# Epikriz AI Architecture Whitepaper

## Document Scope

This whitepaper explains the v1 architecture baseline for Epikriz AI public web and clinical runtime boundaries. The system is verification-first and does not replace clinical judgment.

- Public site and clinical app are isolated deployments.
- Core trust model: citation-linked answers and source verification.
- Primary public language in v1 is Turkish.

## Audience

The intended readers are clinical leaders, IT administrators, security/compliance officers, and institutional buyers evaluating deployment and governance fit.

# 1. System Architecture

## Layer Model

Isolation protects patient-facing operations from crawler or campaign traffic and keeps governance boundaries explicit.

- Public Layer: epikriz.ai landing and documentation.
- Clinical Runtime: demo.epikriz.ai authenticated workflow.
- Internal Infrastructure: DB/vector, queue workers, OCR, provider adapters.

## Operational Flow

- Upload -> secure validation -> extraction/OCR -> chunk/embedding.
- Query -> retrieval -> generation -> citation and verification.
- Audit events persist across critical security and clinical actions.

# 2. Security Model

## Identity and Access

- Role model includes SYSADMIN, DOCTOR, ASSISTANT, SECRETARY.
- MFA and rate limiting controls are part of auth flow.
- Assignment-scoped access limits patient/document visibility.

## Data and File Controls

- Magic-byte validation and filename/path sanitization.
- Queue-based processing with explicit failure states.
- Deployment options support on-prem and single-tenant operation.

# 3. Retrieval and Citation Model

## Grounded Retrieval

Retrieval scopes are intent aware and combine documents, notes, visits, and global context with explicit model transparency.

- Citation metadata includes source context and confidence signals.
- Verify with source is a first-class interaction, not a hidden detail.
- Prompt-safety filters run before generation.

# 4. Audit and Governance Boundaries

## Fail-closed Audit Principle

Critical flows rely on auditable events. If audit persistence fails, the action is treated as blocked for governance integrity.

- Auth, security, patient, document, and query events are captured.
- Admin audit workspace supports filtering and export.
- Risky outputs include additional interaction friction before copy/export.

# 5. Institutional Deployment Posture

## Baseline Requirements

- PostgreSQL + pgvector, Redis + worker queue, OCR and embedding stack.
- Environment-specific app URL and analytics endpoint wiring.
- Backup and recovery plans are mandatory for production readiness.

## Air-gapped Considerations

Air-gapped deployments require provider strategy decisions before rollout and should include operational fallback playbooks.

# 6. Public Experience and Trust Messaging

## Message Guardrails

- Positioning centers on controlled, verifiable, auditable behavior.
- The system is not a diagnostic engine and does not prescribe treatment.
- Claims map to explicit source files to avoid marketing drift.

## Conversion Architecture

CTA hierarchy follows Demo -> Docs -> App open. FAQ and comparison modules answer institutional objections early.

# 7. Appendix and Decision Log

## Version and Change Policy

- Version: v0.1.0 (2026-02-22).
- All public claims should remain traceable to implementation references.
- Future updates should include diff notes for security, retrieval, and governance impacts.

## Contact for Evaluation

Institutional review requests should use the public demo request channel and include deployment preference and security evaluation context.