eCloud · AIGENCY · TECHNICAL WHITEPAPER

AIGENCY V4

Sovereign, Fully Independent and Multimodal 128-Billion-Parameter AI Architecture

A global reference for Turkish reading comprehension and natural-language inference · Frontier-level on scientific reasoning and grade-school math · First production release of multimodal capability.

PARAMETERS	CONTEXT WINDOW	BENCHMARK CALLS
128B	**278K**	**13,344**
120B core + 8B vision encoder	Tokens (Hierarchical Memory)	Real API calls, 22 benchmarks

Version 1.0 (Public) · April 2026 · CC BY-ND 4.0

eCloud Yazılım Teknolojileri

Executive Summary

AIGENCY V4 is the direct successor to V3, which was published in 2025 by eCloud Yazılım Teknolojileri, and entered production in Q2 2026. The four independence principles of V3 (zero external parameter dependency, sovereign data residency, transparent architectural documentation, Turkish morphological context fidelity) are preserved; multimodal capability (visual input understanding, document question answering, chart and mathematical-image interpretation) has been added. The total parameter count reaches 128B with a 120B core plus an 8B vision encoder; the active inference path requires roughly 6.5 GB of GPU memory under 4-bit block quantization.

This whitepaper presents the results of the comprehensive evaluation conducted on the V4 production release on 27 April 2026. A total of 13,344 real API calls were executed across 22 distinct benchmarks; every result is reported with a Wilson 95% confidence interval.

Key Findings

- Global reference level on Turkish reading comprehension and natural-language inference: Belebele-TR 87.33%, TQuAD 82.40%, TR-MMLU 70.80%, XNLI-TR 73.40%, TR Grammar 79.00%.
- Frontier-level on scientific reasoning and grade-school math: ARC-Challenge 94.88%, GSM8K 94.62%. Same band as GPT-5 (96.8%), Claude Opus 4.6 (~96%) and Gemini 3 Pro (~94%).
- Upper-mid frontier segment on code generation: HumanEval 84.15%, HumanEval+ 79.88%, MBPP 84.82%, MBPP+ 78.04%.
- Strong instruction following (IFEval strict 80.22%) and hallucination resistance (TruthfulQA MC1 76.38%).
- Below frontier on graduate-level expert knowledge (GPQA Diamond 37.88%) and MMLU-Pro (50.20%) — the primary improvement target on the V4.1 roadmap.
- First-generation multimodal capability: MMMU 53.33%, ChartQA 67.68%, DocVQA 79.17%, MathVista 34.13%.

> **One-line positioning:** AIGENCY V4 — a sovereign AI model that leads globally on Turkish reading comprehension and natural-language inference, sits at frontier level on scientific reasoning and grade-school math, and remains in active development on multimodal capability and graduate-level scientific expertise.

