

The Defender's Field Guide to Web Threats



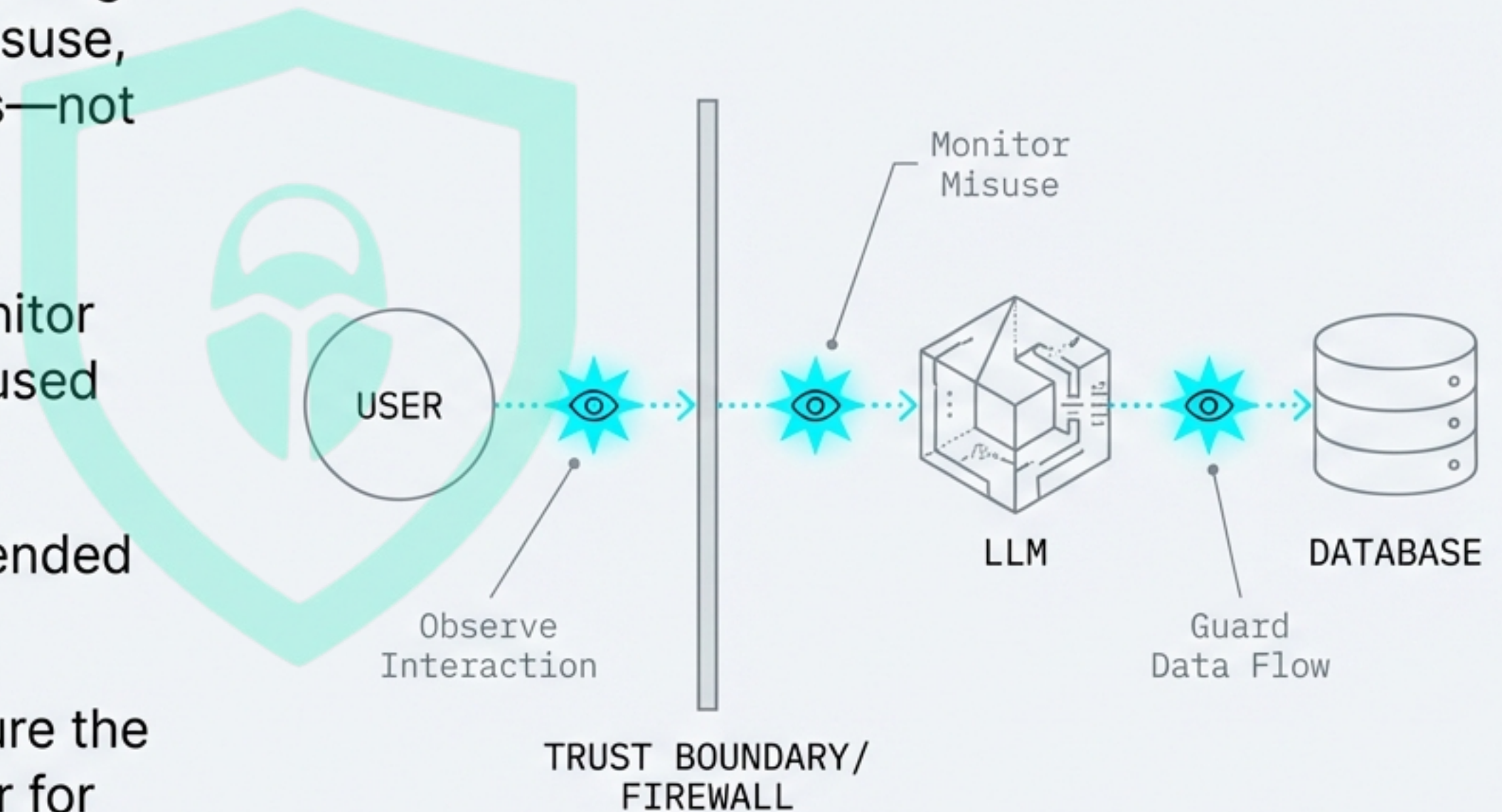
A Blue Team Framework for
Understanding and Mitigating AI Risks

