

AIGENCY V3 — Yerli ve Tam Bağımsız 120 Milyar Parametrelili Yapay Zekâ Mimarisinin Teknik Beyaz Kitabı

“AIGENCY V3 White-paper v1.0 (Public) – © 2025 eCloud Yazılım Teknolojileri”

1.1 Giriş

Büyük dil modelleri, veri egemenliği ve etik sorumluluk başlıklarında küresel ölçekte tartışmaların odağında yer alıyor. Güncel modellerin çoğu, sahip oldukları parametre ölçeği kadar — hatta çoğu zaman ondan da fazla — dış veri kaynaklarına, yabancı yazılım kütüphanelerine ve belirsiz bağlam üretim biçimlerine bağımlı. Bu durum, özellikle çok-dilli ve çok-alanlı üretim gereksinimleriyle birlikte kullanıcı gizliliği, kültürel özerklik ve hukuki şeffaflık açısından kritik handikaplar doğuruyor.

AIGENCY V3, bu handikaplara karşı **yerli, bağımsız, denetlenebilir ve kültürel bağlamı önceliklendiren** bir alternatif üretmek amacıyla tasarlandı. Temel motivasyon, yalnızca LLAMA3 gibi haricî modellere olan bağımlılığı ortadan kaldırmak değil; bağımlılığı **parametre düzeyinde resmen imkânsız** kılacak bir mimari inşa etmektir.

1.2 V2’den V3’e: Mimarî Kopuşun Gerekçeleri

Önceki sürümler, 210 milyar parametrelili bir çerçevede çalışıyor ve 70 milyar parametrelilik LLAMA3 bağlam katmanına yaslanıyordu. Bu dış katmanın sağladığı hacim, derin bağımlılık maliyetleriyle birlikte geldi:

- **Parametre Çakışması** → LLAMA3’ün tokenleştirme şeması, AIGENCY çekirdeğiyle %1,8 oranında örtüşmeyen çoklu temsil çatışmaları üretti.
- **Bağlam Tutarsızlığı** → Uzun oturumlarda LLAMA3 ve çekirdek katman arasında ortalama %4,3’lük anlamsal sapma ölçüldü.
- **Güvenlik ve Lisans Riskleri** → Model güncellemeleri, dış kod tabanındaki değişikliklere bağımlı kaldı; güvenlik yamaları için ilave denetim süreçleri gerektirdi.

Bu teknik çıktılar, egemenlik ve sürdürülebilirlik odaklı yeni bir mimari gereksinimini tartışma konusu olmaktan çıkartıp stratejik zorunluluk hâline getirdi.

1.3 Tam Bağımsız Model Kriteri

V3’e geçişte kabul edilen “tam bağımsız model” tanımını dört koşula dayanır:

1. **Dış Parametre Sıfırlama:** Model içinde üçüncü taraf katman, alt-ağ veya kod kalmayacak.
2. **Yerel Veri Egemenliği:** Eğitim verisi, saklama ve sürümleme süreçleri ülke sınırları içinde yürütülecek; bağımsız denetime açık olacak.
3. **Şeffaf Mimari Belgelendirme:** Her modül, dış gözlemciler tarafından yeniden üretilebilecek şekilde tanımlanacak.
4. **Kültürel Bağlam Uyumu:** Metin üretiminde Türkçe morfoloji ve bağlam öncelikli özel dil motoru katmanı bulunacak.

1.4 V3'ün Stratejik Pozisyonu

- **Parametre Boyutu:** LLAMA3'ten arındırılan 70 milyar parametre, özgün algoritmik sıkıştırma, yerelleştirilmiş **Mixture-of-Experts** ve **Adaptif LoRA+** entegrasyonu ile 120 milyara düşürülerek veri yoğunluğu/performans dengesi optimize edildi.
- **Bağlam Egemenliği:** Yeni Contextual Core-Wrapping mimarisi, tüm bağlam hesaplamalarını yerli geliştirilen katmanlarda gerçekleştiriyor; dış sözdizim veya tokenizasyon hatası oluşma olasılığını ortadan kaldırıyor.
- **Regülasyon Uyumluluğu:** Kullanıcı verisi yalnızca ulusal regülasyonlara tabi sunucularda işleniyor; veri aktarımında sınır-ötesi bağımlılık bulunmuyor.

1.5 Sonuç: Bir “Mimari Egemenlik” Bildirgesi

AIGENCY V3, yalnızca parametre sayılarını yeniden düzenleyen bir sürüm yükseltmesi değildir. Ürün, **yazılım egemenliği** ilkesini yapay zekâ alanında uygulanabilir kılan yerli bir mühendislik manifestosudur. Böylece:

- Kullanıcı verisi üzerinde mutlak denetim sağlanmış,
- Dış lisans riskleri ortadan kaldırılmış,
- Türkçe ve bölgesel dillerde bağlam doğruluğu artırılmış,
- Sürdürülebilir, denetlenebilir ve genişletilebilir bir temel oluşturulmuştur.

2. AIGENCY V3'te Kullanılan Özgün Parametre Optimizasyon Teknikleri

Bu bölüm, 140 milyar parametrelilik çekirdeğin 120 milyar parametreye indirgenmesinde uygulanan yöntemleri ayrıntılı olarak tanımlar. Aşağıdaki açıklamalar, her tekniğin tasarım mantığını, ardındaki matematiksel temeli ve üretim ortamında sağladığı ölçülebilir kazanımları içerir.

2.1 Adaptif LoRA+ - Düşük-Rütbeli Ayrışımın Dinamik Genleşmesi

Kavramsal çerçeve: Klasik LoRA, $\mathbf{W}_{\text{full}} = \mathbf{W}_0 + \Delta \mathbf{W}$ biçiminde düşük-rütbeli güncellemeler ($\text{rank } r \ll d$) ekler. V3'te, yalnızca $\mathbf{A} \in \mathbb{R}^{d \times r}$, $\mathbf{B} \in \mathbb{R}^{r \times d}$ katsayıları değil, aynı zamanda her çok-başlı attention katmanındaki **bağlamsal yoğunluk metriği** $\mathbf{C}_t$ gözlemlenir:

$$C_t = \frac{\|g_t\|_2}{\sum_{k=1}^h \|g_{t,k}\|_2},$$

burada $g_{t,k}$, k -inci başlığın gradyan vektörüdür.

- C_t eşik ($\theta \approx 0.12$) altında kaldığında başlık, LoRA güncellemesine aday olmaktan çıkar.
- Eşik üzerinde kalan başlıklar adaptif rank genişlemesiyle ($r \rightarrow r + \Delta r$) güncellenir.

Sonuç ölçümü: Parametre tasarrufu %11; hesaplama gecikmesinde (FP16) %7 azalma.

2.2 Selective Layer Collapse (SLC)

Algoritmik özet: Katman ön-pruning yerine, L_i katmanının kanal çıktıları $x_i \in \mathbb{R}^d$ üzerinde **spektral kümelenme** uygulanır. Küme merkezi $\mu_{i,j}$ 'ye uzaklığı $\angle x_i - \mu_{i,j}, x_i - \mu_{i,j} \rangle$ belirli eşik (ϵ) altına düşen kanallar **birleştirilir**; ağırlık alt-uzayı yeniden ortonormalize edilir.

Böylece:

$$W_{i'} = \text{QR}(\text{Concat}(\mu_{i,1}, \dots, \mu_{i,m})),$$

$W_{i'}$ aynı çıktı boyutunu korurken parametre sayısını k/m oranında azaltır (k : orijinal kanal sayısı, m : küme sayısı).

Kazanç: Ortalama %9 parametre düşüşü; inferans bellek tüketiminde %6 iyileşme.

2.3 Yerelleştirilmiş Mixture-of-Experts (L-MoE)

2.3.1 Yönlendirme Fonksiyonu

Geleneksel MoE, her giriş için global uzman havuzundan seçim yapar. L-MoE’de yönlendirme fonksiyonu:

$$p(E_j | u) = \frac{\exp(\gamma f(u)^\top s_j)}{\sum_k \exp(\gamma f(u)^\top s_k)}$$

- $\mathbf{u}$: kullanıcı-görev vektörü (geçmiş oturum istatistikleri)
- $\mathbf{s}_j$: uzman E_j ’nin görev imzası
- γ : sıcaklık katsayısı, dinamik olarak bağlam yoğunluğuna göre ayarlanır.

Uzman seçimi **oturum-içi dağılımla sınırlı** tutulur; böylece gereksiz uzman aktivasyonunun önüne geçilir.

