Paper Coding Instructions

1 Artifact Selection Criteria

Papers are selected based on adoption by the following artifacts:

1.1 Tools (Criterion: ≥1,000 GitHub stars)

Tool	Stars	Description
CleverHans	6,401	Adversarial example library (Toronto)
IBM ART	5,789	Adversarial Robustness Toolbox
TextAttack	3,348	NLP adversarial attacks
PyRIT	3,343	LLM red-teaming (Microsoft)
Foolbox	2,936	Adversarial attacks (Bethge Lab)

1.2 Benchmarks (Criterion: Peer-reviewed publication)

Benchmark	Venue	Description
RobustBench	NeurIPS 2021	Adversarial robustness leaderboard
AutoAttack	ICML 2020	Standardized attack evaluation
HarmBench	ICML 2024	LLM jailbreak evaluation

1.3 Regulatory (Criterion: Industry threat framework)

Framework	Description
MITRE ATLAS	Adversarial ML tactics/techniques (like ATT&CK)

1.4 Not Included (Manual Extraction Required)

Framework	Reason
NIST AI 100-2	PDF document, not machine-readable
OWASP LLM Top 10	PDF document, not machine-readable

2 Paper Selection Criteria

71 papers selected for coding:

1. **61 papers** cited by 2+ artifacts (strongest adoption evidence)
2. **10 papers** from MITRE ATLAS only (regulatory adoption)

3 The 12 Coding Columns

3.1 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?
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

3.2 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.

3.3 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

4 Pre-filled Columns (Auto-extracted)

The following columns are already filled:

Column	Description
arxiv_id	arXiv identifier
paper_title	Full paper title
paper_authors	Author names
paper_pub_date	Publication date (YYYY-MM-DD)
found_in_artifacts	Which artifacts cite this paper
num_artifacts	How many artifacts cite it
first_adoption_date	When first artifact added it
first_adoption_artifact	Which artifact adopted first
adoption_lag_months	Months from paper to adoption
selection_reason	Why paper was selected

5 Your Workflow

1. Open `papers_for_coding_71.csv`
2. For each paper, fill in: G1, G2, G3, G4, G5, G6, G7, T1, T2, Q1, Q2, Q3
3. Use arXiv link to read the paper: `https://arxiv.org/abs/[arxiv_id]`
4. Estimated time: 20–30 minutes per paper

6 Files

File	Purpose
papers_for_coding_71.csv	71 papers for manual coding
papers_all.csv	Full 277 papers (for statistics)
papers_by_artifact.csv	Sorted by artifact (for verification)