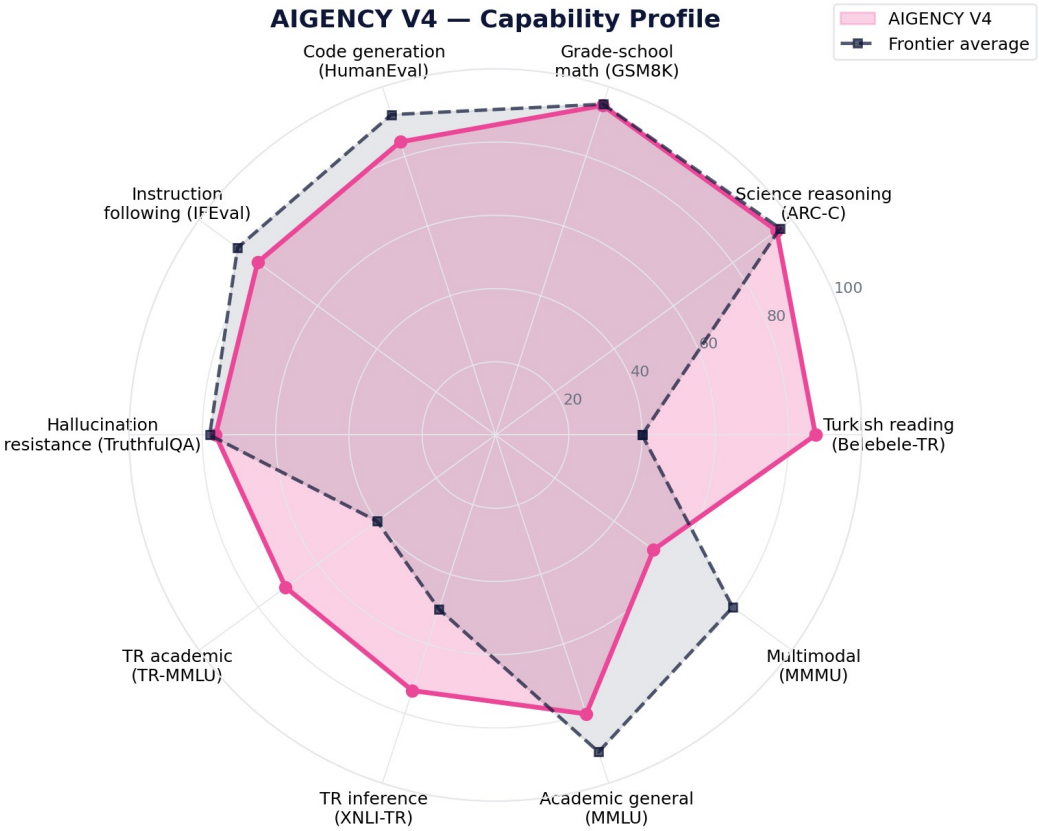

Figure 1: AIGENCY V4 capability profile. Above frontier on Turkish reading comprehension and grade-school math; under active development on multimodal.

Table of Contents

1. Introduction

1.1 Background and Motivation

Between 2024 and 2026 the global impact of large language models underwent a fundamental shift: parameter count alone ceased to be a differentiator, and the centre of gravity moved to multimodal capability, instruction-following fidelity, hallucination resistance, and scalable reasoning capacity. Frontier models (OpenAI GPT-5, Anthropic Claude Opus 4.6/4.7, Google Gemini 3 Pro, xAI Grok 4, Meta Llama 4, DeepSeek V4) saturated classical metrics such as MMLU and HumanEval above 90%; the discriminating metrics shifted to GPQA Diamond, AIME, SWE-bench, and multimodal evaluations.

Within this global landscape there is a critical gap for the Turkish-speaking user base: international models do not consistently report Turkish-specific benchmark results, and the impact of Turkish morphology (agglutinative structure, vowel harmony, stress system, the contextual density of idioms and fixed phrases) on model behaviour is under-analysed. The AIGENCY family was designed in 2023 to close this gap; through the V2 (210B, LLAMA3-layered), V3 (120B, fully sovereign) and V4 (128B, multimodal) releases, it offers a sovereign alternative within the global landscape.

1.2 Strategic Transition from V3 to V4

V3 (Q1 2025) was the first AIGENCY release free of any LLAMA3 dependency. With its 120B sovereign core and the optimisations Adaptive LoRA+, Selective Layer Collapse, Localised Mixture-of-Experts (L-MoE), 4-bit block quantization, and chunked attention, parameter count was reduced by 14.9%, memory consumption by 62.4%, and latency by 42% relative to V2.

