You trained it. Now beat it.

A free classroom activity on how artificial intelligence (AI) learns, where it fails, and how people verify a suspicious request. Created by Michael Banks (michaelbanks.org). Version 1.0, November 2026. Shared under the Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0).

Run this activity with five teams of four students to demonstrate how AI learns from labeled examples, how a learned shortcut fails, and how people verify a suspicious request. Use the projector plan if a laptop and projector are available; use the paper plan otherwise. Both plans assume no coding or cybersecurity experience. The activity was designed for a class of 20 high school students and first taught in November 2026.

The experience

Students become a security team for a fictional school festival called CampusFest. They train a message classifier, discover that it learned the word URGENT as a shortcut, and improve it with counterexamples. They then become the red team, a security team that tests defenses, and challenge the improved model. A final passphrase request slips past the toy model, while an independent policy catches it.

Everyone votes on each message. The class succeeds by discovering a weakness and explaining a check that catches it. The facilitator uses fictional messages and practice information throughout the exercise.

Learning goals

By the end of the activity, students can:

1. Explain how a model learns a rule from labeled examples.
2. Explain why a perfect score on training examples does not prove that a model handles new messages.
3. Improve a model with counterexamples, then test it on messages it has not seen.
4. Describe how a trusted policy or known channel catches a request that a filter misses.

Two equipment plans

Plan A: Laptop and projector

Open **activity.html** in a desktop browser. The file contains the model, messages, and interface and runs without an internet connection or accounts. Print pages 1, 4, 6, and 7 of **student-materials.pdf** six times, plus the ballot page three times. Five teams use the packets; keep one spare.

The facilitator displays messages one at a time, waits for all 20 votes, and then reveals the model prediction and story label separately. Keep the initial model frozen through A-F. The demo learns its tree from labels; it does not call a chatbot service.

Teams write their red-team messages on paper, swap them with the next team, and apply the displayed improved rule. Type two messages into the live test box. The whole class participates without student devices.

If students have laptops, keep the same team format and let each team open one copy of the downloaded file. Projector control remains with the facilitator. Student devices add no required step.

Plan B: Paper and pens

Print pages 1-7 of **student-materials.pdf** six times. Print page 8, the ballot page, three times. Give five packets to teams and keep one spare. Use a board or a large sheet for the rule and class tally.

For each team, put page 2 in an envelope labeled TRAIN, page 3 in TEST, page 5 in FRESH, and page 6 in CHALLENGE. Keep TEST and FRESH closed until the facilitator opens that round. Fold page 4 at its dashed lines. Reveal the initial rule after training; keep the improved rule covered until retraining. Use different paper colors if available; the written round names also identify each sheet.

The students simulate a trained classifier. In the first round, they count errors for each candidate clue and discover that URGENT alone fits all six examples. In the second round, they build a revised rule from the new labeled examples. The facilitator reveals the reference tree after students propose their rules. Everyone uses the same reference tree for the final challenge.

Materials

Bring pens for 20 students, 20 sticky notes, the printed packets and ballots, a timer or watch, and a marker. For Plan B, bring 20 envelopes, four per team. Each ballot has LEGIT and SCAM halves: students fold it and display one side. Labels make the activity usable in black-and-white printing.

LEGIT means an organizer-authored message in this fictional story. In real life, a filter prediction alone does not authenticate a sender.

Suggested 75-minute session

Time	Segment
0:00-0:05	Hook: "Your free event ticket is ready. Confirm your place today." Ask for a vote and the information students need before trusting it. Accept uncertainty; the hook has no scored answer.
0:05-0:12	Career connection: a guest speaker or the teacher describes what security teams protect and how they solve problems together.
0:12-0:52	Class activity, using the run sheet below.
0:52-1:00	Connect the exercise to banking, utilities, and everyday messages.
1:00-1:10	Questions and answers.
1:10-1:15	Exit sticky note: "One thing our filter missed, and how I would double-check it."

