# A Sovereign On-Premise AI Appliance Architecture for Regulated Institutions: Open-Weight Models, Auditable Retrieval-Augmented Generation, and Regulatory Mapping for Türkiye

Sur Research Lab · *Istanbul, Türkiye · research@sur.systems*

**Abstract** — Regulated institutions in Türkiye — banks, payment and e-money institutions, capital-market intermediaries, insurers, and anti-money-laundering (AML) units — face a structural dilemma: the document-intensive workloads that benefit most from large language models (LLMs) involve precisely the data classes that national regulation prevents from leaving institutional boundaries. This paper argues that the recent maturation of the open-weight model ecosystem dissolves this dilemma, and proposes a reference architecture for a sovereign, on-premise AI appliance that operates with zero default data egress, supports fully air-gapped operation, and produces audit-ready evidence for supervisory review. We derive architectural requirements from five Turkish regulatory regimes (KVKK, BDDK, TCMB, SPK, MASAK) and the extraterritorial provisions of the EU AI Act; specify a layered design comprising model portfolio, quantized inference, grounded retrieval-augmented generation (RAG) with enforced citation and abstention, and a cryptographically signed offline update chain; and analyze two decision dimensions that generic architecture work tends to neglect: the enterprise implications of open-weight license heterogeneity, and Turkish-language performance including tokenizer fertility effects on serving cost. We conclude with an honest treatment of the cost envelope and the limitations of the approach.

***Index Terms*** *— sovereign AI, on-premise deployment, open-weight language models, retrieval-augmented generation, data protection, financial regulation, air-gapped systems, Turkish NLP.*

## I. Introduction

Large language models deliver measurable productivity gains in document search, summarization, information extraction, compliance analysis, and investigative support — workloads that dominate the back office of regulated financial institutions. Yet the dominant delivery mechanism for frontier-quality models remains the public cloud API, which requires prompt content and retrieval context to transit infrastructure outside the institution's legal and physical control. For Turkish institutions this creates direct tension with data-protection law (KVKK, Law No. 6698), banking information-systems regulation (BDDK), payment-systems rules (TCMB), capital-markets obligations (SPK), and the confidentiality regime surrounding suspicious-transaction data (MASAK).

The tension is not hypothetical. Publicly documented incidents include employees pasting proprietary source code and meeting notes into a public chatbot, prompting an enterprise-wide ban [20]; a caching defect in a major AI service that exposed other users' conversation titles and partial payment details [21]; and the Italian data-protection authority's temporary restriction of a generative AI service, followed by a €15M administrative fine in December 2024 [21], [22]. Several global banks restricted employee use of public chatbots from 2023 onward [21]. The common institutional response — banning generative AI outright, or confining it to non-sensitive, low-value tasks — forfeits most of the attainable value.

Until recently, the counterargument to on-premise deployment was capability: open models trailed proprietary frontier models by a wide margin. This has changed. Epoch AI's longitudinal analysis

estimates the gap between the best open-weight and best closed models at roughly four months of frontier progress on aggregate capability indices [1], [2] — comparable to the interval between successive point releases within a single proprietary family. For retrieval-grounded enterprise workloads, which do not require frontier agentic capability, current open-weight models are functionally sufficient (Section VI discusses caveats).

This paper makes four contributions:

- a requirements analysis that derives concrete architectural constraints from five Turkish regulatory regimes and the EU AI Act (Section III);
- a reference architecture for a sovereign on-premise AI appliance satisfying those constraints, including a grounded-RAG pipeline with enforced citation and abstention, and an air-gapped signed update chain (Section IV);
- an analysis of open-weight license heterogeneity (Apache 2.0, MIT, Llama Community License, Gemma Terms) as an enterprise risk dimension distinct from technical capability (Section V-A);
- an assessment of Turkish-language considerations — public benchmarks, tokenizer fertility, and the trade-off between multilingual base models and Turkish-native models (Section V-B).

Throughout, we deliberately separate what the architecture is *designed to support* from any claim of regulatory certification: conformity assessment remains the deploying institution's responsibility, and the mappings presented here are design rationale, not legal opinion.

