Paper Coding Instructions

1 How to Select Papers

You do NOT randomly pick papers. You work backwards from industry artifacts.

1.1 Step 1: Go to Each Artifact and Find Papers They Cite

Artifact	Where to Look	What to Extract
CleverHans	GitHub README, docstrings	Paper citations for each attack
IBM ART	Documentation, attacks/ folder	Paper citations for each technique
Foolbox	Docs, source code	Paper citations
TextAttack	Docs, source code	Paper citations
PyRIT	Docs, source code	Paper citations
RobustBench	Leaderboard entries	Defense papers on leaderboard
AutoAttack	The AutoAttack paper	Component attack papers
HarmBench	HarmBench paper/GitHub	Red-teaming method papers
MITRE ATLAS	atlas.mitre.org technique pages	"References" on each page
NIST AI RMF	The PDF document	Bibliography
OWASP LLM Top 10	The PDF	References for each vulnerability

1.2 Step 2: For Each Paper You Find, Record It

If CleverHans implements FGSM and cites Goodfellow et al. 2015, that paper goes in your dataset.

Your ∼100 papers = all papers cited by these artifacts.

2 The 12 Coding Columns

2.1 Group 1: Basic Info (G1–G7)

Col	Question	Options	How to Decide
G1	Is this an attack, defense, or evaluation?	Attack / Defense / Evaluation	Read abstract. What did they build?

Col	Question	Options	How to Decide
G2	What type of attack?	Evasion / Poisoning / Privacy / N/A	**Evasion** = fool model at test time. **Poisoning** = corrupt training data. **Privacy** = steal data/model. **N/A** = defense papers
G3	What domain?	Vision / NLP / Malware / Audio / Tabular / LLM / Cross-domain	What data did they test on? ImageNet = Vision. Text = NLP. ChatGPT = LLM.
G4	Where published?	ML / Security / Journal / arXiv-only	**ML** = NeurIPS, ICML, ICLR, CVPR, ACL. **Security** = S&P, CCS, USENIX, NDSS.
G5	Is code available NOW?	Yes / No	Google "paper name github". Is there code?
G6	When was code released?	At-pub / Post-pub / Never	**At-pub** = within 1 month of paper. **Post-pub** = later.
G7	Publication year	2014–2025	Year of first public version

2.2 Group 2: Threat Model (T1–T2) — Attack Papers Only

Col	Question	Options	How to Decide
T1	How much model access?	White / Gray / Black	**White** = has weights/gradients. **Gray** = surrogate model. **Black** = queries only
T2	Uses gradients?	Yes / No	If they compute ∇L anywhere, it's Yes

Leave T1 and T2 blank for defense papers.

2.3 Group 3: Practical Evaluation (Q1–Q3)

Col	Question	Options	How to Decide
Q1	Tested on real system?	Yes / Partial / No	**Yes** = Google API, Tesla, ChatGPT. **Partial** = realistic sim. **No** = CIFAR/ImageNet only
Q2	Reported cost?	Yes / No	Did they say "X queries" or "Y seconds"?
Q3	Tested against defenses?	Yes / No / N/A	**Yes** = tested vs. adversarial training. **N/A** = for defense papers

3 Adoption Tracking

For each paper, also record:

Column	What It Means
Artifact	Which artifact cites this paper (e.g., "CleverHans", "MITRE ATLAS")
Artifact Type	Tool / Benchmark / Regulatory / Vendor
Adoption Date	When did artifact add this? (Git commit date or document date)
Adoption Lag	Months between paper publication and artifact adoption

4 Concrete Example: FGSM Paper

You find: CleverHans GitHub cites "Explaining and Harnessing Adversarial Examples" (Goodfellow 2015) for FGSM.

You code:

Column	Value	Why
G1	Attack	Paper proposes an attack
G2	Evasion	Fools classifier at test time
G3	Vision	Tested on MNIST, ImageNet
G4	ML	Published at ICLR
G5	Yes	Code exists on GitHub
G6	At-pub	Released with paper
G7	2015	Published 2015
T1	White	Uses target model gradients
T2	Yes	Gradient-based attack
Q1	No	Only tested on benchmarks
Q2	No	No runtime/query count reported
Q3	No	No defense evaluation
Artifact	CleverHans	Where you found it
Artifact Type	Tool	It's a library
Adoption Date	Oct 2016	First CleverHans commit with FGSM
Adoption Lag	22 months	Oct 2016 − Dec 2014

5 Your Workflow

1. **Pick an artifact** (e.g., IBM ART)
2. **Find all papers it cites** (from docs, code, README)
3. **For each paper:** fill in G1–G7, T1–T2, Q1–Q3, plus adoption info
4. **Repeat** for all 11 artifacts
5. **Remove duplicates** (same paper cited by multiple artifacts = one row, multiple adoption events)

6 Spreadsheet 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
```