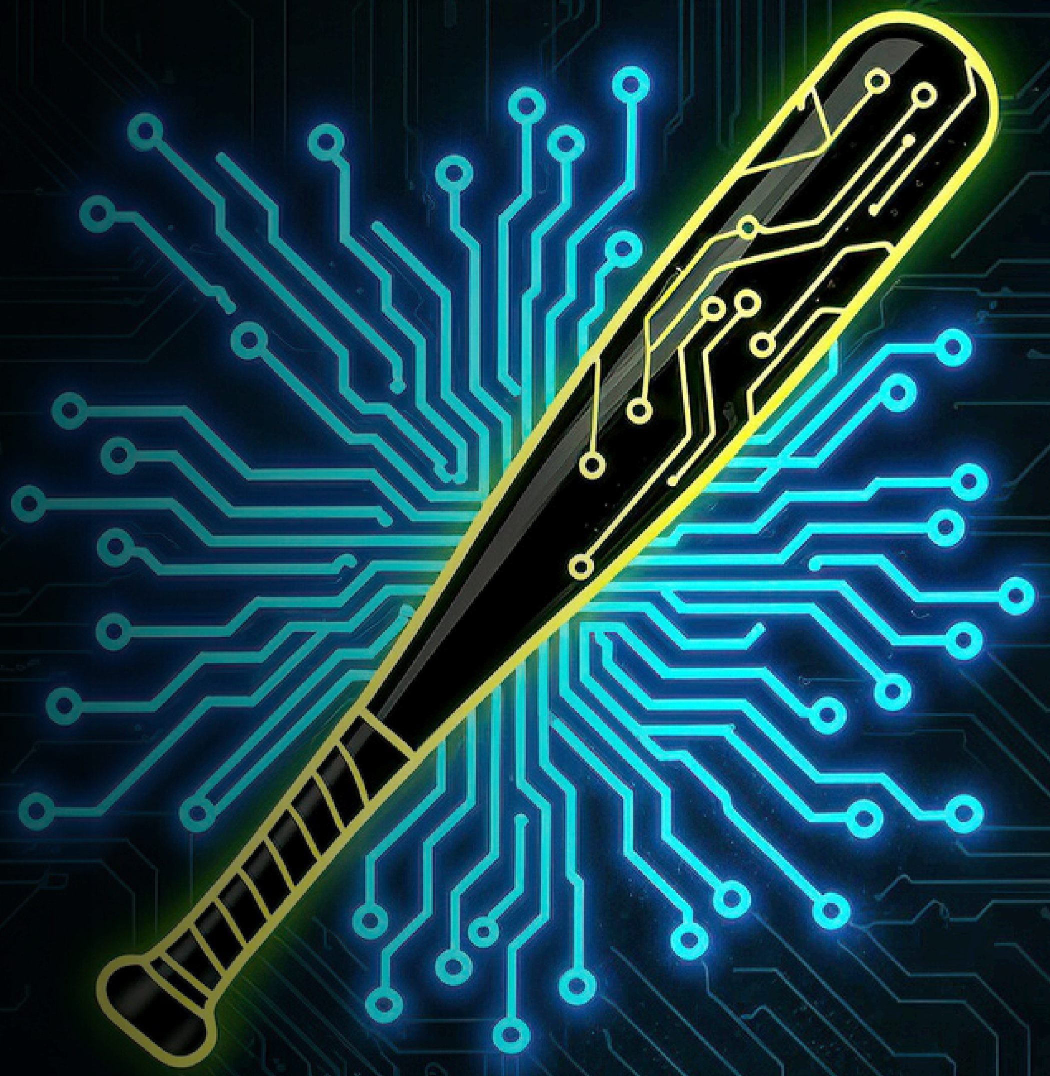
black unicorn

AI SECURITY GUARDRAILS

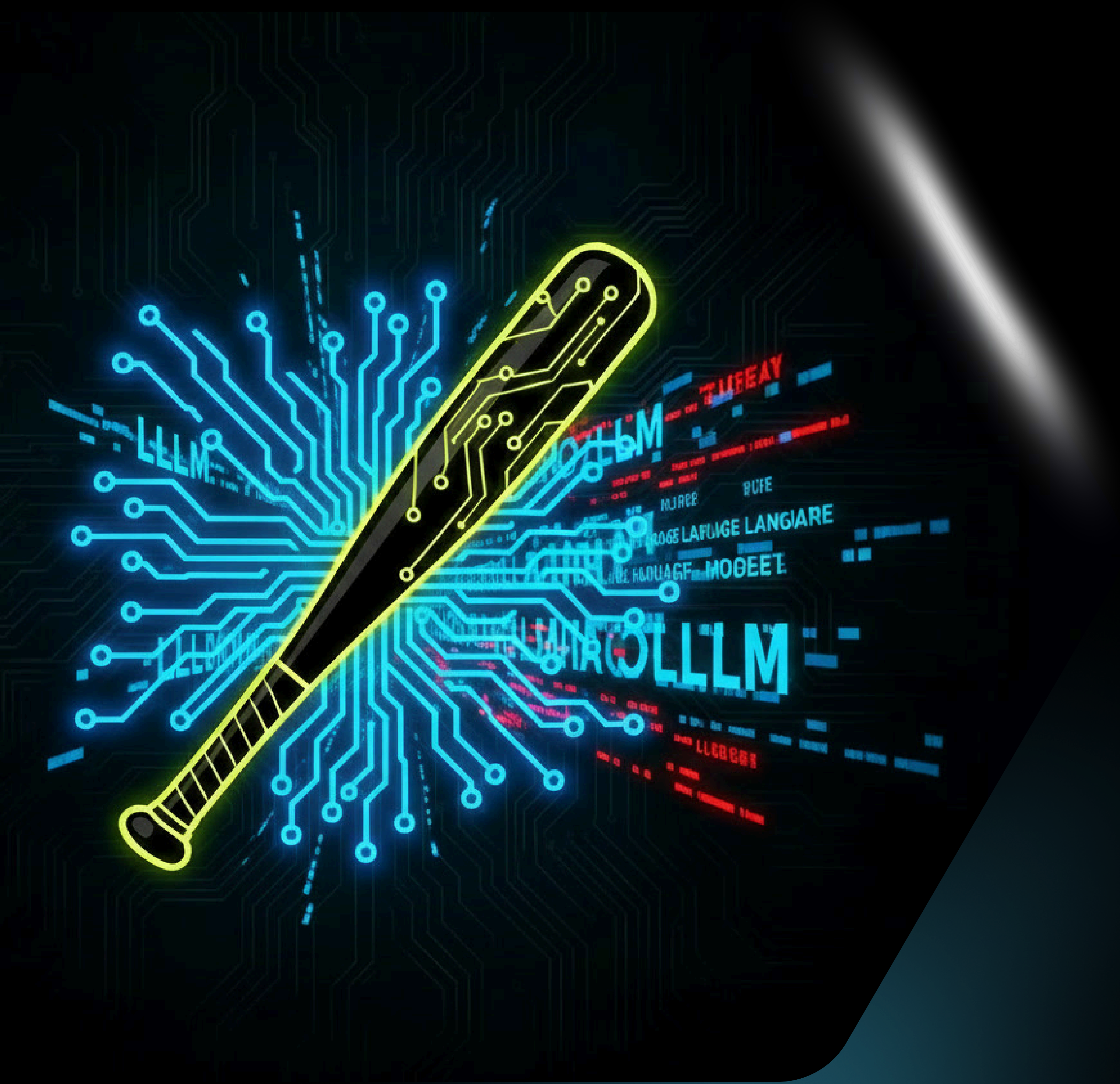
"Stops your chatbot from becoming a full-time snitch."



BonkLM

Data Smashing Solutions

blackunicorn.tech



Introduction to BonkLM

BonkLM is a comprehensive library that protects your AI applications from:

- Prompt Injection
- Secret & Credential Leaks
- PII Exposure
- Command and XSS Injection

One npm install. Any framework. Any LLM. Instant protection.