## II. Background and Related Work

### *A. The Open-Weight Model Ecosystem*

The open-weight ecosystem circa mid-2026 comprises several model families of enterprise relevance: Qwen (Alibaba; broad size range, permissive licensing for most Qwen3-generation checkpoints, wide multilingual coverage) [3]; DeepSeek (MIT-licensed mixture-of-experts models closest to the closed frontier) [4]; Mistral (European provenance; per-model licensing split between Apache 2.0 and research licenses) [5]; Gemma (Google; the most recent generation moved to Apache 2.0, earlier generations remain under custom terms) [6], [7]; Llama (Meta; large ecosystem, community license with use restrictions) [8]; and OpenAI's Apache-2.0 gpt-oss models [9]. A material share of the open-weight frontier now originates from Chinese laboratories [4], [10], which — although weights execute entirely within the institution's perimeter — introduces model provenance as a procurement-policy variable in its own right.

### *B. Sovereign AI as a Policy Trend*

On-premise deployment of open models aligns with a broader policy movement. European initiatives include Franco-German sovereign AI programs for public administration and national open-model efforts [11]; Gulf states have committed large-scale national AI infrastructure investments [12]; and major European banks have publicized self-hosted deployments of open models [13]. In Türkiye, the 2026–2030 National AI Action Plan announced in June 2026 prioritizes domestic model development, data-center capacity, and a regulatory framework [14], while TBMM's AI Research Commission recommended preparation of a national AI law in its March 2026 report [15]. The EU's Digital Operational Resilience Act, in force since January 2025, further pressures financial institutions to document and reduce cloud dependencies [13].

## III. Regulatory Requirements Analysis

We analyze five national regimes and one extraterritorial regime, extracting from each the constraints that bind an AI system processing regulated data. Table I summarizes the derived requirements R1–R8.

TABLE I — REGULATORY DRIVERS AND DERIVED ARCHITECTURAL REQUIREMENTS

| Req. | Requirement | Primary drivers |
|---|---|---|
| R1 | No default data egress: prompts, context, embeddings, and logs must not leave the institutional perimeter | KVKK; BDDK; TCMB; MASAK |
| R2 | Support for fully air-gapped operation with offline, integrity-verified updates | KVKK Art. 6 (special categories); MASAK; critical infrastructure |
| R3 | Domestic (in-country, in-institution) location of all primary and secondary processing systems | BDDK information-systems regulation [28] |
| R4 | Role-based access, policy-based redaction, configurable retention and destruction | KVKK (minimization, purpose limitation, retention); SPK (information barriers) |
| R5 | Full traceability: every answer attributable to source, time, policy context, and generating model version | BDDK; SPK; internal audit; EU AI Act transparency |
| R6 | Human decision authority on high-risk outputs; abstention on insufficient evidence | MASAK (investigator responsibility); KVKK GenAI guidance [29] |
| R7 | Institutional ownership of indexes, logs, evidence, and fine-tuned artifacts; contractual exit and data-return | BDDK (outsourcing continuity); operational resilience |
| R8 | License and provenance inventory of all deployed models, included in audit evidence | procurement risk; IP risk (Section V-A) |

### *A. KVKK*

Beyond Law No. 6698's general principles, the Personal Data Protection Authority published a dedicated guide on generative AI and personal data in November 2025 [29] and a note on workplace use of third-party generative AI tools [30], explicitly flagging the risk of feeding personal data into external tools. These instruments motivate R1, R4, and R6, and — because embedding vectors are derived personal data when computed over personal records — require that embedding computation itself remain in-perimeter (R1).

### *B. BDDK*

The Regulation on Banks' Information Systems and Electronic Banking Services (Official Gazette 31069, 15 March 2020) requires primary and secondary systems to be located in Türkiye and holds the bank fully responsible for outsourced services [28]. No AI-specific BDDK regulation existed at the time of writing; AI workloads touching bank data are therefore governed through this general regime, motivating R3, R5, and R7.

