0. Paper Identity

Type of paper This is a systematization + threat evolution analysis, not a pure survey, not a novel defense paper.

If you try to pretend you invented new attacks, reviewers will eat you alive. Own what this is.

Audience

- Security researchers
- ML engineers shipping products
- Reviewers who are tired of OWASP-style lists

---

1. Title & Abstract (Signals Maturity)

Title Template

From Model Exploits to System Failures: A Five-Year Analysis of Security Risks in Generative AI Systems

No buzzwords. No “next-gen”. No chest-thumping.

Abstract Blueprint

4 paragraphs, no more.

1. Context
    - Rapid deployment of generative AI systems
    - Security models lag behind system complexity
2. Problem
    - Existing threat models focus on model-level risks
    - Fail to capture real-world failures involving interfaces, tools, and human interaction
3. Method
    - Systematic analysis of attacks and failures from 2019–2025

    - Categorized using a layered system model
4. Contribution
    - Show threat evolution over time
    - Identify impact patterns
    - Expose limitations of current defenses

If your abstract can be skimmed and still understood, you did it right.

---

2. Introduction (Set the Argument, Not the Background)

What this section must do

- Frame the shift
- Justify why old thinking fails
- Make the reader care

Structure

2.1 Why Generative AI Is Different

- Probabilistic outputs
- Instruction-following behavior
- Integration with tools and workflows

2.2 Why Traditional Security Models Break

- Deterministic assumptions
- Clear trust boundaries
- Human oversight assumed effective

2.3 Research Questions You need explicit questions, for example:

- How have AI-related threats evolved over the last five years?
- Where does real-world impact actually occur?
- Why do existing defenses fail at scale?

No hand-waving. Questions anchor the paper.

---

3. Related Work (Respectful but Critical)

This is where most papers become boring. Yours shouldn't.

What to cover

- OWASP Top 10 for LLMs
- ML security literature (poisoning, extraction)
- Industry safety docs (OpenAI, Anthropic, etc.)

What to say

- Acknowledge their value
- Point out the fragmentation
- State clearly: none provide a longitudinal, system-level view

This section earns you credibility.

---

4. Methodology (How You're Not Making Things Up)

This is crucial.

4.1 Data Sources

- Public incident reports
- Security blogs
- Academic papers
- Documented failures in production systems

Be transparent. You're not claiming completeness.

4.2 Inclusion Criteria

- Must involve generative AI
- Must show real or plausible impact
- Must be documented between 2019–2025

4.3 Analysis Framework

Introduce your layered system model here.

This is one of your core contributions.

---

5. Layered Threat Model (The Backbone)

This is the heart of the paper.

Each subsection:

- Defines the layer
- Lists relevant attacks/failures
- Explains why they occur

Sections

5.1 Model Layer 5.2 Interface Layer 5.3 Tool & Integration Layer 5.4 Agency & Automation Layer 5.5 Human & Organizational Layer

Do not dump examples. Use 2–3 strong cases per layer.

Your job is explanation, not enumeration.

---

6. Threat Evolution Over Time (Where Insight Lives)

This is where the paper stops being “nice” and becomes useful.

Structure

Split by time periods.

- 2019–2021: Model-centric risks
- 2022–2023: Interface exploitation
- 2024–2025: Agency and integration failures

Use a table or timeline figure here. Reviewers love visuals.

Key Insight

Threats move up the stack as systems become more autonomous.

Repeat that. It’s your thesis.

---

7. Impact Analysis (No Vibes, Just Structure)

Define impact clearly.

Categories

- Technical
- Operational
- Human

For each category:

- Describe common failure patterns
- Map which layers contribute most
- Highlight cascading effects

This is where you show maturity.

---

8. Defense Analysis (Where You’re Skeptical)

Do not present defenses as solutions.

Structure

For each defense class:

- What it assumes
- Where it works

- Where it fails

Examples:

- Prompt sanitization
- Human-in-the-loop
- Rate limiting
- Sandboxing

Your tone here should be calm but ruthless.

---

9. Implications & Future Threats (Forward-Looking)

This is not speculation. It’s extrapolation.

Topics

- Increasing autonomy
- Multi-agent systems
- AI managing AI
- Diminishing human oversight

Frame these as systemic risks, not sci-fi.

---

10. Limitations (Be Honest)

This section builds trust.

- Reliance on public incidents
- No access to proprietary data
- Rapidly evolving field

Strong papers acknowledge limits.

---

11. Conclusion (Tie It Back)

Restate:

- Why model-only security fails
- Why system-level thinking is required
- Why this matters now, not later

End with clarity, not drama.

What This Paper Is *Not*

Let me be explicit so you don’t sabotage yourself.

- Not a defense proposal
- Not an OWASP rewrite
- Not an ethics essay
- Not a “AI is dangerous” rant

It is a structured argument about how risk shifted as AI became a system.