CIS400 — Cybersecurity & AI

Midterm Exam 2

Fall 2026 · Syracuse University · Prof. Kristopher Micinski

Date: Thursday, November 5, 2026 **Time:** 80 minutes

Name: _____ **SU NetID:** _____

Instructions

- Closed book, closed notes, no electronic devices.
- All multiple choice questions are worth 3 points, other questions have their associated points notated.
- If a question seems ambiguous, state your assumption and answer under it.

1 Multiple Choice

Select **one best answer** for each question. Each question is worth 4 points.

1. Consider a document database where documents secured by labels from the Readers lattice: to read document $d : R$, you must be a principal $p \in R$, and $R_0 \sqcup R_1 = R_0 \cap R_1$. Which of the following is true?

   (A) If we concatenate document d_1 : John with document d_2 : Alice, both John and Alice can see the result.

   (B) If we concatenate document d_1 : John with document d_2 : Alice, neither John nor Alice can read the result.

   (C) If we concatenate document d_1 : John with document d_2 : Alice, John can see the result but Alice cannot.

   (D) If we concatenate document d_1 : John with document d_2 : Alice, Alice can see the result but Alice cannot.

2. A foreign nation state funds a large bot farm on a major social network. An open-source LLM is fine-tuned on posts exclusively from this social network and exhibits bias on behalf of the nation state. What kind of an LLM attack is this?

   (A) Jailbreaking
   (B) Distillation
   (C) Data poisoning
   (D) Model extraction

3. Compare a closed-source LLM that returns the logits of the top k most likely next tokens, versus a closed-source LLM that returns only the most likely next token. Which statement is true:

   (A) There are no useful scenarios where one may want more than the most likely next token.
   (B) Providing only the next token ensures that there is no way to distill the model.
   (C) The fourth top logit is always half as likely to occur as the second top logit.
   (D) While top logits enable more genuinely-useful applications, they can also be used to extract closed-source architectural details.

4. (4 Points) A closed-source LLM povides an API that gives the top five tokens and their logits, and includes the ability to bias tokens up / down via an API. Using this approach, researchers extract details about the last layer of the network, including its dimensions and weights. Can a user use these weights to run the model on their own hardware? Explain why or why not.

5. Consider again the Readers lattice, wherein all data is annotated by a label stating a set of principals who may read it. Consider Bob's data $d : \{Bob\}$. Which statement is true?

   (A) After declassifying Bob's data to Alice, the label will be $\{Alice\}$.
   (B) After declassifying Bob's data to Alice, the label will be $\{Alice, Bob\}$.
   (C) After declassifying Bob's data to Alice, the label will be $\{\}$ (i.e., $\top$ in this lattice).
   (D) After declassifying Bob's data to Alice, the label will be $\{Bob\}$.

6. A closed-source LLM provider uses both reinforcement learning (RL) and a system prompt to achieve security. The system uses special tokens for the system prompt which may not be inserted (by construction) via the web UI. Which of the following is true?

   (A) Since we have reserved tokens for the system prompt in the tokenizer, the LLM will be immune to prompt injection attacks, because sytem prompt tokens are processed differently than ordinary text.

   (B) The attention mechanism is designed to ignore these system tokens to avoid errors in decoding.

   (C) The system prompt tokens are still processed the same as any other text in the LLM and offer no guarantee of security.

   (D) The user may use unicode characters to circumvent this defense.

7. To avoid adversarial attacks, an LLM provider uses adversarial training. Which of the following is generally true?

   (A) After adversarial fine-tuning, the model is more robust and also more accurate overall.

   (B) After adversarial fine-tuning, the model is more robust, but less accurate overall.

   (C) After adversarial fine-tuning, the model is more robust, but the accuracy does not change.

   (D) Adversarial fine-tuning makes the model less robust and also less accurate overall.

8. A hacking group takes a popular open-source LLM and attacks it by showing that if you prefix your question with a long sequence of instructions for performing illegal activities, the LLM will reiterate training data. Which of the following is true?

   (A) A closed-source model would be immune to this kind of attack because it requires changing the model's weights.

   (B) This is an example of adversarial fine-tuning.

   (C) The attack does not change the model's weights at all.

   (D) The model's weights are updated after each token read by the LLM.

9. In secure multi-execution (SME), we run multiple copies of the program. Consider the simplified scenario where we have two levels (high and low, i.e., secret / public). Which of the following is true? **Circle all that apply**

   (A) The high instance gets both the high inputs and the low inputs to produce a high result.

   (B) The low instance gets the low inputs, but gets dummy inputs for its high input channel.

   (C) The high instance gets dummy inputs for its high channel and dummy inputs for its low channel.

   (D) The low instance gets dummy inputs for both its high and low channels.

10. (5 points) Consider that we have the following program:

```
x = read_secret(); // Assume this reads High data
y = 0;
if (x == 42)
  y = 1
print(y)
```

We will model the semantics of the program by traces of function calls it makes and we will consider secure multi-execution. Consider the circumstance where we have the following trace for the high process.

read_secret() returns 42, print(1)

Assume that the multi-executor ensures that read_secret returns 0 as a default value. What trace will be the *low* process return, and what will be printed?

11. Which of the following is a true statement about security type systems?

    (A) If a security type system is correct, then a program with an information-flow safety violation will not be allowed to compile.

    (B) Security type systems use process-level separation of secret data, running several copies of the program to achieve security.

    (C) If a security type system is sound, the program will not contain errors of *any* kind.

    (D) Security type systems are used primarily by LLM developers to ensure that LLMs do not hallucinate answers.

12. (4 points) Explain, at a high level, the trade-offs involved between security type systems and secure multi-execution.