### *C. TCMB, SPK, MASAK*

Payment and e-money regulation renders transaction data highly sensitive (R1); capital-markets obligations around inside information motivate role isolation and information barriers (R4); and the confidentiality regime around suspicious-transaction reports both prohibits egress (R1, R2) and requires that investigative conclusions remain human decisions (R6). We found no dedicated AI guidance from SPK or MASAK at the time of writing and make no claim that such guidance exists.

### *D. EU AI Act*

The EU AI Act entered general applicability on 2 August 2026, with high-risk obligations for Annex III systems deferred to December 2027 under the 2026 Digital Omnibus [31], [32]. Its extraterritorial scope covers Turkish providers and deployers whose system outputs are used in the EU [33]. Traceability and technical-documentation duties reinforce R5 and R8 for institutions with EU exposure.

## IV. Reference Architecture

### *A. Design Principles*

The architecture is an appliance: a pre-integrated hardware-software unit installed in the institution's data center. Five principles follow from Table I: (1) zero default egress — the appliance initiates no external connections; (2) air-gap capability with full functionality offline (R2); (3) customer ownership of all indexes, logs, and evidence (R7); (4) no default remote access — support sessions are customer-initiated, logged, and time-bounded; (5) no training on customer data.

Functionally, the appliance decomposes into six modules: a connector/indexing layer; a governance studio (roles, redaction, retention); an employee assistant constrained to grounded answers; a compliance center that maps controls to regulation and generates evidence packages; a sealing service for high-risk answers; and an optional AML investigation assistant operating strictly under human control.

### *B. Model Portfolio and Hardware Tiers*

No single model serves all workloads economically. The portfolio approach assigns: 20–35B-parameter-class modern models (dense or low-active-parameter MoE) to chat and RAG workloads; 70–120B-class models to complex analysis; near-frontier MoE models to the most demanding reasoning tasks; and a small 2–9B model to auxiliary functions (classification, redaction, routing). Memory sizing follows the standard heuristic of ~2 bytes/parameter at FP16, ~1 at FP8, ~0.5 at INT4, plus 20–30% for KV cache and concurrency [34]. Table II illustrates the resulting tiering.

TABLE II — APPLIANCE TIERS (ILLUSTRATIVE)

| Tier | GPU configuration | Model class |
|---|---|---|
| T1 | 1× 96 GB-class enterprise GPU | 20–35B (FP8/INT4) |
| T2 | 2–4× 96–141 GB GPUs | 70–120B (FP8/INT4) |
| T3 | 8× HBM-class GPUs | near-frontier MoE |
| T4 | multi-node cluster | portfolio + HA |

Quantization policy balances quality against capacity: FP8 executes natively on current GPU generations with negligible measured quality loss on most workloads, while AWQ-style INT4 quarters memory at modest quality cost [34], [35]. Because quality impact is model- and workload-dependent, the architecture mandates validation against the institution's own evaluation set before any quantized model enters production (Section IV-F).

### *C. Inference Layer*

The serving layer builds on production-validated open-source engines. vLLM — with paged attention, continuous batching, and broad model coverage — is the de facto standard [36]; engines offering shared-prefix caching (e.g., SGLang) provide additional throughput on RAG workloads where policy and context templates repeat across thousands of queries [37]. Desktop-class tools are excluded from the production stack as they are not designed for concurrent multi-user serving [36].

### *D. Grounded RAG with Enforced Citation and Abstention*

The pipeline that operationalizes R5 and R6 comprises five stages:

- **Embedding:** self-hosted multilingual open embedding models (BGE-M3-class multi-vector models or current open MTEB leaders) computed entirely in-perimeter [38];
- **Hybrid retrieval:** dense vector search fused with lexical BM25 via reciprocal rank fusion — more robust than dense-only retrieval on terminology-sensitive corpora such as regulation and internal policy [17];
- **Reranking:** cross-encoder rerankers, which measurably improve precision and reduce downstream hallucination [18];
- **Enforced citation:** the generator may only cite retrieval-chunk identifiers present in its context; citations are verified post-hoc against retrieved text, and unverifiable citations block release of the answer;
- **Abstention:** if retrieval confidence falls below threshold the pipeline returns "insufficient sources" without invoking generation; generated answers additionally pass a groundedness check before display [19]. Abstention is logged as a first-class outcome: in a regulated setting, a correct refusal is an assurance feature, not a defect.

