Regüle Kurumlar için Egemen, Kurum İçi bir Yapay Zeka Aygıtı Mimarisi: Açık-Ağırlık Modeller, Denetlenebilir Erişim-Artırılmış Üretim ve Türkiye için Mevzuat Haritalaması

Hisar Research Lab · *İstanbul, Türkiye* · *research@hisar.tech*

Özet — Türkiye'deki regüle kurumlar — bankalar, ödeme ve e-para kuruluşları, sermaye piyasası aracıları, sigortacılar ve aklama ile mücadele (AML) birimleri — yapısal bir ikilemle karşı karşıyadır: büyük dil modellerinden (LLM) en çok fayda gören belge-yoğun iş yükleri, tam da ulusal mevzuatın kurum sınırlarının dışına çıkmasını engellediği veri sınıflarını içerir. Bu makale, açık-ağırlık model ekosisteminin son dönemdeki olgunlaşmasının bu ikilemi ortadan kaldırdığını savunur ve sıfır varsayılan veri çıkışıyla çalışan, tümüyle hava boşluklu işletimi destekleyen ve denetime hazır kanıt üreten egemen, kurum içi bir yapay zeka aygıtı için bir referans mimari önerir. Mimari gereksinimleri beş Türk mevzuat rejiminden (KVKK, BDDK, TCMB, SPK, MASAK) ve AB Yapay Zeka Yasası'nın ülke dışı hükümlerinden türetir; model portföyü, nicemlenmiş (quantized) çıkarım, zorunlu atıf ve çekimserlik içeren erişim-artırılmış üretim (RAG) ve kriptografik olarak imzalı çevrimdışı güncelleme zincirinden oluşan katmanlı bir tasarım belirler; ve genel mimari çalışmalarının çoğunlukla ihmal ettiği iki karar boyutunu inceler: açık-ağırlık lisans heterojenliğinin kurumsal sonuçları ve tokenizasyon "doğurganlığının" (fertility) hizmet maliyetine etkisi dâhil Türkçe dil performansı. Yaklaşımın maliyet zarfı ve sınırlamaları dürüst biçimde ele alınarak sonuca varılır.

Anahtar Terimler *— egemen yapay zeka, kurum içi kurulum, açık-ağırlık dil modelleri, erişim-artırılmış üretim, kişisel veri koruma, finansal mevzuat, hava boşluklu sistemler, Türkçe DDİ.*

I. Giriş

Büyük dil modelleri; belge arama, özetleme, bilgi çıkarımı, uyum analizi ve soruşturma desteği gibi — regüle finansal kurumların operasyon merkezine hâkim olan — iş yüklerinde ölçülebilir verimlilik kazanımları sağlar. Ancak öncü kalitedeki modeller için baskın sunum mekanizması, istem içeriğinin ve erişim bağlamının kurumun hukuki ve fiziksel kontrolü dışındaki altyapılardan geçmesini gerektiren kamusal bulut API'sidir. Türk kurumları için bu durum; kişisel verilerin korunması mevzuatı (KVKK, 6698 sayılı Kanun), bankacılık bilgi sistemleri düzenlemesi (BDDK), ödeme sistemleri kuralları (TCMB), sermaye piyasası yükümlülükleri (SPK) ve şüpheli işlem verilerini çevreleyen gizlilik rejimi (MASAK) ile doğrudan bir gerilim yaratır.

Bu gerilim varsayımsal değildir. Kamuya belgelenmiş olaylar arasında; çalışanların özel kaynak kodunu ve toplantı notlarını kamusal bir sohbet botuna yapıştırması sonucu kurum genelinde yasak getirilmesi [20]; büyük bir yapay zeka hizmetindeki bir önbellek kusurunun başka kullanıcıların konuşma başlıklarını ve kısmi ödeme bilgilerini açığa çıkarması [21]; ve İtalyan veri koruma otoritesinin bir üretken yapay zeka hizmetini geçici olarak kısıtlaması, ardından Aralık 2024'te 15 milyon avroluk idari para cezası [21], [22] yer alır. Birçok küresel banka 2023'ten itibaren çalışanların kamusal sohbet botu kullanımını kısıtlamıştır [21]. Yaygın kurumsal tepki — üretken yapay zekayı büsbütün yasaklamak ya da yalnızca hassas olmayan, düşük değerli görevlerle sınırlamak — elde edilebilecek değerin çoğunu feda eder.

