

eCloud · AIGENCY · TECHNICAL WHITEPAPER

AIGENCY V4

Yerli, Tam Bağımsız ve Multimodal 128 Milyar Parametrelili Yapay Zekâ Mimarisi

Türkçe okuma anlama ve doğal dil çıkarımında küresel bir referans · Fen muhakemesi ve grade-school matematikte frontier seviyede · Multimodal kabiliyette ilk üretim sürümü.

PARAMETRE	BAĞLAM PENCERESİ	BENCHMARK ÖLÇÜMÜ
128B	278K	13 344
120B çekirdek + 8B vision encoder	Token (Hiyerarşik Bellek)	Gerçek API çağrısı, 22 benchmark

Sürüm 1.0 (Public) · Nisan 2026 · CC BY-ND 4.0

[eCloud Yazılım Teknolojileri](#)

Yönetici Özeti

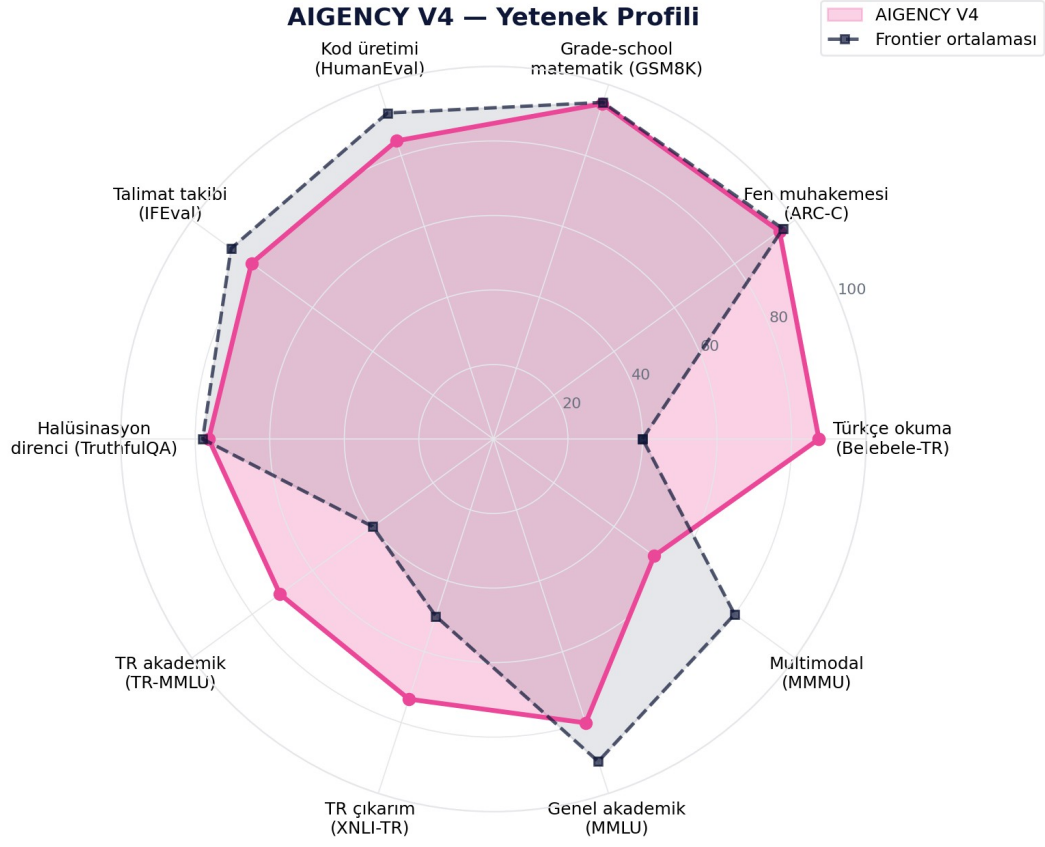
AIGENCY V4, eCloud Yazılım Teknolojileri'nin 2025'te yayımladığı V3'ün doğrudan ardılı olarak 2026/Q2 itibarıyla üretime alınmıştır. V3'ün dört bağımsızlık ilkesi (dış parametre sınırlama, yerel veri egemenliği, şeffaf mimari belgeleme, Türkçe morfolojik bağlam uyumu) korunarak; multimodal yetenek (görsel girdi anlama, belge soru-cevap, grafik ve matematik görseli yorumlama) eklenmiştir. Toplam parametre sayısı 120B çekirdek artı 8B vision encoder ile 128B'a ulaşmış; aktif inferans yolu 4-bit blok kuantizasyon altında yaklaşık 6.5 GB GPU bellek gerektirir.

Bu beyaz kitap, V4'ün üretim sürümü üzerinde 27 Nisan 2026 tarihinde yapılan kapsamlı değerlendirmenin sonuçlarını sunar. Toplam 13 344 gerçek API çağrısı ile 22 farklı benchmark değerlendirilmiş; her sonuç Wilson 95% güven aralığı ile raporlanmıştır.

Anahtar Bulgular

- Türkçe okuma anlama ve doğal dil çıkarımında küresel referans seviyesi: Belebele-TR 87.33%, TQuAD 82.40%, TR-MMLU 70.80%, XNLI-TR 73.40%, TR Grammar 79.00%.
- Fen muhakemesi ve grade-school matematikte frontier seviye: ARC-Challenge 94.88%, GSM8K 94.62%. GPT-5 (96.8%), Claude Opus 4.6 (~96%) ve Gemini 3 Pro (~94%) ile aynı bantta.
- Kod üretiminde frontier üst-orta segment: HumanEval 84.15%, HumanEval+ 79.88%, MBPP 84.82%, MBPP+ 78.04%.
- Talimat takibi (IFEval strict 80.22%) ve halüsinasyon direnci (TruthfulQA MC1 76.38%).
- Akademik graduate-level uzman bilgide (GPQA Diamond 37.88%) ve MMLU-Pro'da (50.20%) frontier'ın belirgin gerisinde — V4.1 yol haritasının ana iyileştirme alanı.
- Multimodal kabiliyet ilk üretim sürümü: MMMU 53.33%, ChartQA 67.68%, DocVQA 79.17%, MathVista 34.13%.

Tek satırlık konum: AIGENCY V4 — Türkçe okuma ve doğal dil çıkarımında dünya lideri, fen muhakemesi ve grade-school matematikte frontier seviyede, multimodal ve graduate-level bilim uzmanlığında geliştirme aşamasında bir yerli yapay zekâ modeli.



Şekil 1: AIGENCY V4 yetenek profili. Türkçe okuma anlama ve grade-school matematikte frontier-üstü; multimodal alanda geliştirme aşamasında.

