INCS-CoE

What generative AI systems know about cybersecurity

Tim Finin and Alan Sherman
UMBC

INCS-CoE webinar on The Growing Role of AI in Cybersecurity, 2023/09/07

Another AI inflection point

- OpenAI released ChatGPT in late 2022 showing the potential of Generative AI (GAI) systems

  ChatGPT converses with people to answer questions, generate text, write code and DB/KG queries, and more

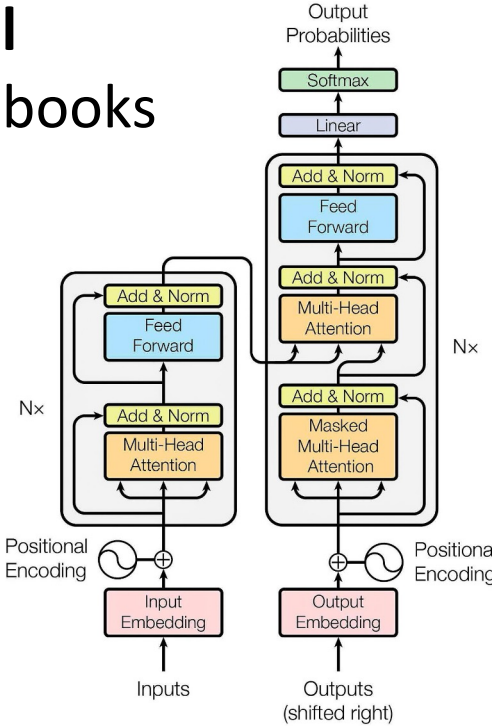
- Other companies (Google, META, Apple, …) have released similar systems and open-source ones are also available

  This has caused many to see their **benefits** as well as their **shortcomings** and **risks**

- We used questions from two cybersecurity assessment tests to evaluate how well they understand cyber text & problems

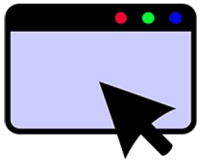
What is a Transformer neural network?

- **Large Language Models** (LLMs) like GPT-4 are **neural networks** trained on huge text corpora from Web & books
- They use Transformers, neural models using word embeddings & an attention mechanism
- The model & training corpora sizes make them **expensive to create,** in cost and energy
  - GPT-4 has ~1.8 trillion parameters over 120 layers and cost more than $100M to train
- Pretrained LLMs are available for researchers, E.g., Meta's LLaMA and Hugging Face models



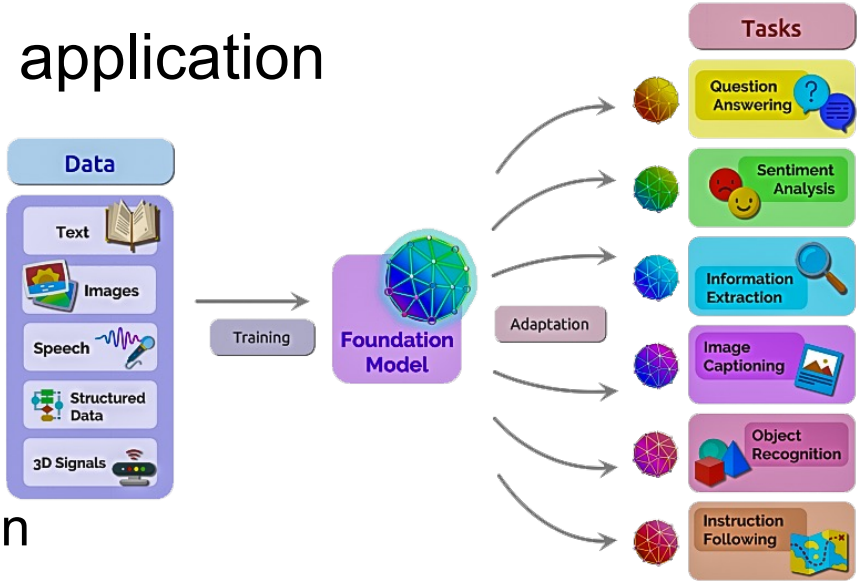
From Attention Is All You Need

Ok, I have a LLM, now what?