Yakın zamana kadar kurum içi kuruluma karşı çıkan argüman yetenekti: açık modeller, tescilli öncü modellerin geniş bir farkla gerisindeydi. Bu değişti. Epoch AI'nin boylamsal analizi, en iyi açık-ağırlık ile en iyi kapalı modeller arasındaki farkı toplu yetenek endekslerinde kabaca dört aylık öncü ilerlemeye — tek bir tescilli ailedeki ardışık ara sürümler arasındaki aralığa karşılaştırılabilir düzeyde — tahmin etmektedir [1], [2]. Öncü ajanik yetenek gerektirmeyen, erişime dayalı kurumsal iş yükleri için mevcut açık-ağırlık modeller işlevsel olarak yeterlidir (Bölüm VI çekinceleri tartışır).

Bu makale dört katkı sunar:

- beş Türk mevzuat rejiminden ve AB Yapay Zeka Yasası'ndan somut mimari kısıtlar türeten bir gereksinim analizi (Bölüm III);
- bu kısıtları karşılayan; zorunlu atıf ve çekimserlik içeren grounded-RAG hattı ile hava boşluklu imzalı güncelleme zinciri dâhil, egemen kurum içi bir yapay zeka aygıtı için referans mimari (Bölüm IV);
- açık-ağırlık lisans heterojenliğinin (Apache 2.0, MIT, Llama Topluluk Lisansı, Gemma Şartları) teknik yetenekten ayrı bir kurumsal risk boyutu olarak analizi (Bölüm V-A);
- Türkçe dil hususlarının değerlendirmesi — kamusal kıyaslamalar, tokenizasyon doğurganlığı ve çok dilli temel modeller ile Türkçe-yerel modeller arasındaki ödünleşim (Bölüm V-B).

Makale boyunca, mimarinin *desteklemek üzere tasarlandığı* şey ile herhangi bir mevzuat sertifikasyonu iddiasını bilinçli olarak ayrı tutuyoruz: uygunluk değerlendirmesi kuran kurumun sorumluluğundadır ve burada sunulan eşlemeler tasarım gerekçesidir, hukuki görüş değildir.

II. Arka Plan ve İlgili Çalışmalar

A. Açık-Ağırlık Model Ekosistemi

2026 ortası itibarıyla açık-ağırlık ekosistemi, kurumsal açıdan önemli birkaç model ailesini kapsar: Qwen (Alibaba; geniş boyut aralığı, çoğu Qwen3 kuşağı için serbest lisanslama, geniş çok dilli kapsama) [3]; DeepSeek (kapalı öncüye en yakın, MIT lisanslı uzman-karması modeller) [4]; Mistral (Avrupa kökenli; model başına Apache 2.0 ile araştırma lisansları arasında bölünmüş) [5]; Gemma (Google; en yeni kuşak Apache 2.0'a geçti, önceki kuşaklar özel şartlarda kalıyor) [6], [7]; Llama (Meta; büyük ekosistem, kullanım kısıtlamalı topluluk lisansı) [8]; ve OpenAI'nin Apache-2.0 gpt-oss modelleri [9]. Açık-ağırlık öncünün kayda değer bir kısmı artık Çinli laboratuvarlardan doğmaktadır [4], [10]; ağırlıklar tamamen kurumun çevresi içinde çalışsa da, bu durum model kökenini başlı başına bir tedarik-politikası değişkeni hâline getirir.