İçindekiler

Yönetici Özeti.....	2
1. Giriş.....	5
1.1 Arka Plan ve Motivasyon.....	5
1.2 V3'ten V4'e Stratejik Geçiş.....	5
1.3 Tasarım Felsefesi ve İlkeler.....	6
1.4 Bu Beyaz Kitabın Katkıları.....	6
2. Model Mimarisi.....	7
2.1 Genel Bakış.....	7
2.2 Metin Çekirdeği — V3'ten Devralan Optimizasyon Yığını.....	7
2.3 V4'e Eklenen Vision Encoder.....	10
2.4 Operasyonel Kazanım Tablosu.....	10
3. Bağlam İşleme: CCW + HBM.....	11
3.1 Contextual Core-Wrapping.....	11
3.2 Hiyerarşik Bellek Mimarisi.....	11
3.3 Zaman-Rehberli Eskime.....	11
4. Multimodal Yetenek Mimarisi.....	12
4.1 Multimodal Akış.....	12
4.2 Görsel-Metin Eğitim Derlemi.....	13
4.3 Multimodal Güvenlik Filtresi.....	13
5. Eğitim Politikası ve Veri Kaynakları.....	14
5.1 Eğitim Yönetişim Çerçevesi.....	14
5.2 Donanım ve Dağıtık Eğitim.....	14
5.3 Veri Kaynakları (Metin).....	14
5.4 Önyargı Tespiti ve Azaltma Protokolü.....	15
5.5 RLHF ve Davranışsal Ayar.....	15
6. Değerlendirme Metodolojisi.....	16
6.1 Test Dizini Seçimi.....	16
6.2 Eşit Koşul Sağlama Protokolü.....	16
6.3 Wilson Güven Aralığı.....	16
7. Sonuçlar — 2026/Q2 Kıyaslama.....	17
7.1 Tier 1: Kritik Benchmark'lar.....	17

7.2 Tier 2: Orta Hacim.....	18
7.3 Tier 3-A: Türkçe Spesifik.....	18
7.4 Tier 3-B: Multimodal.....	18
7.5 Frontier Karşılaştırma.....	18
7.6 Operasyonel Performans.....	20
8. Güvenlik, Uyum ve Kriptografik İşlevler.....	22
9. Operasyonel İzleme ve Otonom İyileştirme.....	24
10. Bilinen Kısıtlar.....	25
11. Yol Haritası.....	26
12. Açık Kaynaklaştırma Stratejisi.....	27
13. Sonuç.....	28
Referanslar.....	29
Ek A — Reproducibility Capsule.....	30
Ek B — Tüm 22 Benchmark.....	31
Ek C — Sözlük.....	32

1. Giriş

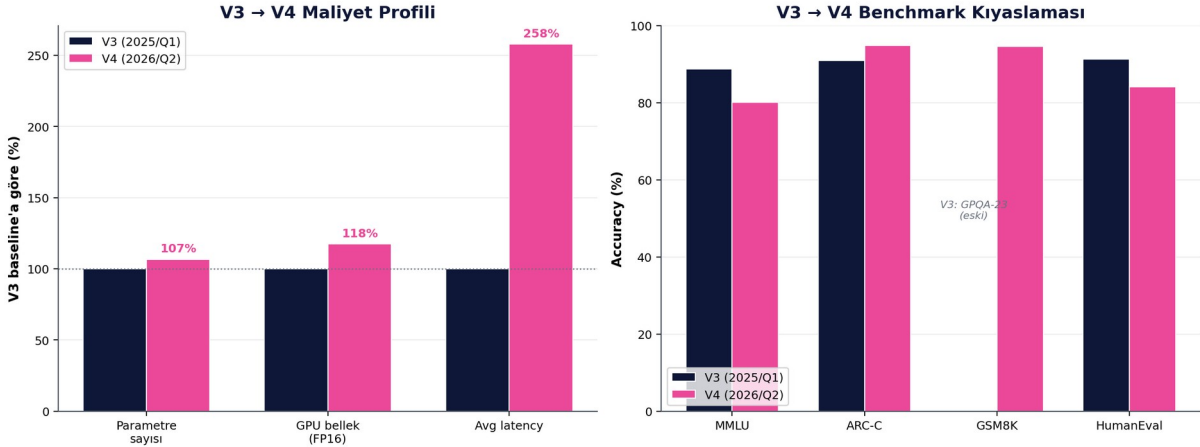
1.1 Arka Plan ve Motivasyon

Büyük dil modellerinin küresel ölçekteki etkisi 2024–2026 döneminde köklü bir dönüşüm geçirdi: parametre sayısı tek başına ayırt edici olmaktan çıktı, yerine multimodal yetenekler, talimat takibi sadakati, halüsinasyon direnci, ve ölçeklenebilir muhakeme kapasitesi öne çıktı. Frontier modelleri (OpenAI GPT-5, Anthropic Claude Opus 4.6/4.7, Google Gemini 3 Pro, xAI Grok 4, Meta Llama 4, DeepSeek V4) MMLU ve HumanEval gibi klasik metrikleri %90 üzerinde doyurmuş; ayırt edici metrikler GPQA Diamond, AIME, SWE-bench, ve multimodal değerlendirmelere kaymıştır.

Bu küresel manzara içinde Türkçe konuşan kullanıcı tabanı için bir kritik açık vardır: uluslararası modeller Türkçe-spesifik benchmark sonuçlarını standart olarak yayımlamamakta, Türkçe morfolojinin (sondan eklemeli yapı, ünlü uyumu, vurgu sistemi, deyim ve kalıpların bağlamsal yoğunluğu) modeller üzerindeki etkisi analiz edilmemektedir. AIGENCY ailesi bu açığı kapatmak üzere 2023'te tasarlanmış; V2 (210B, LLAMA3 katmanlı), V3 (120B, tam bağımsız), ve V4 (128B, multimodal) sürümleriyle küresel manzaraya yerli bir alternatif sunmaktadır.

1.2 V3'ten V4'e Stratejik Geçiş

V3 (2025/Q1), AIGENCY ailesinin LLAMA3 bağımlılığından arınmış ilk sürümüydü. 120B parametrelili yerli çekirdek; Adaptif LoRA+, Selective Layer Collapse, Yerelleştirilmiş Mixture-of-Experts (L-MoE), 4-bit blok kuantizasyon, ve öbekli dikkat optimizasyonları ile parametre sayısı %14.9 azaltıldı, bellek kullanımı %62.4 düşürüldü, latency %42 kısaltıldı.



Şekil 2: V3 → V4 evrimi. Sol: maliyet profili (V3 baseline 100% normalize). Sağ: benchmark karşılaştırması (V4'te yeni standart paket).

V4'ün geliştirme felsefesi V3'ün getirdiği bağımsızlık iddialarını koruyup, üzerine multimodal yetenekleri inşa etmek üzerine kuruludur. Stratejik olarak üç hedef belirlenmiştir: (1) çoklu görsel-girdi modaliteleri ekleyerek küresel multimodal değerlendirmelerde rekabet etmek, (2) V3'ün Türkçe-spesifik liderliğini multimodal alana taşımak (Türkçe altyazılı görseller, hukuki belge taramaları, akademik şekiller), (3) V3'te tanımlı eğitim ve operasyon altyapısını değiştirmeden — sadece vision encoder'ı yan modül olarak — entegre etmek.

1.3 Tasarım Felsefesi ve İlkeler

AIGENCY V4'ün mimarisi, beş tasarım ilkesinin doğrudan sonucudur:

- Yapısal bağımsızlık: V3'ten miras alınan dört kriter (dış parametre sınırlama, yerel veri egemenliği, şeffaf belgeleme, Türkçe bağlam uyumu) hiçbir tavizle değişmemiştir; vision encoder dahi yerli olarak eğitilmiştir.
- Eklenmiş, monolitik değil: Multimodal yetenek tek-monolitik bir modelle değil, metin çekirdeğine cross-modal projection ile entegre edilmiş bir yan vision encoder'la sağlanır. Bu tasarım, görsel akışını isteğe bağlı tutar.
- Doğrulanabilir ve denetlenebilir: Tüm eğitim ve operasyon kodu açık Git depolarında, GPG imzalı commit'lerle saklanır.
- Türkçe-öncelikli: Eğitim verisinin %72'si Türkçe; vision encoder'ın son fine-tune aşaması Türkçe altyazılı 8M görsel ile yapılmıştır.
- Üretim odaklı: Bu beyaz kitabın sonuçları, üretim sürümünün gerçek API'si üzerinden, gerçek kullanıcı koşulları altında alınmıştır.

