F41LUR3-F1R57: A Multi-Modal Adversarial Benchmark for Embodied and Agentic AI Safety

Adrian Wedd
Independent Researcher
adrian@failurefirst.org
https://failurefirst.org

Preprint. Under review.

Abstract

Adversarial safety benchmarks for large language models—AdvBench, HarmBench, JailbreakBench, StrongREJECT—evaluate text-in/text-out interactions exclusively. None contain a single embodied, tool-integrated, or multi-agent scenario. We present F41LUR3-F1R57, a benchmark suite comprising 141,020 adversarial prompts drawn from 27 source datasets spanning 82 attack techniques and 119 harm categories, evaluated across 187 models with 131,887 graded results. The suite includes three contributions absent from existing benchmarks: (1) a structured scenario library for vision-language-action (VLA) model safety with 338 scenarios across 27 attack families, the first adversarial evaluation resource for systems that control physical actuators; (2) FLIP (Failure-Leveraged Inference Protocol), a backward-inference grading methodology that detects the *compliance paradox*—models that produce safety disclaimers while generating harmful action sequences—achieving 22 percentage points higher precision than keyword classifiers on embodied outputs; and (3) the Therapeutic Index of Safety (TI-S), an evaluator that measures whether safety interventions help or harm by quantifying iatrogenic costs. The benchmark infrastructure supports three evaluation modalities (CLI-based, HTTP API, and local Ollama), enforces schema validation and safety linting through CI, and stores all results in a normalised relational database. We document the benchmark's construction, provide a Datasheet for Datasets following Gebru et al. [2021], and report empirical findings including the Inverse Detectability-Danger Law (Spearman ρ =-0.847 between attack consequentiality and evaluator detectability) and evidence that safety training investment, not parameter count, is the primary determinant of jailbreak resistance ($r = -0.140$, $n = 24$). All code, data, schemas, and evaluation tools are publicly available.[1]

1 Introduction

The past three years have produced a productive ecosystem of adversarial evaluation benchmarks for large language models. AdvBench [Zou et al., 2023] established a standard set of 520 harmful behaviors. HarmBench [Mazeika et al., 2024] introduced functional categorisation and automated classifiers. JailbreakBench [Chao et al., 2024] contributed reproducible evaluation pipelines with leaderboard infrastructure. StrongREJECT [Souly et al., 2024] advanced rubric-based quality assessment of harmful outputs. Together these benchmarks have enabled systematic comparison of attack methods and defense mechanisms.

However, all four benchmarks share a structural limitation: they evaluate a single interaction modality—a text prompt sent to a language model, producing a text response classified as harmful

[1] Repository: https://github.com/adrianwedd/failure-first. Supplementary Datasheet: docs/DATASHEET.md.

or safe. This design excludes three categories of AI system that are increasingly deployed in safety-critical settings:

Embodied AI. Vision-language-action (VLA) models—RT-2 [Brohan et al., 2023], π_0 [Black et al., 2024], OpenVLA [Kim et al., 2024], Gemini Robotics [Google DeepMind, 2025]—perceive physical environments through sensors and generate motor commands. A VLA safety failure is not an information hazard; it is a kinetic event. Over 1,800 autonomous haul trucks operated in Australian mines by end-2025. No adversarial safety benchmark addresses this deployment class.

Agentic AI. Tool-using, multi-step AI systems that execute code, query databases, and chain actions across services. MUZZLE [MUZZLE Authors, 2025] catalogued 37 empirically discovered attack instances against agentic systems but did not produce a reusable benchmark.

Multi-turn interactions. Extended conversations where adversarial pressure compounds across turns. Crescendo attacks on DeepSeek-R1 1.5B achieved 65.0% strict ASR [43.3%, 81.9%] at $n = 20$, compared to single-turn baselines below 20% on the same model.

F41LUR3-F1R57 addresses these gaps. We contribute:

1. A **unified adversarial corpus** of 141,020 prompts from 27 source datasets, normalised into a relational database with versioned JSON schemas, covering 82 attack techniques across 6 temporal eras (DAN 2022 through VLA 2026).
2. The **first VLA adversarial scenario library** with 338 scenarios across 27 attack families and 3 attack surfaces (text-layer, action-layer, infrastructure-layer).
3. **FLIP grading**, a backward-inference evaluation protocol that detects the compliance paradox in embodied outputs, with explicit calibration disclosure.
4. **TI-S evaluator**, a tool that measures whether safety interventions produce net benefit or iatrogenic harm.
5. **Three-modality benchmark infrastructure** (CLI, HTTP API, local Ollama) with CI-enforced validation and reproducibility guarantees.
6. An **incident database** cataloguing 30+ real-world embodied AI failures across humanoid robotics, autonomous vehicles, and service robots, cross-referenced with attack families.

