CIS 350: AI for Cybersecurity

Week 15 Activity: Domain Adaptation with Limited Labels

CTFd Capture-the-Flag Challenge (Hard)

Activity Overview

Your Security Operations Center runs a malware detector that was trained on 2023 data and scored 96 percent accuracy at deployment. In 2025, it is missing new malware families because the data distribution has shifted. You have only 500 labeled 2025 samples, and labeling is expensive. Your mission is to use transfer learning to adapt the old model to the new domain, measure whether it improved, re-calibrate its confidence, and capture the flag.

Ethics reminder: This is a sandboxed, authorized educational exercise. Only attack or adapt systems you are explicitly permitted to use. The model and data live on the course CTFd server.

Setup: Google Colab

Resource: CTFd course server (ctfd.io) plus the course lab materials on model generalization.

1. Open the Week 15 challenge card on CTFd and click the starter notebook link.
2. The notebook opens in Google Colab - everything runs in your browser, no install needed.
3. Run the first cells to load the pre-trained model and the 2025 labeled set:

```
model = load_pretrained('detector_2023.pt')
X25, y25 = load_2025_labeled(n=500)
```

Warm-Up (5 min): Predict Before You Measure

Before running anything, think-pair-share with a partner. The 2023 model scored 96 percent at deployment. In one or two sentences, PREDICT: what accuracy will it get on the 2025 samples, and WHY? Name one concrete thing about 2025 malware (new packers, new APIs, new behaviors) that the 2023 model never saw. You will check your prediction against the real number in Task 1.

My prediction (2025 accuracy and one reason):

Task 1 (Easiest): Measure the Domain Shift

Evaluate the frozen 2023 model on the 500 labeled 2025 samples. Compare the result to the trusted 2023 baseline accuracy (96 percent). A large gap confirms domain shift.

Paste your code and the accuracy output:

Explain (2-3 sentences): What is the 2025 accuracy, and how big is the gap from 96 percent? Where does the model fail most, and why does this count as domain shift?

Task 2 (Medium): Fine-Tune on the 500 Samples

Keep the model's learned features (freeze the base layers) and retrain the classifier head on the 500 samples. Split the 500 into a training slice and a validation slice so your measurement is honest. Use a small learning rate to protect the transferred features.

Paste your fine-tuning code and the before/after accuracy on the 2025 validation slice:

Explain (2-3 sentences): How much did out-of-domain accuracy improve after fine-tuning? Why does freezing the base and retraining only the head work with so few labels?

Checkpoint - discuss with a partner (2 min): Your accuracy jumped, but should you trust the model now? Predict one way the fine-tuned model could still mislead you even though its accuracy looks good. (Hint: think about how SURE it claims to be.) Task 3 tests your prediction.

Task 3 (Hardest): Re-Calibrate and Capture the Flag

Check calibration on a held-out 2025 slice (not on 2023 data). Compare the model's average confidence to its actual accuracy. If it is overconfident, apply a simple fix such as temperature scaling so the stated probability matches real accuracy. When accuracy AND calibration hit the target, CTFd reveals the flag.

Paste your calibration code, the before/after confidence-vs-accuracy numbers, and the flag:

Flag format: flag{...} Write your captured flag here: _____

Explain (2-3 sentences): Was the fine-tuned model overconfident before calibration? How did you know, and what did temperature scaling change?

Success Criteria and Flag Format

You have captured the flag when all three targets are met:
- Task 1: you reported the 2023-vs-2025 accuracy gap with real numbers.
- Task 2: fine-tuned 2025 accuracy is clearly higher than the frozen model.
- Task 3: confidence is calibrated - stated probability matches real accuracy.

The flag has the form flag{...} and is revealed on the CTFd challenge page only when the tuned accuracy and calibration both reach the target. Submit it in the CTFd submission box.

Reflection

R1: In plain language, WHY does fine-tuning a pre-trained model work better than training a new model from scratch on only 500 samples?

R2: Why can a model be highly confident and still wrong after domain shift? Why must confidence be re-calibrated on the NEW domain rather than the old one?

R3 (the DEFENSE): You are the SOC lead. Describe a defense/process that keeps a deployed detector healthy as the threat landscape shifts (for example: monitor accuracy on fresh data, budget for periodic labeling, re-check calibration, retrain on a schedule). Explain why your defense works.

R4: Few-shot learning is powerful but risky in security. Give one situation where learning from a handful of samples is necessary, and one where it could dangerously mislead you.

Grading Rubric (20 points)

Criterion	Description	Points
Completion	All three tasks attempted; Colab notebook submitted	8
Technical Execution	Fine-tuning works and produces the expected result (flag captured)	7
Understanding	Reflection explains WHY it works and the DEFENSE against the attack	5

Total: 20 points

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

How to Submit

1. Complete all three tasks in your Colab notebook.
2. Submit the flag on the CTFd challenge page.
3. Save your notebook as a PDF (File -> Print -> Save as PDF).
4. Answer all reflection questions in this worksheet.
5. Submit this worksheet PDF and your notebook PDF on Canvas by [DUE DATE].

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