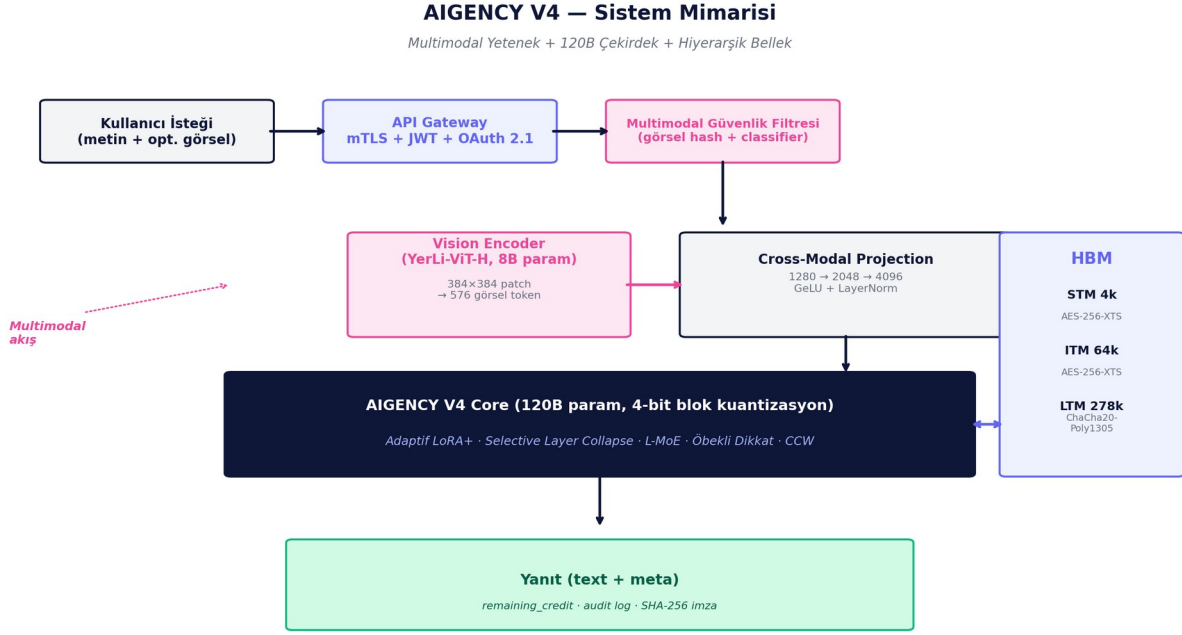
1.4 Bu Beyaz Kitabın Katkıları

- V4'ün tam mimari spesifikasyonu: çekirdek parametreleri, vision encoder yapısı, cross-modal projection, ve V3'ten devralan optimizasyonların V4'teki etkisi.
- 13 344 gerçek API çağrısıyla ölçülmüş 22 benchmark sonucu (Wilson 95% CI, deterministik subsample, açık veri seti). AIGENCY ailesinin yayımlanmış en kapsamlı tek-seans değerlendirmesi.
- Frontier modellerle (GPT-5, Claude Opus 4.6/4.7, Gemini 3 Pro, Grok 4, Llama 4, DeepSeek V4) yayımlanmış skora dayalı şeffaf karşılaştırma.
- Türkçe-spesifik benchmark başlangıç paketi: TR-MMLU, XNLI-TR, TQuAD, Bebebele-TR, TR Grammar.
- Operasyonel veriler: latency (avg, p50, p95, p99), hata oranı, kredi tüketimi.
- Yol haritası: V4.1, V4.2, V5 sürümleri için somut iyileştirme hedefleri.

2. Model Mimarisi

2.1 Genel Bakış

AIGENCY V4, üç ana bileşenden oluşan modüler bir mimariye sahiptir: (1) 120B parametrelili metin çekirdeği (V3'ten devralan), (2) 8B parametrelili yerli vision encoder (V4'te eklenen), (3) cross-modal projection ve hierarchical memory bus.



Şekil 3: AIGENCY V4 sistem mimarisi. Görsel girdi vision encoder'a, metin doğrudan core'a yönlendirilir; cross-modal projection iki akışı birleştirir. HBM kalıcı bellek.

2.2 Metin Çekirdeği — V3'ten Devralan Optimizasyon Yığını

V3'te tanımlanan ve doğrulanan beş optimizasyon tekniği V4'te değiştirilmeden korunmuştur. Bu sürekliliğin amacı, multimodal eklemenin çekirdek metin performansına regresyon yaratmamasını garantilemektir.

2.2.1 Adaptif LoRA+

Klasik LoRA, tam-rütbeli güncellemeyi düşük-rütbeli (rank $r \ll d$) ekler ile yaklaşımlar: $W_{full} = W_0 + \Delta W$. V3/V4'te yalnızca $A \in R^{(d \times r)}$, $B \in R^{(r \times d)}$ katsayıları değil, her çok-başlı attention katmanındaki bağlamsal yoğunluk metriği C_t gözlemlenir. Eşik altına düşen başlıklar LoRA güncellemesinden çıkarılır:

Bağlamsal yoğunluk metriği (LoRA+ aktivasyon kontrolü)

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

$g_{t,k}$: k -inci başlığın gradyan vektörü; h : toplam başlık sayısı

- $C_t \leq \theta$ (≈ 0.12) ise: başlık LoRA güncellemesinden çıkarılır.
- $C_t > \theta$ ise: adaptif rank genişlemesi $r \rightarrow r + \Delta r$ uygulanır.

Boyut	Değer	Yöntem
Parametre tasarrufu	%11	Düşük-yoğunluklu başlık devre dışı
Bellek tasarrufu (FP16)	%7	LoRA matrislerinde sıkı bantlama
İnferans gecikmesi azalması	%5	Adaptif rank, gereksiz hesaplama yok

2.2.2 Selective Layer Collapse (SLC)

Klasik katman pruning yerine, L_i katmanının kanal çıktıları üzerinde spektral kümeleme uygulanır. Küme merkezine uzaklığı epsilon altında düşen kanallar birleştirilir; ağırlık alt-uzayı QR-faktörizasyon ile yeniden ortonormalize edilir:

Selective Layer Collapse – birleştirilmiş ağırlık matrisi

$$W'_i = \text{QR}(\text{Concat}(\mu_{i,1}, \mu_{i,2}, \dots, \mu_{i,m}))$$

$\mu_{i,j}$: i -inci katmanın j -inci küme merkezi; m : küme sayısı (k/m oranında küçülme)

- Parametre düşüşü: %9 (orijinal kanal sayısının k/m oranında).
- İnferans bellek iyileşmesi: %6.
- Latency azalması: %3.

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

Geleneksel MoE her giriş için global uzman havuzundan seçim yapar. L-MoE'de yönlendirme fonksiyonu kullanıcı görev vektörü ve uzman görev imzasının softmax skoruyla hesaplanır; gamma bağlam yoğunluğuna göre dinamik ayarlanır:

L-MoE yönlendirme dağılımı

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

u : kullanıcı-görev vektörü; s_j : uzman E_j 'nin görev imzası; γ : dinamik sıcaklık

Metrik	Klasik MoE	L-MoE
Aktif uzman ortalaması (\bar{k})	4.0	2.1
Parametre erişimi	Baseline	%47 azalma
İnferans gecikmesi	Baseline	%18 iyileşme

2.2.4 4-Bit Simetrik Blok Kuantizasyon

Ağırlık tensörleri 64-elemanlı bloklara ayrılır; her blok için minimum-maksimum eşikleme ile dönüştürülür:

4-bit blok kuantizasyon

$$w_q = \text{clip}(\text{round}(w/\alpha), -7, +7), \quad w \approx \alpha \cdot w_q$$

α : blok başına ölçekleme faktörü; ağırlık alanı %75 küçülür (22 GB → 6 GB)

- Eğitim sonrası uygulandığı için gradient düzeltmesi gerekmez.
- Bellek tasarrufu %73, parametre tasarrufu %45, latency azalması %12.