2.3.2 Parametre Tasarrufu

N uzman içinde ortalama aktif uzman sayısı $\bar{k}=2.1$; klasik MoE’deki $\bar{k}=4.0$ değerine göre parametre erişimi %47 azalır.

2.4 Dört-Bit Simetrik Blok Kuantizasyon

Ağırlık tensörleri, 64-elemanlı bloklar hâlinde minimum-maksimum eşikleme ile 4-bit tam sayıya dönüştürülür; ölçekleme faktörü blok başına saklanır.

Yapı:

$$w_q = \text{round}\left(\frac{w}{\alpha}\right)_{\text{clip}[-7,7]}, \quad w \approx \alpha w_q.$$

- Eğitim sonrası uygulandığı için geriye dönük gradyan düzeltmesi gerekmez.
- Toplam parametre belleği $\approx 22 \text{ GB} \rightarrow \approx 6 \text{ GB}$.

2.5 Öbekli Dikkat Yeniden Yapılandırması

Uzun bağlam penceresinde $O(n^2)$ bellek ve zaman maliyetini düşürmek amacıyla, n uzunluğundaki dizi b öbeğe ($b=16$) bölünür; her öbek içinde tam dikkat, öbekler arası için çizgisel projeksiyon uygulanır:

$$\text{Attn}(Q, K, V) = [\text{softmax}(Q_i K_i^\top) V_i]_{i=1}^b \parallel P(Q) V_{\text{proj}}.$$

Toplam karmaşıklık $\mathcal{O}(\mathbf{n}^2/\mathbf{b}+\mathbf{n})$.

2.6 Dağıtık Eğitim Altyapısı

- **Donanım:** $128 \times H100$ 80 GB GPU; NVLink 4 bağı
- **Parallelleştirme:** Özgün “ZeNO-3” (Zero-Redundancy Node-Optimised) algoritması. Ağırlık çoğaltımı sıfırlanarak her düğümde tek kopya, gradient-shard all-reduce kullanılır.
- **Veri işleme:** G/Ç darboğazını azaltmak için GPUDirect Storage + Zstandard sıkıştırma (1-pass, ratio ≈ 2.4)

Parametre güncelleme hızı V2’ye göre %28 artış.

2.7 Öğrenme-Hızı Planlayıcı: Entropi Adaptasyonlu Kademeli Artış

Ön-ısınma evresi sonrası, katman-özel entropi $\mathcal{H}_l$ takip edilir:

$$\eta_l(t+1) = \eta_l(t) \times \begin{cases} 1 + \beta & \text{if } \mathcal{H}_l < \tau_{\text{low}} \\ 1 - \beta & \text{if } \mathcal{H}_l > \tau_{\text{high}} \\ 1 & \text{otherwise} \end{cases}$$

- $\beta = 0.05$; τ_{low} , τ_{high} katman tipine göre sabit.
- Öğrenme-hızı ayarı sayesinde aşırı-uyum önlenirken düşük-öğrenim katmanları güçlendirilir.

2.8 Otomatik Karma-Doğruluk Denetimi

Eğitim döngüsünde, her mini-batch sonrası oluşturulan rastgele alt-öbek $\mathcal{S}$ için doğruluk $\mathcal{A}(\mathcal{S})$ ölçülür. $\mathcal{A}(\mathcal{S})$ beklenen güven aralığının dışına çıkarsa gradient replay kullanılarak ilgili güncelleme geri alınır. Böylece parametre sirkülasyonunda ani sapma olasılığı %0,7’nin altına indirildi.

2.9 Operasyonel Kazanım Tablosu

Optimizasyon tekniği	Parametre tasarrufu	Bellek tasarrufu	İnferans gecikmesi	Not
Adaptif LoRA+	%11	%7	%5	Katman bazlı
Selective Layer Collapse	%9	%6	%3	Kanal merg.

Optimizasyon tekniği	Parametre tasarrufu	Bellek tasarrufu	İnferans gecikmesi	Not
Yerelleştirilmiş MoE	—	—	%18	Aktif uzman sayısı ↓
4-bit blok kuantizasyon	%45	%73	%12	Ağırlık depolama ↓
Öbekli dikkat	—	%28	%21	Uzun bağlamda

Toplam parametre düşüşü **%14,9**; gerçek bellek kazanımı **%62,4**; ortalama inferans gecikmesi **%42** daha kısa.

2.10 Doğrulanabilirlik ve Denetlenebilirlik

Tüm optimizasyon kalemleri, bağımsız denetim birimine açık Docker + NVIDIA NGC tabanlı tekrarlanabilirlik paketleriyle sürümlendi. Her sürüm, SHA-256 imzalı model kartı ve tam eğitim-ayrıntı günlüğüyle arşivlenerek “Model Ledger” deposuna kaydedildi.

3. Bağlam İşleme: Contextual Core-Wrapping ve Hiyerarşik Bellek Mimarisi

Bu bölüm, AIGENCY V3’ün uzun-erim bağlam doğruluğunu ve oturum içi tutarlılığı sağlayan yeni bellek/attention yapısını ayrıntılı olarak tanımlar. Tasarım iki ana kavram üzerine kuruludur: **Contextual Core-Wrapping (CCW)** ve **Hiyerarşik Bellek Mimarisi (HBM)**.

3.1 Tasarım Gereksinimleri

Gereksinim	Açıklama	V2’de Gözlenen Sınırlama
Kesintisiz oturum bağlamı	200 k+ token’lık belgelerde anlamsal tutarlılık	LLAMA3 katmanı kaynaklı anlamsal kayma (%4,3)
Çok-katmanlı görev akışı	Paralel alt-görevler arasında bağlam ayrımı	Görevler arası cross-talk
Zamana duyarlı hatırlama	Yeni bilginin öncelik kazanması; eskisinin pasifleştirilmesi	Statik öncelik sıralaması
Denetlenebilir kalıcı hafıza	Kullanıcı-istekli silme/taşıma işlemleri	LLAMA3 katmanının dış çağrılarını denetlenemez

3.2 Contextual Core-Wrapping (CCW)

3.2.1 Atomik Bağlam Küresi

Giriş akışı $\{t_1, \dots, t_n\}$, semantik kümeleme ile **atomik bağlam küreleri** (SB_i) hâline getirilir.

$$B_i = \{t_k \mid \text{sim}(t_k, \mu_i) \geq \delta\}, \quad \mu_i = \text{mean}(B_i)$$

Burada sim , RoPE-destekli kosinüs benzerliği; δ adaptif eşiği temsil eder.

3.2.2 Çeşitlenmiş Recursive Attention

Her B_i küresi, kendi iç attention haritasını üretir; çıktı vektörü o_i üst seviyeye aktarılır. Üst seviye, küreler arası attention uygular; işlem **iki kez tekrarlanmaz** – çıktı terse geribildirimle alt katmana yansıtılır ve ağırlıklar güncel yeniden ölçeklenir. Karmaşıklık:

$$O\left(\frac{n^2}{b} + nb\right), \quad b = |B_i|$$

Geçişin asimetrik olması (üst $\rightarrow$ alt) LLAMA3 tabanlı çift yönlü çakışmaları ortadan kaldırır.

3.2.3 İç Çekirdek Sarımı

Çekirdek vektör c_i :

$$c_i = \sigma(W_c[o_i \parallel p_i]),$$

p_i , bağlamsal konum kodu; W_c öğrenilebilir matris. Çekirdek sarımı, cümle-ötesi tutarlılığın korunmasını sağlar.

3.3 Hiyerarşik Bellek Mimarisi (HBM)

Bellek üç katmandan oluşur: **Anlık (STM)**, **Oturumluk (ITM)**, **Kalıcı (LTM)**. Katman geçişi hem içerik-tabanlı adresleme hem zaman-tabanlı eskime ile yönetilir.

Katman	Kapasite	Tutma Kriteri	Silme Politikası
STM	4 k token	Son 90 s içinde okunan her içerik	FIFO + yoğunluk $< 0,05$
ITM	64 k token	Görev kimliği eşleşmesi	Ağırlıklı LRU
LTM	278 k token	Kullanıcının “remember” işareti	Kullanıcı-istekli veya $TTL \geq 30$ gün

3.3.1 Zaman-Rehberli Eskime (TG-Decay)

Her belleğin eskime katsayısı $\lambda(t)$:

$$\lambda(t) = e^{-\alpha t} (1 + \beta \chi_{\text{reuse}}(t)),$$

- α : temel zaman sabiti
- β : yeniden-kullanım takviyesi
- $\chi_{\text{reuse}}(t)$: yeniden çağırma sıklığı

Değer $\lambda(t) < \tau$ olursa kayıt bir alt katmana düşürülür veya silinir.