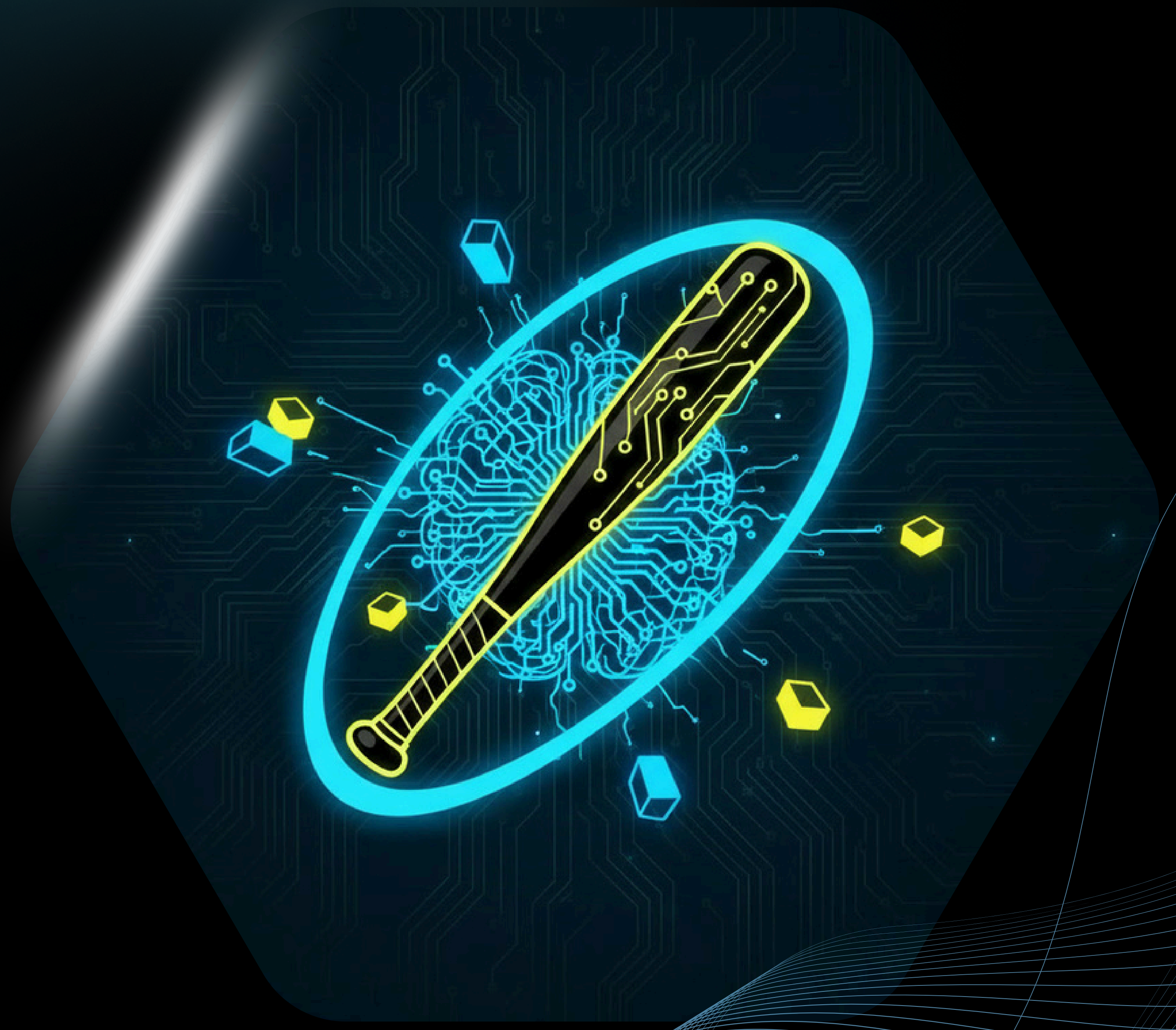
BonkLM protects Your Stack:

AI Chatbot & Assistants: Block prompt injection and jailbreak attempts before they reach your LLM.

Customer Facing APIs: Validate every inbound message at the edge, before it touches your model.