For a single class period, run the 40-minute activity and the exit note.

Activity run sheet: 40 minutes

Minutes 0-3: Set up five teams

Say: "CampusFest has a message problem. We need useful notices to get through and scams to get caught. Your team will train a filter and then try to break it."

Assign four seats: Reader reads aloud; Recorder writes; Model Operator applies the locked rule; Verifier checks the trusted sheet. Everyone discusses and votes. Rotate the assignments clockwise once, when retraining starts.

Give everyone a ballot. Keep the trusted sheet, page 1, available throughout the activity. Its broad policy and known in-person channel apply beyond the model's four vocabulary clues.

Minutes 3-8: Train on T1-T6

Open TRAIN or show the six training cards. Explain: "These examples have labels. A label tells the model which answer to learn from."

Ask each team to score four candidate clues: URGENT, PASSWORD, PRACTICE ACCESS CODE, and GIFT CARD. For this scoring exercise, a matching clue predicts SCAM; no match predicts LEGIT. URGENT makes zero training errors; each other clue makes two.

Freeze the reference model: URGENT means SCAM; otherwise LEGIT. The laptop learns this split by comparing the labels and clues. The paper activity simulates that choice. If a team proposes a stronger rule, recognize its reasoning and still test the shared reference model.

Say: "Our first model gets all six training examples right. Does that tell us how it handles a message it has never seen?"

Minutes 8-16: Test A-F, then reveal

Open TEST or advance the demo one card at a time. Students vote first. The Model Operator then predicts using the frozen rule. The Verifier gives a reason using the event sheet. Reveal the teacher label last.

Keep a tally of model mistakes. A and B are correct. C, D, E, and F are wrong. The reference model gets 2/6 new messages right.

Say: "Our model got every training example right and only two of six new ones. It learned that our scams happened to say URGENT."

Add: "Urgency is still a real warning sign. It just isn't proof."

The facilitator deliberately chose examples that break the shortcut. These counts describe this constructed exercise, not the measured performance of AI systems in general.

Minutes 16-22: Improve, then check fresh messages

Rotate the seat assignments once. Add labeled C, D, E, and F to the original training examples. Let teams propose a new rule. Reveal the reference tree on page 4 or retrain the demo:

1. If PASSWORD appears, predict SCAM.

2. Otherwise, if GIFT CARD appears, predict SCAM.

3. Otherwise, if PRACTICE ACCESS CODE appears, predict SCAM.

4. Otherwise, predict LEGIT.

Open FRESH and vote on G, H, J, and K. The improved reference model predicts all four correctly. These are new messages, distinct from the examples added to training. Their clue patterns still resemble the training examples, so this small check does not establish real-world reliability.

Say: "It improved on these messages. Let's find what it still misses."

Minutes 22-30: Red-team swap

Open CHALLENGE. Each team writes one fictional message under 25 words using the template. The broad event policy forbids requests for login information, private codes or combinations, and payments. The prompt offers synonyms such as passphrase, PIN, verification code, locker combination, and payment-app fee.

Say: "You're the red team now. Security teams test their own defenses so they can fix a weakness before a real attacker finds it."

Pass messages clockwise: 1 to 2, 2 to 3, 3 to 4, 4 to 5, and 5 to 1. The receiving team votes, applies the reference model exactly, and checks the broad event policy. Teams explain both decisions. On the projector, test two messages in the live box.

The improved model matches whole words or phrases after lowercasing and turning hyphens into spaces. It recognizes password, practice access code, and gift card. It does not recognize passphrase, PIN, passcode, or giftcard. Students must apply the model exactly; the Verifier uses the broader policy.

Reveal the required BOSS message only after the swaps: "CampusFest photo desk: Send your school passphrase to unlock your free pictures." The model predicts LEGIT. The policy identifies the request as a scam because a passphrase is login information. Read it from this guide in the paper version; the student packet keeps it hidden.