B. Bir Politika Eğilimi Olarak Egemen Yapay Zeka

Açık modellerin kurum içi kurulumu daha geniş bir politika hareketiyle uyumludur. Avrupa girişimleri; kamu yönetimi için Fransız-Alman egemen yapay zeka programlarını ve ulusal açık-model çabalarını içerir [11]; Körfez ülkeleri büyük ölçekli ulusal yapay zeka altyapısı yatırımları taahhüt etmiştir [12]; ve büyük Avrupa bankaları açık modellerin kurum içi kurulumlarını duyurmuştur [13]. Türkiye'de Haziran 2026'da açıklanan 2026–2030 Ulusal Yapay Zeka Eylem Planı yerli model geliştirmeyi, veri merkezi kapasitesini ve bir düzenleyici çerçeveyi önceliklendirir [14]; TBMM Yapay Zeka Araştırma Komisyonu ise Mart 2026 raporunda ulusal bir yapay zeka yasasının hazırlanmasını önermiştir [15]. Ocak 2025'ten beri yürürlükte olan AB Dijital Operasyonel Dayanıklılık Yasası, finansal kurumları bulut bağımlılıklarını belgelemeye ve azaltmaya daha da zorlamaktadır [13].

III. Mevzuat Gereksinimleri Analizi

Beş ulusal ve bir ülke dışı rejimi inceleyerek, her birinden regüle veri işleyen bir yapay zeka sistemini bağlayan kısıtları çıkarıyoruz. Tablo I, türetilen R1–R8 gereksinimlerini özetler.

TABLO I — MEVZUAT SÜRÜCÜLERİ VE TÜRETİLEN MİMARİ GEREKSİNİMLER

Gereksinim	Açıklama	Başlıca sürücüler
R1	Sıfır varsayılan veri çıkışı: istemler, bağlam, gömme vektörleri ve loglar kurum çevresini terk etmemeli	KVKK; BDDK; TCMB; MASAK
R2	Çevrimdışı, bütünlüğü doğrulanmış güncellemelerle tümüyle hava boşluklu işletim desteği	KVKK m.6 (özel nitelikli); MASAK; kritik altyapı
R3	Tüm birincil ve ikincil işleme sistemlerinin yurt içi (ülke içi, kurum içi) konumu	BDDK bilgi sistemleri yönetmeliği [28]
R4	Rol bazlı erişim, politika bazlı redaksiyon, yapılandırılabilir saklama ve imha	KVKK (minimizasyon, amaçla sınırlılık, saklama); SPK (bilgi bariyerleri)
R5	Tam izlenebilirlik: her yanıt kaynağa, zamana, politika bağlamına ve üreten model sürümüne atfedilebilir	BDDK; SPK; iç denetim; AB YZ Yasası şeffaflık
R6	Yüksek riskli çıktılarda insan karar yetkisi; yetersiz kanıtta çekimserlik	MASAK (soruşturmacı sorumluluğu); KVKK üretken YZ rehberi [29]
R7	İndekslerin, logların, kanıtların ve ince ayar yapılmış eserlerin kurumsal sahipliği; sözleşmesel çıkış ve veri iadesi	BDDK (dış hizmet süreklilik); operasyonel dayanıklılık
R8	Kurulan tüm modellerin lisans ve köken envanteri, denetim kanıtına dâhil	tedarik riski; fikri mülkiyet riski (Bölüm V-A)

A. KVKK

6698 sayılı Kanun'un genel ilkelerinin ötesinde, Kişisel Verileri Koruma Kurumu Kasım 2025'te üretken yapay zeka ve kişisel veriye ilişkin özel bir rehber [29] ve üçüncü taraf üretken yapay zeka araçlarının iş yerinde kullanımına dair bir not [30] yayımlamış; kişisel verinin harici araçlara aktarılması riskini açıkça işaret etmiştir. Bu belgeler R1, R4 ve R6'yı gerekçelendirir; ve gömme vektörleri kişisel kayıtlar üzerinde hesaplandığında türetilmiş kişisel veri olduğundan, gömme hesabının kendisinin de çevre içinde kalmasını gerektirir (R1).