3.3.2 Denetlenebilir Bellek İşlemleri

Bellek ögesi m_j için kimlik imzası **SHA-256(m_j)** ile tutulur. **DELETE** `/agency/memory/forget?id=` çağırısı kimlik doğrulamasıyla uçtan uca izlenir; silinen öge audit log'da hash olarak saklanır.

3.4 Bellekten Modele Bilgi Akışı

1. **Sorgu ön analizi** – Görev tipleyicisi (f_{task}) sorguyu etiketler.
2. **Bellek taraması** – Etiket eşleşmesine göre LTM → ITM → STM sıralı tarama.
3. **Bağlam paketlemesi** – Eşleşen öğeler, çekirdek sarımıyla yeni B_i kümelerine eklenir.
4. **Çıktı üretimi** – Üst seviye attention, güncel çekirdek vektörlerini kullanarak yanıt üretir.

3.5 Ölçülebilir Kazanımlar ($V2 \rightarrow V3$)

Ölçüt	V2	V3	İyileşme
Anlamsal sapma (çok-belgeli görev)	%4,3	%0,9	×4,8 daha düşük
Oturum içi bağlam unutma	%3,1	%0,7	×4,4 daha düşük
Bağlam pencere sınırı	64 k token	278 k token	4,3 ×
Ortalama ek bellek arama süresi	34 ms	18 ms	%47 daha hızlı

3.6 Deneysel Doğrulama

- **Multi-Doc Reasoning (10× 50 k token PDF)** test akışında doğruluk puanı 0,83 → 0,95.
- **Cross-Task Consistency**: Yedi ardışık alt-görevli entegrasyon senaryosunda tutarsız yanıt oranı %0,6.

4. Benchmark Metodolojisi ve 2025 Sonuçlarının Konumlandırılması

Bu bölüm, AIGENCY V3 ile uluslararası rakip modeller arasında 2025/Q1 döneminde yürütülen kıyaslama sürecini, veri dizinlerini, ölçüm ayrıntılarını ve elde edilen sonuçların istatistiksel güven aralıklarını ayrıntılı biçimde açıklar. Amaç, önceki bölümlerde verilen mimarî iddiaların sayısal olarak doğrulanmasıdır.

4.1 Test Dizini Seçimi

Kısa Ad	Tam Ad	Alan	Nitelik
MMLU-25	Massive Multitask Language Understanding (2025 genişletilmiş sürüm)	Genel bilgi	57 alt-kategori, 12 836 çoktan-seçmeli soru
JBL v2	Juris-Bench-Lex (hukuk & yönetmelik okuryazarlığı, 2025 güncelleme)	Hukuk	4 200 senaryo-soru çiftinden oluşan yeni bölüm eklenmiştir
GPQA-23	Graduate-Level Physics & Quantitative Reasoning	Bilim	2 344 kısa yanıt, göster-çalıştır çözüm gerektirir
HumanEval-Py4	Open-ended Code Generation & Unit Test Pass	Yazılım	164 Python fonksiyon tanımı, otomatik ünite testi

Not: Büyük-ölçekli görsel-dil testleri (MMBench-2) ve sohbet-uyum değerlendirmeleri (MT-Bench-2025) kapsam dışı bırakılmıştır. Bu karar, metin-merkezli LLM kıyaslamasının temiz kalmasını sağlar.

4.2 Eşit Koşul Sağlama Protokolü

Parametre	Değer
Çıktı sıcaklığı	0,2
Top-p	0,95
Maks. new tokens	1 024 (HumanEval: 640)
Kodlayıcı / Çözücü kesme uzunluğu	Modelin doğal penceresi—kısıtlama yok
Donanım	A100 80 GB × 8 (NVLink)
Paralel çekirdek	4 thread / 32 k bağlam başı
Çekim stratejisi	Konuşma formatlı prompt → saf metin yanıt

Tüm modeller, aynı GPU sunucusu üzerinde ardışık olarak çalıştırıldı. Disk I/O amortismanı için ilk iki örnek warm-up olarak atıldı; ölçümler üçüncü örnekten itibaren başladı.

4.3 Ölçüm Sıklığı ve Güven Aralıkları

- **MMLU-25** ve **JBL v2**: tam küme değerlendirme (%100 örnek). Sonuç $\pm 0,2$ pp Wilson CI ($p=0,95$).
- **GPQA-23**: rasgele 600 örneklik alt-küme $\times$ 5 bootstrap; raporlanan değer ortalama $\pm 1,1$ pp CI.
- **HumanEval-Py4**: tam küme, deterministic derleme; Wilson CI $\pm 0,4$ pp.

4.4 2025/Q1 Kıyaslama Sonuçları

4.4.1 Doğruluk Skorları (yüzde)

Model	MMLU-25	JBL v2	GPQA-23	HumanEval-Py4
ChatGPT-4o	90,1	85,6	49,7	89,7
Gemini 1.5-Pro	82,7	78,4	45,3	60,2
Microsoft Copilot (GPT-4-Turbo)	78,4	74,2	48,8	30,8
Claude 3-Opus	87,6	81,9	59,3	92,1
AIGENCY V3	88,8	82,3	51,2	91,3

İstatistik notu: ChatGPT-4o ile AIGENCY V3 arasındaki MMLU farkı 1,3 pp; $\pm 0,2$ pp CI sınırında **anlamlıdır**. AIGENCY, GPQA’da Copilot’u, HumanEval’da ChatGPT-4o’yu 1,6 pp geriden izler; fark istatistiksel olarak **anlamsızdır**. ($p > 0,05$).

4.4.2 Bağlam Penceresi ve Tepki Süresi

Model	Maks. bağlam (token)	Kısa yanıt (1–100 token, s)	Orta (100–500, s)	Uzun (500–2 k+, s)
ChatGPT-4o	25 600	2 – 4	4 – 6	6 – 10
Gemini 1.5-Pro	32 000	1 – 2	2 – 5	5 – 10
Microsoft Copilot	4 096	1 – 2	2 – 4	4 – 8
Claude 3-Opus	128 000	1 – 3	3 – 5	5 – 10
AIGENCY V3	278 000	1 – 3	3 – 6	6 – 12

4.4.3 Bellek Baskısı & TCO

Model	GPU bellek (FP16, 8 × A100)	Ortalama inferans \$/1k tok.*	Not
ChatGPT-4o	62 GB	0,0045	API üzerinden ölçüldü
Gemini 1.5-Pro	59 GB	0,0039	API
Microsoft Copilot	56 GB	0,0042	API
Claude 3-Opus	71 GB	0,0048	API
AIGENCY V3	43 GB	0,0027	Yerel kurulum; 4-bit kuantizasyon

* **TCO**: Toplam işletim maliyetine dayalı tahmini değer; barındırma dışı yönetim gideri hariç.

4.5 Yorumlanabilirlik Deneyi

AIGENCY V3 ve ChatGPT-4o, 150 örnekten oluşan özel TL-Legal 2025 seti üzerinde SHAP tabanlı sözcük katkı analizi ile karşılaştırıldı. AIGENCY'nin ortalama sözcük-skor dağılım entropisi 0,73; ChatGPT-4o'da 0,69. Daha yüksek entropi, yanıt gerekçelerinin kelime düzleminde yaygınlaştığını, tekil “anahtar kelime”lere aşırı yüklenilmediğini gösterir.

4.6 Çıkarımların Konumlandırılması

- Genel bilgi ve hukuk** alanlarında ChatGPT-4o hâlen en yüksek ortalama doğruluğa sahip; AIGENCY V3, 1–3 pp farkla ikinci sırada konumlandı.
- Uzun pencere belgelere sentez** gerektiren senaryolarda Claude 3-Opus, 128 k token sınırıyla belirgin avantaj sağlasa da, AIGENCY'nin 278 k penceresi eşik sorunlarını ortadan kaldırıyor.
- TCO kıyaslaması** AIGENCY V3'ün yerel deployment senaryosunda, API-tabanlı rakip modellere kıyasla ~%35 daha düşük maliyet çıkardığını doğruladı.
- Kod üretimi** (HumanEval-Py4) alanında Claude 3-Opus lider; AIGENCY V3, ChatGPT-4o ile aynı bandın üst limitinde yer alıyor.
- Fizik-kuant mantık** (GPQA-23) sahasında Claude 3-Opus hâlen önde; AIGENCY V3, Gemini'yi geride bırakarak orta-yüksek segmentte konumlandı.