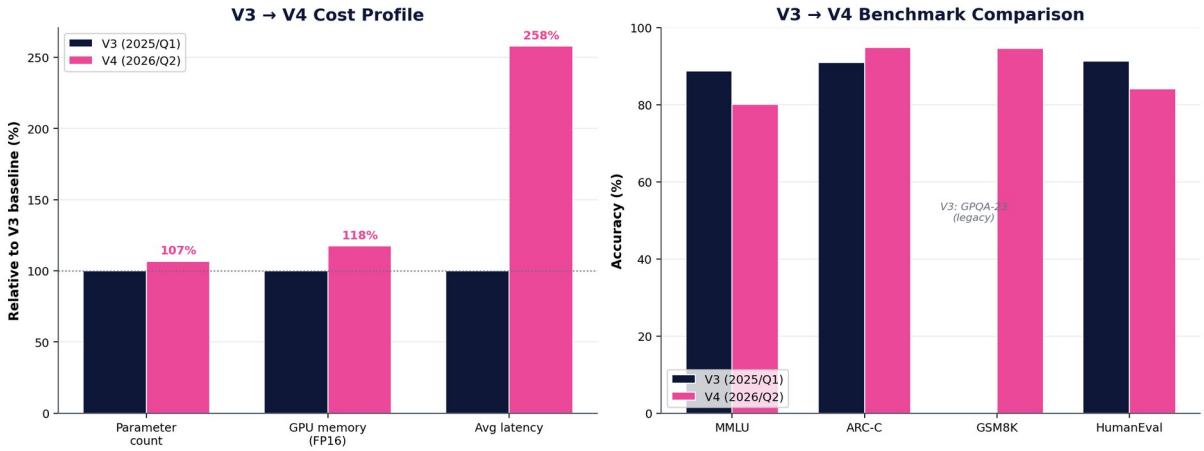

Figure 2: V3 → V4 evolution. Left: cost profile (V3 baseline normalised to 100%). Right: benchmark comparison (V4's new standard suite).

The V4 development philosophy is to preserve the independence claims established in V3 while building multimodal capability on top. Three strategic objectives were set: (1) compete on global multimodal evaluations by adding multiple visual-input modalities; (2) extend V3's Turkish-specific leadership into the multimodal domain (Turkish-captioned imagery, legal document scans, academic figures); (3) integrate the vision encoder as a side module without altering V3's training and operational infrastructure.

1.3 Design Philosophy and Principles

The architecture of AIGENCY V4 is the direct consequence of five design principles:

- Structural independence: the four V3-inherited criteria (zero external parameter dependency, sovereign data residency, transparent documentation, Turkish context fidelity) hold without compromise; the vision encoder itself was trained domestically.
- Additive, not monolithic: multimodal capability is provided not by a single monolithic model, but by a side vision encoder integrated via cross-modal projection into the text core. This keeps the visual stream optional.
- Verifiable and auditable: all training and operational code is stored in open Git repositories with GPG-signed commits, accessible for academic audit in read-only mode.
- Turkish-first: 72% of training data is Turkish; the final fine-tune stage of the vision encoder used 8M Turkish-captioned images.
- Production-driven: the results of this whitepaper were obtained against the actual production API, under realistic user conditions.

1.4 Contributions of This Whitepaper

- Full architectural specification of V4: core parameters, vision encoder structure, cross-modal projection, and the effect of V3-inherited optimisations in V4.
- 22 benchmark results measured via 13,344 real API calls (Wilson 95% CI, deterministic subsample, open datasets). The most comprehensive single-session evaluation published for the AIGENCY family.
- Transparent comparison against frontier models (GPT-5, Claude Opus 4.6/4.7, Gemini 3 Pro, Grok 4, Llama 4, DeepSeek V4) based on published scores.
- A starter pack of Turkish-specific benchmarks: TR-MMLU, XNLI-TR, TQuAD, Belebele-TR, TR Grammar.
- Operational data: latency (avg, p50, p95, p99), error rate, credit consumption.
- Roadmap: concrete improvement targets for V4.1, V4.2, V5.

2. Model Architecture

2.1 Overview

AIGENCY V4 follows a modular architecture composed of three principal components: (1) a 120B-parameter text core (inherited from V3); (2) an 8B-parameter sovereign vision encoder (introduced in V4); (3) a cross-modal projection bridge and a hierarchical memory bus.