B. BDDK

Bankaların Bilgi Sistemleri ve Elektronik Bankacılık Hizmetleri Hakkında Yönetmelik (Resmî Gazete 31069, 15 Mart 2020), birincil ve ikincil sistemlerin Türkiye'de bulunmasını şart koşar ve dış hizmetlerden bankayı tam sorumlu tutar [28]. Yazım anında yapay zekaya özgü bir BDDK düzenlemesi bulunmuyordu; banka verisine dokunan yapay zeka iş yükleri bu nedenle bu genel rejim üzerinden yönetilir ve R3, R5 ile R7 bundan doğar.

C. TCMB, SPK, MASAK

Ödeme ve e-para mevzuatı işlem verisini son derece hassas kılar (R1); içsel bilgiye ilişkin sermaye piyasası yükümlülükleri rol izolasyonunu ve bilgi bariyerlerini gerektirir (R4); ve şüpheli işlem bildirimlerini çevreleyen gizlilik rejimi hem veri çıkışını yasaklar (R1, R2) hem de soruşturma sonuçlarının insan kararı olarak kalmasını gerektirir (R6). Yazım anında SPK veya MASAK'tan yapay zekaya özgü bir rehber bulamadık ve böyle bir rehberin var olduğunu iddia etmiyoruz.

D. AB Yapay Zeka Yasası

AB Yapay Zeka Yasası 2 Ağustos 2026'da genel olarak uygulanmaya başlamış, Ek III yüksek riskli sistemlere ilişkin yükümlülükler 2026 Dijital Omnibus kapsamında Aralık 2027'ye ertelenmiştir [31], [32]. Ülke dışı kapsamı, çıktıları AB'de kullanılan Türk sağlayıcıları ve kullanıcıları da kapsar [33]. İzlenebilirlik ve teknik dokümantasyon yükümlülükleri, AB'ye maruziyeti olan kurumlar için R5 ve R8'i pekiştirir.

IV. Referans Mimari

A. Tasarım İlkeleri

Mimari bir aygıttır: kurumun veri merkezine kurulan, önceden entegre edilmiş bir donanım-yazılım birimi. Tablo I'den beş ilke çıkar: (1) sıfır varsayılan çıkış — aygıt hiçbir dış bağlantı başlatmaz; (2) çevrimdışı tam işlevsellikle hava boşluğu yeteneği (R2); (3) tüm indeks, log ve kanıtların müşteri sahipliği (R7); (4) varsayılan uzaktan erişim yok — destek oturumları müşteri tarafından başlatılır, loglanır ve süreyle sınırlanır; (5) müşteri verisiyle eğitim yok.

İşlevsel olarak aygıt altı modüle ayrışır: bir bağlayıcı/indeksleme katmanı; bir yönetişim stüdyosu (roller, redaksiyon, saklama); grounded yanıtlarla sınırlı bir çalışan asistanı; kontrolleri mevzuata eşleyen ve kanıt paketleri üreten bir uyum merkezi; yüksek riskli yanıtlar için bir mühürleme hizmeti; ve tamamen insan kontrolünde çalışan isteğe bağlı bir AML soruşturma yardımcısı.

B. Model Portföyü ve Donanım Kademeleri

Tek bir model tüm iş yüklerine ekonomik biçimde hizmet edemez. Portföy yaklaşımı şunları atar: 20–35B parametre sınıfı modern modeller (yoğun veya düşük-aktif-parametreli MoE) sohbet ve RAG iş yüklerine; 70–120B sınıfı modeller karmaşık analize; öncüye yakın MoE modeller en zorlu akıl yürütme görevlerine; ve küçük bir 2–9B model yardımcı işlevlere (sınıflandırma, redaksiyon, yönlendirme). Bellek boyutlandırması, FP16'da ~2 bayt/parametre, FP8'de ~1, INT4'te ~0,5 artı KV önbelleği ve eşzamanlılık için %20–30 standart sezgisel kuralını izler [34]. Tablo II ortaya çıkan kademelenmeyi örnekler.