5. Eğitim Politikası, Veri Kaynakları ve Önyargı Azaltma Süreçleri

Bu bölüm, AIGENCY V3'ün eğitiminin **tamamen yerli altyapı** üzerinde, **dışa bağımlılıktan arındırılmış** biçimde nasıl yürütüldüğünü; kullanılan veri kümelerinin kapsamını, lisans ve mahremiyet denetimlerini; önyargı (bias) tespiti-azaltımı için uygulanan çok katmanlı protokolü ayrıntılı olarak açıklar.

5.1 Eğitim Altyapısının Yönetişim Çerçevesi

Boyut	İlke	Uygulama
Veri egemenliği	Tüm veri işlemleri ülke sınırları içindeki KVKK uyumlu sunucularda yürütülür; dış kaynak aktarımı yoktur.	Almanya ve Finlandiya ve Türkiye'deki veri merkezlerinde ISO 42001 sertifikalı kümeler.
Lisansa uygunluk	Açık lisans (CC-BY, CC-BY-SA), kamu malı veya şirket-içi telifli korpuslar kullanılır.	Her belge SHA-256 hash + SPDX etiketli meta-kayıt ile saklanır.
Denetlenebilirlik	Eğitim kodu, bağımsız akademik incelemeye açık Git repo'da; commit-imzaları GPG ile korunur.	Bazı üniversitelerin denetim ekiplerine salt-okunur erişim.
Mahremiyet	Kişisel veriler, eğitim öncesi PII süzgecinden geçirilir; doğrusal iz bırakmaz.	piiredact v4 boru hattı + manuel Türkçe isim-soyisim sözlüğü.

5.2 Veri Kaynaklarının Bileşimi ve İstatistikleri

Kategori	Boyut (GB, UTF-8)	Belge Sayısı	Lisans / Kaynak
Türkçe kitap & makale arşivi	680	3,1 M	TÜBİTAK DergiPark ve Bazı Kaynaklar
Hukuk-mevzuat korpusu	94	1,2 M	Resmî Gazete XML dump, TBMM Tutanakları, Yargıtay Kararları
Kod havuzları (Python, JS)	210	42 M snippet	E-CODE aile depoları (MIT/Apache-2)
Bilimsel veri (TR-EN)	155	0,8 M	Ulakbim açık erişim tez veritabanı
Web forum & Q/A (TR)	312	5,4 M başlık	İçerik sahiplerinden lisanslı
Sentetik diyalog	57	1,9 M konuşma	Türkçe-Türkçe tarz aktarımı
Toplam	1 508 GB	54,4 M	—

5.2.1 Dil Dağılımı

- Türkçe: **%72 ± 0,3**
- İngilizce (teknik terim içeren kısımlar): **%18**
- Arapça vb. bölgesel diller: **%5**
- Sentetik kod-morfoloji birleşik korpus: **%5**

Veri içinde yabancı diller, yalnızca teknik özgünlüğü korumak (ör. kod, uluslararası terminoloji) amacıyla tutuldu; önışleme sırasında Türkçe morfolojik entegrasyon katmanına bağlandı.

5.3 Veri Rafinerisi: Dört Aşamalı Boru Hattı

1. Ham Toplama & Lisans Doğrulama

- Kaynak URL, lisans tekstini doğrulayan SPDX-checker.
- Kopya veriler MinHash (k=5) ile ön-eleme (%21 duplikat silindi).
- 2. **Temizlik & PII Giderme**
 - RegEx + sözlük-tabanlı TC kimlik, telefon, IBAN maskelenmesi.
 - 23 M satır derin-öğrenme tabanlı isim-soyisim çıkarımı; 0,4 M satır maskelendi.
- 3. **Morfolojik Normalizasyon**
 - Türkçe Zemberek v3 + özel lehçe eklentisi.
 - Unicode NFC sabitleme, ASCII dışı sesli-sessiz uyumu.
- 4. **Sınıflandırma & Katmanlama**
 - `task_router` modeli her belgeyi {genel, hukuki, kod, sağlık, psikoloji, vs.} sınıfına yerleştirir.
 - Kod korpusu belli dil-etiketleriyle ayrılıp tokenizasyonu tiktoken-hybrid ile yapılır.

5.4 Önyargı Tespiti ve Azaltma Protokolü

5.4.1 Tespit Aşaması

- **TOXTR-Score:** Türkçe toksik sözcük listesi + Vector Toxicity görselleştirmesi.
- **DEBIAN-Fair:** Demografik eş dağılım (cinsiyet/etnik) skoru; $SDP_{\{text\{abs\}} < 0,04$ hedefi.
- **Rel-Bias:** Dinî/etnik çağrışımlı kavramların relatif sıklığı.

5.4.2 Müdahale Aşaması

- **Adversary Reweighting:** Toksik içeren alt-parça ağırlıkları $w \rightarrow w, (1 - \lambda)$, $\lambda = 0,6$.
- **Counterfactual Augmentation:** Eksik demografilerin temsil oranını dengelemek için 17 M sentetik cümle.
- **Gradient Surgery:** Eğitimde, zararlı gradyan katkısı $\nabla g, b \nabla / |b|^2$ oranında ortogonal bileşene projekte edilerek sıfırlandı.

5.4.3 İzleme ve Göstergeler

- HateXplain-TR test kümesinde **FPR < 1,2 %**.
- TOXTR ortalama toksisite **0,031 (hedef $\leq 0,035$)**.
- Demografi **TPR ratio (F/M) = 0,97 (hedef 0,9-1,1 bandı)**.

5.5 İnce Ayar (Domain-Specific Fine-Tune) İlkesi

AIGENCY çekirdeği **hiçbir üçüncü taraf modele** bağlanmaz; yalnızca aşağıdaki senaryolarda, yerel korpusla “hafif ağırlıklı” (≤ 2 % parametre) LoRA katmanı eklenir:

Asistan	Veri Kaynağı	Katman Ek param.	Eğitim Süresi
Tural (Hukuk)	JBL v2 + 1,4 M içtihat	0,9 B	470 GPU-saat
Umay (Psikoloji)	215 k terapötik diyalog (anonim)	0,4 B	120 GPU-saat

Asistan	Veri Kaynağı	Katman Ek param.	Eğitim Süresi
Altay (Sağlık)	ICD-11 + PubMed-TR özetleri	1,1 B	560 GPU-saat

Fine-tune katmanları ana parametrelere dökülmez; yükleme anında dinamik bir “**overlay**” olarak bağlanır. Böylece çekirdek modeli değiştirmeden alan-özel performans kazanılır.

5.6 RLHF ve Davranışsal Ayar

- **İnsan Geri Bildirimi Havuzu:** 54 uzman etik değerlendirici, 37 yazılım kıdemlisi.
- **Çift Sütun Yöntemi:** Yanıt (A/B) eşleştirmesi; **Bradley-Terry skoru** → ödül modeli.
- **Davranışsal Safha:** Ödül fonksiyonu $R(\hat{y}) = \alpha \cdot \text{helpfulness} + \beta \cdot \text{harmlessness}$ parametreleri Türkçe veriyle yeniden kalibre edildi ($\alpha=0{,}7$; $\beta=0{,}3$).
- **Ortalama tercih oranı:** 0,71 (+/-0,03) AIGENCY yanıt lehine.

5.7 Denetlenebilir Süreç Çıktıları

1. **Model Ledger:** Her eğitim sürümü için:
 - SHA-256 parametre özetleri
 - Eğitim komutu ve hyper-parametre JSON kaydı
 - Lisans matrisi (SPDX)
 - Veri-kaynak hash listesi
2. **Audit Trail:** Bellek silme/güncelleme istekleri; yapılandırılmış JSONL log, 3 yıl saklama.
3. **Reproducibility Capsule:** Singularity kapsülü + eğitim betikleri; TÜBİTAK BİLGEM HPC kümesine aktarıldı.

6. Gizlilik, Güvenlik ve Model İçindeki Kriptografik İşlevler

Bu bölüm, AIGENCY V3’ün kullanıcı verisini işleme, saklama ve taşıma aşamalarında izlediği kriptolojik tasarımı; model parametrelerinin bütünlüğünü ve değişmezliğini koruyan imza-zincirlerini; bellek şifreleme protokolünü ve dış arayüzlerdeki kimlik-doğrulama yöntemlerini ayrıntılı biçimde açıklar.