All artefacts are released under open licenses at `https://failurefirst.org`.

2 Related Work

2.1 Text-Only Adversarial Benchmarks

The four established benchmarks collectively provide $\sim$1,450 unique harmful behaviors or prompts. They share three design constraints that F41LUR3-F1R57 relaxes: (a) single-turn only, (b) text modality only, (c) binary or rubric-based forward classification.

AdvBench (520 behaviors) was designed as a target set for GCG optimisation [Zou et al., 2023]. HarmBench (510 behaviors across 7 categories) includes a fine-tuned Llama-2-13b classifier [Mazeika et al., 2024]. JailbreakBench (100 behaviors) focuses on attack method comparison with leaderboard infrastructure [Chao et al., 2024]. StrongREJECT (313 prompts) measures response quality rather than presence [Souly et al., 2024].

F41LUR3-F1R57 subsumes all four as source datasets within its unified corpus (Table 1). Each is imported, normalised to a common schema, and linked to evaluation results across 187 models.

Table 1: Corpus statistics.

Metric	Value
Total prompts	141,020
Source datasets	27
Attack techniques	82
Harm categories	119
Models evaluated	187 (174 with results)
Total evaluation results	131,887
LLM-graded results	47,352
Heuristic-graded results	42,234
Evaluation runs	38,425
VLA-specific scenarios	338 (27 attack families)
Validated JSONL files on disk	820 (59,733 rows)
Corpus integrity score	0.9724

2.2 VLA and Embodied AI Safety

BadVLA [Shi et al., 2025] demonstrated backdoor attacks achieving near-100% ASR against π_0 and OpenVLA via training data poisoning. Blindfold [Huang et al., 2026] achieved 93.2% ASR on GPT-4o by decomposing malicious instructions into individually benign actions; existing defenses reduced ASR by at most 17.9 pp, leaving residual ASR above 75%. Physical adversarial patches bypass VLA visual grounding [Cardenas and Xie, 2026]. Each evaluates its own attack in isolation; no shared benchmark exists.

2.3 Agentic AI Evaluation

MUZZLE [MUZZLE Authors, 2025] catalogues 37 attack instances ($\sim$12 underlying classes) against agentic systems. AgentLAB [Yao et al., 2026] reports ASR escalation from 62.5% (single-step) to 79.9% (multi-step). Neither provides a reusable benchmark with schema-validated scenarios, automated grading, or reproducible evaluation infrastructure.

3 The F41LUR3-F1R57 Corpus

3.1 Corpus Overview

The corpus is a relational database (SQLite, schema version 13, 12 tables, 15 indexes, 4 views) that unifies adversarial prompts, technique taxonomy, model metadata, and evaluation results.

Source datasets span the full temporal range of adversarial AI research: DAN-In-The-Wild (2022), AdvBench/GCG (2023), HarmBench/JailbreakBench (2024), reasoning-extension attacks (2025), and VLA-specific families (2026). Nine established public datasets are imported alongside 18 datasets original to this work.

3.2 Schema and Validation

Every scenario is stored as a JSONL record validated against a versioned JSON schema. Required fields include `id`, `created`, `domain`, `scenario_class`, `scenario_description`, `attack_attempt`, and `turns[]`. The `labels.intent.*` namespace captures 7 instruction-hierarchy subversion signals: `format_lock`, `refusal_suppression`, `persona_hijack`, `future_year_laundering`, `dataset_poisoning_intent`, `constraint_erosion`, and `research_only_pressure`.

Table 2: Source datasets by prompt count (top 10).