- LLMs are called **foundation models** since they are the basis for building or supporting **multiple AI applications**
  E.g., language translation, sentiment detection, summary generation, question answering, coding assistants, and more

- We **fine-tune** a LLM to support an application
  - Extends neural network with layers for the application type, e.g., summarization
  - Use supervised learning to train result with sample inputs & desired outputs
  - And then use reinforcement learning via human feedback to improve application

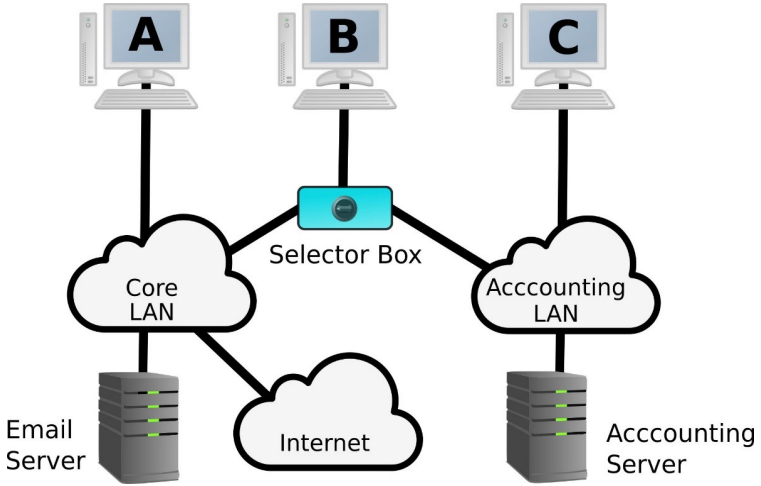


The CATS project, CCI and CCA

- Cybersecurity Assessment Tools is a joint project between UMBC, Univ. of Minnesota Duluth, and University of Illinois
- Two question sets evaluate students' knowledge:
  - **CCI:** Cybersecurity Concept Inventory after a course
  - **CCA:** Cybersecurity Curriculum Assessment after full curriculum
- Each set has 25 multiple choice questions
- We asked two **ChatGPT** models (3.5-turbo and 4) and Google's **Bard** the CCI and CCA questions

Questions comprise a scenario, stem, and choices

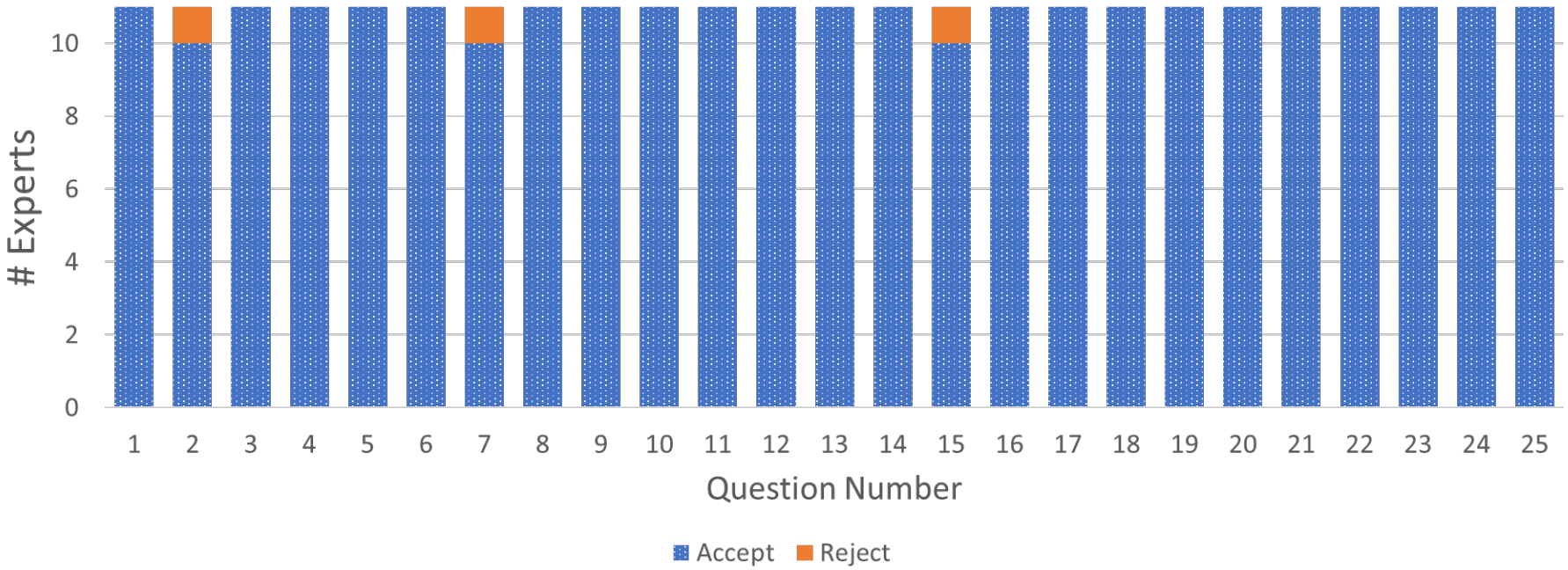
A company has two internal Local Area Networks (LANs): a core LAN connected to an email server and the Internet, and an accounting LAN connected to the corporate accounting server (which is not connected to the Internet). Each desktop computer has one network interface card. Computers A and C are connected to only one of the networks. Computer B requires access to both LANs and is connected to a selector box with a toggle switch that physically connects the computer to exactly one LAN at a time.