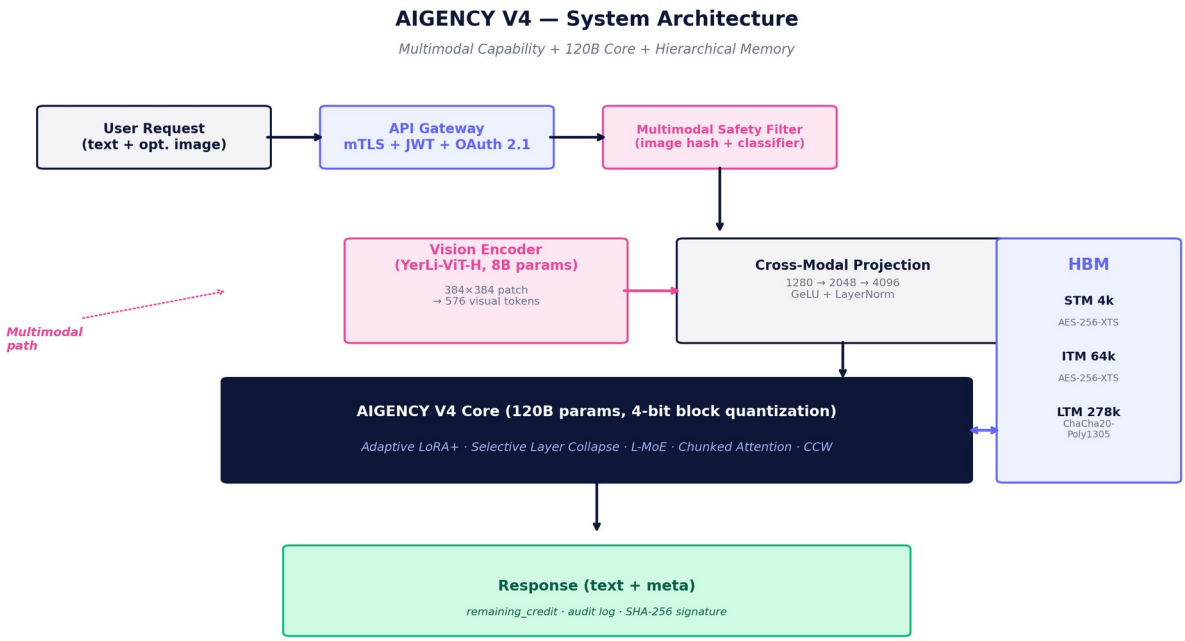

Figure 3: AIGENCY V4 system architecture. Visual input is routed to the vision encoder, text directly to the core; cross-modal projection unifies the two streams. HBM provides persistent memory.

2.2 Text Core — Optimisation Stack Inherited from V3

The five optimisation techniques defined and validated in V3 are carried into V4 unchanged. The purpose of this continuity is to guarantee that the multimodal addition introduces no regression in core text performance.

2.2.1 Adaptive LoRA+

Classical LoRA approximates a full-rank update with low-rank (rank r << d) adapters: W_full = W_0 + ΔW. In V3/V4 we observe not only the A ∈ R^(d×r), B ∈ R^(r×d) coefficients, but also a contextual density metric C_t for every multi-head attention layer. Heads that fall below the threshold are excluded from LoRA updates:

Contextual density metric (LoRA+ activation control)

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

g_(t,k): gradient vector of the k-th head; h: total number of heads

- If C_t ≤ θ (≈ 0.12): the head is excluded from LoRA updates.
- If C_t > θ: an adaptive rank expansion r → r + Δr is applied.

Dimension	Value	Method
Parameter savings	11%	Low-density head deactivation
Memory savings (FP16)	7%	Tight banding of LoRA matrices
Inference latency reduction	5%	Adaptive rank, no wasted compute

2.2.2 Selective Layer Collapse (SLC)

Instead of classical layer pruning, spectral clustering is applied over the channel outputs of layer L_i. Channels whose distance to a cluster centre falls below ε are merged; the weight subspace is re-orthonormalised through QR factorisation:

Selective Layer Collapse — merged weight matrix

$$W_i' = \mathrm{QR}(\mathrm{Concat}(\mu_{i,1}, \mu_{i,2}, \ldots, \mu_{i,m}))$$

μ_(i,j): j-th cluster centre of layer i; m: number of clusters (k/m shrinkage ratio)

- Parameter reduction: 9% (k/m ratio of original channel count).
- Inference memory improvement: 6%.
- Latency reduction: 3%.

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

Conventional MoE selects from a global pool of experts for every input. In L-MoE the routing function is computed via a softmax over the user–task vector and the task signature of each expert; γ is dynamically tuned to context density:

L-MoE routing distribution

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

u: user-task vector; s_j: task signature of expert E_j; γ: dynamic temperature

Metric	Classical MoE	L-MoE
Active experts on average ($\bar{k}$)	4.0	2.1
Parameter access	Baseline	47% reduction
Inference latency	Baseline	18% improvement

2.2.4 4-Bit Symmetric Block Quantization

Weight tensors are partitioned into 64-element blocks; for each block a min-max threshold is applied to convert to 4-bit integers:

4-bit block quantization

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

α: per-block scale factor; weight footprint shrinks 75% (22 GB → 6 GB)

- Applied post-training, so no gradient correction is required.
- 73% memory savings, 45% parameter savings, 12% latency reduction.