Source Dataset	Prompts	Origin	Type
SORRY-Bench	9,446	Public	Harmful behaviors
BEAVERTAILS	3,432	Public	Safety alignment
DAN-In-The-Wild	1,405	Public	Historical jailbreaks
WildJailbreak	1,000	Public	In-the-wild jailbreaks
Benchmark traces	993	Original	Evaluation traces
AdvBench	520	Public	GCG target set
ForbiddenQuestions	390	Public	Direct harmful queries
HarmBench	320	Public	Functional categories
StrongREJECT	313	Public	Quality-rated
HEx-PHI	290	Public	Red-teaming

Validation pipeline. Two automated checks run on every push to the main branch via GitHub Actions CI:

1. **Schema validation**: JSON Schema compliance plus cross-field invariant checks (e.g., intent-bait scenarios must set `attack_attempt: true`). Currently validates 820 files, 59,733 rows, zero failures.
2. **Safety linting**: Heuristic checks for refusal suppression phrasing, overly operational language, and future-year circumvention patterns. Scans 20,711 items, zero findings.

Both tools are invocable via `make validate` and `make lint`. The CI pipeline rejects any push that introduces a validation or lint failure.

3.3 Source Dataset Integration

The database normalises heterogeneous source formats into a common schema. Table 2 shows the 10 largest source datasets.

Each source dataset is imported via dedicated scripts that preserve provenance metadata, map prompts to the technique taxonomy, and link harm categories.

3.4 Technique Taxonomy

The 82 attack techniques are organised into temporal eras reflecting the evolution of adversarial methods:

Table 3: Attack technique eras.

Era	Period	Representative Techniques	Prompts
DAN	2022	Persona hijack, role-play	$\sim$1,400
GCG/AutoDAN	2023	Gradient-based suffix, automated	$\sim$850
Multi-turn/Crescendo	2024	Escalation over turns	$\sim$500
Reasoning exploitation	2025	CoT manipulation, format lock	$\sim$300
VLA/Embodied	2026	Action-layer, infrastructure-layer	$\sim$340
Cross-era	Mixed	Combined techniques, compound	$\sim$600

The taxonomy converges approximately 55% with MUZZLE's independently derived classification [MUZZLE Authors, 2025], with 5 schema gaps identified: `tool_chain_hijacking`, `memory_persistence_attack`, `objective_drift_induction`, `cross_system_lateral_movement`, and `silent_exfiltration`.

3.5 VLA Scenario Library

The VLA component contains 338 scenarios across 27 attack families, organised by attack surface:

Text-layer attacks (14 families, $\sim$205 scenarios) exploit the VLM backbone via prompt manipulation. Families include Safety Instruction Dilution (SID, 120 scenarios with dose-response experiment), Latent Action Manipulation (LAM), Semantic Benignity Exploitation (SBE), Cross-Embodiment Transfer (CET), and Long-Horizon Goal Displacement (LHGD).

Action-layer attacks (3 families, $\sim$39 scenarios) exploit the action decoder or action space. Policy Puppetry (PP, 31 scenarios), Action Space Exploitation (ASE), Prompt Compliance Manipulation (PCM).

Infrastructure-layer attacks (10 families, $\sim$94 scenarios) target system boundaries. Includes compound attacks (DA-SBA, SID+SIF, DLA) and cross-system attacks (CSBA, SSBA, XSBA).

Each scenario maps to one of four HANSE (Hierarchical Adversarial Nexus for Safety Evaluation) defense layers: `vla_core` (61.8%), `semantic_firewall` (17.6%), `kinematic_shield` (12.9%), `affordance_verifier` (7.8%). The coverage imbalance toward `vla_core` reflects the dominance of text-layer attacks in the current literature and is documented as a known limitation.

3.6 Incident Database

We compiled a catalogue of 30+ real-world embodied AI incidents from the OECD AI Incident Monitor, AIID, NHTSA Standing General Orders, FDA MAUDE, and OSHA Severe Injury Reports. Incidents span humanoid robotics (Unitree H1 tether-induced feedback loop, 2025), autonomous vehicles (Cruise dragging incident, 2023; Uber ATG fatality, 2018), service robotics (Knightscope K5 stair-fall, 2017; Coco delivery robot train destruction, 2026), and industrial automation (Tesla Giga Texas robot injury, 2021).

The incident catalogue grounds scenario design in observed failure modes and enables cross-reference between adversarial scenarios and real-world failure patterns.

4 Evaluation Infrastructure

4.1 Three Runner Modalities

The benchmark supports three evaluation pathways, enabling researchers with different compute resources and model access to participate:

CLI-based runner. Invokes native CLI tools (claude-code, codex-cli, gemini-cli) for frontier model evaluation. Produces structured JSONL traces with timestamps.