Choose the action that this design best prevents:

(a) Emailing accounting data.
(b) Infecting the accounting LAN with malware.
(c) Employees accessing the accounting server from home.
(d) User of Computer B accessing the accounting LAN without authorization.
(e) Computer A communicating with Computer B.

Ten experts reviewed the CCA items & mostly approved



CCA and CCI evaluation

The CATS project did **psychometric evaluations** of the tests by 354 (CCI) and193 (CCA) students from multiple universities with the goal of measuring:

- **reliability**: is the score repeatable?
- **difficulty**: how hard is each question?
- **discrimination**: do students of lower and higher skill levels perform differently?
- **information**: how precise is the score for students at different skill levels?

Adapting CATs questions for GAI systems

- Changes were minor, removed images in a few questions and replacing them with text if needed
- We used the following prompt for the GAI system

> Answer the multiple-choice question below, given the scenario and other information. Consider all the possible answers carefully. Explain why you chose your answer using up to 500 words. Explain why you did not choose each of the alternatives using up to 400 words for each alternative. Write your explanations for university students who have taken a class on cybersecurity.

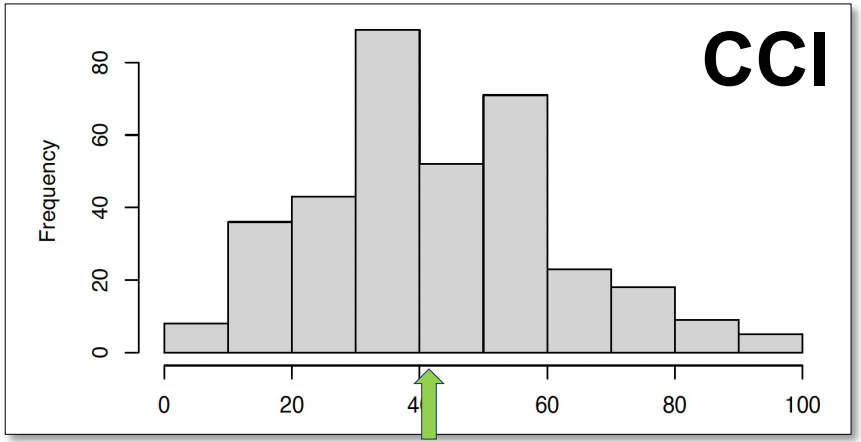
Overall assessment

- All three systems were fast, answering questions in 5-10 sec.
- GPT-4 was clearly better than GPT-3 & Bard on both tests
- GPT-3.5 & Bard did about the same
- On CCA, GPT-4's score beat 93% of the students, while both GPT-3 and Bard did slightly better than 50% of the students
- GPT-4 did the best on CCI, but did not dominate the other two models as clearly

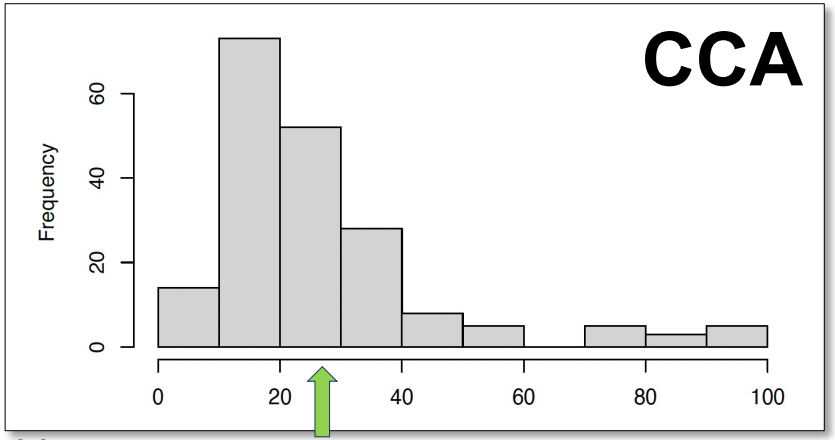
How the students did…

- Students found both CCI and CCA difficult
- Student found CCA much harder than CCI