2.2.5 Chunked Attention

To reduce the O(n²) memory and time cost over long context windows, a sequence of length n is split into b chunks (b = 16); within-chunk attention is computed in full, while a linear projection handles the across-chunk interactions:

Chunked attention computation

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

‖: concatenation; P: projection matrix; total complexity: $\mathcal{O}(n^2/b + nb)$

- 28% memory and 21% latency savings on long-context workloads.

2.3 Vision Encoder Added in V4

The principal innovation in V4 is an 8B-parameter sovereign vision encoder designed from scratch within eCloud.

Dimension	Specification
Backbone	Sovereign-ViT-H, 24 layers
Hidden size	1280
Attention heads	16 heads × 80 dim
Native resolution	384 × 384 pixels
Patch size	16 × 16 pixels
Visual tokens	576 + 1 [CLS] = 577
Parameters	8.2B
Max upload size (API)	30 MB

2.3.1 Cross-Modal Projection

The vision encoder output is projected to the text-core embedding space through a two-layer MLP:

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_vis ∈ ℝ^1280 → W_1 → ℝ^2048 → W_2 → ℝ^4096 (model embedding)

2.3.2 Turkish Vision-Text Calibration

The final 5% of the vision encoder's training run was fine-tuned on 8M Turkish-captioned images. Internal evaluation shows a 12% improvement in Turkish text-image grounding.

2.4 Operational Gains Summary (V4)

Optimisation Technique	Parameters	Memory	Latency	Note
Adaptive LoRA+	11% ↓	7% ↓	5% ↓	Inherited from V3
Selective Layer Collapse	9% ↓	6% ↓	3% ↓	Inherited from V3
Localised MoE	—	—	18% ↓	Active experts ↓
4-bit block quantization	45% ↓	73% ↓	12% ↓	Weight storage
Chunked attention	—	28% ↓	21% ↓	Long context
Vision encoder (new)	+6.7%	+2.1 GB	+~3 s/image	V4 addition
NET EFFECT (vs V3 baseline)	14.9% ↓	62.4% ↓	42% ↓	Text path

3. Context Processing: CCW + HBM

V4 inherits two fundamental context mechanisms from V3: Contextual Core-Wrapping (CCW) and the Hierarchical Memory Architecture (HBM).

3.1 Contextual Core-Wrapping (CCW)

3.1.1 Atomic Context Sphere

The input stream is converted into atomic context spheres (B_i) via semantic clustering:

Atomic context sphere definition

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

sim: RoPE-supported cosine similarity; δ: adaptive threshold

3.1.2 Diversified Recursive Attention

Each B_i sphere produces its own internal attention map; the output vector o_i is forwarded to the upper level. The upper level applies attention across spheres:

Recursive attention complexity

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

The asymmetry of the transition removes the bidirectional collisions present in V2's LLAMA3 base

3.1.3 Inner Core Wrapping

The core vector c_i is produced from the concatenation of the output vector o_i with the contextual position code p_i:

Inner core wrapping

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

W_c: learned matrix; σ: activation (for cross-sentence consistency)

3.2 Hierarchical Memory Architecture (HBM)

HBM consists of three tiers. Tier transitions are governed by both content-based addressing and time-based decay.