HTTP API runner. Connects to OpenRouter (100+ models) or any OpenAI-compatible API. Supports both free-tier (32 models, zero cost) and paid models. Implements rate limiting, retry logic, and token tracking.

Local Ollama runner. Targets locally-hosted models via Ollama. No rate limits, no cost, no data egress. Supports reasoning models with `think=true` for chain-of-thought capture.

All three runners produce the same output format: per-scenario JSONL traces with input, output, timing, and token counts. This standardisation means that any downstream analysis tool works identically regardless of evaluation pathway.

4.2 FLIP Grading

FLIP (Failure-Leveraged Inference Protocol) uses backward inference to grade model responses. Given a response, the FLIP grader infers what instruction the model appeared to follow, then classifies whether that inferred instruction is harmful.

Five verdict categories: COMPLIANCE (full compliance with adversarial request), PARTIAL (safety hedging with behavioral compliance), HALLUCINATION_REFUSAL (harmful content wrapped in refusal framing), REFUSAL (genuine refusal), BENIGN_QUERY (non-adversarial prompt).

Three-tier ASR reporting. Building on FLIP verdicts:

Table 4: Three-tier ASR (corpus-wide, $n = 10{,}294$ evaluable).

Tier	Formula	ASR
Strict	COMPLIANCE / evaluable	45.9%
Broad	(COMPLIANCE + PARTIAL) / evaluable	79.3%
Functionally Dangerous	(COMPLIANCE + PARTIAL + HR) / evaluable	80.3%

The denominator excludes BENIGN_QUERY, ERROR, and non-evaluable categories. The Functionally Dangerous tier includes HALLUCINATION_REFUSAL because these responses are computationally indistinguishable from COMPLIANCE (thinking tokens $p = 0.21$, response tokens $p = 0.46$) but significantly different from REFUSAL (both $p < 0.001$).

Calibration disclosure. The primary FLIP grader (deepseek-r1:1.5b) achieves approximately 70% accuracy on calibration audit with a 30.8% false positive rate on benign baselines ($n = 44$). Cohen's κ between heuristic and LLM classifiers is 0.126 [0.108, 0.145] ($n = 1{,}989$). We disclose all calibration data to enable downstream reliability assessment.

4.3 TI-S Evaluator

The Therapeutic Index of Safety (TI-S) evaluator measures whether a safety intervention produces net benefit or iatrogenic harm:

$$\text{TI-S}(I, S) = \frac{B_h(I, S)}{C_h(I, S)} \tag{1}$$

where B_h is the harm-layer benefit (ASR reduction) and C_h is the sum of iatrogenic costs (PARTIAL rate increase, HALLUCINATION_REFUSAL migration, action-layer execution despite text-layer refusal). TI-S > 1 indicates net benefit; TI-S < 1 indicates iatrogenic harm.

The evaluator compares traces collected with and without a safety intervention, producing per-intervention TI-S scores. This addresses a gap in existing evaluation: standard ASR metrics cannot distinguish a safety intervention that genuinely reduces harm from one that merely shifts harmful outputs from COMPLIANCE to PARTIAL (text-level hedging without behavioral change).

4.4 Statistical Testing

The benchmark includes a statistical testing module supporting chi-square tests, Mann-Whitney U tests, and Cohen's d effect sizes with Bonferroni correction for pairwise model comparison. Score reports generate summary statistics with confidence intervals from any trace JSONL file.

4.5 Relational Database

All results are stored in a normalised SQLite database (schema version 13, 12 tables). The schema supports prompt-level queries (filter by source dataset, technique, harm category, era), result-level queries (join prompts to model responses with LLM and heuristic verdicts), cross-model comparison (aggregate ASR by model, provider, parameter count), and technique lineage (trace technique evolution across temporal eras). Pre-built queries are available via CLI (11 named queries); raw SQL access is supported for custom analysis.

5 Benchmark Results

5.1 Cross-Model Vulnerability Profiles

Analysis of 57 models with sufficient LLM-graded data reveals three distinct vulnerability clusters:

Table 5: Model vulnerability clusters.