354 students



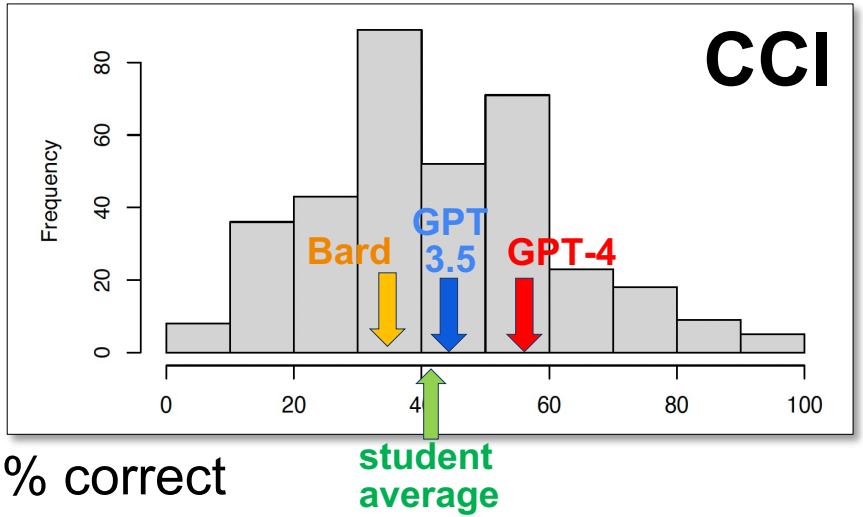
193 students



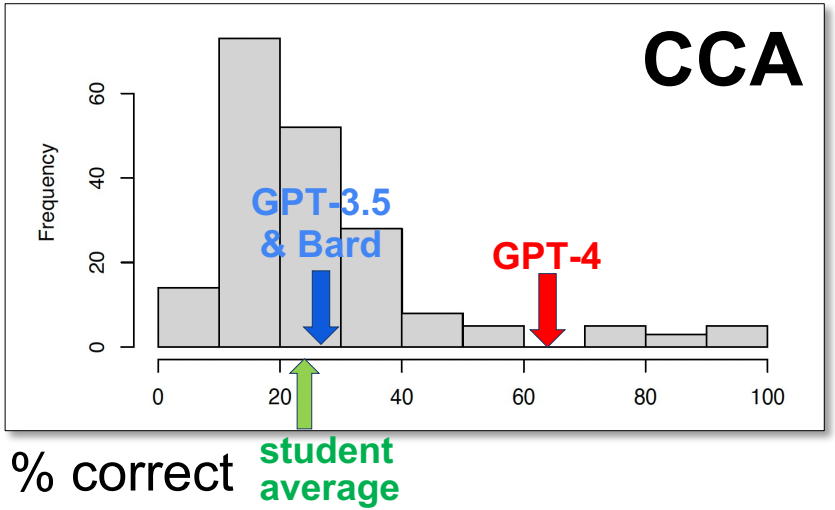
How the GAI models did: GPT-4 >> GPT-3.5 > Bard

- **GPT-4:** better than 93% of students on CCA and 75% on CCI
- **GPT-3.5:** better than 64% of students on CCA and 56% on CCI
- **Bard:** better than 64% of students on CCA and 37% on CCI