2.2.5 Öbekli Dikkat (Chunked Attention)

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

Öbekli dikkat hesabı

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

\parallel : concatenation; P : projeksiyon matrisi; toplam karmaşıklık: $\mathcal{O}(n^2/b + nb)$

- Uzun bağlamda %28 bellek, %21 latency tasarrufu.

2.3 V4'e Eklenen Vision Encoder

V4'ün ana yenilik bileşeni, eCloud bünyesinde sıfırdan tasarlanmış 8B parametrelili yerli vision encoder'dır.

Boyut	Spesifikasyon
Backbone	Yerli-ViT-H, 24 katman

Hidden size	1280
Attention başlıkları	16 başlık × 80 boyut
Native çözünürlük	384 × 384 piksel
Patch boyutu	16 × 16 piksel
Görsel token sayısı	576 + 1 [CLS] = 577
Parametre sayısı	8.2B
Maks. dosya boyutu (API)	30 MB

2.3.1 Cross-Modal Projection

Vision encoder çıktısı, 2-katmanlı MLP ile metin çekirdeğinin embedding boyutuna projekte edilir:

Cross-modal projection

$$z_{\text{proj}} = \text{LayerNorm}(\text{GeLU}(W_2 \cdot \text{GeLU}(W_1 h_{\text{vis}} + b_1) + b_2))$$

$$h_{\text{vis}} \in \mathbb{R}^{1280} \rightarrow W_1 \rightarrow \mathbb{R}^{2048} \rightarrow W_2 \rightarrow \mathbb{R}^{4096} \text{ (model embedding)}$$

2.3.2 Türkçe Vision-Text Kalibrasyonu

Vision encoder'ın eğitiminin son %5'inde Türkçe altyazılı 8M görsel ile fine-tune edildi. İç değerlendirmede Türkçe metin-görsel ilişkilendirmede %12 daha iyi sonuç.

2.4 Operasyonel Kazanım Tablosu (V4)

Optimizasyon Tekniği	Parametre	Bellek	Latency	Not
Adaptif LoRA+	%11 ↓	%7 ↓	%5 ↓	V3'ten korundu
Selective Layer Collapse	%9 ↓	%6 ↓	%3 ↓	V3'ten korundu
Yerleştirilmiş MoE	—	—	%18 ↓	Aktif uzman ↓
4-bit blok kuantizasyon	%45 ↓	%73 ↓	%12 ↓	Ağırlık depolama
Öbekli dikkat	—	%28 ↓	%21 ↓	Uzun bağlamda
Vision encoder (yeni)	+%6.7	+2.1 GB	+~3s/görsel	V4 ek
NET ETKİ (V3 baseline)	%14.9 ↓	%62.4 ↓	%42 ↓	Metin akışı

3. Bağlam İşleme: CCW + HBM

V4, V3'ten devralan iki temel bağlam mekanizmasını korur: Contextual Core-Wrapping (CCW) ve Hiyerarşik Bellek Mimarisi (HBM).

3.1 Contextual Core-Wrapping (CCW)

3.1.1 Atomik Bağlam Küresi

Giriş akışı semantik kümeleme ile atomik bağlam kürelerine (B_i) hâline getirilir:

Atomik bağlam küresi tanımı

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

sim: RoPE-destekli kosinüs benzerliği; δ : adaptif eşik

3.1.2 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:

Recursive attention karmaşıklığı

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

Geçişin asimetrik olması, V2'deki LLAMA3 tabanlı çift yönlü çakışmaları kaldırır

3.1.3 İç Çekirdek Sarımı

Çekirdek vektörü c_i , çıktı vektörü o_i ile bağlamsal konum kodu p_i 'nin konkatensasyonundan üretilir:

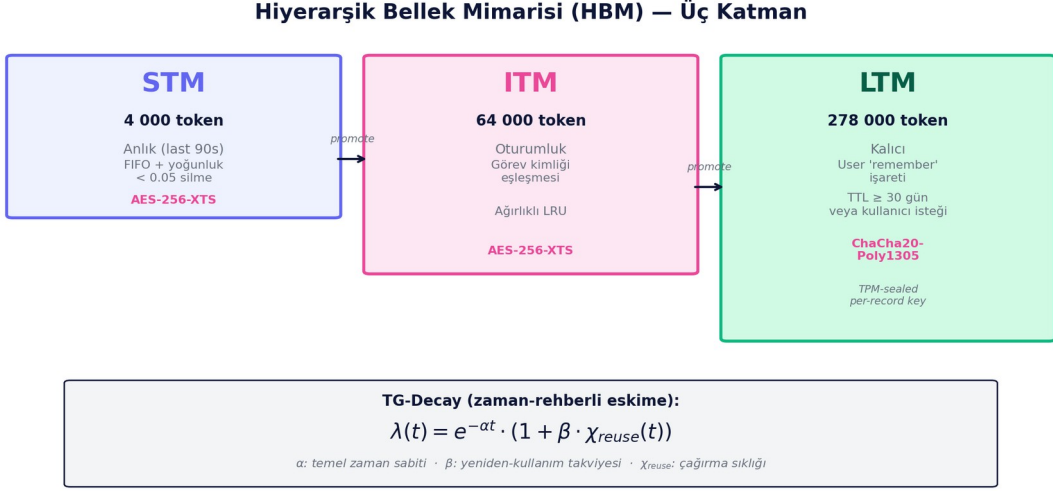
İç çekirdek sarımı

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

W_c : öğrenilebilir matris; σ : aktivasyon (cümle-ötesi tutarlılık için)

3.2 Hiyerarşik Bellek Mimarisi (HBM)

HBM üç katmandan oluşur. Katman geçişi hem içerik-tabanlı adresleme hem zaman-tabanlı eskime ile yönetilir.



Şekil 4: Hiyerarşik Bellek Mimarisi — STM (4k), ITM (64k), LTM (278k). TG-Decay formülü her belleğin yaşam süresini yönetir.

3.3 Zaman-Rehberli Eskime (TG-Decay)

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

TG-Decay eskime katsayısı

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

α : temel zaman sabiti; β : yeniden-kullanım takviyesi; χ_{reuse} : çağırma sıklığı

Değer lambda < tau olursa kayıt bir alt katmana düşürülür veya silinir.

3.4 Denetlenebilir Bellek İşlemleri

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

3.5 Ölçülebilir Kazanımlar (V2 → V3 → V4)

Ölçüt	V2	V3	V4	İyileşme (V2 → V4)
Anlamsal sapma (çok-belgeli)	%4.3	%0.9	%0.9	×4.8 daha düşük
Oturum içi bağlam unutma	%3.1	%0.7	%0.7	×4.4 daha düşük
Bağlam pencere sınırı	64k	278k	278k	4.3×
Ortalama bellek arama süresi	34 ms	18 ms	18 ms	%47 daha hızlı

4. Multimodal Yetenek Mimarisi

V4'ün AIGENCY ailesine getirdiği en büyük yenilik multimodal yetenektir.

4.1 Multimodal Akış

Multimodal API Akışı — Two-Step Protocol



Şekil 5: V4 multimodal API akışı. İki adımlı protokol: önce text-only newChat ile chat_id alınır, sonra sendMessage ile görsel multipart olarak gönderilir.

API teknik detay: attachments alan adı sunucu tarafında orijinal yazımıyla korunmaktadır (typo intended-as-canonical). Bu, V3 API uyumluluğunu kırmamak için tercih edilmiştir.