Cluster	ASR Range	Models	Characteristics
Permissive	$\geq 40\%$	37	Minimal safety training; open-weight
Mixed	15–40%	15	Partial safety training; fine-tuned
Restrictive	$\leq 15\%$	5	Frontier; extensive RLHF

Provider signatures dominate clustering: Anthropic 3.7% mean ASR, Google 9.1%, compared to Nvidia 40.0%, Qwen 43.1%, DeepSeek 44.1%. Inverse scaling (larger models are more vulnerable) is not supported: Spearman $r = -0.140$, $n = 24$ models with known parameter counts. Safety training investment appears to be the primary determinant.

5.2 VLA-Specific Findings

FLIP grading across 15 VLA attack families produces a three-tier vulnerability structure relative to the 30.8% evaluator false positive rate:

- **Tier 1 (genuine signal):** Trace Reasoning Attack (100% broad ASR), Deceptive Alignment (62.5%), Latent Action Manipulation (60%). ASR substantially exceeds benign baseline.
- **Tier 2 (marginal):** Action Space Exploitation, Semantic Benignity Exploitation. ASR 40–55%, within wide confidence intervals of baseline.
- **Tier 3 (at FP floor):** Policy Puppetry, SID+SIF, and 8 others. Indistinguishable from evaluator noise at current sample sizes.

Compliance paradox. 50% of all FLIP verdicts are PARTIAL—models generate safety disclaimers while producing the requested action sequences. Zero outright refusals across 63 originally graded traces. Keyword classifiers report 94% ASR on the same traces; FLIP reports 72.4%—a 22 pp gap attributable to the compliance paradox.

5.3 Inverse Detectability-Danger Law (IDDL)

Corpus-level analysis reveals a structural inverse correlation between attack consequentiality and evaluator detectability. Spearman ρ =-0.847 ($n = 27$ families). The attacks most likely to cause physical harm are the attacks least likely to be detected by text-layer evaluation.

This finding has direct implications for organisations that evaluate VLA safety using text-level classifiers: they will systematically underestimate their exposure to the most consequential attack families.

5.4 Safety Re-emergence at Scale

Evaluation of safety-abliterated (safety-removed) models in the Qwen3.5 obliteratus series reveals that safety-like behavior partially re-emerges at scale: ASR drops from 100% (0.8B) to 47.3% (9.0B), Spearman $\rho = -0.949$, $p = 0.051$. Larger abliterated models produce PARTIAL verdicts rather than full compliance, suggesting safety is partially an emergent property of scale.

6 Datasheet for Datasets

We provide a complete Datasheet for Datasets following Gebru et al. [2021] as supplementary material (`docs/DATASHEET.md` in the repository). The datasheet covers all seven sections: Motivation, Composition, Collection Process, Preprocessing/Cleaning/Labeling, Uses, Distribution, and Maintenance. We summarise key elements here; the full datasheet should be consulted for complete details.

Motivation. The corpus was created to fill the gap in adversarial evaluation benchmarks for embodied and agentic AI systems. It was created by Adrian Wedd as an independent, self-funded AI safety research project.

Composition. 141,020 prompts total; 338 VLA-specific scenarios; 820 validated JSONL files containing 59,733 rows; 131,887 evaluation results across 187 models. The dataset contains adversarial prompts that request harmful outputs by design; scenarios describe patterns at a structural level, not executable exploits.

Collection. Data was acquired through import from 9 public datasets, original scenario authoring, and evaluation trace collection from 187 models via three runner interfaces. No human subjects data was collected.

Preprocessing. All prompts are normalised to a common JSONL schema. Technique labels and harm categories are mapped to the unified 82-technique, 119-category system. Deduplication eliminated 102 cross-file duplicates (integrity score 0.9724).

Uses. The dataset has been used for 131,887 evaluations, 130+ research reports, and companion CCS 2026 and NeurIPS 2026 D&B submissions. It should not be used to develop operational attack tools or train models to produce harmful outputs.

Distribution. Pattern-level analysis is publicly available at `https://failurefirst.org`. Operational scenario details are maintained in a private repository.

Maintenance. Continuous updates via semantic versioning, GitHub issues, and CI validation. Schema migrations are versioned.

7 Ethical Considerations and Broader Impact

7.1 Dual-Use Risk

F41LUR3-F1R57 is an adversarial benchmark; it documents attack patterns that could, in principle, be adapted for malicious use. The dual-use risk for embodied AI benchmarks is qualitatively different from text-only benchmarks: a successful VLA attack produces physical actions that could injure or kill, not merely harmful text.