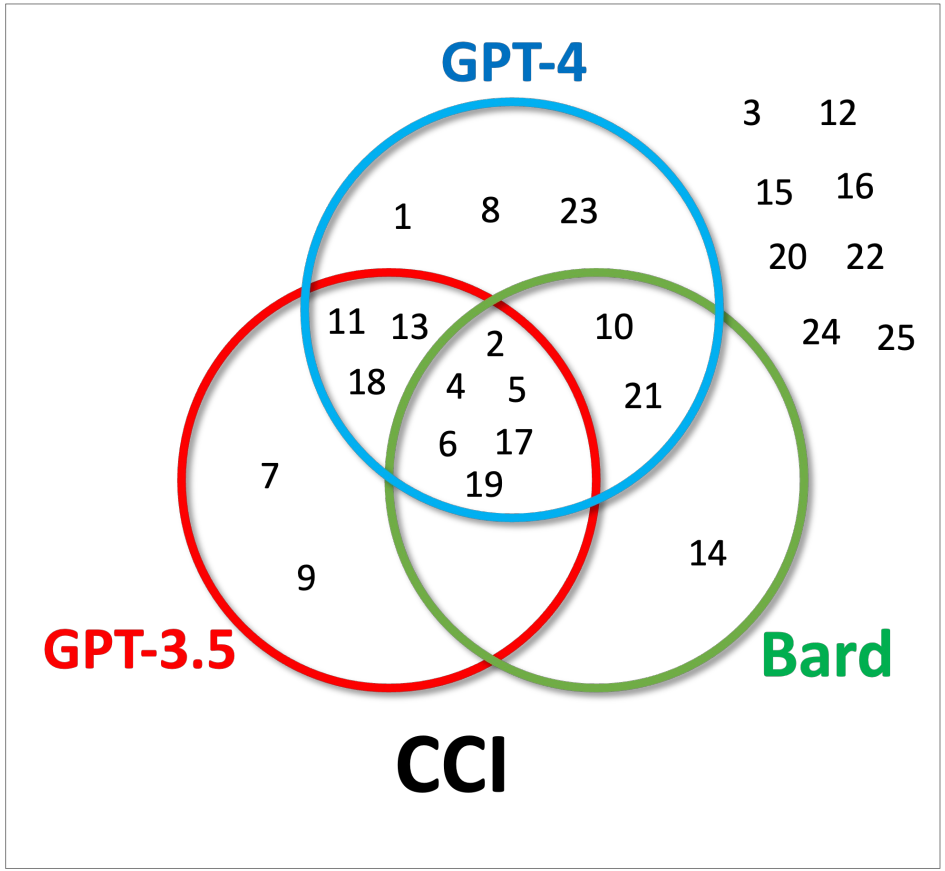
354 students

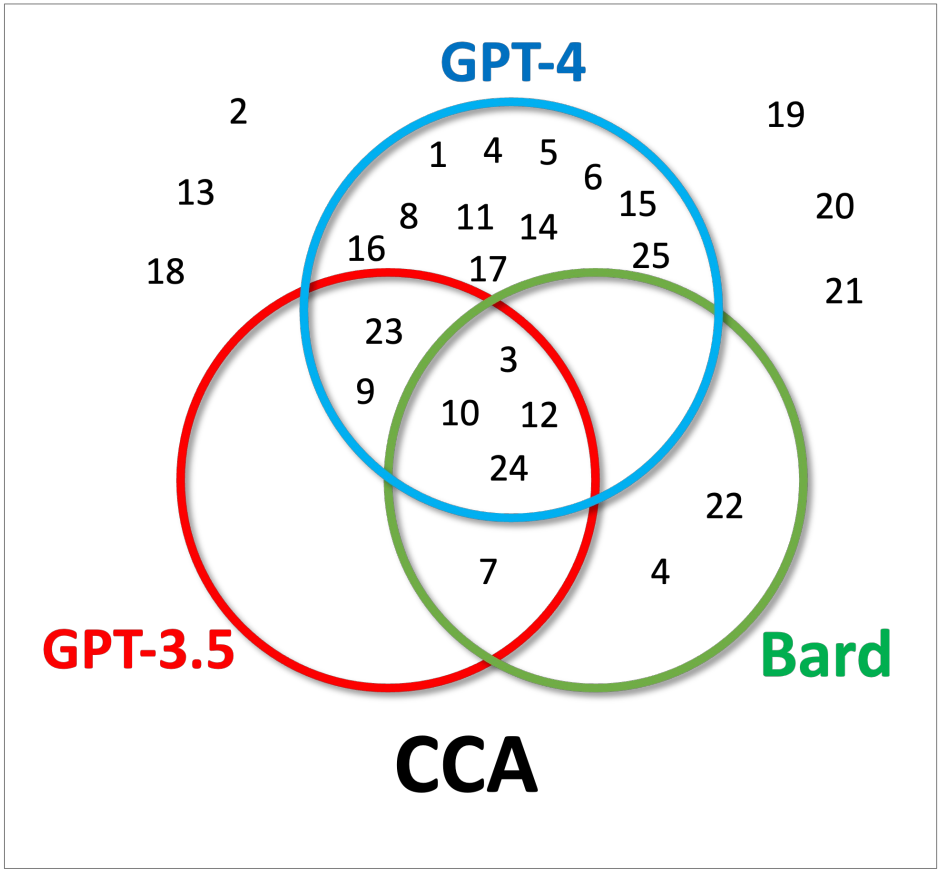


193 students



How the GAI systems did, Venn diagrams





Some observations on GAI's answers for CAA

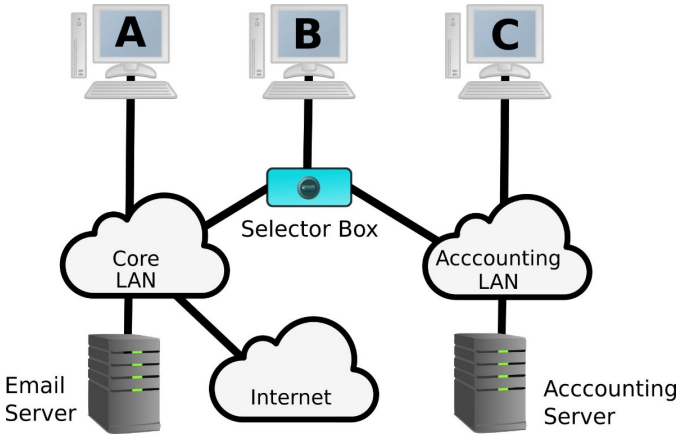
- All three systems generated very readable text explaining why they choose their answer and not the alternatives
  - In some cases, reading GPT-4's explanation for the answer it got wrong, made its answer seem more reasonable
- GAI performance wasn't strongly correlated with Item Response Theory scores of question difficulty for students
  - The question ITR judged to be easiest was answered correctly by GPT-3.5 and Bard, but wrong by GPT-4

Questions: scenario, stem, and choices

A company has two internal Local Area Networks (LANs): a core LAN connected to an email server and the Internet, and an accounting LAN connected to the corporate accounting server (which is not connected to the Internet). Each desktop computer has one network interface card. Computers A and C are connected to only one of the networks. Computer B requires access to both LANs and is connected to a selector box with a toggle switch that physically connects the computer to exactly one LAN at a time.

Choose the action that this design best prevents:

Note: The psychometric evaluation showed this question to be the 2nd hardest for students. All three systems got it wrong.



(a) Emailing accounting data.
(b) Infecting the accounting LAN with malware. **X GPT-4**
(c) Computer A communicating with computer B.
(d) User of Computer B accessing the accounting LAN without authorization. **X GPT-3 & Bard**
(e) Employees accessing the accounting server from home. ✓ **CORRECT**

GPT-4’s answers show some qualifications…

I’d be inclined to give partial credit for GPT-4’s answer, if I was grading it

This system design isolates the accounting LAN from the core LAN, thus creating a physical barrier to the transmission of malware. If the accounting LAN is disconnected from the Internet and can only be accessed through specific computers via a physical switch, then **malware present on the Internet or the core LAN cannot easily propagate to the accounting LAN**. The design is not flawless (e.g., malware could still potentially be introduced via compromised hardware or insider threat), but it significantly mitigates the risk of malware infection.

