

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	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.

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)

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

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 verify all 71 papers. Do NOT discuss until both are done.

1. Open `extraction_runs/run1/papers_coded_verified.csv`
2. For each paper, verify/correct: 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: 5–10 minutes per paper (faster since pre-coded)
5. After both coders finish: compare codes, resolve disagreements, compute Cohen's κ

6 Files

All files are in `extraction_runs/run1/`:

File	Purpose
<code>papers_for_coding_71.csv</code>	71 papers (blank coding columns)
<code>papers_coded_verified.csv</code>	GPT-4o coded + manually verified
<code>papers_coded_gpt4o.csv</code>	Original GPT-4o coding
<code>coding_corrections.csv</code>	39 corrections made to GPT-4o output
<code>papers_all.csv</code>	Full 277 AML papers (for statistics)

Recommendation: Use `papers_coded_verified.csv` as your starting point. It has been pre-coded by GPT-4o and verified/corrected for known errors. You should still manually verify each coding.