An architecture you can inspect.

Separate the public demonstration from customer workloads. Apply identity and document permissions before any content reaches the model.

The controlled retrieval path

Identity Authenticated user and authorized tenant	**Retrieval** Document permissions checked before ranking	**Inference** Approved context, bounded generation, no tools	**Review** Citations, limitations, and audit event

Ingestion starts with approved sources, bounded parsing, provenance, classification, and indexing. Re-index or revoke access when documents or permissions change.

Trust boundaries

Boundary	Required control
Browser to application	TLS, authenticated session, input validation, origin checks, rate limits
Application to data	Private network, least-privilege credentials, tenant and document filters
Application to inference	Fixed configured endpoint, timeouts, bounded tokens, isolated model network
Operations to system	MFA, role review, audit trails, controlled maintenance and recovery

Models are replaceable; evidence is required

Adapters support local serving without binding the product to one model. Select candidates against approved tasks and licensing requirements. Evaluate Bulgarian and English separately, including tables, terminology, missing evidence, and citation accuracy.

Retrieval-only mode shows source passages without claiming model generation. A configured model changes the behavior and must be evaluated before production use.

What this does not prove

Local hosting does not remove prompt injection, poisoning, access-control mistakes, or operational risk. Retrieved text is untrusted input. Model output must be treated as text and reviewed before consequential action. This overview is not a certification or a customer-specific security assessment.