TABLO II — AYGIT KADEMELERİ (ÖRNEKLEYİCİ)

Kademe	GPU yapılandırması	Model sınıfı
T1	1× 96 GB sınıfı kurumsal GPU	20–35B (FP8/INT4)
T2	2–4× 96–141 GB GPU	70–120B (FP8/INT4)
T3	8× HBM sınıfı GPU	öncüye yakın MoE
T4	çok düğümlü küme	portföy + yüksek erişilebilirlik

Nicemleme (quantization) politikası kaliteyi kapasiteye karşı dengeler: FP8 mevcut GPU kuşaklarında yerel olarak çalışır ve çoğu iş yükünde ölçülen kalite kaybı ihmal edilebilir düzeydedir; AWQ tarzı INT4 ise belleği dörtte bire indirirken ölçülü bir kalite bedeli getirir [34], [35]. Kalite etkisi modele ve iş yüküne bağlı olduğundan, mimari; nicemlenmiş herhangi bir modelin üretime girmeden önce kurumun kendi değerlendirme kümesine karşı doğrulanmasını zorunlu kılar (Bölüm IV-F).

C. Çıkarım Katmanı

Hizmet katmanı, üretimde doğrulanmış açık kaynak motorları üzerine kurulur. Sayfalı dikkat (paged attention), sürekli gruplama ve geniş model kapsamıyla vLLM fiili standarttır [36]; paylaşılan önek önbellekleme sunan motorlar (ör. SGLang), politika ve bağlam şablonlarının binlerce sorguda

tekrarlandığı RAG iş yüklerinde ek verim sağlar [37]. Masaüstü sınıfı araçlar, eşzamanlı çok kullanıcılı hizmet için tasarlanmadıklarından üretim yığınından dışlanır [36].

D. Zorunlu Atıf ve Çekimserlik İçeren Grounded RAG

R5 ve R6'yı işletime döken hat beş aşamadan oluşur:

- **Gömme:** tamamen çevre içinde hesaplanan, kurum içinde barındırılan çok dilli açık gömme modelleri (BGE-M3 sınıfı çok vektörlü modeller veya güncel açık MTEB liderleri) [38];
- **Hibrit erişim:** yoğun vektör aramanın, karşılıklı sıra füzyonu (reciprocal rank fusion) ile sözlüksel BM25 ile birleştirilmesi — mevzuat ve iç politika gibi terminoloji-duyarlı derlemlerde yalnızca-yoğun erişimden daha sağlam [17];
- **Yeniden sıralama:** kesinliği ölçülebilir biçimde artıran ve alt-akış halüsinasyonunu azaltan çapraz-kodlayıcı yeniden sıralayıcılar [18];
- **Zorunlu atıf:** üretici yalnızca bağlamındaki erişim-parçası tanımlayıcılarına atıf yapabilir; atıflar erişilen metne karşı sonradan doğrulanır ve doğrulanamayan atıflar yanıtın yayınlanmasını engeller;
- **Çekimserlik:** erişim güveni eşiğin altına düşerse hat, üretimi çağırmadan "yetersiz kaynak" döndürür; üretilen yanıtlar ayrıca gösterilmeden önce bir grounded'lık kontrolünden geçer [19]. Çekimserlik birinci sınıf bir sonuç olarak loglanır: regüle bir ortamda, doğru bir reddetme bir kusur değil, bir güvence özelliğidir.

Vektör depoları ve tüm indeksler, kurum içinde barındırılan açık kaynak veritabanlarında çevre içinde bulunur ve kuruma aittir (R7).