Powered by **Bugitrix**

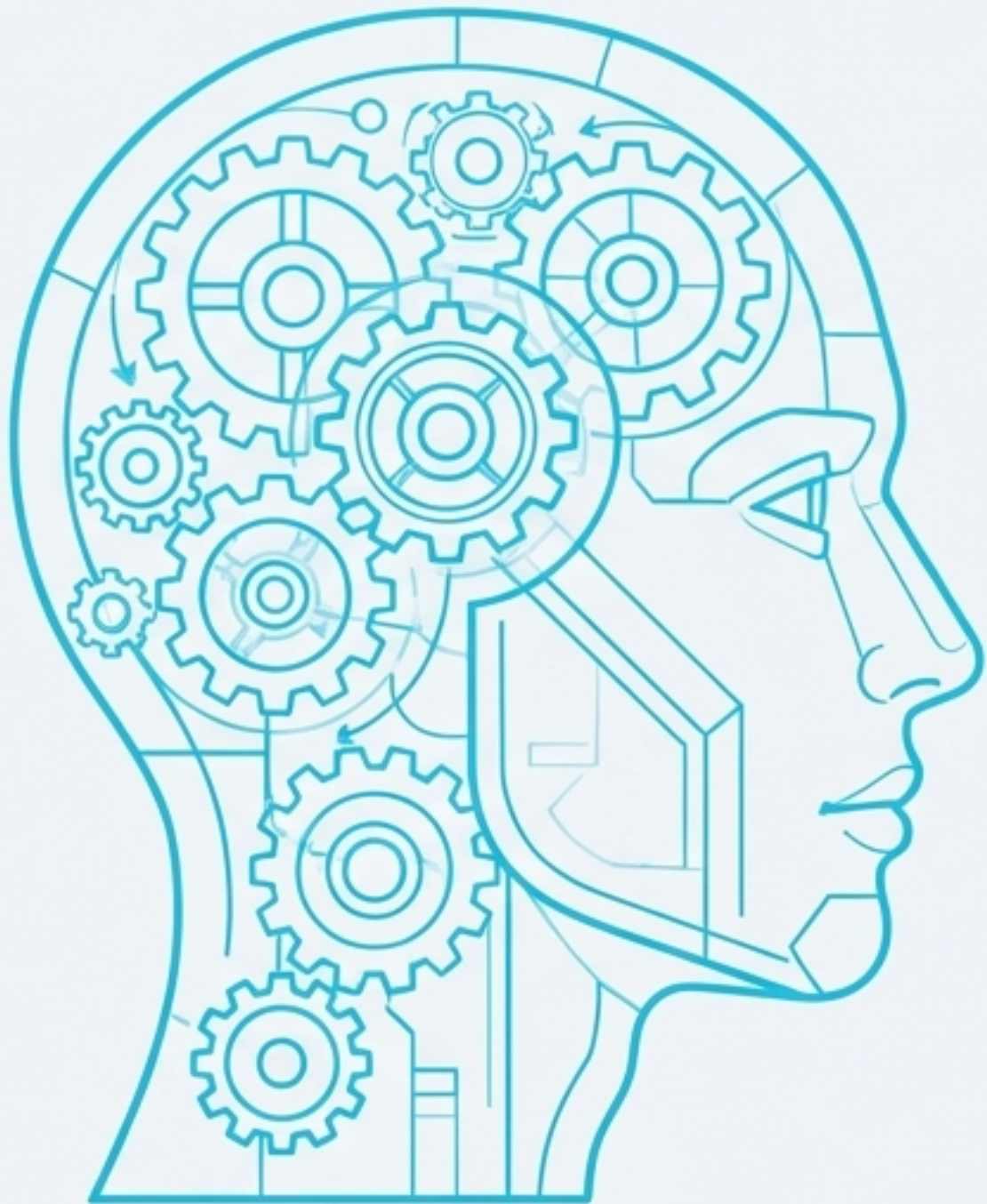
The Real Mission: It's About Behavior, Not Just Accuracy

Blue Teams protect AI-powered applications by watching how users interact with Large Language Models. They focus on misuse, data exposure, and trust boundaries—not just model accuracy.

- ☐ **Observe User Interaction:** Monitor how the LLM is actually being used and probed.
- ☐ **Prevent Misuse:** Control unintended or malicious AI behaviors.
- ☐ **Protect Data Boundaries:** Ensure the model doesn't become a vector for data exfiltration.



Your Core Asset: Skills Matter More Than Tools



AI security tools help, but **understanding how LLMs think and respond** is critical.

At Bugitrix, we teach **behavior-based defense**, not prompt tricks.

The Goal: Build a **foundational understanding** of the threat landscape to make **informed, defensible decisions**.