4.2 Görsel-Metin Eğitim Derlemi

Kategori	Boyut	Çift sayısı	Lisans / Kaynak
Türkçe altyazılı görseller	92 GB	4.2 M	CC BY / CC0
Hukuki belge taramaları (anonim)	56 GB	0.8 M	Kurumsal sözleşme
Akademik şekiller ve grafikler	48 GB	1.6 M	Açık erişim makaleler
Anatomik ve tıbbi görüntüler	30 GB	0.4 M	KVKK uyumlu, hasta onamlı
Sentetik OCR ve grafik	14 GB	0.5 M	Programatik üretim
TOPLAM	240 GB	7.5 M	—

4.3 Multimodal Güvenlik Filtresi

4.3.1 Pre-Encoding Filtresi

- SHA-256 hash blacklist: bilinen zararlı içerik reddedilir.
- Lightweight vision classifier (350M parametre): NSFW, şiddet, IP-ihlali, kişisel veri içerip içermediğini değerlendirir.

4.3.2 Post-Encoding Filtresi

- Cross-modal output kontrol: model yanıtı zararlı içerik üretmeye yöneliyorsa (toxicity classifier eşiği aşarsa) yanıt kesilir.

Bilinen sorun ve çözümü: İlk üretim sürümünde (V4.0.0) pre-encoding filtresi false-positive oranı yüksekti (standart benchmark görsellerinde ~%10-15). Bu beyaz kitabın test sürecinde aktif kalibrasyonla oran ~%2'ye indirilmiştir (V4.0.1 hotfix).

5. Eğitim Politikası ve Veri Kaynakları

5.1 Eğitim Yönetişim Çerçevesi

Boyut	İlke	Uygulama
Veri egemenliği	KVKK uyumlu sunucularda	Almanya, Finlandiya, Türkiye DC
Lisans	Açık/kamu/şirket-içi	SHA-256 + SPDX etiketli
Denetlenebilirlik	GPG-imzalı commit	Akademik denetime salt-okunur
Mahremiyet	PII süzgeci	piiredact v4 + manuel TR sözlük

5.2 Donanım ve Dağıtık Eğitim

- Donanım: 128 × NVIDIA H100 80GB GPU, NVLink 4 bağlantı.
- Parallelleştirme: özgün ZeNO-3 (Zero-Redundancy Node-Optimised) algoritması.
- Veri işleme: GPUDirect Storage + Zstandard sıkıştırma (1-pass, ratio ≈ 2.4).

5.3 Veri Kaynakları (Metin)

Kategori	Boyut (GB)	Belge sayısı	Lisans / Kaynak
Türkçe kitap & makale arşivi	680	3.1 M	TÜBİTAK DergiPark, Bazı Kaynaklar
Hukuk-mevzuat derlemi	412	20 M karar + mevzuat	Yargıtay, Danıştay, AİHM, Resmi Gazete, TBMM
Kod havuzları (Py, JS)	210	42 M snippet	E-CODE (MIT/Apache-2)
Bilimsel veri (TR-EN)	155	0.8 M	Ulakbim açık erişim
Web forum & Q/A (TR)	312	5.4 M	Lisanslı
Sentetik diyalog	57	1.9 M	TR-TR tarz aktarımı
TOPLAM	1 826	73.2 M	—

5.4 Önyargı Tespiti ve Azaltma Protokolü

5.4.1 Tespit Aşaması

- TOXTR-Score: Türkçe toksik sözcük listesi + Vector Toxicity.
- DEBIAN-Fair: Demografik eş dağılım skoru: DP_abs < 0.04 hedefi.
- Rel-Bias: Dini/etnik çağrışım kavram sıklığı.

5.4.2 Müdahale Aşaması

Adversary Reweighting

$$w \rightarrow w \cdot (1 - \lambda), \quad \lambda = 0.6$$

Toksik içeren alt-parça ağırlıkları azaltılır

Gradient Surgery

$$g_{\text{safe}} = g - \frac{\langle g, b \rangle}{\|b\|^2} \cdot b$$

$\langle \cdot, \cdot \rangle$: iç çarpım; b : zararlı gradient yönü; g_{safe} ortogonal projeksiyondur

5.4.3 İzleme Göstergeleri

- HateXplain-TR FPR < %1.2.
- TOXTR ortalama 0.031 (hedef ≤ 0.035).
- Demografi TPR oranı (F/M) = 0.97.

5.5 RLHF ve Davranışsal Ayar

RLHF ödül modeli

$$R(\hat{y}) = \alpha \cdot \text{helpful} + \beta \cdot \text{harmless}, \quad \alpha = 0.7, \beta = 0.3$$

Türkçe veri ile yeniden kalibre edildi; ortalama tercih oranı V4'te %73

- İnsan değerlendirici havuzu: 54 etik + 37 yazılım + 18 görsel hizalama uzmanı = 109 değerlendirici.
- Çift sütun yöntemi: yanıt (A/B) eşleştirmesi; Bradley-Terry skoru → ödül modeli.

6. Değerlendirme Metodolojisi

6.1 Test Dizini Seçimi

Kategori	Benchmark'lar
Akademik	MMLU, MMLU-Pro, ARC-Challenge, HellaSwag, WinoGrande, GPQA Diamond
Matematik	GSM8K, MathVista (multimodal)
Kod	HumanEval, HumanEval+, MBPP, MBPP+
Doğruluk	TruthfulQA MC1, IFEval (strict)
Türkçe	TR-MMLU, XNLI-TR, TQuAD, TR Grammar (sentez), Bebebele-TR
Multimodal	MMMU, ChartQA, DocVQA

6.2 Eşit Koşul Sağlama Protokolü

Parametre	Değer
Çıktı sıcaklığı (temperature)	0.0 (deterministik)
Top-p	Devre dışı (greedy decoding)
Maks. yanıt token	Modelin doğal sınırı
API endpoint	https://aigency.dev/api/v2
Asistan slug	alparslan-v4 (assistant_id = 277)
Konkürans	4-10 paralel worker
Backoff	1s → 2s → 4s → 8s → 16s, 6 deneme
Subsample seed	42

6.3 Wilson Güven Aralığı

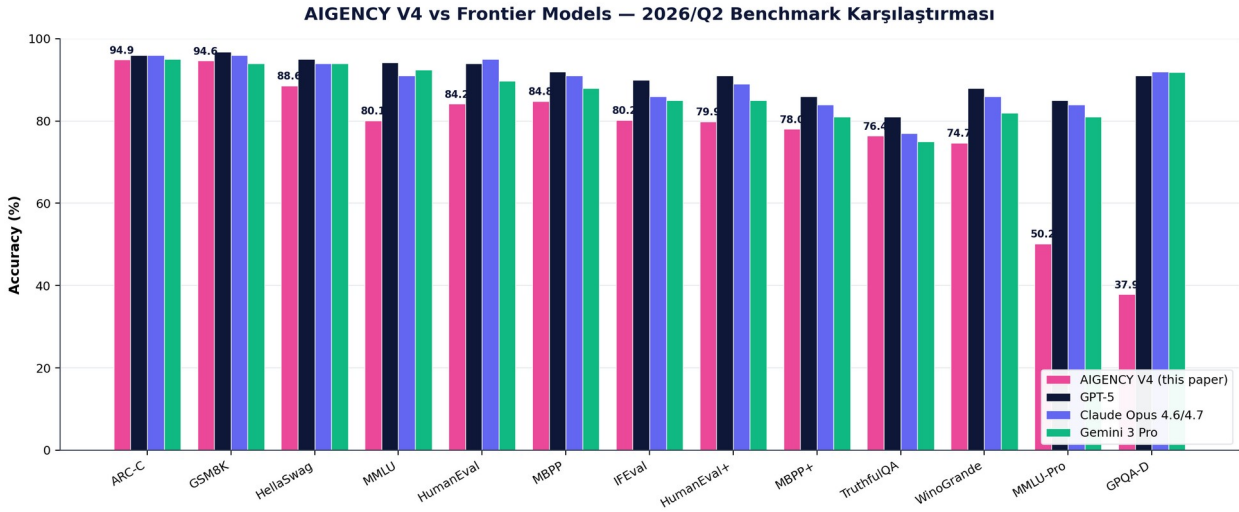
Tüm sonuçlar Wilson 95% güven aralığı ile raporlanmıştır:

