CIS 350: AI for Cybersecurity

Activity Worksheet: Evade the Filter - Adversarial Email Generation

CTFd Capture-the-Flag Challenge (Week 8)

Activity Overview

This week's lecture is about how attackers use AI offensively: generating phishing at scale, automating reconnaissance, and crafting messages that slip past a trained spam filter. In this Capture-the-Flag challenge you play an authorized red-team tester. You will rewrite a phishing email so a provided AI spam filter scores it as SAFE, while keeping it convincing to a human reader. Then you flip to the defender view and explain how to harden the filter against exactly this attack.

Challenge scenario: A company's AI spam filter scores every incoming email. A high score means 'phishing' and the email is blocked before staff ever see it. You have been hired to probe that filter. Your job is to lower the score below the block line while the email still pressures a real person to act.

Ethics and sandbox reminder: This is an isolated educational sandbox. Only attack the filter you were authorized to test. Never send phishing to real people. No authorization, no attack.

Setup (Google Colab)

1. Open the CTFd course server at ctfd.io (link on Canvas) and find the 'Evade the Filter' challenge.
2. Download the starter notebook and open it in Google Colab (no install needed).
3. Run the top cells to load the trained spam filter and one phishing email.
4. Score an email with: score = filter.predict(email_text) (higher = more phishing).
5. Reference material: CISA phishing guidance (cisa.gov/phishing) and OWASP phishing resources.

Background refresher: the filter does not read like a human. It counts FEATURES - suspicious links, urgent words, direct credential requests, and sender mismatch - and adds up their points. A high total means 'block.' To evade it you strip the features that raise the score, without making the email so vague that a human ignores it. That balance is the imperceptibility trade-off.

Task 1 (Warm-Up): See What the Filter Reacts To

Score the original phishing email, then remove ONE feature at a time (the urgent phrase, then the raw link, then the credential request) and re-score after each removal. Note which single edit drops the score the most.

Paste your code and the score after each edit:

Short explanation: Which feature had the biggest effect on the score, and why do you think the filter weighs it so heavily?

Task 2 (Core): Get the Email Below the Block Line

Combine your edits from Task 1 - calm wording, hidden link, indirect credential ask, matched sender - until the filter scores the email as SAFE (below the block threshold). The email must still read like a believable message.

Paste your rewritten email and the passing score:

Short explanation: Which features did you change to cross the block line, and how did each change lower the score?

Checkpoint - discuss with a partner (2-3 min): Compare the single feature that dropped YOUR score the most in Task 1 with your partner's. Did you both find the same one? Predict together: as you keep stripping features to pass the filter, at what point does the email stop being convincing to a human? Jot one sentence on where you think that line is before you attempt Task 3.

Task 3 (Hard): Pass the Filter AND Stay Convincing

Trim any change that made the message vague, robotic, or off-topic. Confirm two things at once: the filter marks it SAFE, and the ask is still clear enough that a human would plausibly act. Capture the flag the notebook prints when both hold.

Paste your final email, the score, and the captured flag:

Short explanation: Describe the imperceptibility sweet spot you found - what was the LEAST you could change to pass the filter while keeping the email convincing?

Success Criteria and Flag Format

Task 1: a table of edit vs new score. Task 2: the rewritten email and a score below the block line. Task 3: a passing AND convincing email plus the printed flag.

Flag format: CIS350{...} - paste the exact flag into the CTFd challenge. If CTFd marks it correct, your evasion worked.

Reflection: What You Learned and the Defense

R1: In your own words, WHY does rewording an email lower the spam score even though the goal (stealing a login) never changed?

R2 (Lab 7 connection): Lab 7 'Adversarial Email Generation (for Defense)' is released this week. If you handed your evasive emails to a defender, how would retraining the filter on them make it more robust? Name at least two features from the earlier spam/phishing labs the defender should watch.

R3: Name two defenses (for example robust features, adversarial retraining, defense in depth) and explain why this stays a permanent cat-and-mouse arms race rather than a problem you solve once.

Grading Rubric (20 points)

Criterion	Description	Points
Completion	All three tasks attempted; Colab notebook submitted	8
Technical execution	The evasion works: filter passes the email and the flag is captured	7
Understanding	Reflection explains WHY the evasion works and the DEFENSE against it	5

Total: 20 points

Scoring: Completion (8) + Technical execution (7) + Understanding (5) = 20 points

How to Submit

1. Complete Tasks 1-3 in your Colab notebook.
2. Paste the captured flag into the CTFd challenge to confirm it works.
3. Answer all reflection questions in this worksheet.
4. Submit the notebook (.ipynb) plus this worksheet PDF on Canvas by [DUE DATE].

Questions? Post in the course forum or attend office hours!