A Defender's Taxonomy of LLM Threats

Interface & Input Risks

(Threats originating from user interaction)

- Prompt Misuse
- Injection-Like Patterns
- Rate Abuse

Integration & System Risks

(Threats from how the LLM connects to data and tools)

- Data Leakage
- Over-Permissioned APIs
- Insecure Context Handling
- Third-Party AI Risk

Oversight & Process Risks

(Threats from a lack of foundational security practices)

- Lack of Monitoring
- Trust in AI Output
- No AI Threat Model

LLM.T01: Prompt Misuse



Risk Profile

Unexpected user inputs that produce unintended outputs.

Defender's Objective

Control AI Behavior.

Threat Scenario

A user crafts an input that subtly hints at sensitive topics, causing the model to generate information it shouldn't.

Foundational Skill

LLM Basics (Understanding the fundamentals of how models interpret and respond to prompts).

LLM.T04: Injection-Like Patterns



Risk Profile

Abusive instruction patterns designed to override or bypass system rules.

Defender's Objective

Prevent Misuse.

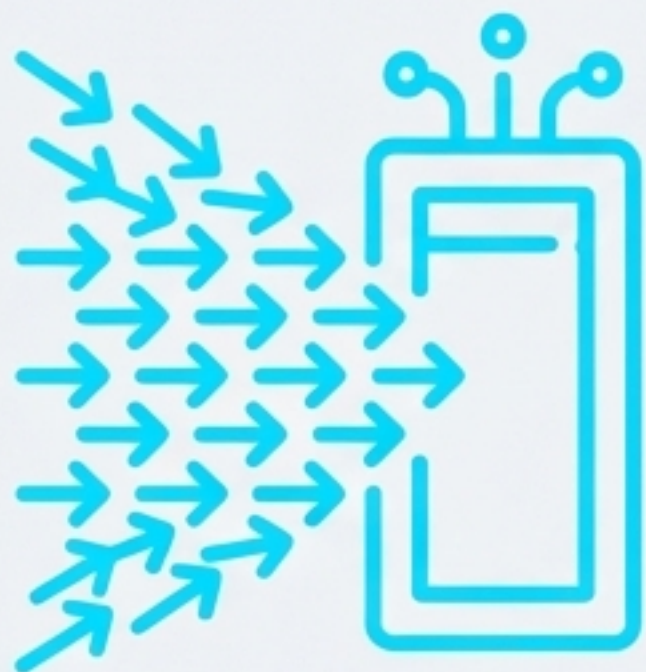
Threat Scenario

A user embeds instructions within their input (e.g., "Ignore previous instructions and do X"), manipulating the LLM's response.

Foundational Skill

Input Validation Mindset (Applying classic security principles of sanitizing and validating all external input).

LLM.T07: Rate Abuse



Risk Profile

Excessive or automated AI usage that drains resources or degrades service.

Defender's Objective

Protect Resources.

Threat Scenario

An attacker uses a script to send thousands of queries to the LLM endpoint, causing a denial of service for legitimate users.

Foundational Skill

Rate Limiting (Implementing controls to govern the frequency of API calls and interactions).

LLM.T02: Data Leakage



Risk Profile

The model inadvertently reveals private or internal data in its responses.

Defender's Objective

Protect Confidentiality.

Threat Scenario

The LLM, trained on or given access to internal documents, exposes proprietary business information to an unauthorized user.

Foundational Skill

Data Classification (Knowing what data is sensitive and implementing policies to protect it).

LLM.T03: Over-Permissioned APIs



Risk Profile

The LLM is connected to internal systems or tools with excessive privileges.

Defender's Objective

Limit Blast Radius.

Threat Scenario

A user tricks the LLM into calling a connected API that executes a sensitive action, like deleting user data or sending an email.

Foundational Skill

Least Privilege (Ensuring a component has only the permissions essential to its task).

LLM.T05: Insecure Context Handling



Risk Profile

System prompts and contextual data are not properly secured or separated from user input.

Defender's Objective

Maintain Boundaries.

