# Current Security State

Platform security controls implemented by development team | Reviewed March 2026

Version 2.0 | Supersedes March 9, 2026 release | Updated for CIL v1.6

**ASSESSMENT: LAUNCH-READY | Solid security posture for pre-revenue SaaS. No critical gaps blocking March 25. AI governance and memory security controls added in v2.**

New in v2.0: Control 11 — AI Governance & Memory Security (CIL v1.6). Cross-reference to CIL v1.6 added. Governance Event Log elevated to implemented control. Roadmap updated with CIL v1.6 implementation items.

## Implemented Controls

**1 Strong Auth Basics**

Passwords hashed with bcrypt (12 rounds). JWTs scoped (user vs participant) and expire (7d/24h).

`userService.ts`

**2 Secret Hardening**

Server refuses to start if JWT_SECRET missing or under 32 chars. Fail-safe, not fail-open.

`env.ts`

**3 Password Reset Protections**

Reset tokens SHA-256 hashed, expire in 1 hour. Forgot-password always returns success to prevent email enumeration.

`auth.ts`

**4 Admin Authorization (Server-Enforced)**

Admin routes require valid bearer token, DB role check, and not-banned status. No client-side-only checks.

`adminAuth.ts`

**5 Strict Input Validation**

REST and WebSocket payloads validated with Zod and .strict(). Rejects unknown fields automatically.

`auth.ts validation.ts`

**6 API Hardening**

Helmet enabled, strict CORS allowlist (blocks missing Origin), request size cap (16kb).

`index.ts cors.ts`

**7 Rate Limiting & Abuse Control**

Auth/reset limits, room/general API limits, WS upgrade limits per IP, per-socket message/AI cooldowns.

`auth.ts rateLimiter.ts wsRateLimiter.ts message-handlers.ts`

**8 Room/File Access Control**

Uploads/downloads require auth, ban check, room membership. Filenames path-sanitized. Attachments force download + nosniff.

`upload.ts`

**9 Upload Safety Controls**

MIME + extension allowlist, 10MB max file size, UUID storage names. Prevents path traversal and executable uploads.

`upload.ts`

**10 Frontend XSS/Clickjacking Defenses**

Markdown sanitized with DOMPurify. CSP defined. Vercel sets security headers (CSP, HSTS, X-Frame-Options, nosniff).

`markdown.pipe.ts index.html vercel.json`

**11 AI Governance & Memory Security** NEW v2.0

Room-scoped data isolation enforced at architecture layer — no cross-room aggregation or global user profiling. Governance Event Log: tamper-evident, append-only audit trail for all governance setting changes. Psychological profiling guardrail active. Inference flagging detects sensitive persistence events. Full spec: CIL v1.6 Sections 14-15.

`cil.ts governanceLog.ts memoryPolicy.ts`

**AI Governance cross-reference:** AI behavioral governance, deliberation integrity, and memory policy are specified in CIL v1.6 at theaisymposium.net/governance — the companion document to this assessment.

# Security Hardening Roadmap

Prioritized next steps organized by timeline | March 2026

## Before Launch (March 25)

CRITICAL **Auth-Specific Rate Limiting**
5 failed login/reset attempts per IP per 15 min, then hard block. Credential stuffing is the #1 attack vector for new SaaS.

CRITICAL **Structured Logging on Auth Events**
Log all auth failures, rate limit hits, upload sanitization failures with IP, timestamp, user agent.

HIGH **JWT Refresh Token Rotation**
Implement short-lived access tokens (15-30 min) with longer-lived refresh token. Rotate refresh token on each use.

HIGH **CORS Allowlist Audit**
Verify CORS is not wildcard in any environment (dev, staging, prod). Audit all Vercel environment configs.

HIGH **WebSocket Connection Limits**
Add total concurrent connection limit per user to prevent socket exhaustion. Per-socket cooldowns already exist.

## Post-Launch (April 2026)

MEDIUM **Security Headers Audit**
Run through securityheaders.com and Mozilla Observatory. Verify CSP, HSTS, Referrer-Policy in production.

MEDIUM **Dependency Vulnerability Scanning**
Run npm audit weekly. Automated alerts for critical CVEs. Consider Snyk or GitHub Dependabot.

MEDIUM **RBAC Enforcement Layer**
Deploy roles table, usage_tracking table, and middleware. Every API call checks role (403) and rate limit (429).

MEDIUM **Two-Factor Authentication (2FA)**
TOTP-based 2FA for Pro/Enterprise tiers before Entra migration. Adds immediate value for early paying customers.

MEDIUM **Admin Action Audit Logs**
Log all admin actions (role changes, bans, billing) with timestamp, actor, and target. Enterprise customers will ask.

MEDIUM **CIL Governance Event Log — Implementation** NEW v2.0
Append-only log per CIL v1.6 Section 14. Every governance setting change creates a permanent, timestamped room event.

MEDIUM **Memory Policy Controls — Implementation** NEW v2.0
Per CIL v1.6 Section 15: Persist Context toggle, Forget This Exchange, Purge Room Memory, consent prompt, inference flagging.

LOW **HTTPS Certificate Pinning (Mobile)**
When Android app ships, implement certificate pinning to prevent MITM on mobile API traffic.

## Future (Q2-Q3 2026)

PLANNED **Penetration Testing**
Engage pen-test firm once paying customers justify cost. Focus: API security, WebSocket surface, multi-tenant isolation.

PLANNED **Entra Identity Migration**
Replace JWT with Entra B2C (consumer), then B2B (enterprise SSO). RBAC layer is auth-agnostic.

PLANNED **Compliance Module + API Key Rotation** UPDATED
Audit logging with reasoning traces, tiered access controls, data residency routing. CIL v1.6 integrates directly.

PLANNED **Secrets Management (Azure Key Vault)**
Migrate from environment variables to Azure Key Vault for all secrets. Pairs with Entra migration. Critical at scale.

PLANNED **Clean Room Sovereign Deployment** NEW v2.0
Air-gapped Sovereign tier with local Llama fallback, data residency routing, full CIL v1.6 Clean Room enforcement.

**Document relationship:** Platform infrastructure security is documented here. AI governance integrity and memory policy are in CIL v1.6 at theaisymposium.net/governance. Both documents are required for enterprise and regulated-industry security review.