

# Paper Coding Instructions

## 1 Artifact Selection Criteria

Papers are selected based on adoption by the following artifacts:

### 1.1 Tools (Criterion: $\geq 1,000$ 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)

## 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 9 Coding Columns

#### 3.1 Group 1: Research Characteristics (G1–G6)

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	<b>Evasion</b> = fool model at test time. <b>Poisoning</b> = corrupt training data. <b>Privacy</b> = steal data/model. <b>N/A</b> = 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	<b>ML</b> = NeurIPS, ICML, ICLR, CVPR, ACL. <b>Security</b> = 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	<b>At-pub</b> = within 1 month of paper. <b>Post-pub</b> = later.

#### 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	<b>White</b> = has weights/gradients. <b>Gray</b> = surrogate model. <b>Black</b> = queries only
T2	Uses gradients?	Yes / No	If they compute $\nabla L$ anywhere, it’s Yes

Leave T1 and T2 blank for defense papers.

#### 3.3 Group 3: Practical Evaluation (Q1)

Col	Question	Options	How to Decide
Q1	Tested on real system?	Yes / Partial / No	<b>Yes</b> = Google API, Tesla, ChatGPT. <b>Partial</b> = realistic sim. <b>No</b> = CIFAR/ImageNet only

## 4 Pre-filled Columns (Auto-extracted)

The following columns are already filled:

Column	Description
selection_reason	Why paper was selected (2+ artifacts / MITRE ATLAS)
is_aml_paper	Always YES for this dataset
arxiv_id	arXiv identifier
found_in_artifacts	Which artifacts cite this paper
num_artifacts	How many artifacts cite it
paper_title	Full paper title
paper_authors	Author names
paper_pub_date	Publication date (YYYY-MM-DD)
first_adoption_date	When first artifact added it
first_adoption_artifact	Which artifact adopted first
all_adoptions	All adoption events (artifact:date pairs)
adoption_lag_months	Months from paper to first adoption

## 5 Your Workflow

**Dual coding:** Both coders independently code all 71 papers. Do NOT discuss papers until both are done.

1. Open `extraction_runs/run1/papers_for_coding_71.csv`
2. For each paper, fill in: G1, G2, G3, G4, G5, G6, T1, T2, Q1
3. Use arXiv link to read the paper: [https://arxiv.org/abs/\[arxiv\\_id\]](https://arxiv.org/abs/[arxiv_id])
4. Estimated time: 15–20 minutes per paper
5. After both coders finish: compare codes, resolve disagreements, compute Cohen’s  $\kappa$

## 6 Files

All files are in `extraction_runs/run1/`:

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