Paper Coding Instructions

Quick Reference Guide

Overview

You are coding papers that have been adopted by industry artifacts (tools, benchmarks, regulations). Each paper takes 20–30 minutes.

Step 1: Paper Selection

Papers come from these artifacts (you don't select randomly):

1. **Tools:** IBM ART, CleverHans, Foolbox, TextAttack, PyRIT
2. **Benchmarks:** RobustBench, AutoAttack, HarmBench
3. **Regulatory:** MITRE ATLAS, NIST AI RMF, OWASP LLM Top 10
4. **Vendor:** AWS, Azure, GCP, OpenAI, Anthropic security docs

For each artifact, find techniques with paper citations → those papers are your sample.

Step 2: Record Metadata

For each paper, record:

- Title
- Authors
- DOI or arXiv ID
- Publication date (earliest of arXiv or venue)
- Venue name
- Technique name (e.g., "FGSM", "GCG", "PGD")

Step 3: Code the Paper

Research Characteristics (G1–G7)

Code	Values	How to Decide
G1	Attack / Defense / Evaluation	What is the main contribution?
G2	Evasion / Poisoning / Privacy / N/A	N/A if defense. Evasion = test-time. Poisoning = training-time. Privacy = data leakage.
G3	Vision / NLP / Malware / Audio / Tabular / LLM / Cross-domain	What data/models are tested? Cross-domain = 2+ domains.
G4	ML / Security / Journal / arXiv-only	See venue list below.
G5	Yes / No	Does code exist NOW? Check paper + GitHub.
G6	At-pub / Post-pub / Never	At-pub = within 1 month of paper date.
G7	2014–2025	Year of earliest public version.

Venue Classification:

- **ML:** NeurIPS, ICML, ICLR, CVPR, ECCV, ICCV, ACL, EMNLP, NAACL
- **Security:** IEEE S&P, ACM CCS, USENIX Security, NDSS, IEEE SaTML
- **Journal:** TPAMI, TIFS, TDSC, Pattern Recognition, etc.
- **arXiv-only:** Never published at peer-reviewed venue

Threat Model (T1–T2) – Attack Papers Only

Code	Values	How to Decide
T1	White / Gray / Black	White = uses target model gradients. Gray = uses surrogate gradients. Black = queries only, no gradients.
T2	Yes / No	Are gradients computed at ANY stage?

Leave T1 and T2 blank for defense papers.

Practical Evaluation (Q1–Q3)

Code	Values	How to Decide
Q1	Yes / Partial / No	Yes = tested on production system (API, deployed car). Partial = realistic simulation. No = CIFAR/ImageNet only.
Q2	Yes / No	Does paper report FLOPs, runtime, or query count?
Q3	Yes / No / N/A	N/A for defenses. Yes = tests against adaptive defenses or AutoAttack.

Step 4: Record Adoption Events

For each artifact where this paper appears, record:

- Artifact name (e.g., "IBM ART")
- Artifact type (Tool / Benchmark / Regulatory / Vendor)
- Adoption date (Git commit date, or document date)

Step 5: Calculate Adoption Lag

$$AdoptionLag(months) = \frac{AdoptionDate - PublicationDate}{30}$$

For papers in multiple artifacts, record:

- First adoption lag (minimum)
- Regulatory adoption lag (if MITRE/NIST/OWASP)

Quick Decision Rules

1. **Paper has both attack and defense?** Code G1 based on which has more experiments.
2. **Can't find threat model section?** Look for "Adversary Capabilities" or "Assumptions."
3. **Unsure if White/Gray/Black?** If paper says "transfer attack," it's Gray. If "query-based," it's Black.
4. **LLM jailbreak with no gradients?** T1 = Black, T2 = No.
5. **Paper tested on "Google Cloud Vision API"?** Q1 = Yes (production system).
6. **Paper tested on "ImageNet"?** Q1 = No (standard benchmark).

Example: Coding FGSM Paper

Paper: Goodfellow et al., "Explaining and Harnessing Adversarial Examples," ICLR 2015

Variable	Value
G1: Type	Attack
G2: Threat	Evasion
G3: Domain	Vision
G4: Venue	ML
G5: Code	Yes
G6: Timing	At-pub
G7: Year	2015
T1: Access	White
T2: Gradient	Yes
Q1: Real-world	No
Q2: Cost	No
Q3: Defense-aware	No
Adoption	CleverHans (Oct 2016)
Lag	22 months

Data Recording Template

Create a spreadsheet with these columns:

```
paper_id, title, authors, venue, pub_date, technique_name,
G1, G2, G3, G4, G5, G6, G7, T1, T2, Q1, Q2, Q3,
artifact_name, artifact_type, adoption_date, adoption_lag_months
```