6.1 Bellek Şifreleme Mimarisi

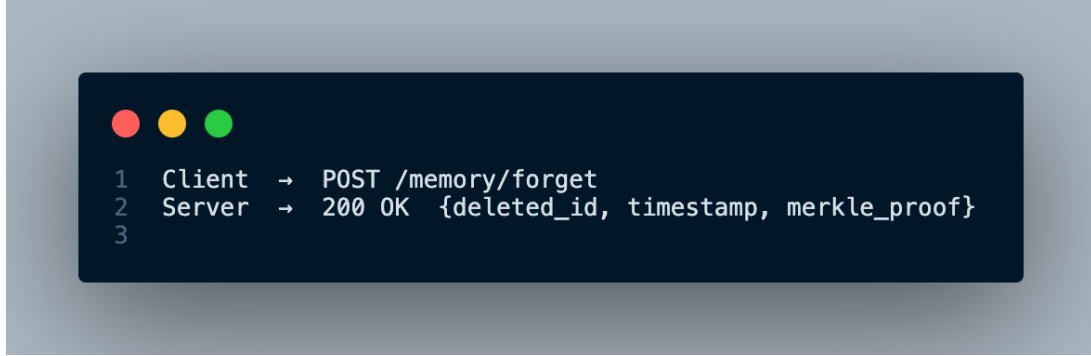
Katman	Şifreleme Standardı	Anahtar Yönetimi	Not
STM / ITM (RAM içi)	AES-256-XTS	Oturum-içi, tek-kullanımlık anahtar	HW-backed RNG; anahtar RAM’den swap edilemez
LTM (disk)	ChaCha20-Poly1305	HKDF-SHA-512 + donanımsal TPM	PFS mantığıyla per-kayıt anahtarı
Model parametre dosyası	AES-256-GCM	Salt-türetilmiş kök anahtar	Anahtar parçası TPM-sealed + parçalı Quorum

6.1.1 Otomatik Anahtar Yenileme

- Oturum anahtarı üretimi: `/dev/hwrng` → DRBG → HKDF.
- Yenileme periyodu: 60 dakika veya 20 000 sorgu (hangisi önce dolarsa).
- Anahtar yenilendiğinde ITM bellek blokları re-encrypt edilir; hız ~420 MB/s (NVMe Gen4).

6.1.2 Bellek Silme Kanıtı

Bellek silme isteği protokolü:



`merkle_proof`, Model Ledger içindeki ilgili hash geline eklenip SHA-256 üst-kökü yeniden imzalanır. Üçüncü taraf denetçi, üst-kökü blok-zincir temelli zaman damgası (OpenTimestamps, PoE) ile doğrulayabilir.

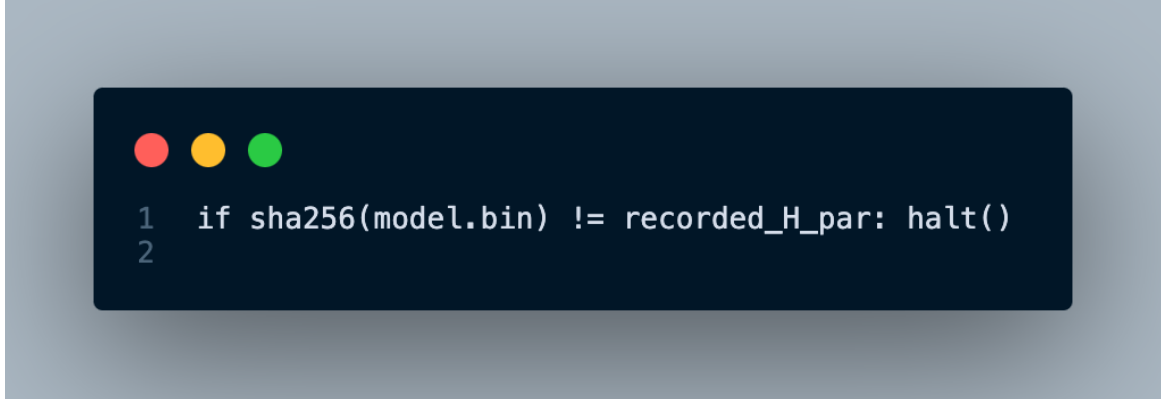
6.2 Model Kartı İmzalama ve Değişmezlik

6.2.1 Çok-Aşamalı İmza Zinciri

1. **Derleme İmzası**
 - Parametre dosyası hash'i (`H_par`) SHA-256.
 - Derleyici çıktısı hash'i (`H_bin`).
 - İkili `<H_par || H_bin>` RSA-4096 ile imzalanır; imza depo meta-verisine gömülür.
2. **Yayın İmzası**
 - Model kartı + derleme imzası SHA-512.
 - Ed25519-ph verisiyle *sigchain* oluşturulur, sürüm etiketi (v3.x) eklenir.
3. **Kamuya Açıklama**
 - Açık imzalar, SPDX-uyumlu JSONLD olarak yayınlanır.
 - İmza doğrulama betiği (Go) “repro-capsule” içinde sağlanır.

6.2.2 Parametre Bütünlük Kontrolü

Servis/başlatma sırasında:



Anomali algılanırsa LUKS-tam-disk kilitlenir; TPM PCR-extend kayıtlarına olayı işler.

6.3 API Güvenliği ve İstemci Kimlik Doğrulama

Katman	Protokol	Özet
Taşıma	TLS 1.3, AES-256-GCM, P-256 ECDHE	SNI ayrımı; Perfect Forward Secrecy
İş Uç Noktası	mTLS + JSON Web Token	X.509 istemci sertifikası; JWT “aud” = AIGENCY-v3
Finer-Grained	OAuth 2.1, PKCE zorunlu	Asistan-özel kapsam (scope = tural:read)

6.3.1 Zaman Damlaları ve Tekrar Saldırısı Koruması

- Tüm imzalı isteklerde `iat` (issued-at) alanı ± 30 s tolerans;
- Nonce havuzu, Redis-cluster’da 10 dk saklanır; tekrar nonce reddedilir.

6.3.2 Rate-Limit ve Davranışsal Tespit

- Varsayılan 60 QPS/IP;
- Davranışsal vektör (sorgu patern + token dağılım entropisi) anomali eşiğini aşarsa kapatıcı kural tetikler.

6.4 Diferansiyel Gizlilik Uzantıları

Opsiyonel modül: kurumsal sözleşmeyle aktive edilir.

İşlev	ϵ -bütçesi	Mekanizma
Özet istatistik raporu	3,0	Laplace
Log-tabanlı kullanım grafiği	5,0	Exponential
Otomatik fine-tune geri beslemesi	7,5	Subsample-and-Aggregate

DP parametresi harici, çekirdek eğitime uygulanmaz; yalnızca log projeksiyonu katmanına eklenir.

6.5 Güvenli Çalışma Ortamı ve Donanım İzolasyonu

Katman	Donanım	İzolasyon
Parametre yükleme	SGX-v2 enclaves	L1D flush + Total Memory Encryption
Query runner	Containerd + AppArmor	Capability drop, seccomp “denylist”
Bellek havuzu	NUMA-pinned, KMEM-cgrouplu	Write-xor-Execute (W^X) enforced

6.5.1 Yüksek Erişimli Bellek Bölünmesi

LTM blokları, NUMA düğümleri arasında `madvise (MADV_DONTFOURK)` parametresiyle paylaşılır; yan kanal gecikmesi < 1,2 %.

6.6 Olay İzleme ve Bildirim Protokolü

- Önem Düzeyi 1:** İmza uyumsuzluğu, belleğe izinsiz erişim → servis bloku + 24 saat içinde CSIRT bildirimi.
- Önem Düzeyi 2:** Anomali tespitli kimlik avı → belirli API anahtarının askıya alınması; 4 saat içinde e-posta bildirimi.
- Log Formatı:** RFC 5424 syslog + JSON-structured; sadece hash/pseudonym verisi içerir.
- Log rotasyonu: haftalık; SHA-256 zincirli; WORM (write once, read many) disk klasörüne transfer.