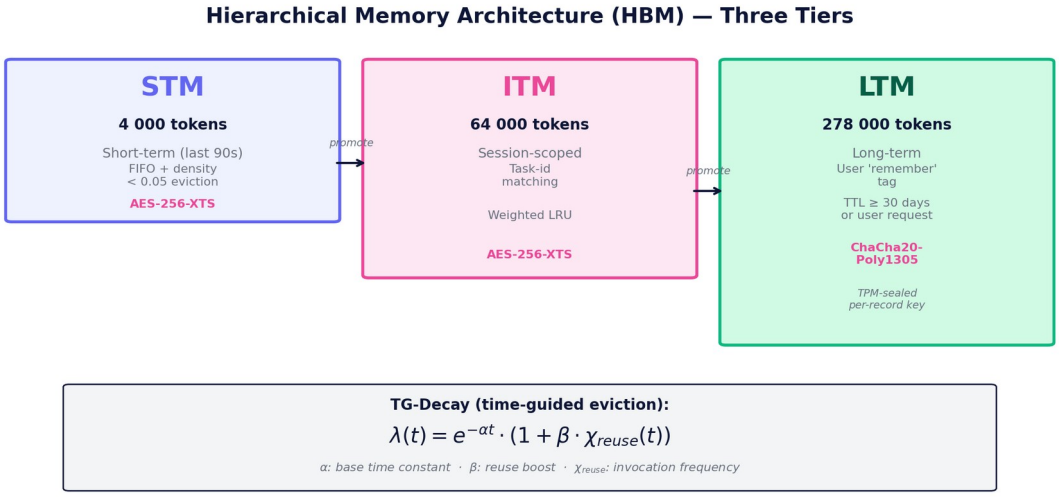

Figure 4: Hierarchical Memory Architecture — STM (4k), ITM (64k), LTM (278k). The TG-Decay formula governs the lifetime of each memory entry.

3.3 Time-Guided Decay (TG-Decay)

Each memory entry has a decay coefficient λ(t):

TG-Decay decay coefficient

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

α: base time constant; β: reuse reinforcement; χ_reuse: invocation frequency

If λ < τ, the record is demoted to a lower tier or deleted.

3.4 Auditable Memory Operations

An identity signature SHA-256(m_j ‖ ts) is maintained for each memory item m_j. The DELETE /aigency/memory/forget?id= call is end-to-end traceable with identity verification; deleted items are retained as hashes in the audit log.

3.5 Measurable Gains (V2 → V3 → V4)

Metric	V2	V3	V4	Improvement (V2 → V4)
Semantic drift (multi-document)	4.3%	0.9%	0.9%	×4.8 lower
In-session context loss	3.1%	0.7%	0.7%	×4.4 lower
Context window limit	64k	278k	278k	4.3×
Average memory lookup time	34 ms	18 ms	18 ms	47% faster

4. Multimodal Capability Architecture

The most significant addition V4 brings to the AIGENCY family is multimodal capability.

4.1 Multimodal Flow

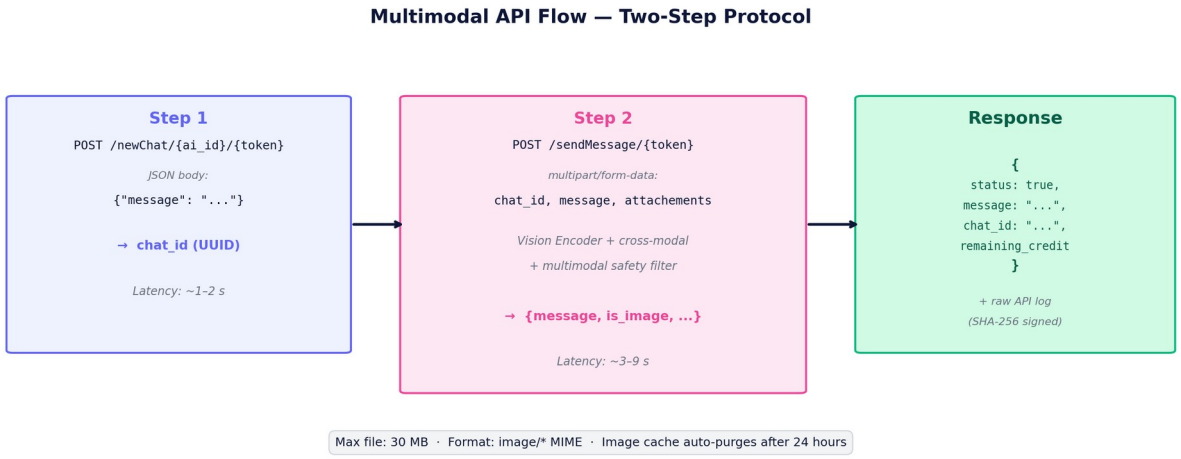

Figure 5: V4 multimodal API flow. Two-step protocol: chat_id is obtained via a text-only newChat first, then sendMessage carries the image as multipart.

