The True Nature of Most Threats Behind AI-ML

A data-driven look at six AI-ML threat catalogues

OWASP AI Exchange · OWASP ML Top 10 · PLOT4AI · OWASP LLM · OWASP Agentic AI · CSA Maestro

---

Principal Cybersecurity Architect, JPMC **Abdullah Garcia**

Disclosure – Opinions are my own and not representative of the firm I work for.

What if most “AI-specific” threats aren’t about “AI” at all?

That’s the question this analysis was built to answer.

The Story We Keep Telling

Every week, AI-ML security is framed as a new species of "risk" —

- **Prompt Injection**
- **Model Poisoning**
- **Adversarial Examples**
- **Model Inversion**
- **Training Data Extraction**
- **Membership Inference**
- **Jailbreaks**
- **Agent Runaway**
- **Supply Chain Backdoors**

— each one "breathlessly novel", each one demanding a new defence stack.

The Finding

2/3

of **general AI-ML threats** are extensions of long-standing **data security** problems.

27 of 42 distinct threats · OWASP AI Exchange, ML Top 10, PLOT4AI

The pattern is then tested across generative and agentic AI.

How We Counted

Six industry threat catalogues. Every individual threat mapped to its primary attack vector.

OWASP AI Exchange

12 threats

OWASP ML Top 10

10 threats

PLOT4AI

20 threats

OWASP LLM & GenAI

16 threats

OWASP Agentic AI

15 threats

CSA Maestro

55 threats

What we counted as “data” → next slide.

What We Counted As “Data” And Why

If it exists as bits, and its content shapes how the model behaves, we counted it as data.

DATA

- Training and fine-tuning corpora
- Model weights (serialised state)
- System and user prompts
- Retrieved context (RAG, tools)
- Runtime inputs, outputs, logs
- Embeddings and vector stores

NOT DATA

- Model architecture (layers, attention)
- Training algorithms (loss, optimiser)
- Compute and network infrastructure
- Identity and access systems
- Human processes and governance

The rule: bits that shape behaviour are data. Configuration is not.

Three Classes Of AI-ML Threat

Each one tells the same story from a different angle.

General

Threats applicable to any AI-ML system: poisoning, leakage, inference manipulation.

Generative

Threats specific to LLMs and generative pipelines: prompt injection, content extraction.

Agentic

Threats arising from autonomy: self-directed planning, action chaining, tool misuse.

1 · General AI-ML Threats

Attacks that target any AI-ML system, irrespective of model family or deployment.

CATALOGUE	DATA-CENTRIC	SHARE
PLOT4AI · Security	13 / 20	**65%**
OWASP ML Top 10	7 / 10	**70%**
OWASP AI Exchange	7 / 12	**58%**

Three independent catalogues. One direction.

General AI-ML · Aggregate

64%

of distinct general AI-ML threats are data-centric.

27 of 42 threats catalogued

Whatever the model, the attacker's first move is the data.

2 · Generative AI-ML Threats

The headline-grabbing class. Prompt injection, training-data extraction, output manipulation.

OWASP LLM + GenAI Data Security Best Practices

16 discrete threats analysed

10 of 16 (~63%)

weaponise data as the primary attack surface

AND THE TWIST

0 of 16

threats proved unique to generative AI-ML.

Every threat has an analogue in classical AI-ML pipelines.

Not a single threat is truly new.

Every catalogued Gen AI-ML threat is a familiar data-security failure dressed in new clothes: poor input sanitisation, weak provenance, insufficient confidentiality.

3 · Agentic AI-ML threats

Self-governing systems that plan, act, and adapt without continuous oversight.

OWASP Agentic AI

15 threats catalogued

2 of 15 Truly agent-specific

3 of 15 Data-centric (~20%)

CSA Maestro

55 threats catalogued

2 of 55 Truly agent-specific

17 of 55 Data-centric (~31%)

Two catalogues. The same scarcity of genuinely agentic threats.

Agentic AI-ML · Aggregate

6%

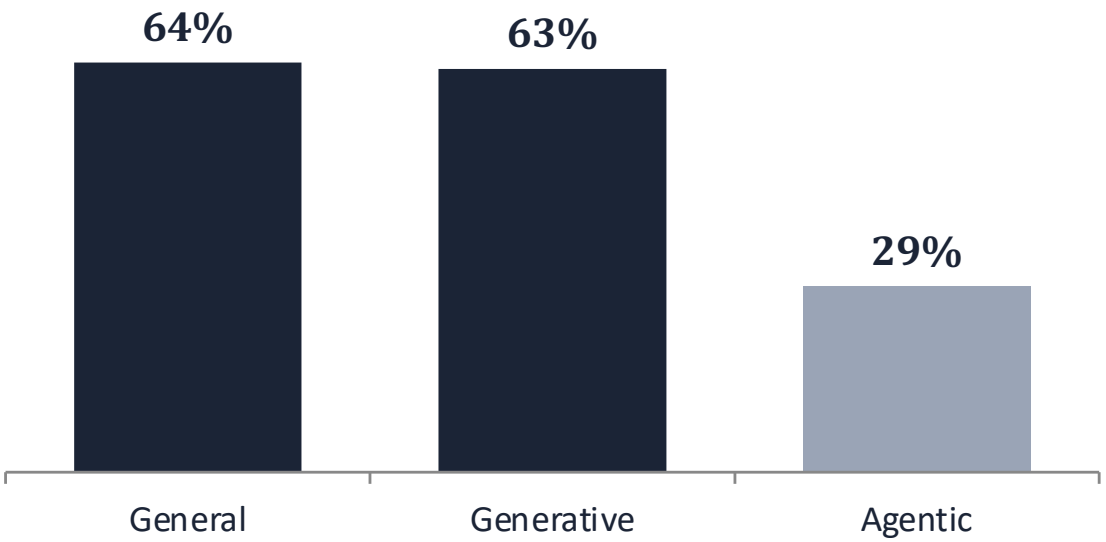
of the 70 agentic threats are genuinely agentic in nature.

29%

hinge on data as the primary attack vector. The rest mirror non-agentic AI-ML.

Three classes. One pattern.

Share of catalogued threats with data as the primary attack vector.



Different model families.
Same attack surface.

Data.

So, if most threats are data threats, where should the budget go?

Four Foundations For AI-ML Data Security

These are not new disciplines. They are existing ones, now load-bearing.

Provenance

Verifiable origin of every dataset, model weight, and input feed.

Lineage

Traceability of how data moves and transforms through the pipeline.

Integrity

Tamper detection across training, evaluation, and inference data.

Access Control

Least-privilege over data, models, and agent effectors alike.

Why Guardrails Aren’t Enough

Guardrails defend the surface. Provenance defends the substance.

MODEL-CENTRIC DEFENCES

- Output filters and guardrails
- Red-team evaluations
- Alignment fine-tuning
- Prompt firewalls

Necessary, but not sufficient on its own.

DATA-CENTRIC FOUNDATION

- Provenance and lineage tracking
- Integrity attestation
- Access governance over data + effectors
- Runtime monitoring of inputs

Catches what the model layer never sees.

Monday Morning

Three actions every AI-ML initiative can take this week.

01

Inventory the data supply chain.

Map every training corpus, retrieval source, and live feed feeding your AI-ML systems. Provenance starts with knowing.

02

Reuse your existing data controls.

Bring AI-ML data flows under the same classification, encryption, and access policies as the rest of the enterprise.

03

Budget for data, then model.

Make the data foundation the non-negotiable line item; let model-centric defences layer on top.

Secure the data.
The model follows.

Thank you.

Annex

Annex – Supply Chain Security