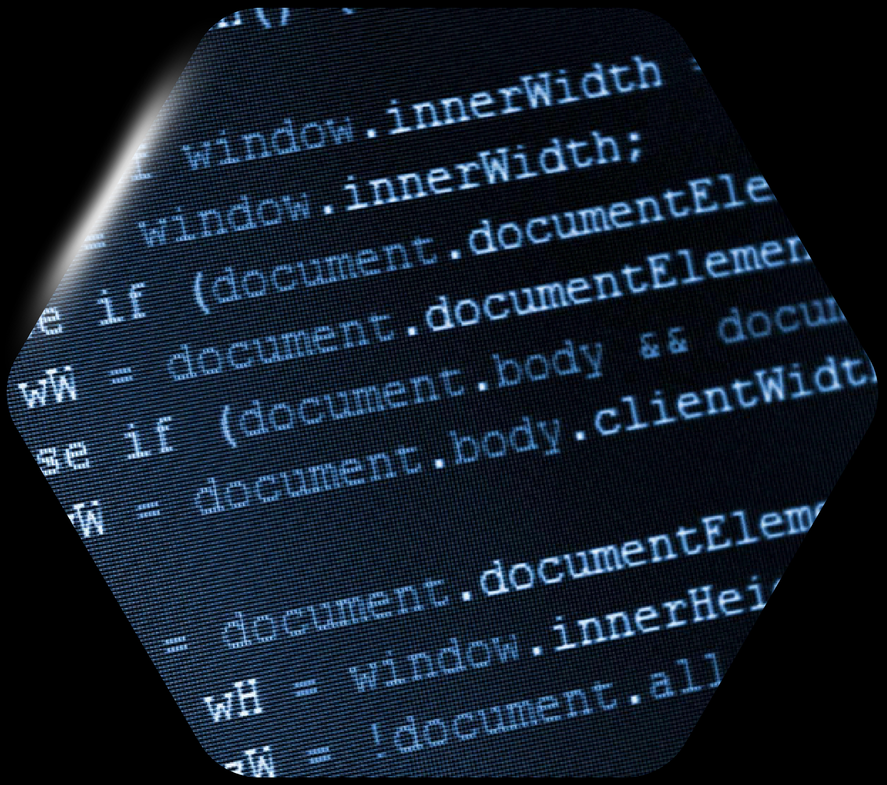
RAG Pipeline: Sanatize users queries and prevent data exfiltration through Retrieval-Augmented Systems.

Code Generation and Dev Tools: Catch credential leaks and PII exposure in developer-facing AI features.

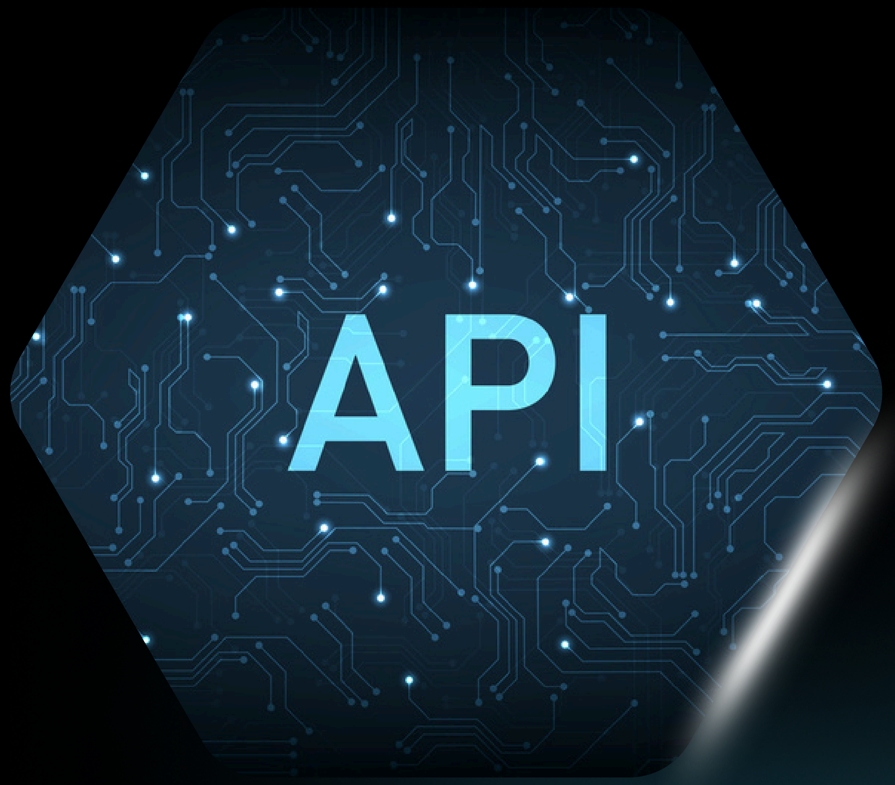


USE CASES

Developer Copilot: Secret Guard, Bash Safety, Guard Catch & dangerous shell commands.
OpenClaw: Secure Agent Orchestration, block instructions hijacking & prompt injection.
Stack Integration: 19 Native connectors and typescript API with a single npm install



```
window.innerWidth
window.innerWidth;
e if (document.documentEle
wW = document.documentElemen
se if (document.body && docu
W = document.body.clientWidt
= document.documentElem
wH = window.innerHei
wW = !document.all
```



Benefits for Users

BonkLM gives your AI applications a security layer that's fast to ship, easy to maintain, and built for production. BonkLM drops into your existing Node.js stack in minutes with zero custom configuration required. It's framework-agnostic, provider-agnostic, and TypeScript native, so it fits your architecture, not the other way around.