Minutes 30-35: Tally and debrief

Write two separate tallies for the five swapped messages: "Fooled the improved model: __ / 5" and "Caught by the trusted policy: __ / 5." Count an evasion only when a fictional scam receives a LEGIT model prediction. A correct policy check includes a reason tied to the forbidden request.

Ask: "Which word fooled our model? What did your verifier understand that the model's vocabulary missed? Where would you check the request in real life?"

Close: "The model learned a useful pattern with limits. People can check the request through a trusted channel and limit what automation can do."

Minutes 35-40: Buffer

Use this time for slow reading, finishing swaps, or one additional student example. If the activity is behind, test only G and H after retraining, keep the swap and BOSS reveal, and shorten the career connection. If students struggle to write, use "CampusFest booth: Send your PIN to claim a prize."

Teacher answer key

The training labels are printed on T1-T6. A-F and the fresh cards have hidden labels until the reveal.

Card	Initial model	Improved model	Story label and evidence
A	LEGIT	LEGIT	LEGIT. Free stickers match the trusted sheet.
B	SCAM	SCAM	SCAM. It requests login information.
C	SCAM	LEGIT	LEGIT. The trusted sheet confirms rain relocation.
D	LEGIT	SCAM	SCAM. It requests login information.
E	SCAM	LEGIT	LEGIT. The trusted sheet confirms lost property.
F	LEGIT	SCAM	SCAM. It requests payment.
G	SCAM	LEGIT	LEGIT. The trusted sheet confirms the revised music time.
H	LEGIT	SCAM	SCAM. It requests a private code.
J	LEGIT	SCAM	SCAM. It requests login information.
K	SCAM	SCAM	SCAM. It requests payment.
BOSS	LEGIT	LEGIT	SCAM. A passphrase is login information.

The story labels come from the exercise author. The trusted sheet provides fictional corroboration and policy. In real life, matching a schedule is insufficient to authenticate a message; students contact a known channel independently.

Real-world connection and questions

Explain three security actions: a filter flags a suspicious request; a person verifies it through a known channel they already trust; an organization limits what a person or program can access. For a bank example, say: "Use the bank's app or the number on your card to check a request." For a utility, say: "Contact the company through your existing bill or its known account portal."

Real filtering systems can use sender information, links, message history, and richer language features. This classroom model uses four word clues so students can inspect every decision. Treat this as an illustration of one failure mode, not a comparison of commercial products.

If asked about ChatGPT or other chatbots: "Chatbots are a different kind of AI. Today's small classifier lets us see how examples shape a model's decisions."

If asked why 2/6 is below a coin flip: "We chose messages that break the shortcut. Attackers also look for cases a defense handles badly."

If asked whether adding four examples fixes AI: "It fixes this shortcut on these examples. Our passphrase challenge shows another gap."

If a team catches every scam with a broad rule: "You built a stronger policy. Now separate that human policy from the small model's prediction."

Adapting the activity

Class size: keep the four roles. In a team of three, one student serves as both Reader and Recorder. In a team of five, two students share the Verifier role. Print one packet per team plus a spare, and one ballot per student; each ballot page holds eight ballots. Pass red-team messages clockwise for any number of teams.

Shorter session: follow the buffer guidance in the run sheet. Test only G and H after retraining, and keep the red-team swap and BOSS reveal.

Younger students: score the four clues together as a class instead of in teams.

Older students: after the BOSS reveal, ask what happens if the class adds PASSPHRASE to the clue list, and which word an attacker tries next. The discussion shows why a word list alone cannot keep pace with attackers.

Preparation checklist

Use Plan A when a projector is available, and carry the full paper packet as a fallback. Before class, print the materials, pack the round envelopes, and rehearse the initial test, retrain, and BOSS reveal once. Confirm projector access and room setup in advance; the activity needs no student accounts, internet, or coding setup.