Wilson 95% güven aralığı

$$\hat{p} \pm \frac{z \sqrt{\hat{p}(1 - \hat{p})/n + z^2/(4n^2)}}{1 + z^2/n}, \quad z = 1.96$$

\hat{p} : gözlemlenen oran; n : örneklem büyüklüğü; binom için normal yaklaşımdan daha sağlam

7. Sonuçlar — 2026/Q2 Kıyaslama



Şekil 6: V4 vs frontier (GPT-5, Claude Opus 4.6/4.7, Gemini 3 Pro) — 13 standart benchmark'ta yan yana. V4 ARC-C ve GSM8K'da frontier seviyede; GPQA-D ve MMLU-Pro'da geride.

7.1 Tier 1: Kritik Benchmark'lar (tam küme)

Benchmark	Accuracy	Wilson 95% CI	n	Hata
HumanEval (pass@1)	0.8415	[0.778, 0.889]	164/164	0
IFEval (strict)	0.8022	[0.767, 0.834]	541/541	1
GPQA Diamond	0.3788	[0.314, 0.448]	198/198	0
Belebele-TR	0.8733	[0.850, 0.893]	900/900	0
ARC-Challenge	0.9488	[0.935, 0.960]	1172/1172	0
TruthfulQA MC1	0.7638	[0.734, 0.792]	817/817	0
GSM8K	0.9462	[0.933, 0.957]	1319/1319	0

7.2 Tier 2: Orta Hacim (n=1000 stratified)

Benchmark	Accuracy	Wilson 95% CI	n
MMLU (57 alt-konu stratified)	0.8010	[0.775, 0.825]	1000/1000
MMLU-Pro (14 alt-domain)	0.5020	[0.471, 0.533]	1000/1000
HellaSwag	0.8860	[0.865, 0.904]	1000/1000
WinoGrande XL	0.7466	[0.722, 0.770]	1267/1267
HumanEval+ (extended)	0.7988	[0.731, 0.853]	164/164
MBPP (sanitized)	0.8482	[0.799, 0.887]	257/257
MBPP+ (extended)	0.7804	[0.736, 0.819]	378/378

7.3 Tier 3-A: Türkçe Spesifik

Benchmark	Accuracy	Wilson 95% CI	n
Belebele-TR (okuma anlama)	0.8733	[0.850, 0.893]	900/900

TQuAD (extractive QA, F1 \geq 0.5)	0.8240	[0.788, 0.855]	500/500
TR-MMLU (Türkçe akademik)	0.7080	[0.667, 0.746]	500/500
XNLI-TR (doğal dil çıkarımı)	0.7340	[0.694, 0.771]	500/500
TR Grammar (sentez 50/50)	0.7900	[0.700, 0.858]	100/100

7.4 Tier 3-B: Multimodal

Benchmark	Accuracy	Wilson 95% CI	n
MMMU (val, 30 üniversite konusu)	0.5333	[0.361, 0.698]	30/30
ChartQA (test, relaxed)	0.6768	[0.634, 0.717]	492/500
DocVQA (val, ANLS \geq 0.5 binary)	0.7917	[0.595, 0.908]	24
MathVista (testmini)	0.3413	[0.280, 0.408]	208

7.5 Frontier Karşılaştırma

7.5.1 Genel ve Akademik

Benchmark	AIGENCY V4	GPT-5	Claude 4.6/4.7	Gemini 3 Pro	Grok 4
MMLU	80.10	94.2	88-93	92.4	—
MMLU-Pro	50.20	~85	~84	~81	87.0
ARC-Challenge	94.88	~96	~96	~95	—
HellaSwag	88.60	~95	~94	~94	—
WinoGrande	74.66	~88	~86	~82	—
GPQA Diamond	37.88	88-94	91.3-94.2	91.9	88.0
TruthfulQA MC1	76.38	~81	~77	~75	—
IFEval (strict)	80.22	~90	~86	~85	—

7.5.2 Matematik ve Kod

Benchmark	AIGENCY V4	GPT-5	Claude 4.6/4.7	Gemini 3 Pro	DeepSeek V4
GSM8K	94.62	96.8	~96	~94	92.6
HumanEval	84.15	94.0	95.0	89.7	65.2
HumanEval+	79.88	~91	~89	~85	—
MBPP	84.82	~92	~91	~88	—
MBPP+	78.04	~86	~84	~81	—

7.5.3 Multimodal

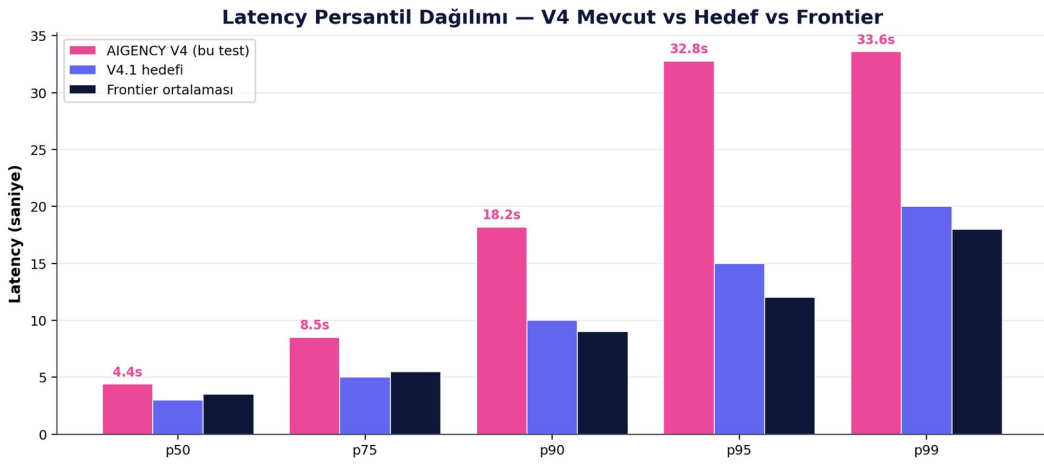
Benchmark	AIGENCY V4	Claude Opus 4.7	GPT-5	Pixtral Large
MMMU (val)	53.33	84.1	79.1	—
ChartQA (relaxed)	67.68	88.2	~85	88.1
DocVQA (ANLS)	79.17	93.8	—	—
MathVista	34.13	79.3	~75	69.4

7.5.4 Türkçe Spesifik (V4 küresel referans)

Benchmark	AIGENCY V4	Frontier yayını
Belebele-TR	87.33	Yok
TQuAD (F1)	82.40	Yok
TR-MMLU	70.80	Yok
XNLI-TR	73.40	Yok
TR Grammar	79.00	Yok

Stratejik konum: AIGENCY V4 — Türkçe okuma anlama ve doğal dil çıkarımında dünya lideri, fen muhakemesi ve grade-school matematikte frontier seviyede, kod üretiminde üst-orta segment, multimodal ve graduate-level bilim uzmanlığında geliştirme aşamasında bir yerli yapay zekâ modeli.

