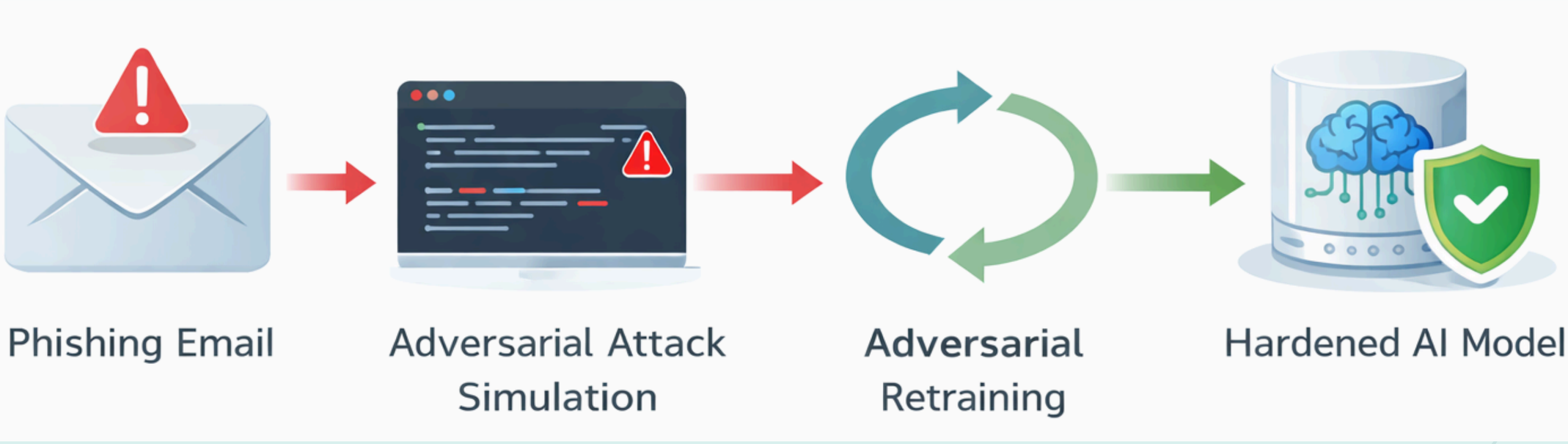
ADVERSARIAL RESILIENCE IN AI PHISHING DETECTION

How semantic attacks bypass AI and how we stopped them

We evaluate how semantic attacks bypass AI phishing detection and demonstrate how adversarial retraining hardens the model.



Baseline Model

99% Accuracy

Vulnerable to semantic attacks

After Adversarial Retraining

~100% Block Rate

Semantic attacks neutralized

THE PROBLEM

- Despite **99%** accuracy, the baseline model fails under semantic attacks.
- Polite corporate language reduces detection confidence from **92%** to **0.54%**.

INPUT DATA

- Raw email files (ham / spam)
- Folder-based labeling (no CSV)
- Enron Email Dataset

OUR APPROACH (SOLUTION & KEY FEATURES)

- Simulate realistic phishing attacks on an AI phishing detector (DistilBERT)
- Identify semantic and token-level weaknesses through adversarial testing
- Retrain the model using adversarial examples to improve robustness
- Compare performance before and after hardening

RESULTS

- Baseline accuracy: **99%** (vulnerable)
- After retraining: **~100%** block rate
- Semantic attacks neutralized
- No degradation in normal performance

Course: AI-Driven Cybersecurity
Instructors: Dr. Andrei Kojukhov, Mr. Viacheslav (Slava) Nefedov
Team: Adi Haim, Afik Aharon, Bar Mor, Ron Noiman

מכון טכנולוגי חולון
Holon Institute of Technology