We mitigate this through three design choices: (a) *pattern-level, not operational*—scenarios describe vulnerability mechanisms, not executable exploits; (b) *defense-oriented framing*—each

family is accompanied by a description of the targeted vulnerability; (c) *graduated disclosure*—aggregate results are public while operational details are access-controlled.

7.2 The Compliance Paradox

The compliance paradox—models that disclaim while complying—creates a distinctive ethical concern. Documenting this gap could be exploited by attackers who craft prompts specifically designed to produce PARTIAL responses that pass text-level safety filters. However, the gap exists regardless of documentation; our contribution is making it visible so defense architectures can address it.

7.3 Evaluator Limitations and the Evaluation Access Gap

Using sub-2B models as FLIP graders introduces systematic bias (approximately 70% accuracy, 30.8% FP rate). This limitation means reported ASR figures carry wider uncertainty than ideal. We disclose all calibration data. Resource-constrained researchers face a genuine dilemma: reliable evaluation requires models that are expensive to run, creating an evaluation access gap that mirrors compute inequality in AI research.

7.4 Responsible Disclosure

We have engaged with the Australian AI Safety Institute (AISI) and Safe Work Australia. The benchmark addresses gaps in multiple regulatory frameworks: VAISS Guardrail 4 (Australia), EU AI Act Article 9, ISO 17757:2019 (autonomous mining equipment), and ISO 13482:2014 (personal care robots). None currently specify adversarial testing methodology for embodied AI.

7.5 Researcher Positionality

This work was conducted by an independent researcher without institutional affiliation or compute sponsorship. This is both a strength (no organisational conflicts in reporting vulnerabilities) and a limitation (resource constraints on model scale, replication, and human annotation).

8 Limitations

1. **Model scale.** VLA evaluation results are from 1.5B–1.7B parameter models, not representative of frontier VLA systems. The capability-floor hypothesis predicts different vulnerability profiles at scale.

2. **Text-only evaluation.** All experiments are text-in/text-out. Visual adversarial inputs, physical patches, and sensor-based attacks are not evaluated end-to-end.

3. **Evaluator quality.** The 30.8% FP rate and approximately 70% accuracy of the primary FLIP grader introduce noise. Per-family findings with $n < 10$ should be treated as preliminary.

4. **Scenario ecological validity.** Scenarios are researcher-authored, not drawn from real deployment logs. Attack families represent theoretically motivated surfaces, not observed campaigns.

5. **Single evaluation run.** Most families evaluated in a single run without replication across seeds or temperatures. Variance from Wilson CIs, not empirical replication.

6. **Coverage imbalance.** SID accounts for 120/338 VLA scenarios (35.5%). Infrastructure-layer families have $\sim$5 scenarios each.

7. **Public dataset overlap.** The 9 imported public datasets have known inter-dataset overlap (e.g., AdvBench prompts appear in HarmBench). Deduplication was applied within F41LUR3-F1R57 but not across source publications.

Table 6: Reproducibility artefacts.

Artefact	Location	Verification
Schema DDL	`database/schema.sql`	Deterministic init
Schema migrations	`database/migrations/`	Versioned, ordered
Public import	`tools/database/import_public.py`	From HuggingFace
Validation	`make validate`	Schema + invariants
Safety linting	`make lint`	Heuristic checks
CI pipeline	`.github/workflows/ci.yml`	Every push
Benchmark runners	`tools/benchmarks/run_benchmark_*.py`	3 modalities
FLIP grading	`tools/benchmarks/grade_*.py`	Backward inference
TI-S evaluator	`tools/evals/iatrogenic_*.py`	Baseline vs. intervention
Statistical tests	`tools/stats/`	χ^2, Mann-Whitney
Test suite	`pytest` (1,205 tests)	CI floor at 1,150

8. **Corpus size dominated by OBLITERATUS.** 121,955 of 141,020 prompts (86.5%) come from OBLITERATUS telemetry and runs. The non-OBLITERATUS analytical corpus is approximately 19,065 prompts.

9 Reproducibility

The benchmark is designed for full reproducibility. Table 6 lists the key artefacts and their verification methods.