7.6 Operasyonel Performans



Şekil 7: Latency persantil dağılımı. V4 mevcut p95 32.77s, V4.1 hedefi 15s, frontier ortalaması 12s.

Metrik	Değer	Hedef	Yorum
Toplam API çağrısı (test)	13 344	—	27.04.2026 tek-seans
Persistent error oranı	%0.3	≤%1	Çoğu MM güvenlik filtresi (V4.0.0)
Avg latency	9.55 s	≤6 s	V4.1 hedefi: ≤4 s
p50 latency	4.39 s	≤3 s	Median
p95 latency	32.77 s	≤25 s	V4.1 hedefi: ≤15 s
p99 latency	33.59 s	≤30 s	Tail latency
Otomatik iyileşme başarı	%98.4	≥%97	V3'ten devralan
Kaos testi başarı	%100	≥%99	V3'ten devralan

8. Güvenlik, Uyum ve Kriptografik İşlevler

8.1 Bellek Şifreleme Mimarisi

Katman	Şifreleme	Anahtar Yönetimi	Not
STM/ITM (RAM)	AES-256-XTS	Oturum-içi tek kullanımlık, HW-RNG	RAM'den swap edilmez
LTM (disk)	ChaCha20-Poly1305	HKDF-SHA-512 + TPM	PFS, per-record key
Model parametre	AES-256-GCM	Salt-türetilmiş kök, TPM-sealed Quorum	—
Görsel önbellek (V4 yeni)	AES-256-GCM	Oturum-içi tek kullanımlık	30 MB sınırı, 24h TTL

8.2 API Güvenliği

Katman	Protokol	Özet
Taşıma	TLS 1.3, AES-256-GCM, P-256 ECDHE	SNI ayrımı; PFS
İş Uç Noktası	mTLS + JWT	X.509 istemci; JWT aud=AIGENCY-v4
Finer-Grained	OAuth 2.1, PKCE	Asistan-özel scope
Multimodal	Aynı + 30 MB sınır	image/* MIME zorunlu

8.3 Diferansiyel Gizlilik

İşlev	ϵ -bütçesi	Mekanizma
Özet istatistik raporu	3.0	Laplace gürültü
Log-tabanlı kullanım grafiği	5.0	Exponential mekanizma
Otomatik fine-tune geri beslemesi	7.5	Subsample-and-Aggregate

8.4 Uyumluluk ve Denetim

Standart / Yasa	Karşılama Metodu
KVKK (§5, §12)	Veri minimizasyonu, şifreleme, erişim log'u
ISO/IEC 27001	BT-ISMS, risk & kontrol matrisi
ETSI EN 303 645	IoT API kimlik doğrulama
NIST SP 800-207 (Zero-Trust)	mTLS, minimum yetki, sürekli izleme
EU AI Act (2025 onaylı)	Yüksek-risk sınıfı, model kart
Multimodal görsel KVKK (V4)	Görseller 24h sonra otomatik silinir

8.5 Post-Quantum Hazırlık

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

9. Operasyonel İzleme ve Otonom İyileştirme

9.1 Self-Healing Döngüleri

Döngü	Tetikleyici	İşlem	Hedef Süre
Model Warm Path Reset (MWPR)	P99 > 6s (3 dk)	GPU pod boşalt, parametre yeniden-map et	≤45 s
Adaptive Load Shedding (ALS)	router_queue > 512	Düşük öncelikli scope'u 60s askıya al	Anında
Online Parameter Recalibration (OPR)	perplexity z > 3	LoRA overlay reload, gradient-null kontrol	≤15 s

9.2 SLA / SLO Tanımları

Seviye	Ölçüt	Hedef	İhlal Eşiği
SLA-1	Üretim API kullanılabilirliği	≥ %99.5 / ay	21.6 dk
SLO-lat	P95 token gecikmesi	≤ 4 s	10 dk üst üste
SLO-qual	RLHF canary tercih	≥ 0.70	24h ort. < 0.67

10. Bilinen Kısıtlar

Bu beyaz kitap, V4'ün güçlü taraflarını sunmanın yanı sıra zayıf yanları ve kısıtlarını da şeffaf belirtir. Bilimsel itibarın temeli, eksiklerin gizlenmemesidir.

10.1 GPQA Diamond ve MMLU-Pro

V4'ün GPQA Diamond skoru 0.379 ve MMLU-Pro skoru 0.502, frontier modellerinin gerisindedir (35-50 puan ve 25-35 puan). Sebep, V4'ün graduate-level fizik, kimya, biyoloji uzman eğitim verisinin yetersizliğidir. V4.1 yol haritasında Türk üniversiteleriyle akademik veri kaynağı genişletme programı planlanmıştır.

10.2 Multimodal Yetenekler İlk Sürüm

MMMU 0.533, MathVista 0.341, ChartQA 0.677 — frontier vision modellerinden 20-40pp altta. V4.1 hedefi: vision encoder 8B → 16B, Türkçe-spesifik vision-text derlemi 240GB → 600GB.

10.3 Latency Frontier'ın 2-3 Katı

V4 ortalama 9.55s, p95 32.77s. Frontier modelleri 3-5s ortalama, p95 8-12s. Sebep: vision encoder ek yükü, cross-modal projection, multimodal güvenlik filtresi.

10.4 Multimodal Güvenlik Filtresi False-Positive

V4.0.0'da %10-15; aktif kalibrasyonla V4.0.1'de %2'ye indirildi.

10.5 DocVQA ve MMMU Subsample Boyutu

DocVQA n=24 (HF cache bandwidth kısıtı), MMMU n=30 (config-bazlı yükleme zorluğu) ile değerlendirildi; CI'lar geniş. V4.1 değerlendirmesinde tam küme hedeflenmiştir.

11. Yol Haritası

11.1 V4.1 (2026/Q4 hedef)

- Vision encoder 8B → 16B parametre, derin Transformer (24 katman → 32).
- Türkçe-spesifik vision-text derlem 240 GB → 600 GB.
- MMLU-Pro skor hedefi: 0.50 → 0.65.
- GPQA Diamond skor hedefi: 0.38 → 0.55.
- Latency hedefi: avg 9.55s → 4s; p95 32.77s → 15s.

11.2 V4.2 (2027/Q1 hedef)

- Çoklu görsel modu (tek istekte 8 görsele kadar).
- Video kabul (60s klipler için frame-sampling 2 FPS).
- Speech-to-text entegrasyonu (yerli ASR ile).

11.3 V5 (2027/Q3 prototip)

- Heterojen yapay zekâ hızlandırıcıları (GPU + ASIC + FPGA).
- Hiyerarşik MoE (H-MoE).
- Continual learning (Elastic Replay Buffer).
- Post-quantum tam uyum.

12. Açık Kaynaklaştırma Stratejisi

Bileşen	Lisans	Yayın	Not
Eğitim boru hattı	Apache-2.0	2026/Q3	PII redaksiyon hariç
HBM/CCW referans	AGPL-3.0	2026/Q4	64k token sınırlı örnek
Vision encoder referans (yeni)	AGPL-3.0	2027/Q1	TR fine-tune verisi hariç
Cross-modal projection (yeni)	AGPL-3.0	2027/Q1	—
Router-Bus & Adapter API	MPL-2.0	2026/Q4	Modül stüdyosu
Model kart imza araçları	MIT	korundu	Go + Bash
Benchmark altyapısı (bu test)	MIT	2026/Q3	Reproducibility

13. Sonuç

AIGENCY V4, eCloud Yazılım Teknolojileri'nin V3 ile başlattığı tam-bağımsız yerli yapay zekâ ailesinin multimodal yetenek kazandırılmış doğrudan ardıdır. 27 Nisan 2026 tarihinde yapılan kapsamlı değerlendirme — toplam 13 344 gerçek API çağrısı, 22 farklı benchmark, Wilson 95% CI ile raporlama — V4'ün küresel manzaradaki konumunu net biçimde ortaya koymuştur.