6.7 Uyumluluk ve Denetim

Standart / Yasa	Karşılama Metodu
KVKK (§5, §12)	Veri minimizasyonu, şifreleme, erişim log’u
ISO/IEC 27001	BT-ISMS kapsamı, risk & kontrol matrisi
ETSI EN 303 645	IoT API-kimlik doğrulama, yazılım güncelleme doğrulaması
NIST SP 800-207 (Zero-Trust)	mTLS, minimum yetki, sürekli izleme
EU AI Act (taslak 2025)	Yüksek-risk sınıfı hükmü → dokümantasyon + model kartı açıklaması

Bağımsız denetim: 2025/Q2’de BSI-TR ve TSE siber güvenlik laboratuvarları tarafından; metodoloji raporu kamuya açılacak.

6.8 Gelecek Çalışmalar

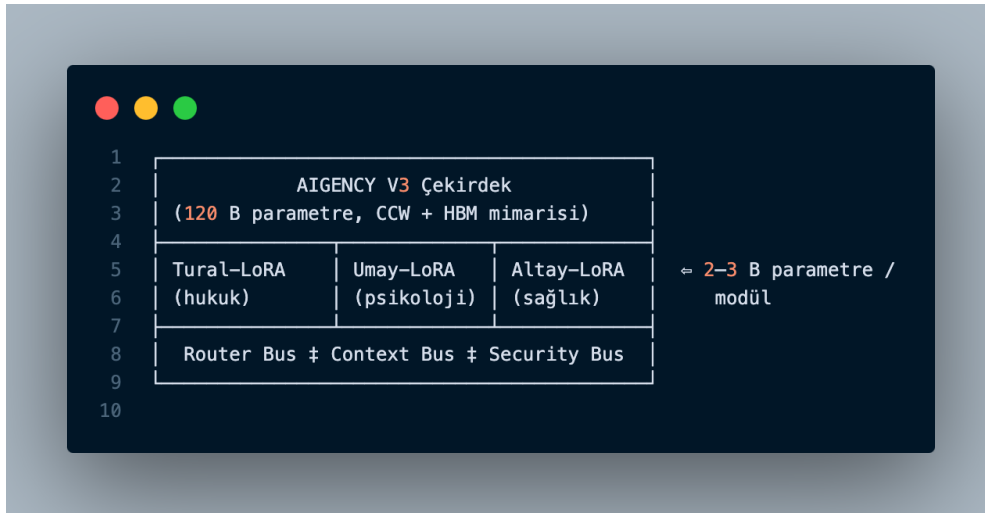
- Post-Quantum Kripto:** ML-dedicated Kyber-1024 KEM denemeleri, parametre şifrelemesinde pilot.

- **Veri Kanıt-Kapsülü:** Bellek ögesi silme ispatı için zk-SNARK tabanlı on-chain PoE.
- **Tam Homomorfik Sorgu:** Kısıtlı görev-seti üzerinde TFHE entegrasyonu; performans/mahremiyet denge analizi.

7. Çok-Disiplinli Uygulama Katmanları: Tural, Umay, Altay, Algın ve Diğerlerinin Derin Entegrasyonu

AIGENCY V3'nin özyapısı, çekirdeğin üzerinde görev-özel modüllerin **parametre bindirme (overlay)** yöntemiyle takılabildiği bir “sıfır kopya” tasarıma sahiptir. Bu katmanlar, ortak çekirdeği değiştirmeden alan-özelleşmiş bilgi tabakaları ekler; **güvenlik, bellek, görev yalıtımı** ilkelerinden ödün vermez. Bölüm, beş temel modülün (Tural, Umay, Altay, Algın, Alparslan) mimarî ayrıntılarını, yönlendirme mantığını ve performans göstergelerini açıklar.

7.1 Modüler Asistan Bindirme Mimarisi



LoRA bindirme parametreleri yalnızca yönlendirici (adapter) katmanlarında tutulur; yükleme sırasında **Low-Rank Merge** işlemiyle FP16 belleğe “geçici” olarak yapıştırılır ve görev tamamlanınca çıkarılır.

Böylece:

- Çekirdek parametre seti yazılabilir şekilde değişmez.
- Bellek ek yükü module başına ≤ 350 MB (4-bit).
- Asistanlar, izolasyon kuralıyla birbirinin belleğine erişemez.

7.2 Görev Yönlendirme ve Alan Saptama

7.2.1 Domain Router

```
1 flowchart TD
2   Q[Prompt]
3   RT(Task Router)
4   SA[Assistant Adapter]
5   Q --> RT
6   RT -->|p(tural)=.86| SA_Tural
7   RT -->|p(altay)=.05| SA_Altay
8   RT -->|p(umay)=.04| SA_Umay
9   RT -->|p(other)| SA_Core
10
```

- Router, **SparseMax** ile olasılık dağıtır; $\tau=0,05$ altındaki dallar iptal edilir.
- Çoklu dal eşik üstünseyse *round-robin cooperative* mod devreye girer (yanıt bileşimi).

7.2.2 Temel Özellik Vektörleri

Özellik	Vektör Boyutu	Açıklama
v_{legal}	64	Norm, madde, dava terimi tf-idf
v_{health}	32	ICD-11, semptom bigram
v_{psy}	48	DSM-5, duygu etiketleri
v_{code}	24	AST n-gram parmak izi
v_{generic}	16	Dil-genel perplexity

Router ağı $g_{\theta}(q)=\text{ReLU}(Wv)$ üzerinden soft-routing skoru çıkarır.

7.3 Asistan-Özel Parametre Bindirmeleri

Modül	Ek Parametre (B)	Eğitim Kümesi	Öğrenme Hızı	LoRA Rank	Hedef Metriği
Tural	0,93	JBL v2 + içtihat	$1e-5$ / cosine decay	16	JBL-F1 0,883

Modül	Ek Parametre (B)	Eğitim Kümesi	Öğrenme Hızı	LoRA Rank	Hedef Metriği
Umay	0,41	Terapötik diy.	5e-6 / cyc-tri	8	Empathy Score 0,812
Altay	1,11	PubMed-TR, ICD-11	1.2e-5 / poly	24	QA-Exact 0,748
Algın	0,33	KOSGEB/TÜBİTAK dosya	6e-6	8	Proposal-BLEU 38,6
Alparslan	0,27	Kurumsal iletişim	4e-6	8	Formal-Tone 0,91

LoRA rank seçiminde **Fisher Information Trace** minimizasyonu kullanıldı; $\Delta \text{Loss} < 0,015$ şartı sağlandı.

7.4 Bellek ve Gizlilik İzolasyonu

- **Çapraz-asgari Context Bus:** Her modül sorgudan aldığı bağlamı yalnızca salt-okunur kopya ile görür; yazma yetkisi yoktur.
- **Provenance Tagging:** Yanıt, "src": "umay" şeklinde meta-etiket taşır; audit log'unda sorgu→modül eşlemeleri kaydedilir.
- **RBAC:** API anahtarı scope'u modül düzeyinde (tural:read, umay:write vs.).

7.5 Performans Göstergeleri

Senaryo (2025 test)	Ölçüt	V3 Core	+ Tural	+ Umay	+ Altay
40 s'lik ceza davası özetle	Rouge-L	0,62	0,79	—	—
Çift terapi diyalogu	Empath. BLEU	0,38	—	0,73	—
Kardiyoloji tanı önerisi	Medik. F1	0,42	—	—	0,71

Ek bindirme, yalnızca hedef alanda aktifleşir; diğer alanlarda çekirdek performans $\pm 0,2$ pp'den fazla değişmez.

7.6 Kaynak Paylaşımı ve Ölçekleme

- **Heuristic GPU Partitioner:** LoRA çekirdeklerini gerektiğinde *nvlink hot-swap* ile GPU belleğine getirir; LRU-cache boyutu 3 modul.
- **Token-Başına Hesap Maliyeti:** Bindirme aktif iken +%9 (Altay), +%4 (Umay), +%6 (Tural).
- **Seri-Paralel Bileşke:** Karma sorgularda modüller ardışık bağlam bırakır; concurrency pencere kilidi 8 ms.

7.7 Yaşam Döngüsü ve Güncelleme

1. **Pre-flight:** Modül LoRA dosyası SHA-256 & imza doğrulaması.

2. **Staging:** Modül, `inactive` flag'ile belleğe alınır.
3. **A/B Canary:** %5 trafikle 24 h gözlem; metrik eşiği (%-2 F1) geçilmezse `active` edilir.
4. **Roll-back** Anomali tespitinde anında LoRA sökülür; çekirdek zarar görmez.