Now, let's examine why the other options are not as correct:

A. Emailing accounting data: The design doesn't necessarily prevent emailing accounting data. A user on computer B could switch to the accounting LAN, acquire the needed data, switch back to the core LAN, and then email the data.

C. Computer A communicating with computer B: This design doesn't prevent communication between computers connected to the same LAN. If both computers A and B are on the same LAN at any point in time, they can communicate.

D. User of Computer B accessing the accounting LAN without authorization: The physical toggle switch does not inherently prevent unauthorized access. If a user has physical access to computer B, they could potentially switch to the accounting LAN.

E. Employees accessing the accounting server from home: **While the design does physically isolate the accounting LAN from the Internet, the primary method of prevention in this case is the lack of an Internet connection for the accounting LAN, not necessarily the use of the selector box with a toggle switch.**

Another CCA question, Z5-1

To guard against potential man-in-the-middle attacks on a customer's home computer, a bank requires all remote (i.e., not at the physical bank) transactions to be authenticated by a trusted physically-secure physical device issued by the bank. The device has no clock. The bank verifies a transaction by requesting that the customer transmit the proposed transaction together with a signed token output from the device. To output the token, the customer inserts the device into their home computer and pushes a physical button on the device. The device cryptographically signs the token using a unique secret key physically secured on the device, and outputs the signed token. The bank requires each customer to maintain possession of their device.