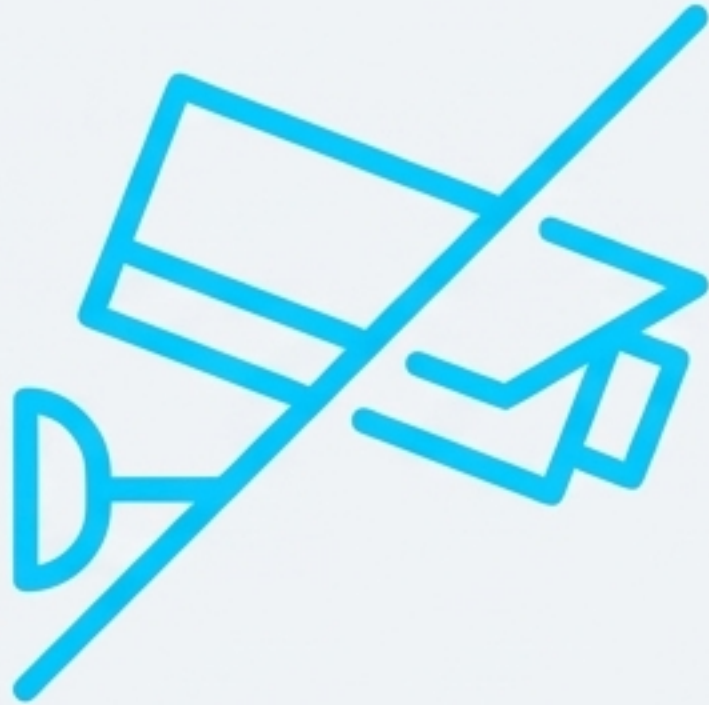
Threat Scenario

A user's input is crafted to make the LLM confuse its instructions with user data, leading to 'role confusion' and unintended behavior.

Foundational Skill

Secure Design (Architecting the system to enforce strong separation between instruction, context, and user input).

LLM.T06: Lack of Monitoring



Risk Profile

No sufficient logging or monitoring of AI interactions is in place.

Defender's Objective

Detect Abuse.

Threat Scenario

A subtle, low-and-slow abuse pattern goes completely unnoticed because there are no logs to analyze LLM queries and responses.

Foundational Skill

Logging Basics (Understanding what to log, how to store it, and how to analyze it for security events).

LLM.T08: Trust in AI Output



Risk Profile

Blindly trusting the LLM's responses without verification, leading to flawed decisions.

Defender's Objective

Prevent Business Errors.

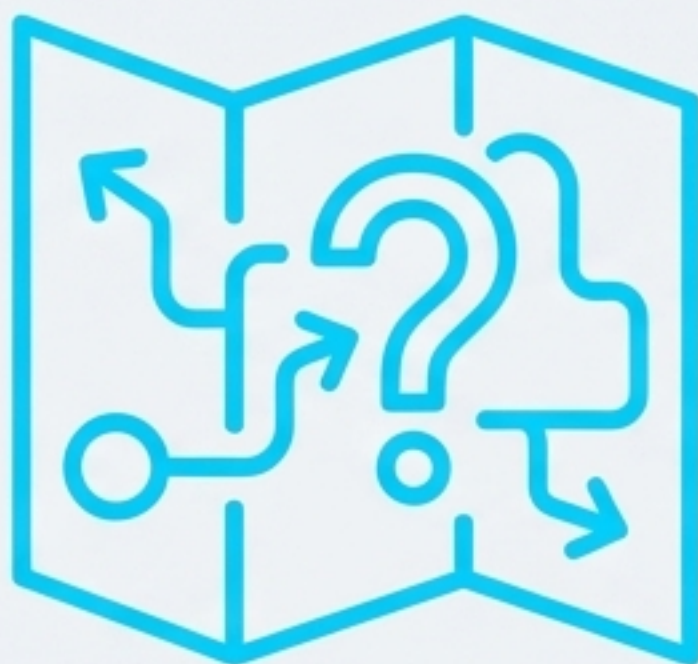
Threat Scenario

An internal tool uses an LLM to summarize financial data; a hallucinated or incorrect summary is trusted, leading to a wrong business action.

Foundational Skill

Human-in-the-Loop (Designing processes that require human oversight for critical decisions).

LLM.T10: No AI Threat Model



Risk Profile

AI-specific risks have not been formally identified, documented, or mapped to defenses.

Defender's Objective

Prepared Defenses.

Threat Scenario

A novel attack path that is unique to the LLM's integration is exploited because the team never brainstormed or considered it.

Foundational Skill

Threat Modeling (The systematic process of identifying and evaluating potential threats and vulnerabilities).

The Defender's Playbook: Five Core Principles



Apply Least Privilege: Strictly limit the permissions and data access granted to AI integrations.



Separate Data Tiers: Clearly classify and segregate public, private, and sensitive data to prevent leakage.



Monitor All Interactions: Implement comprehensive logging and monitoring for all AI queries and responses to detect abuse.



Keep Humans in the Loop: Ensure critical decisions based on AI output are verified by a human.



Threat-Model AI Features: Treat AI components like any other part of your application and proactively model their unique risks.

A Final Mandate: Operate Ethically and Responsibly

AI security testing must follow legal and ethical guidelines. Practice only on authorized systems. Bugitrix supports responsible AI and cyber security education.