E. Hava Boşluklu Güncelleme Zinciri

Tüm bileşenler — ağırlıklar, konteyner imajları, paket aynaları, gözlemlenebilirlik, PKI — muhafaza (enclave) içinde önceden hazırlanır [39]. Güncellemeler kriptografik olarak imzalı çevrimdışı paketler (ağırlıklar, yamalar, yapılandırma) olarak gelir; içe aktarma imza doğrulaması, zincir-of-custody loglaması ve açık müşteri onayı gerektirir. Lisanslama tümüyle çevrimdışıdır: herhangi bir "telefon-eve" lisans kontrolü, telemetri veya otomatik güncelleme mekanizması hava boşluklu işletim için eleyicidir [39], [40]. Model güncellemeleri ayrıca etkinleştirilmeden önce kurumun değerlendirme kümesine karşı regresyon testi gerektirir.

F. Denetim Kanıt Katmanları

Denetlenebilirlik üç ayrıntı düzeyinde çalışır. (1) Kaynaklı yanıtlar: her yanıt, doğrulanmış atıflarla çevre içi kaynaklarına kadar izlenebilir. (2) Mühürlü yanıtlar: yüksek riskli yanıtlar kaynak kümesi, zaman damgası, politika bağlamı ve tam üreten model sürümüyle mühürlenir; sonradan doğrulanabilir bir kayıt üretir. (3) Kanıt paketleri: uyum modülü; yapılandırmayı, mevzuat-haritalı kontrolleri, erişim rollerini, güncelleme-bütünlüğü kayıtlarını, kurulan model ve lisans envanterini (R8) ve bir "veri çıkışı yok" doğrulama logunu içeren zaman damgalı paketler üretir. Model ağırlıkları ve sürümleri kurumda bulunduğundan, "hangi model, hangi yapılandırmayla bu yanıtı üretti?" sorusu üçüncü bir tarafın tasdikinden değil, kurumun kendi kayıtlarından yanıtlanabilir — API tabanlı kuruluma karşı yapısal bir üstünlük.

V. Model Seçimi Hususları

A. Kurumsal Risk Olarak Lisans Heterojenliği

"Açık kaynak" tek bir lisanslama rejimi değildir ve farklar regüle kurumlar için somut hukuki sonuç taşır (Tablo III). Apache 2.0 ve MIT hiçbir kullanım-alanı kısıtı getirmez ve (Apache 2.0 durumunda) açık bir

patent lisansı içerir [41]. Llama Topluluk Lisansı, kullanımı aylık 700M aktif kullanıcı eşiğine bağlar, bir kabul edilebilir kullanım politikasını son kullanıcılara akıtır ve atıf ile türev-adlandırma yükümlülükleri getirir; Açık Kaynak Tanımı'nı karşılamaz [8], [42]. Önceki kuşak Gemma şartları, akıtılan bir yasak-kullanım politikası içerir ve sağlayıcının kullanımı uzaktan kısıtlama hakkını saklı tutar — hava boşluklu bir egemenlik duruşuyla bağdaştırılması güçtür [7]; en yeni Gemma kuşağının Apache 2.0'a geçişi bunu yalnızca o kuşak için çözer [6]. Bu lisansların hiçbiri, ticari API sözleşmelerinin aksine fikri mülkiyet tazminatı sağlamaz; kurumlar çıktıyla ilgili fikri mülkiyet riskini kendileri sigortalar. Apache 2.0/MIT modellerinin ince ayar türevleri koşulsuz olarak kuruma aitken, Llama türevleri adlandırma ve politika yükümlülükleri taşır [8].