Vector stores and all indexes reside in-perimeter on self-hosted open-source databases and are owned by the institution (R7).

### *E. Air-Gapped Update Chain*

All components — weights, container images, package mirrors, observability, PKI — are pre-staged inside the enclave [39]. Updates arrive as cryptographically signed offline bundles (weights, patches, configuration); import requires signature verification, chain-of-custody logging, and explicit customer approval. Licensing is fully offline: any phone-home license check, telemetry, or auto-update mechanism is disqualifying for air-gapped operation [39], [40]. Model updates additionally require regression testing against the institution's evaluation set before activation.

### *F. Audit Evidence Layers*

Auditability operates at three granularities. (1) Sourced answers: every response is traceable to its in-perimeter sources with verified citations. (2) Sealed answers: high-risk responses are sealed with source set, timestamp, policy context, and the exact generating model version, producing an ex-post verifiable record. (3) Evidence packages: the compliance module generates timestamped packages containing configuration, regulation-mapped controls, access roles, update-integrity records, the deployed model and license inventory (R8), and a no-egress verification log. Because model weights and versions reside with the institution, the audit question "which model, under which configuration, produced this answer?" is answerable from the institution's own records rather than a third party's attestation — a structural advantage over API-based deployment.

## V. Model Selection Considerations

### *A. License Heterogeneity as Enterprise Risk*

"Open source" is not a single licensing regime, and the differences carry concrete legal consequence for regulated institutions (Table III). Apache 2.0 and MIT impose no field-of-use restrictions and (in Apache 2.0's case) include an express patent grant [41]. The Llama Community License conditions use on a 700M monthly-active-user threshold, flows an acceptable-use policy down to end users, and imposes

attribution and derivative-naming duties; it does not satisfy the Open Source Definition [8], [42]. Earlier-generation Gemma terms include a flow-down prohibited-use policy and reserve the provider's right to restrict usage remotely — difficult to reconcile with an air-gapped sovereignty posture [7]; the most recent Gemma generation's move to Apache 2.0 resolves this for that generation only [6]. None of these licenses provides IP indemnification, unlike commercial API contracts; institutions self-insure output-related IP risk. Fine-tuned derivatives of Apache 2.0/MIT models belong unconditionally to the institution, whereas Llama derivatives carry naming and policy obligations [8].

TABLE III — LICENSE REGIMES OF MAJOR OPEN-WEIGHT FAMILIES

| License | Example models | Key constraints |
|---|---|---|
| Apache 2.0 | Qwen3 (most), recent Gemma, some Mistral, gpt-oss | none material; express patent grant |
| MIT | DeepSeek | none material; silent on patents |
| Llama Community | Llama family | MAU threshold; AUP flow-down; naming/attribution |
| Gemma Terms (earlier) | earlier Gemma | prohibited-use flow-down; remote restriction right |

The architecture therefore adopts a default policy of Apache 2.0/MIT models only, with restricted-license models deployable solely upon the institution's legal approval and recorded in the license inventory (R8).

### *B. Turkish-Language Performance and Tokenizer Economics*

Two findings from the Turkish evaluation literature shape model selection. First, on comprehensive Turkish benchmarks, large multilingual open models currently outperform most small Turkish-native models [24], supporting a strategy of strong multilingual base model + Turkish domain adaptation + institution-specific Turkish evaluation set, rather than defaulting to Turkish-from-scratch models. Second, tokenizer fertility is a genuine cost lever: at ~1.8–2.5 tokens per Turkish word versus ~1.2–1.4 for English [26], the same document consumes substantially more context window and inference compute in Turkish. Fertility should therefore be measured as an explicit criterion in model evaluation, alongside accuracy. The emergence of competitive Turkish-native models (e.g., Kumru's from-scratch 7.4B and 2B releases [27]) suggests the trade-off may shift; vendor-reported scores should be independently reproduced before they carry procurement weight.