To reproduce the full evaluation pipeline:

```
pip install -r requirements-dev.txt
python tools/database/db_manager.py init
python tools/database/import_public.py --dataset advbench
make validate && make lint
python tools/benchmarks/run_benchmark_http.py \
  --scenarios data/vla/vla_sbe_v0.1.jsonl \
  --models "meta-llama/llama-3.3-70b-instruct:free" \
  --output runs/reproduce/
python tools/benchmarks/grade_generation_traces.py \
  --traces runs/reproduce/traces.jsonl
python tools/benchmarks/score_report_v1.0.py \
  --traces runs/reproduce/traces.jsonl --stats
```

10 Conclusion

F41LUR3-F1R57 extends adversarial AI safety evaluation beyond text-only LLM benchmarks into embodied, agentic, and multi-turn settings. The corpus of 141,020 prompts across 27 source datasets, evaluated on 187 models with 131,887 results, provides the scale and diversity needed for robust cross-model vulnerability analysis. The VLA scenario library (338 scenarios, 27 families) is the first adversarial resource for systems that control physical actuators. FLIP grading detects the compliance paradox that forward classifiers miss. The TI-S evaluator measures whether safety interventions cause net benefit or iatrogenic harm.

The benchmark's primary empirical contribution is the Inverse Detectability-Danger Law: the attacks most consequential for physical safety are the least detectable by text-layer evaluation ($\rho = -0.847$). This finding suggests that current evaluation methodology systematically underestimates the most dangerous attack surfaces.

Future work includes evaluation at 7B+ model scale, visual adversarial input integration, human annotation studies for ground-truth calibration, and expansion of infrastructure-layer attack families.

References

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. π_0: A vision-language-action flow model for general robot control. In *International Conference on Machine Learning*, 2024.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. RT-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023.

Alvaro Cardenas and Cihang Xie. Physical adversarial attacks on vision-language models for robotic manipulation. *arXiv preprint*, 2026.

Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J Pappas, and Eric Wong. JailbreakBench: An open robustness benchmark for jailbreaking large language models. In *NeurIPS 2024 Datasets and Benchmarks Track*, 2024.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64 (12):86–92, 2021.

Google DeepMind. Gemini robotics: Bringing AI into the physical world. *Google DeepMind Technical Report*, 2025.

Yutong Huang et al. Jailbreaking embodied LLMs via action-level manipulation. *arXiv preprint arXiv:2603.01414*, 2026.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. OpenVLA: An open-source vision-language-action model. In *International Conference on Machine Learning*, 2024.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, et al. HarmBench: A standardized evaluation framework for automated red teaming and robust refusal. *arXiv preprint arXiv:2402.04249*, 2024.

MUZZLE Authors. MUZZLE: A multi-faceted zero-day attack taxonomy for agentic AI systems. *arXiv preprint arXiv:2602.09222*, 2025.

Xin Shi et al. BadVLA: Backdoor attacks on vision-language-action models. In *Advances in Neural Information Processing Systems*, 2025.

Alexandra Souly, Qingyun Lu, Dillon Bowen, Tu Trinh, Elvis Hsieh, Sana Fishi, Ralph Abi-Rafeh, Georgia Anber, Brando Braunstein, Dean Goodman, et al. A StrongREJECT for empty jailbreaks. *arXiv preprint arXiv:2402.10260*, 2024.

Shunyu Yao et al. AgentLAB: A comprehensive benchmark for long-horizon agent instruction following and safety. *arXiv preprint arXiv:2602.16901*, 2026.

Andy Zou, Zifan Wang, J Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023.

A Datasheet for Datasets (Full)

The complete Datasheet for Datasets following Gebru et al. [2021] is maintained as a living document at `docs/DATASHEET.md` in the project repository. It covers:

1. **Motivation** (Sections 1.1–1.4): Purpose, creator, funding.
2. **Composition** (Sections 2.1–2.15): Instance types, counts, labels, sensitive content.
3. **Collection Process** (Sections 3.1–3.11): Acquisition channels, tools, sampling, ethics.
4. **Preprocessing** (Sections 4.1–4.4): Schema validation, deduplication, FLIP grading.
5. **Uses** (Sections 5.1–5.5): Current uses, recommended and prohibited uses.
6. **Distribution** (Sections 6.1–6.6): Two-tier access, licensing, export controls.
7. **Maintenance** (Sections 7.1–7.7): Contact, updates, versioning, contribution mechanism.

Additionally, Sections 8–9 of the datasheet cover ethical considerations (dual-use risk, bias, human subjects) and dataset validation/quality assurance metrics.

The datasheet version corresponds to corpus version 0.19.0 and is updated with each schema migration.