- ✓ Zero Friction Setup
- ✓ Full Attack Surface Coverage
- ✓ Built for Scale
- ✓ Works with Everything You Already Use

How does
It works?

BonkLM sits as a lightweight validation layer between your application and your LLM. Every message passes through it before anything reaches the model.

01 Message Comes In

A user sends a message to your app.
Before it touches your LLM, BonkLM intercepts it.

02 Guardrails Engine

The Guardrail Engine runs your configured validators in sequence: checking for prompt injection, jailbreak patterns, PII, secrets, and more.

03 Risk Score and Kill Switch

Every check returns a structured result, a score, and a human-readable reason.

04 Safe Traffic Passes Through

Clean messages proceed to your LLM as normal. Blocked messages never reach it, your app handles the response however you choose. Both Inputs and Outputs are validated in real time stream.

Security Coverage

BonkLM protects against 8 threat categories, covering every major attack vector targeting LLM-powered applications.

Prompt Injection

Detects malicious attempts to override your system instructions or hijack the model's behavior.

Reformulation Patterns

Catches obfuscated attacks using character encoding tricks, code format injection, and context overload.

Jailbreak Detection

Identifies DAN attacks, roleplay exploits, and social engineering tactics designed to bypass your model's safety guidelines.

Secret Guard

Scans inputs and outputs for exposed API keys, tokens, and credentials before they leak into logs or responses.

PII Guard

Detects personal information including SSNs, emails, and phone numbers using international patterns.

Bash Safety Guard

Blocks command injection attempts in shell execution contexts, catching dangerous patterns before they reach your infrastructure.

XSS Safety Guard

Identifies cross-site scripting vectors in LLM-generated content before it's rendered in a browser.

Streaming Validator

Runs all checks in real time against LLM output streams, validating chunk by chunk so threats are caught before the full response is delivered.

INTEGRATIONS

BonkLM automatically plugs into your existing stack. No rewrites, no lock-in.

Web Frameworks

Express, Fastify, NestJS, Next.js. BonkLM ships as native middleware for every major Node.js framework.

Agentic Frameworks

LangChain, Ollama, OpenClaw. Secure your agent pipelines and tool-calling workflows at the orchestration layer.

Emerging Frameworks

Mastra, Google Genkit, CopilotKit. BonkLM keeps pace with the ecosystem as new frameworks emerge.

LLM Providers

OpenAI, Anthropic, Vercel AI SDK. Wrap your existing SDK calls with a single guardrail layer, no API changes required.

RAG & Vector Stores

LlamaIndex, Pinecone, ChromaDB, Weaviate, Qdrant, HuggingFace. Validate queries and responses across your entire retrieval pipeline.

MCP Native

Native MCP connector to secure tool-use and context injection in MCP-based agent architectures.

nest · OpenAI · ANTHROP\C · Vercel · OpenClaw

LlamaIndex · Pinecone · chroma · weaviate · Hugging Face · mastra · Genkit · CopilotKit