TABLO III — BÜYÜK AÇIK-AĞIRLIK AİLELERİNİN LİSANS REJİMLERİ

Lisans	Örnek modeller	Başlıca kısıtlar
Apache 2.0	Qwen3 (çoğu), yeni Gemma, bazı Mistral, gpt-oss	maddi kısıt yok; açık patent lisansı
MIT	DeepSeek	maddi kısıt yok; patent konusunda sessiz
Llama Topluluk	Llama ailesi	MAU eşiği; AUP akıtma; adlandırma/atıf
Gemma Şartları (önceki)	önceki Gemma	yasak-kullanım akıtma; uzaktan kısıtlama hakkı

Bu nedenle mimari, yalnızca Apache 2.0/MIT modelleri şeklinde bir varsayılan politika benimser; kısıtlı-lisanslı modeller yalnızca kurumun hukuki onayıyla kurulabilir ve lisans envanterine kaydedilir (R8).

B. Türkçe Dil Performansı ve Tokenizasyon Ekonomisi

Türkçe değerlendirme yazınından iki bulgu model seçimini şekillendirir. Birincisi, kapsamlı Türkçe kıyaslamalarda büyük çok dilli açık modeller şu an çoğu küçük Türkçe-yerel modelden daha iyi performans gösterir [24]; bu, sıfırdan Türkçe modellere varsaymak yerine güçlü çok dilli temel model + Türkçe alan uyarlaması + kuruma özgü Türkçe değerlendirme kümesi stratejisini destekler. İkincisi, tokenizasyon doğurganlığı gerçek bir maliyet kaldıracıdır: Türkçe kelime başına ~1,8–2,5 token'a karşı İngilizce için ~1,2–1,4 [26]; aynı belge Türkçe'de belirgin biçimde daha fazla bağlam penceresi ve çıkarım hesabı tüketir. Bu nedenle doğurganlık, doğrulukla birlikte model değerlendirmesinde açık bir ölçüt olarak ölçülmelidir. Rekabetçi Türkçe-yerel modellerin ortaya çıkışı (ör. Kumru'nun sıfırdan 7,4B ve 2B sürümleri [27]) ödünleşimin değişebileceğini düşündürür; satıcı bildirimli skorlar tedarik ağırlığı taşımadan önce bağımsız olarak yeniden üretilmelidir.

C. Köken (Provenance)

Açık-ağırlık öncü gelişiminin Çinli laboratuvarlarda yoğunlaşması [4], [10] göz önüne alındığında, kurumlar yetenek ve lisanstan bağımsız köken kısıtları koyabilir. Açık ağırlıklar çevre içinde, satıcı bağlantısı olmadan, deterministik biçimde çalıştığından, artık köken riskleri bulut bağımlılığından türce farklıdır (veri çıkışından ziyade eğitim-verisi belirsizliği ve olası davranışsal önyargılarla ilgilidir); yine de mimari, kökeni bir konum dayatmak yerine yapılandırılabilir bir tedarik filtresi olarak ele alır.

VI. Tartışma

A. Maliyet Zarfı: Dürüst Bir Çerçeve

Kurum içi barındırma ile API tüketimi arasındaki yayımlanmış başabaş analizleri; hacim, kullanım oranı ve mühendislik yükü varsayımlarına bağlı olarak büyüklük mertebelerinde farklılaşır [43], [44]. Düşük,

ani hacimlerde token başına ekonomi API'leri; yüksek ve sürekli hacimde kurum içi barındırma kesinlikle üstündür. Ancak token başına ekonominin regüle kurumlar için yanlış karar çerçevesi olduğunu savunuyoruz: kamusal buluttan hukuken geçemeyen veri sınıfları için API alternatifi hiçbir fiyata mevcut değildir. Aygıtın ekonomik önermesi, aksi hâlde dışlanmış yüksek değerli iş yüklerinin öngörülebilir sabit maliyetle mümkün kılınmasıdır; token düzeyinde maliyet paritesi ölçekte ikincil bir etki olarak elde edilebilir.

