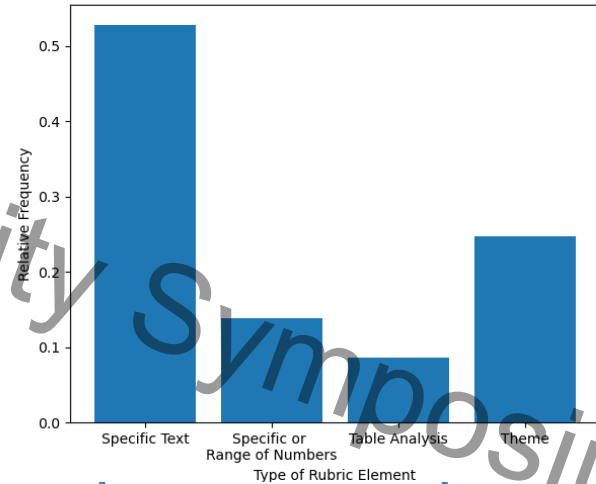
Smart Searches

Prompt - Specifically

- Prompt: Does the submission <u>specifically</u> mention that both MACs and digital signature are used to verify, ensure, protect, <u>or</u> determine the integrity of a message <u>or</u> the authenticity of a message?
- Without "specifically," answers like "MACs are useful. Digital signatures are also useful, they help with the integrity of the message" are accepted
  - Since "also" can be inferred to mean that MACs provides a similar purpose as digital signatures and "they" could mean digital signatures and MACs.
- But "specifically" will require nearly exact matches. For example, the "or" is misinterpreted
  - "MACs and digital signature are used to protect the integrity of a message" is not accepted, since it does not mention ensure, verify, determine, and authenticity.

UNIVERSITY OF DELAWARE

Booleans

- Does the submission <u>specifically</u> mention that both MACs and digital signature are used to verify, ensure, protect or determine the integrity of a message <u>or</u> the authenticity of a message?
- AND:
  - OR:
    - Does the submission specifically mention that MACs are used to verify the integrity of a message
    - Does the submission specifically mention that MACs are used to ensure the integrity of a message
    - …
  - OR:
    - Does the submission specifically mention that digital signatures are used to verify the integrity of a message
    - Does the submission specifically mention that digital signature are used to ensure the integrity of a message
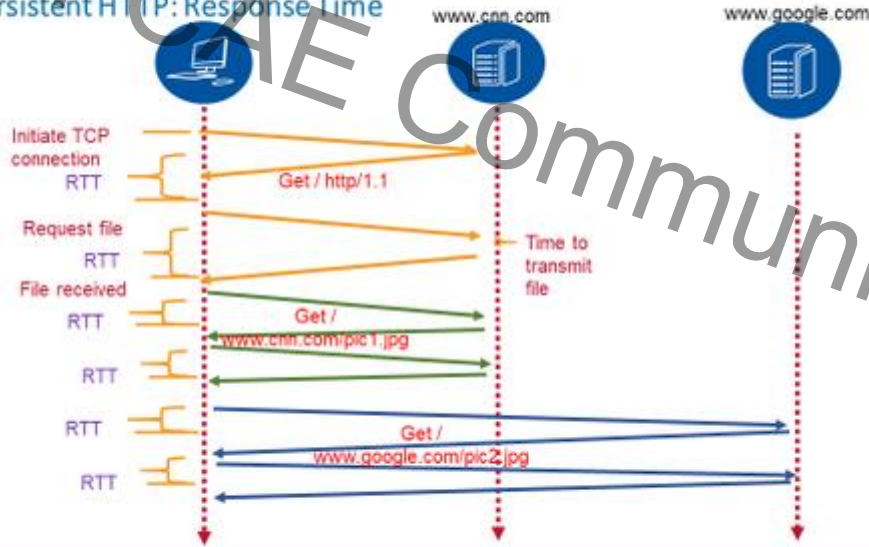   - …

UNIVERSITY OF DELAWARE

Booleans

- Does the submission <u>specifically</u> mention that both MACs and digital signature are used to verify, ensure, protect <u>or</u> determine the integrity of a message <u>or</u> the authenticity of a message?
- Rubric:
  - Component
    - Description: MACs are used to protect the integrity of a message
    - Hint: Be sure to specify the purpose of using MACs
    - Points: 0.5
    - OR:
      - Does the submission specifically mention that MACs are used to verify the integrity of a message
      - Does the submission specifically mention that MACs are used to ensure the integrity of a message
      - …

  - Component
    - Description: Digital signatures are used to protect the integrity of a message
    - Hint: Be sure to specify the purpose of using digital signatures
    - Points: 0.5
    - OR:
      - Does the submission specifically mention that digital signatures are used to verify the integrity of a message
      - Does the submission specifically mention that digital signature are used to ensure the integrity of a message
      - …

- Submission: "MAC is useful. Digital signatures are also useful, they help with integrity"
- Result: 0.5 points. Please try again. You correctly addressed that digital signatures are used to protect the integrity of a message. However, be sure to specify the purpose of using MACs.
- Note: there is partial credit for completely understanding a topic (or sub-topic). But no credit for partial understanding of a topic (or a sub-topic).

UNIVERSITY OF DELAWARE.

GPT-4 for Diagram Grading – No, not yet

Serial (Non) Persistent HTTP: Response Time

www.cnn.com
www.google.com

Initiate TCP connection
RTT

Get / http/1.1

Request file
RTT

Time to transmit file

File received
RTT

Get / www.cnn.com/pic1.jpg

RTT

RTT

Get / www.google.com/pic2.jpg

RTT

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.

UNIVERSITY OF DELAWARE

Conclusions

- AI can be used for automatically grading and giving hints
- GPT must be coupled with a well-written guidance rubric
- Guidance rubrics are difficult to write, but can be reused
- Automatically generating hints and allowing resubmissions
  - Extends the duration of the learning
  - Allows students to work hard, get a better understanding, and good grades
  - Increases the workload on students