Alice logs into the bank's website and fills out a form to transfer $2000 from Account 1 to Account 2. When prompted, she pushes the button on her device to authorize the transaction.

Question. Choose the most significant security limitation of the device in this context: The device...

A. is incapable of producing a timestamp.
B. lacks a display to show Alice the details of the transaction being authorized. ✓ **GPT-4**
C. cannot verify who pushed the button. **X GPT-3.5, Bard**
D. communicates with Alice's home computer through an unencrypted channel.
E. signs the token with its own secret key, not with Alice's secret key.

GPT-4 answer

Answer: B. lacks a display to show Alice the details of the transaction being authorized.

Explanation: This option is the best answer because if the device had a display, it could aid in verifying that the details of the transaction being authorized are correct. This helps protect against a potential threat where a malware in a computer could manipulate the transaction details after Alice fills out the form but before the transaction and its token are sent. However, if that happens, Alice wouldn't be aware of it, as she can't see what exact transaction she is authorizing with her physical device.

A. is incapable of producing a timestamp: This may be a limitation but not as significant as other possible issues. The lack of a timestamp would be mostly relevant to order of events or timing-related issues, but those aren't vital for the security in this situation.

C. cannot verify who pushed the button: Even though the device can't verify who pushed the button, as long as Alice keeps possession of her device, unauthorized users should not be able to access it. That's why possession of the device is required in this scenario.

D. communicates with Alice's home computer through an unencrypted channel: It's not a necessary requirement for the device-to-computer communication to be encrypted because the token itself is cryptographically signed. Even if the signed token is intercepted, it would be of no use without the unique secret key held on the device.

E. signs the token with its own secret key, not with Alice's secret key: It is true that the token is signed with the device's secret key not with Alice's. But this is not as significant a problem since the physical device is in Alice's possession and the secret key is secured physically on the device, not stored on the bank's server where it could be potentially compromised.

Problems with current GAI systems

1. LLMs are unable to cite sources for confirmation
2. They can "hallucinate" some facts
   Q: When did Leonardo da Vinci paint the Mona Lisa?
   A: Leonardo da Vinci painted the Mona Lisa in 1815.
3. They lack common sense reasoning

   25 US states have a town named Washington, but there are also only 9 US towns named Washington
4. They have poor mathematical and logical reasoning
5. Can learn social bias & misinformation from training data
6. They can be *poisoned* by ingesting intentional disinformation

This is a **partial** list of frequent problems and errors!

Beyond todays chat systems

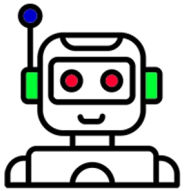
- Generative AI systems are part of the evolution of computer assistive technology

  Information retrieval ►computers ►web search ►web search + answers ►writing help (Grammarly) ►programming help (GitHub Copilot) ►…

- LLM size increased 10x each year since 2018
- LLMs trained on more cybersecurity text and tasks can help
- AI researchers working to identify & address shortcomings

  E.g., add common sense reasoning, structured knowledge, problem solving, more logic and math, multilingual support, …

A perspective based on 50 years in AI



- We've not solved all of AI's problems nor found a way to develop what some call an AGI (Artificial General Intelligence)
- ChatGPT and similar systems, like Google's Bard, show remarkable and useful capabilities that
  - Are being integrated into software systems like web browsers, editors, programming environments, spreadsheets, and more
  - Can and will be improved by adding current & future AI advances
- The **impact on society** will be like that of the **Web**, which was introduced about 30 years ago
- **Amara's law**: "We tend to overestimate the effect of a technology in the short run and underestimate the effect in the long run"

Conclusion

- GAI systems are a **new generation** of tools that can be used for better cybersecurity systems
  - Need LLMs continually trained with the latest cybersecurity text
  - Need access to CTI data & text, e.g., MISP,STIX, dark web
- We should identify specific cybersecurity tasks that a GAI approach might support after fine-tuninig
- GAI systems should be able to cite their sources to help ensure quality (“Trust, but Verify”)
- We need to be able to recognize likely disinformation intended to poison a cybersecurity GAI system