SECURITY REFERENCE CARD

AI Security Guardrails Checklist for Development Teams

A dense, actionable reference for every team integrating AI into their development workflow. Cover all three domains before shipping any AI-powered feature.

🔒 Data Privacy

- [ ] Never send secrets, PII, or proprietary algorithms to external APIs
- [ ] Use environment variables for API keys — never hardcode credentials
- [ ] Check provider's data processing terms — opt out of training where available
- [ ] Consider self-hosted models for sensitive or regulated data

🛡️ Prompt Injection Defenses

- [ ] Separate data from instructions using clear delimiters
- [ ] Validate and sanitize all user input before passing to the model
- [ ] Validate AI output before acting on it — never trust blindly
- [ ] Apply least privilege: restrict what the AI agent can access or execute
- [ ] Require human-in-the-loop approval for sensitive or irreversible operations

⚖️ IP & Team Policy

- [ ] AI-generated code via commercial APIs: you own it — but always review it
- [ ] Flag AI-generated code explicitly in PRs and commit messages
- [ ] Team policy **must** cover: approved tools, data classification rules, and review requirements

> **"Security is not optional. Every AI integration needs guardrails."**

Approved Tools

Maintain a vetted list of AI tools permitted for use in production workflows.

Data Rules

Define what data classifications may be sent to which AI providers.

Review Requirements

Mandate peer review for all AI-generated code before merge.