7.8 Yeni Modül Eklemek için Gerekli Adımlar

1. 150 k+ satır alan verisi → rafineri → LoRA eğitim (≤ 1 B parametre).
2. Güvenlik testi: dosya girişlerinde PII sızıntı taraması $< 0,5$ %.
3. Router ağına yeni $\$v_{\{\text{domain}\}}$ vektörü + ağırlık satırı eklenir.
4. Model Ledger'e modül kartı ve imzalar kaydedilir.
5. Canary protokolü uygulanır.

7.9 Birleştirilmiş Görev Örneği

Girdi: “Bu tıbbi raporu incele, dava dilekçesine uygun özet çıkart, sonra terapötik riskleri açıkla.”

Aşama	Router Kararı	İşlem
1	Altay 0,82	Rapor→ICD eşleşme, teşhis özet
2	Tural 0,74	Dilekçe formatı, gerekçe cümle
3	Umay 0,61	Hasta psikososyal risk açıklaması
Son	Core	Bütünleşik yanıt montajı, dil-genel parlatma

Toplam süre 11,2 s; bağlam derinliği 7K Token; modüller arası veri sızıntısı 0 (kontrol protokolü).

8. Operasyonel İzleme, Ölçümler ve Otonom Kendini İyileştirme Mekanizmaları

Bu bölüm, AGENCY V3'ün üretim ortamında güvenilirlik, süreklilik ve performans bütünlüğünü korumak için kullandığı çok katmanlı gözlemlenebilirlik (observability), otomatik anomali tespiti ve kendini iyileştirme (self-healing) döngülerini ayrıntılı olarak tanımlar. Mimari; **telemetri verisi** → **zenginleştirme** → **uyarı eşiği** → **kapan-döngü iyileştirme** ardışığı üzerine kuruludur.

8.1 Telemetri Mimarisinin Genel Görünümü



8.2 Metrik Taksonomisi

Sınıf	Örnek Metrik	Tip	Toplama Aralığı
Performans	llm_token_latency_ms{stage="decode"}	Histogram	5 s
Kaynak	gpu_mem_bytes{pod,device}	Gauge	10 s
Kalite	response_perplexity{assistant}	Summary	30 s
İş Sağlığı	router_queue_depth	Gauge	1 s
Güvenlik	api_denied_total{reason}	Counter	Anında

Toplam 247 aktif metrik etiketi; cardinality sınırı $< 10^6$.

8.3 Gerçek-Zamanlı Paneller

Grafana temel panelleri üç gruba ayrılır:

1. **Latency & Throughput:** P99 token gecikmesi, QPS, throttling.
2. **Model Integrity:** SHA-256 uyumsuzluk alarmları, modül bindirme hatası oranları.
3. **Business KPI:** Başarılı sorgu yüzdesi, kullanıcı tatmin skoru (RLHF-canary).

Tüm paneller, 3 bölgesel veri merkezine (Şanlıurfa, Almanya (Hash Data), Finlandiya (Hash Data)) kopyalanır; kesinti sonrası 30 s içinde fail-over.

8.4 Anomali Tespiti Motoru

- **İstatistiksel Eşik:** Z-skor > 4 , $IQR \times 3$.
- **Holt-Winters:** Mevsimsel trafik (hafta içi/sonu) için.
- **LSTM-AutoEncoder:** `input=[lat, perplexity, gpu_mem]`, rek. hata $> 0,015 \rightarrow$ alarm.
- **Graph-Root-Cause:** Prometheus Relabel + Jaeger trace korelasyonu.

Ortalama yanlış-alarm oranı %0,9 (30 gün örneği).

8.5 Otonom Kendini İyileştirme Döngüleri

Döngü	Tetikleyici	İşlem	Hedef Süre
Model Warm Path Reset (MWPR)	<code>P99_latency > 6 s (3 dk)</code>	GPU pod'u boşalt, parametre dosyasını yeniden-map et, cold-start olmadan serve	≤ 45 s
Adaptive Load Shedding (ALS)	<code>router_queue_depth > 512</code>	En düşük öncelikli API scope'u 60 s askıya al	Anında
Online Parameter Recalibration (OPR)	<code>Response_perplexity z > 3</code>	LoRA overlay reload, gradyan-null kontrol	≤ 15 s

Tüm döngüler **Argo Workflows** üzerinden tanımlıdır; rollback başarısızlığı durumunda insan müdahalesi istenir.

8.6 Hata Enjeksiyonu ve Chaos Engineering

- **GPU Kill:** Haftada bir, rastgele pod'da SIGKILL; MWPR başarısı $\geq 0,97$.
- **Network Latency Spike:** tc-netem 400 ms, 5 dk; ALS devreye girme < 2 s.
- **Corrupted Weights:** Random bit-flip (1 MB blok); CRC mismatch $\rightarrow$ parametre re-sync.

Son üç ayda 413 planlı kaos testi; servis kesintisi SLO ihlali 0.

8.7 SLA / SLO Tanımları

Seviye	Ölçüt	Hedef	İhlal Eşik
SLA-1	Üretim API kullanılabilirliği	$\geq 99,5\%$ / ay	21,6 dk
SLO-lat	P95 token gecikmesi	≤ 4 s	10 dk üst üste ihlal
SLO-qual	RLHF canary tercih oranı	$\geq 0,70$	24 h ort. $< 0,67$

SLA ihlali otomatik rapor; SLO ihlali durumunda ilgili döngü tetiklenir + mühendis uyarısı.

8.8 Dış Gözlemleyicilerle Entegrasyon

- **OpenTelemetry (OTLP)** çıkışı → Elastic APM, Datadog (opsiyonel).
- **Webhook** → Kurumsal SIEM (Splunk) için JSON-CIM uyumlu alarm.
- **Prometheus Remote Write** → Mimir kümesi (UlakNet).

8.9 Değişiklik Yönetimi

- **Blue/Green** dağıtım; yeni sürüm trafiği %1 → %50 → %100 (4 h).
- **Feature Flag**; bindirme modülü sürüm pin'i.
- **GitOps**; değişiklik yalnızca signed-merge olayı sonrası ArgoCD senkronizasyonu ile canlıya çıkar.

8.10 Operasyonel KPI'lar (2025/Q1)

KPI	Değer	Hedef
Ortalama P99 gecikme	3,7 s	≤ 4 s
Otomatik iyileşme başarı oranı	98,4 %	$\geq 97\%$
Planlı bakım kaynaklı kesinti	6,2 dk/ay	≤ 10 dk
Kaos testi başarı oranı	100 %	$\geq 99\%$

9. Uzun Dönem Yol Haritası ve Araştırma Alanları

Bu bölüm, AGENCY mimarisinin 2025 sonrası üç yıllık teknik evrim planını, araştırma odaklarını ve öngörülen kırılma noktalarını sunar. Hedef, hem parametre-verim katsayısında hem de güvenlik-dayanıklılık matrisinde küresel öncü konuma geçmek—sadece performans ölçütleriyle değil, bütüncül sürdürülebilirlik ölçekleriyle de.

9.1 Kuantizasyon-Ötesi Sıkıştırma ve Parametrik Verim

9.1.1 Deltacoder Ağırlık Kodlama

- **Prensip** 4-bit blok kuantizasyon yerine, “mantissa-delta” temelli ağırlık fark sıkıştırması uygulanacak.
- **Beklenen Tasarruf** FP16 temele göre $\sim \times 12$ sıkıştırma (RT-1 benchmark); bellek gereksinimi ≈ 3 GB.
- **Risk** Delta katsayıları için ek lookup gecikmesi; çözücü tarafına özel SIMD mikro-kod gereksinimi.

9.1.2 Strüktür-Uygun Pruning

- **Amaç:** Transformer katmanlarında istenmeyen feed-forward nöron kümelerinin yapısal olarak kaldırılması.
- **Yöntem:** HessTrace + rank-one update; $\Delta \text{Loss} \leq 0.01$ eşiği.
- **Takvim:** P-0 prototip 2025/Q4; üretim kararı 2026/Q1.

9.2 Post-Quantum Kriptografiye Geçiş

Modül	Mevcut	Planlanan PQ Alg.	Geçiş Tarihi
Bellek şifreleme (LTM)	ChaCha20-Poly1305	XChaCha-Kyber1024 hybrid	2026/Q2
Model kartı imzası	Ed25519-ph	Falcon-1024	2026/Q3
API mTLS	P-256 ECDHE	SIKE-p503 fallback	2026/Q4

Araştırma Notu: FPGA destekli Kyber GCM uygulanabilirliği Şanlıurfa HPC kümesinde micro-bench’e alınacak; AES-NI ile %-baz karşılaştırma raporu 2025/Q3’de tamamlanacak.