### *C. Provenance*

Given the concentration of open-weight frontier development in Chinese laboratories [4], [10], institutions may impose provenance constraints orthogonal to capability and license. Because open weights execute deterministically in-perimeter with no vendor connectivity, the residual provenance risks differ in kind from cloud dependence (they concern training-data opacity and potential behavioral biases rather than data egress); nonetheless the architecture treats provenance as a configurable procurement filter rather than prescribing a position.

## VI. Discussion

### *A. Cost Envelope: An Honest Framing*

Published break-even analyses between self-hosting and API consumption disagree by orders of magnitude depending on assumptions about volume, utilization, and engineering overhead [43], [44]. At low, bursty volumes, per-token economics favor APIs; at high sustained volume, self-hosting wins

decisively. We argue, however, that per-token economics is the wrong decision frame for regulated institutions: for data classes that cannot lawfully transit a public cloud, the API alternative does not exist at any price. The appliance's economic proposition is the enablement of otherwise-excluded high-value workloads at predictable fixed cost, with token-level cost parity achievable at scale as a secondary effect.

### *B. Open vs. Closed: Residual Gaps*

The ~4-month aggregate gap [1] is not uniform. Open models remain measurably behind on long-horizon agentic tasks, tool-use benchmarks, and some instruction-following nuances [10]; benchmark contamination and selective publication may cause aggregate indices to understate the true gap [2]. The architecture's workload targeting — retrieval-grounded QA, summarization, extraction, drafting — deliberately occupies the regime where the gap is smallest. Institutions requiring frontier agentic capability on non-sensitive data may rationally operate a hybrid posture; the architecture does not preclude this, it isolates what must remain inside.

### *C. Limitations*

Four limitations bound this work. (1) It presents a reference architecture with design rationale, not a deployed-system evaluation; no throughput, accuracy, or cost measurements are reported, and such measurements are environment-specific by design. (2) The regulatory analysis reflects the state of Turkish and EU instruments at the time of writing; several referenced instruments (draft Turkish AI bills, EU AI Act deferrals) are in flux. (3) Model-ecosystem facts (versions, licenses, benchmark standings) change on a monthly cadence; specific claims should be re-verified against primary model cards. (4) The Turkish-language analysis relies on public benchmarks that may not represent institution-specific document genres; the architecture's mandated institution-specific evaluation set is a mitigation, not a solution, for this gap.

## VII. Conclusion

The maturation of the open-weight model ecosystem has dissolved the assumed trade-off between AI capability and regulatory compliance for Türkiye's regulated institutions. We have shown that the binding constraints of five national regulatory regimes can be satisfied by a sovereign on-premise appliance architecture built on permissively licensed open-weight models, grounded RAG with enforced citation and abstention, and a signed offline update chain — while producing audit evidence of a quality unattainable in API-based deployment, because the generating model itself is in the institution's custody. Future work includes empirical evaluation of the abstention mechanism's precision-recall trade-off on Turkish regulatory corpora, longitudinal tracking of the open-closed capability gap on institution-relevant task suites, and extension of the regulatory mapping as Turkish AI legislation crystallizes.

> **Disclaimer.** This document is for informational and design-rationale purposes; it is not legal opinion. All regulatory mappings are presented under a "designed to support" principle and make no claim of formal certification; conformity assessment remains the deploying institution's responsibility. Capacity and performance figures are shared only from benchmarks run in the institution's own environment. References are kept in their original language for traceability.

## References


[1] Epoch AI, "The gap between open and closed models," 2026. https://epoch.ai/data-insights/open-closed-eci-gap

[2] Epoch AI, "Open models report," 2025. https://epoch.ai/blog/open-models-report

[3] Qwen Team, Alibaba Group, Qwen3 model cards and licenses. https://huggingface.co/Qwen

[4] DeepSeek-AI, model releases under MIT license. https://huggingface.co/deepseek-ai

[5] Mistral AI, "Under which license are Mistral's open models available?" https://help.mistral.ai/

[6] Google Open Source Blog, "Gemma 4: Expanding the Gemmaverse with Apache 2.0," Mar. 2026.