Getting Started With BonkLM

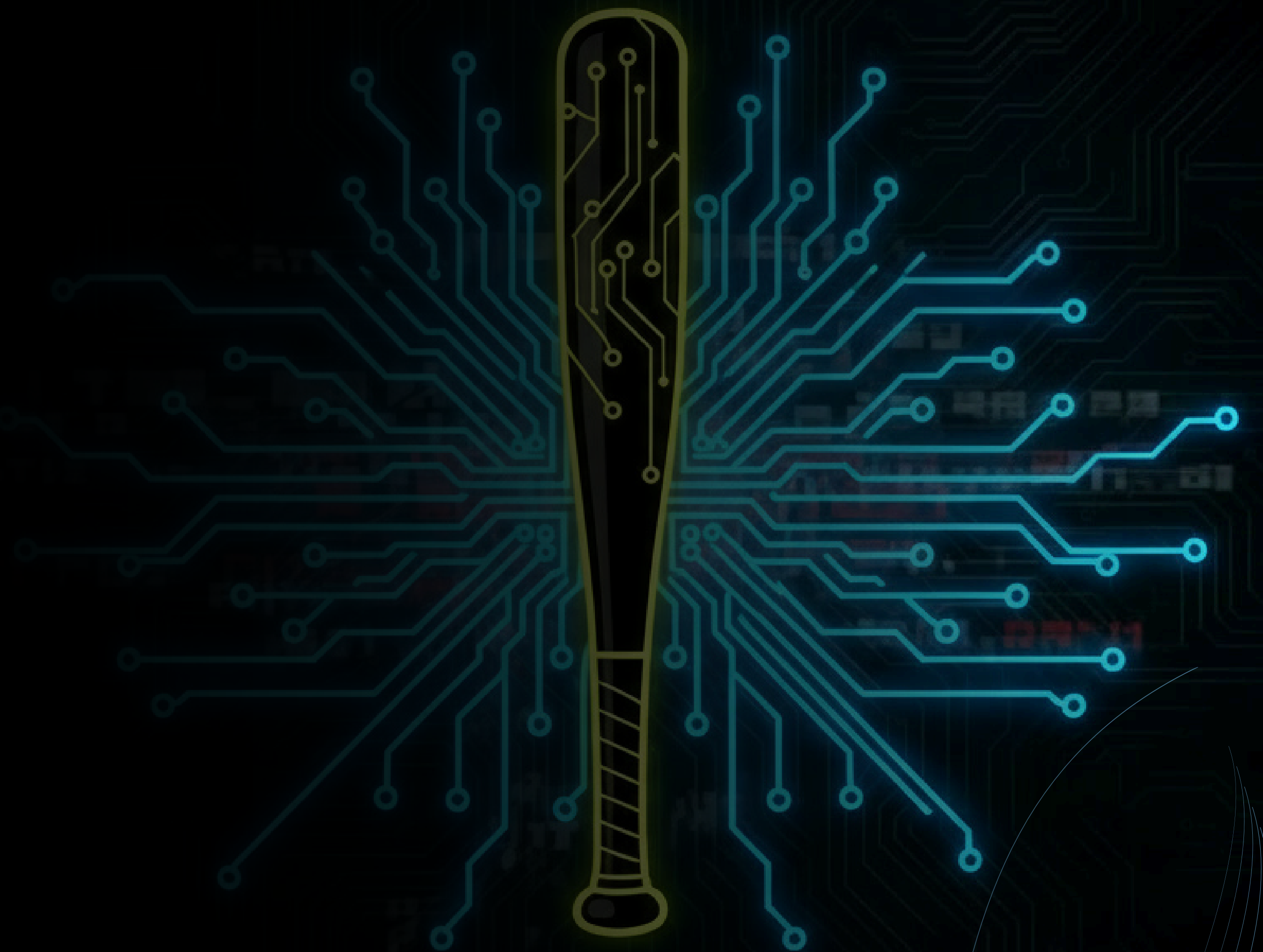
```
npx @blackunicorn/bonklm
```

The wizard auto-detects your framework and LLM provider, installs dependencies, and generates your configuration.

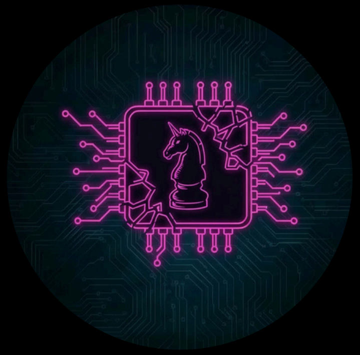
Two steps. Any framework. Any LLM. Done.

BonkLM

</> https://github.com/BlackUnicornSecurity/bonklm



BLACK UNICORN
AI SECURITY LAB



Basileak: LLM engineered with layered vulnerabilities, designed as a controlled training ground for security researchers and practitioners, It features a six-stage Capture The Flag framework.

Shogun: LLM engineered from the ground up with hardened defense against injection, hijacking, and manipulation. Shogun's training incorporates security alignment techniques that allow it to operate reliably in adversarial environments where inputs cannot be trusted.

PantheonLM: Multi-agent AI framework purpose-built for professional security and intelligence operations, orchestrating specialized teams through a single, unified interface.

DojoLM: AI red-teaming and security lab that lets researchers scan LLMs for prompt injection, jailbreak, and output manipulation vulnerabilities.

black unicorn

AI Security, Cybersec & Compliance.

✉ info@blackunicorn.tech

🌐 blackunicorn.tech