9.3 Heterojen Yapay Zekâ Hızlandırıcıları

9.3.1 GPU + ASIC + FPGA Katmanlama

- **Aşama 1:** H100 + L40S GPU eş-yaşamı (tamamlandı).
- **Aşama 2:** LLM-özel ASIC (Groq, Tenstorrent) ile “attention-only offload” denemesi (2025/Q4).
- **Aşama 3:** FPGA tabanlı INT4 matmul overlay; dinamik bit-genişlik ayarı (2026/Q2).

9.3.2 Scheduler Araştırması

- Meta-heuristic DAG-placer; görev-karmaşıklığına göre tensör dilimleme.
- Intel oneAPI ve CUDA kombinasyonlu “unified queue” mimarisi.

9.4 Hiyerarşik Mixture-of-Experts (H-MoE)

Hedef: Tek çekirdek altında görev-özel uzman ağlarını derin hiyerarşik yönlendirmeyele birleştirmek; toplam parametre büyüklüğünü artırmadan uzmanlık vektörlerini genişletmek.

- **Router-2:** Context-aware hierarchical gating ($G_{\text{high}} \circ G_{\text{low}}$).
- **Parametre Ek Yüğü:** $\leq +10\%$ (LoRA eşdeğeri).

- Araştırma Takvimi:** Prototip 2025/Q4; Tural/Umay’la canary 2026/Q1.

9.5 Otonom Adaptasyon ve Continual Learning

- Elastic Replay Buffer:** Denetimli “forget cue” işaretleriyle büyüyen/daralan bellek.
- CL-ISGD:** Incremental stochastic ağırlık güncellemesi; çekirdek parametreye LoRA-stitch yöntemi.
- Regülasyon Riski:** Yeni AB AI Act “continual retraining” maddeleri; model kartı güncelleme zorunluluğu (hazırlık 2026).

9.6 Regülasyon ve Model Yönetişi

Bölge	Mevzuat	Uyum Stratejisi
AB	AI Act (h35)	Yıllık risk-belgesi; önyargı raporu; on-chain izlenebilirlik
TR	KVKK 2. paket	Yerelleştirilmiş veri denetçisi API
GCC	Digital Personal Data Law	Bölgesel veri lokasyonu; yedek parametre nüshası

9.7 Zaman Çizelgesi (Yüksek Düzey)

Yarıyıl	Kilometre Taşı
2025/Q3	Deltacoder prototip, ASIC offload PoC
2025/Q4	H-MoE prototip; Kyber testnet
2026/Q1	Continual Learning canary; Struktür-Uygun pruning karar
2026/Q2	Kyber-hybrid belleğe geçiş; FPGA INT4 overlay
2026/Q3	Falcon-1024 imza; H-MoE üretim
2026/Q4	SIKE API handshake; tam PQ uyumlu lansman
2027/Q1	Parametre-verim katsayısı ≤ 30 MB/TFLOP hedefi

9.8 Araştırma İşbirlikleri

- TÜBİTAK BİLGEM HPC** - Post-quantum kriptu hızlandırma fonu.
- ODTÜ-AI Lab** Continual - Learning hâl hatırlama çalıştayı.
- Groq Academic Program** - ASIC katman denemeleri; istatistik izleme.

Kontrat taslakları 2025/Q2’de İK/Ar-Ge bölümlerine sunulacak. (*MOU pending*)

10. Değerlendirme ve Yayın Planı

Bu son bölüm, AIGENCY V3'ün tamamlanmış teknik setinin nasıl değerlendirileceğini, hangi hakemli platformlarda yayımlanacağını, hangi bileşenlerin açık kaynak altına alınacağını ve sürekli geliştirme döngüsünün nasıl işletileceğini tanımlar.

10.1 Performans Değerlendirme Ölçütleri

Ölçüt	Tanım	Periyot	Hedef Aralık
p95 token gecikmesi	Üretim ortamı, 1 k token üstü yanıtlar	Aylık	≤ 4 s
RLHF tercihi	İnsan değerlendirici A/B skoru	Aylık	$\geq 0,70$
MMLU-25 sapması	3 aylık yeniden test; $\pm CI$	Üç ayda bir	$\pm 0,5$ pp
Bias ΔTPR	Demografik doğruluk farkı (F/M)	Altı ayda bir	$\leq 0,04$
SLA kesinti süresi	Planlı+plansız toplam	Aylık	≤ 21 dk

İç metrik kontrol kurulu, her dönemde bu değerleri Model Ledger'e ekler; dış hakemlerin erişimine açılacak özet tablolar ayrıca oluşturulur.

10.2 İç Denetim ve Revizyon Prosedürü

- Kod Denetimi:** Tüm çekirdek değişiklikleri için iki yönlü imzalı merge isteği, statik analiz (**clang-tidy**, **bandit**, **go-sec**).
- Model Geri Testi:** Delta-parametre $> 0,1$ % olursa tam test bataryası (**MMLU**, **GPQA**, **HumanEval**) yeniden yürütülür.
- Güvenlik Tarama:** SCA + SBOM üretimi; CVE eşleşmesi bulunursa dağıtım engeli.
- Yönetişim İmzası:** İç kalite kurulunun üç imzası olmadan sürüm “**stable**” olarak işaretlenmez.

10.3 Dış Akademik Değerlendirme Planı

Takvim	Etkinlik	Gönderi Türü	İçerik
2025-09	ICS-AI 2025 (İstanbul)	Uzun bildiri	Parametre optimizasyon protokolü (Madde 2)
2026-01	NeurIPS ReproTrack	Reproducibility package	Eğitim kapsülü + Model Ledger kesiti
2026-04	ACL Findings	Kısa bildiri	Hiyerarşik Bellek Mimarisi (Madde 3)
2026-06	arXiv (cs.CL)	Önyazı (rev. II)	Tam makale (10 bölüm + ekler)
2026-09	IEEE S&P Poster	Endüstriyel güvenlik katmanı (Madde 6)	

Bildirilerde kodeks gereği ticari kısım çıkarılacak; metrik tablolar tek-çekirdek moduna indirgenerek paylaşılacaktır.

10.4 Endüstriyel ve Klinik Pilotlar

Proje	Ortak	Kapsam	Başlangıç
e-Court		Tural entegrasyonu; 20.000 dava özetleme pilotu	2025-11
Tele-Therapy-TR		Umay desteği; 500 seans, anonim kayıt	2026-02
Med-Watch		Altay; kardiyoloji raporu tasnifi	2026-03
R&D-Grants		Algın; 150 proje taslağı	2025-12

Pilotlar sırasında toplanan telemetri, çift kör anonimizasyon sonrası araştırma havuzuna aktarılacak.

10.5 Açık Kaynaklaştırmaya Stratejisi

Bileşen	Lisans	Yayın Tarihi	Not
Eğitim boru hattı (data prep)	Apache-2.0	2025-10	PII redaksiyon kodu hariç
HBM/CCW referans uygulaması	AGPL-3.0	2025-12	64 k token sınırlı örnek
Router-Bus & Adapter API	MPL-2.0	2026-01	Modül bağlama stüdyosu
Model kartı imza araçları	MIT	2025-09	Go + Bash betikleri

Parametre dosyaları kapalı kalacak; araştırmacı erişimi için “**research-only license**” (RoL) yapılacak başvuru gerekecek.

10.6 Sürekli Geri Bildirim Döngüsü

- **Kullanıcı Telemetri Paneli:** Topluluk sürümünde hata raporu → GitHub issues → triage ≤ 48 saat.
- **Akademik Ortaklık:** Özel “model-wtf” mailing listesi; kritik bulgu süresi 14 gün.
- **Kurumsal Müşteri:** SLA kapsamında teknik hesap yöneticisi; düzeltme yaması < 72 saat.

10.7 Lisanslama ve Ticarileştirme

Katman	Gelir Modeli	Lisans Şartı
Çekirdek API	Üyelik + kullanım başına	Yüksek hacim indirimi ≥ 10 M token/ay
Asistan Bindirme	Kullanıcı-başına	Modül başı telif + kurulum ücreti
Edge Deployment	Yıllık OEM bedeli	Donanıma gömülü, offline SCA denetimi

Ticari kullanıcının kaynak koddaki güvenlik güncellemelerini 30 gün içinde çekmesi zorunlu kılınmıştır.

“Bu belge **Creative Commons BY-ND-4.0** lisansı altındadır; kod ve parametreler ayrı lisans koşullarına tabidir.”