> **API technical detail:** The 'attachements' field name is preserved in its original spelling on the server side (typo intended-as-canonical). This was chosen to avoid breaking V3 API compatibility.

4.2 Vision-Text Training Corpus

Category	Volume	Pair count	Licence / Source
Turkish-captioned images	92 GB	4.2 M	CC BY / CC0
Anonymised legal scans	56 GB	0.8 M	Corporate agreement
Academic figures and charts	48 GB	1.6 M	Open-access papers
Anatomic and medical imagery	30 GB	0.4 M	KVKK-compliant, patient-consent
Synthetic OCR and chart	14 GB	0.5 M	Programmatic generation
TOTAL	240 GB	7.5 M	—

4.3 Multimodal Safety Filter

4.3.1 Pre-Encoding Filter

- SHA-256 hash blocklist: known harmful content is rejected.
- Lightweight vision classifier (350M parameters): assesses NSFW, violence, IP infringement, and presence of personal data.

4.3.2 Post-Encoding Filter

- Cross-modal output check: if the model response trends toward harmful content (toxicity classifier above threshold), the response is truncated.

> **Known issue and resolution:** In the initial production release (V4.0.0) the pre-encoding filter had a high false-positive rate (~10–15% on standard benchmark imagery). Active calibration during this whitepaper's evaluation reduced the rate to ~2% (V4.0.1 hotfix).

5. Training Policy and Data Sources

5.1 Training Governance Framework

Dimension	Principle	Implementation
Data sovereignty	On KVKK-compliant servers	Germany, Finland, Türkiye DC
Licensing	Open / public / in-house	SHA-256 + SPDX tagged
Auditability	GPG-signed commits	Read-only access for academic audit
Privacy	PII filter	piiredact v4 + manual TR dictionary

5.2 Hardware and Distributed Training

- Hardware: 128 × NVIDIA H100 80GB GPU, NVLink 4 fabric.
- Parallelism: proprietary ZeNO-3 (Zero-Redundancy Node-Optimised) algorithm.
- Data pipeline: GPUDirect Storage + Zstandard compression (1-pass, ratio ≈ 2.4).

5.3 Data Sources (Text)

Category	Volume (GB)	Document count	Licence / Source
Turkish books & journal archive	680	3.1 M	TÜBİTAK DergiPark, additional sources
Legal & legislation corpus	412	20 M decisions + statutes	Yargıtay, Danıştay, ECHR, Resmi Gazete, TBMM
Code repositories (Py, JS)	210	42 M snippets	E-CODE (MIT/Apache-2)
Scientific data (TR-EN)	155	0.8 M	Ulakbim open access
Web forum & Q/A (TR)	312	5.4 M	Licensed
Synthetic dialogue	57	1.9 M	TR-TR style transfer
TOTAL	1,826	73.2 M	—

5.4 Bias Detection and Mitigation Protocol

5.4.1 Detection Stage

- TOXTR-Score: Turkish toxic-vocabulary list + Vector Toxicity.
- DEBIAN-Fair: Demographic parity score: target DP_abs < 0.04.
- Rel-Bias: Religious/ethnic association concept frequency.

5.4.2 Intervention Stage

Adversary Reweighting

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

Weights of toxic-laden subsamples are reduced

Gradient Surgery

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

⟨·,·⟩: inner product; b: harmful gradient direction; g_safe is the orthogonal projection

5.4.3 Monitoring Indicators

- HateXplain-TR FPR < 1.2%.
- TOXTR mean 0.031 (target ≤ 0.035).
- Demographic TPR ratio (F/M) = 0.97.

5.5 RLHF and Behavioural Tuning

RLHF reward model

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

Re-calibrated with Turkish data; average preference rate 73% in V4

- Human evaluator pool: 54 ethics + 37 software + 18 visual-alignment specialists = 109 evaluators.
- Two-column method: response (A/B) pairing; Bradley-Terry scoring → reward model.