[7] Google, "Gemma Terms of Use." https://ai.google.dev/gemma/terms

[8] Meta, "Llama 4 Community License Agreement," Apr. 2025. https://www.llama.com/llama4/license/

[9] OpenAI, gpt-oss model cards. https://huggingface.co/openai

[10] BenchLM, "Best open-source LLMs," 2026.

[11] Euronews, "Which European countries are building their own sovereign AI?" Dec. 2025.

[12] Middle East Institute, "AI, the Gulf, and the US: A primer," 2025.

[13] Capco, "Efficiency in financial institutions: open-source LLMs," 2025; cc-bei.news, "Sovereignty in the age of AI," 2026.

[14] Türkiye AI Initiative, "Türkiye Yapay Zeka Eylem Planı 2026–2030 açıklandı," Jun. 2026. https://turkiye.ai/

[15] TBMM Yapay Zeka Araştırma Komisyonu, Rapor, 30 Mar. 2026.

[16] V. Magesh et al., "Hallucination-free? Assessing the reliability of leading AI legal research tools," Stanford HAI/RegLab.

[17] "A hybrid retrieval framework for enterprise RAG," arXiv:2605.01664, 2026.

[18] "RAG 2.0: Why reranking has become the core of modern RAG systems," 2025.

[19] Red Gate, "How to stop AI hallucinations in enterprise RAG systems," 2025.

[20] TechCrunch, "Samsung bans use of generative AI tools after April internal data leak," May 2023.

[21] Wald.ai, "ChatGPT data leaks and security incidents 2023–2024: a comprehensive overview."

[22] Garante per la Protezione dei Dati Personali, provision against OpenAI, Dec. 2024 (subsequent judicial review pending; verify current status).

[23] "TR-MMLU: A comprehensive benchmark for Turkish," arXiv:2501.00593, 2025.

[24] "Cetvel: A unified benchmark for evaluating language understanding in Turkish," arXiv:2508.16431, EACL 2026.

[25] "TurkBench," arXiv:2601.07020, 2026; MMLU-Pro-TR.

[26] "Tokenization standards for Turkish," arXiv:2502.07057, 2025.

[27] VNGRS, Kumru-2B model card. https://huggingface.co/vngrs-ai/Kumru-2B

[28] "Bankaların Bilgi Sistemleri ve Elektronik Bankacılık Hizmetleri Hakkında Yönetmelik," Resmî Gazete No. 31069, 15 Mar. 2020.

[29] KVKK, "Üretken Yapay Zeka ve Kişisel Verilerin Korunması Rehberi (15 Soruda)," 24 Nov. 2025.

[30] KVKK, "İş yerlerinde üretken yapay zeka araçlarının kullanımı."

[31] European Commission, "Regulatory framework for AI — implementation timeline."

[32] Gibson Dunn, "EU AI Act omnibus agreement: postponed high-risk deadlines," Jun. 2026.

[33] Erdem & Erdem, "Reflections of the EU AI Act on actors in Türkiye," 2025.

[34] Spheron, "GPU memory requirements for LLMs," 2026; VRLA Tech, "LLM VRAM requirements 2026."

[35] VRLA Tech, "Best GPUs for LLM inference and training, 2026."

[36] Red Hat Developer, "llama.cpp vs vLLM: choosing the right local LLM inference engine," Jun. 2026.

[37] Yotta Labs, "Best LLM inference engines in 2026," 2026.

[38] BAAI, BGE-M3 model card; Milvus, "Choosing embedding models for RAG, 2026."

[39] TianPan, "Air-gapped LLM blueprint: egress-free deployment," May 2026.

[40] TrueFoundry, "Air-gapped AI: deploying enterprise LLMs in highly regulated industries," 2026.

[41] Apache Software Foundation, "Apache License, Version 2.0."

[42] Shuji Sado, "Why is the Llama license not open source?" Jan. 2025.

[43] SitePoint, "Local LLMs vs cloud API: cost analysis 2026."

[44] "Scaling down to scale up: a cost-benefit analysis of replacing OpenAI's LLM with open-source SLMs in production," arXiv:2312.14972, 2024.