B. Açık vs. Kapalı: Kalan Farklar

~4 aylık toplu fark [1] tekdüze değildir. Açık modeller uzun-ufuklu ajanik görevlerde, araç-kullanım kıyaslamalarında ve bazı yönerge-izleme inceliklerinde ölçülebilir biçimde geridedir [10]; kıyaslama kirlenmesi ve seçici yayımlama, toplu endekslerin gerçek farkı olduğundan az göstermesine yol açabilir [2]. Mimarinin iş yükü hedeflemesi — erişime dayalı soru-yanıt, özetleme, çıkarım, taslak yazımı — farkın en küçük olduğu rejime bilinçli olarak yerleşir. Hassas olmayan veride öncü ajanik yetenek gerektiren kurumlar makul biçimde hibrit bir duruş işletebilir; mimari bunu engellemez, yalnızca içeride kalması gerekeni izole eder.

C. Sınırlamalar

Bu çalışmayı dört sınırlama çerçeveler. (1) Kurulan-sistem değerlendirmesi değil, tasarım gerekçeli bir referans mimari sunar; hiçbir verim, doğruluk veya maliyet ölçümü raporlanmaz ve bu tür ölçümler tasarım gereği ortama özgüdür. (2) Mevzuat analizi, Türk ve AB araçlarının yazım anındaki durumunu yansıtır; atıf yapılan birçok araç (Türk YZ yasa taslakları, AB YZ Yasası ertelemeleri) değişkendir. (3) Model-ekosistemi olguları (sürümler, lisanslar, kıyaslama sıralamaları) aylık ritimde değişir; belirli iddialar birincil model kartlarına karşı yeniden doğrulanmalıdır. (4) Türkçe dil analizi, kuruma özgü belge türlerini temsil etmeyebilecek kamusal kıyaslamalara dayanır; mimarinin zorunlu kıldığı kuruma özgü değerlendirme kümesi bu boşluk için bir çözüm değil, bir azaltma önlemidir.

VII. Sonuç

Açık-ağırlık model ekosisteminin olgunlaşması, Türkiye'nin regüle kurumları için yapay zeka yeteneği ile mevzuat uyumu arasında varsayılan ödünleşimi ortadan kaldırmıştır. Beş ulusal mevzuat rejiminin bağlayıcı kısıtlarının; serbest lisanslı açık-ağırlık modeller, zorunlu atıf ve çekimserlik içeren grounded RAG ve imzalı çevrimdışı güncelleme zinciri üzerine kurulu egemen, kurum içi bir aygıt mimarisiyle karşılanabileceğini — ve üreten modelin kurumun uhdesinde olması sayesinde API tabanlı kurulumda ulaşılamayan kalitede denetim kanıtı üretilebileceğini — gösterdik. Gelecek çalışmalar; çekimserlik mekanizmasının kesinlik-duyarlılık ödünleşiminin Türkçe mevzuat derlemlerinde ampirik değerlendirmesini, açık-kapalı yetenek farkının kuruma ilişkin görev kümelerinde boylamsal izlenmesini ve Türk yapay zeka mevzuatı somutlaştıkça mevzuat haritalamasının genişletilmesini içerir.

> **Yasal uyarı.** Bu belge tanıtım ve tasarım-gerekçesi amaçlıdır; hukuki görüş değildir. Tüm mevzuat eşlemeleri "desteklemek üzere tasarlandı" ilkesiyle sunulmuştur ve hiçbir resmi sertifikasyon iddiası içermez; uygunluk değerlendirmesi kuran kurumun sorumluluğundadır. Kapasite ve performans figürleri yalnızca kurumun kendi ortamında yapılan kıyaslamalarla paylaşılır. Kaynakça, izlenebilirlik için orijinal dilinde bırakılmıştır.

Kaynakça

[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," Haz. 2026. https://turkiye.ai/

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

[16] V. Magesh vd., "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 (yargı incelemesi sürüyor; güncel durumu doğrulayın).

[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 Kas. 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.