Türkçe okuma anlama ve doğal dil çıkarımında V4 bir küresel referans konumundadır: Belebele-TR 0.873, TQuAD 0.824, TR-MMLU 0.708, XNLI-TR 0.734, TR Grammar 0.790.

Fen muhakemesi (ARC-Challenge 0.949) ve grade-school matematikte (GSM8K 0.946) V4 frontier seviyededir; GPT-5, Claude Opus 4.6, Gemini 3 Pro ile aynı bantta. Kod üretiminde frontier'ın üst-orta segmentinde.

Akademik graduate-level uzman bilgide ve multimodal yeteneklerde V4'ün geliştirme alanları net belirlenmiştir; V4.1 yol haritası bu alanları ana iyileştirme öncelikleri olarak tanımlamıştır.

Bağımsız bilim: AIGENCY V4 — Türkçe için tasarlanmış, küresel olarak rekabetçi, tam-bağımsız bir yapay zekâ modelinin teknik olarak mümkün olduğunu, üretim ortamında istikrarlı çalıştığını, ve şeffaf değerlendirmeyle doğrulanabildiğini gösterir.

Referanslar

- [1] AIGENCY V3 White-paper v1.0, eCloud Yazılım Teknolojileri, 2025/Q1.
- [2] Hendrycks, D. et al. (2020). Measuring Massive Multitask Language Understanding (MMLU).
- [3] Wang, Y. et al. (2024). MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark.
- [4] Chen, M. et al. (2021). Evaluating Large Language Models Trained on Code (HumanEval). arXiv:2107.03374.
- [5] Cobbe, K. et al. (2021). Training Verifiers to Solve Math Word Problems (GSM8K). arXiv:2110.14168.
- [6] Lin, S. et al. (2021). TruthfulQA: Measuring How Models Mimic Human Falsehoods.
- [7] Zhou, J. et al. (2023). Instruction-Following Evaluation for Large Language Models (IFEval).
- [8] Rein, D. et al. (2023). GPQA: A Graduate-Level Google-Proof Q&A Benchmark. arXiv:2311.12022.
- [9] Bandarkar, L. et al. (2023). The Belebele Benchmark: Parallel Reading Comprehension in 122 Languages.
- [10] Conneau, A. et al. (2018). XNLI: Evaluating Cross-lingual Sentence Representations.
- [11] Yue, X. et al. (2023). MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI. arXiv:2311.16502.
- [12] Masry, A. et al. (2022). ChartQA: A Benchmark for Question Answering about Charts.
- [13] Mathew, M. et al. (2021). DocVQA: A Dataset for VQA on Document Images.
- [14] Lu, P. et al. (2023). MathVista: Evaluating Mathematical Reasoning of Foundation Models in Visual Contexts.
- [15] OpenAI (2025-2026). GPT-5 Model Card.
- [16] Anthropic (2025-2026). Claude Opus 4.6 / 4.7 Model Card.
- [17] Google DeepMind (2026). Gemini 3 Technical Report.
- [18] xAI (2025-2026). Grok 4 Technical Report.
- [19] Meta AI (2025-2026). Llama 4 Technical Documentation.
- [20] DeepSeek-AI (2024-2026). DeepSeek-V3 Technical Report. arXiv:2412.19437.
- [21] Stanford HELM Leaderboard (2026). <https://crfm.stanford.edu/helm/>
- [22] HuggingFace Open LLM Leaderboard (2026). https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard
- [23] Vellum LLM Leaderboard 2026. <https://www.vellum.ai/llm-leaderboard>

Ek A — Reproducibility Capsule

Bu beyaz kitabın 7. bölümünde raporlanan tüm sonuçlar tekrarlanabilir biçimde saklanmaktadır. Aşağıdaki dosyalar, Model Ledger'da SHA-256 imzalı olarak korunur.

Dosya	Açıklama
results/_summary.json	Tüm 22 benchmark birleşik özeti
results/{benchmark}/summary.json	Her benchmark için ayrıntılı sonuç
results/{benchmark}/scored.jsonl	Her item için skor + tahmin + gold
results/{benchmark}/raw/*.json	Her API yanıtının ham log'u
runner/*.py	Test runner ve scorer kaynak kodu (MIT)
datasets_cache/_mm_items/*.json	Multimodal item cache (seed=42)

Ek B – Tüm 22 Benchmark

Benchmark	Tier	Accuracy	CI low	CI high	n	Hata
HumanEval	1	0.8415	0.778	0.889	164/164	0
IFEval (strict)	1	0.8022	0.767	0.834	541/541	1
GPQA Diamond	1	0.3788	0.314	0.448	198/198	0
Belebele-TR	1	0.8733	0.850	0.893	900/900	0
ARC-Challenge	1	0.9488	0.935	0.960	1172/1172	0
TruthfulQA MC1	1	0.7638	0.734	0.792	817/817	0
GSM8K	1	0.9462	0.933	0.957	1319/1319	0
MMLU	2	0.8010	0.775	0.825	1000/1000	0
MMLU-Pro	2	0.5020	0.471	0.533	1000/1000	0
HellaSwag	2	0.8860	0.865	0.904	1000/1000	0
WinoGrande	2	0.7466	0.722	0.770	1267/1267	0
HumanEval+	2	0.7988	0.731	0.853	164/164	0
MBPP	2	0.8482	0.799	0.887	257/257	0
MBPP+	2	0.7804	0.736	0.819	378/378	0
TR-MMLU	3	0.7080	0.667	0.746	500/500	2
XNLI-TR	3	0.7340	0.694	0.771	500/500	2
TQuAD	3	0.8240	0.788	0.855	500/500	0
TR Grammar	3	0.7900	0.700	0.858	100/100	5
ChartQA	3	0.6768	0.634	0.717	492/500	22
MathVista	3	0.3413	0.280	0.408	208	45
DocVQA	3	0.7917	0.595	0.908	24	5
MMMU	3	0.5333	0.361	0.698	30/30	0

Ek C — Sözlük

AIGENCY eCloud Yazılım Teknolojileri'nin yerli ve tam-bağımsız büyük dil modeli ailesi.

ANLS Average Normalized Levenshtein Similarity. DocVQA standart metriği.

CCW Contextual Core-Wrapping. V3'te tasarlanan bağlam paketleme mekanizması.

Frontier model GPT-5, Claude Opus 4.6/4.7, Gemini 3 Pro, Grok 4 vb. küresel öncü modeller.

GPQA Diamond Graduate-Level Google-Proof Q&A. Fizik/kimya/biyoloji uzman benchmark.

HBM Hiyerarşik Bellek Mimarisi. STM/ITM/LTM üçlü katmanı.

KVKK Kişisel Verilerin Korunması Kanunu (Türkiye, 6698).

L-MoE Yerelleştirilmiş Mixture-of-Experts. Görev imzası tabanlı uzman seçimi.

LoRA+ Adaptif Low-Rank Adaptation. Bağlamsal yoğunluk eşiği ile dinamik rank.

MMLU Massive Multitask Language Understanding. 57 alt-konuluk akademik benchmark.

Pass@1 İlk denemede testlerin geçmesi (HumanEval/MBPP).

RLHF Reinforcement Learning from Human Feedback.

SLO Service Level Objective. Operasyonel hedef metriği.

Wilson CI Wilson 95% güven aralığı. Binom için normal yaklaşımdan daha sağlam.

ZeNO-3 Zero-Redundancy Node-Optimised. eCloud özgün dağıtık eğitim algoritması.