codemate.ai

TECHNICAL & SECURITY OVERVIEW

AI Guardrails & Platform Security

Prepared for **HP** · Confidential · 2026

Overview

CodeMate is an AI-native SDLC agent that understands an entire codebase and automates research, prototyping, development, testing, and code review across the software lifecycle. It runs local-first inside the developer's own environment and supports on-premises, VPC, and cloud deployment. As a strategic ecosystem partner, HP hardware helps power CodeMate's on-device, hardware-optimized inference.

Because CodeMate works with sensitive codebases and proprietary logic, security is engineered into every interaction through a two-layer model that spans the model itself and the infrastructure around it.

LAYER 1

AI Inference Guardrails

Real-time controls at the model layer that intercept harmful, manipulative, or policy-violating inputs and outputs before they reach users.

LAYER 2

Software Security Systems

Hardened backend and infrastructure controls governing data storage, access, network traffic, and service continuity.

01 AI Inference Guardrails

Four systems operate directly on the inference pipeline, in real time, before and after every model output is delivered.

1 Moderation Handling

An ensemble of independently trained classifiers evaluates every request and response across multiple safety dimensions, so no single model failure can reach the user.

- **Ensemble voting —** multiple classifiers must agree before a response passes through.
- **Graduated severity —** low-risk outputs are flagged, medium-risk are rewritten, high-risk are blocked and logged.
- **Fast and auditable —** classifiers run in parallel with no perceivable latency and feed a central audit log.

2 Injection & Jailbreak Handlers

Detection uses a combination of classifier models and small language models (SLMs). Rule-based approaches fail to catch the nuances of long, well-structured, and well-worded prompts, so the models interpret intent in full context rather than matching fixed patterns.

- **Classifiers and SLMs —** classifiers score risk at speed while purpose-built SLMs, fine-tuned on adversarial corpora, reason over the whole prompt and generalize to novel, unseen attacks.
- **Full coverage —** separate detection for direct, indirect (via retrieved content), and role-play jailbreaks.
- **Policy manifest —** outputs are compared against a safety policy manifest; deviations above threshold are blocked.
- **Continuous retraining —** newly discovered attack patterns enter the next training cycle within 48 hours.

3 PII Protection

Sensitive data is detected and masked before it reaches the model, so private values are never stored, logged, or exposed.

- **Broad detection —** names, emails, phone numbers, national IDs, financial data, IP addresses, API keys, and custom patterns.
- **Vaulted masking —** values are masked pre-inference and held in an encrypted vault, rehydrated only when required.
- **Enterprise controls —** configurable sensitivity profiles for regulated environments, plus optional zero-retention sessions.

4 Context Boundary Protection

Strict isolation ensures one user's, project's, or organization's data can never leak into another's context on the shared model.

- **Hard session isolation —** each context is scoped to a unique, cryptographically signed session token.
- **RAG guards —** retrieval-augmented content is validated against the requesting user's permitted scope.
- **Tenant namespaces —** all lookups are scoped by organization ID, preventing cross-tenant access.

02 Software Security Systems

Operating independently of the AI pipeline, these controls harden the backend, databases, and network infrastructure.

5 Network Abuse Systems

Active detection and mitigation protect the platform from malicious traffic and automated exploitation at both the edge and application layer.

- **Rate limiting —** adaptive per-user, per-IP, and per-organization throttling based on behavioural baselines.

- **Threat defense —** anomaly-based intrusion detection, IP reputation scoring, DDoS mitigation, and bot detection.

6 AES Encryption

All sensitive data is encrypted at rest and in transit, leaving it unreadable to unauthorized parties even in a breach.

- **Standards —** AES-256 at rest for projects, session logs, and the PII vault; TLS 1.3 enforced in transit.
- **Key management —** keys are managed separately with automatic rotation and never stored beside the data they protect.

7 Role-Based Database Access

Strict RBAC limits the blast radius of any credential compromise; nothing accesses more data than its function requires.

- **Least privilege —** scoped roles with separated read and write permissions.
- **Gated access —** human access via just-in-time PAM with session recording; queries logged to an immutable store.

8 LLM Kill-Switches

Operationally tested controls let human operators partially or fully disable inference with minimal latency during any incident.

- **Global & granular —** a global switch halts all inference platform-wide; granular switches disable individual models.
- **Automated response —** automatic triggers, incident-ticket logging, and a graceful-degradation fallback mode.

03 Deployment & Data Handling

CodeMate is designed to keep enterprise code and knowledge inside the customer's boundary.

- **Flexible deployment —** local-first execution plus on-premises, VPC, Kubernetes, machine-image, and managed-cloud options — the knowledge base can be fully self-hosted.
- **Hardware-optimized —** ONNX runtime, memory-resident models, and dynamic NPU/GPU routing enable sub-100ms, offline, on-device inference on HPE and partner hardware.
- **Data minimization —** prompts and responses are held encrypted for roughly 30 seconds and are never used to train models; usage data is opt-out.

Security at CodeMate is a layered, continuously evolving architecture that protects users, their code, and the platform at every point in the request lifecycle — from the raw network packet to the semantics of an AI-generated response.

codemate.ai · contact@codemate.ai