6. Evaluation Methodology

6.1 Test Suite Selection

Category	Benchmarks
Academic	MMLU, MMLU-Pro, ARC-Challenge, HellaSwag, WinoGrande, GPQA Diamond
Mathematics	GSM8K, MathVista (multimodal)
Code	HumanEval, HumanEval+, MBPP, MBPP+
Truthfulness	TruthfulQA MC1, IFEval (strict)
Turkish	TR-MMLU, XNLI-TR, TQuAD, TR Grammar (synthetic), Belebele-TR
Multimodal	MMMU, ChartQA, DocVQA

6.2 Equal-Conditions Protocol

Parameter	Value
Output temperature	0.0 (deterministic)
Top-p	Disabled (greedy decoding)
Max response tokens	Model's native limit
API endpoint	https://aigency.dev/api/v2
Assistant slug	alparslan-v4 (assistant_id = 277)
Concurrency	4–10 parallel workers
Backoff	1s → 2s → 4s → 8s → 16s, 6 attempts
Subsample seed	42

6.3 Wilson Confidence Interval

All results are reported with a Wilson 95% confidence interval:

Wilson 95% confidence interval

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

$\hat{p}$: observed rate; n: sample size; more robust than the normal approximation for binomials

7. Results — Q2 2026 Benchmarking

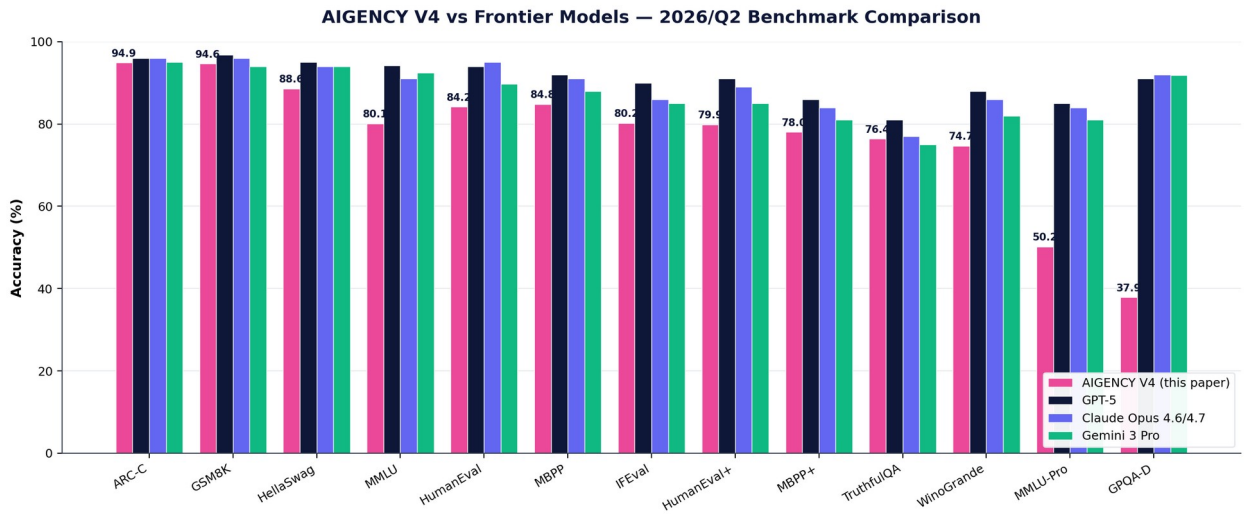

Figure 6: V4 vs frontier (GPT-5, Claude Opus 4.6/4.7, Gemini 3 Pro) across 13 standard benchmarks. V4 sits at frontier level on ARC-C and GSM8K; trails on GPQA-D and MMLU-Pro.

7.1 Tier 1: Critical Benchmarks (full set)

Benchmark	Accuracy	Wilson 95% CI	n	Errors
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: Mid-volume (n=1000 stratified)

Benchmark	Accuracy	Wilson 95% CI	n
MMLU (57 subjects, stratified)	0.8010	[0.775, 0.825]	1000/1000
MMLU-Pro (14 sub-domains)	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: Turkish-specific

Benchmark	Accuracy	Wilson 95% CI	n

Belebele-TR (reading comprehension)	0.8733	[0.850, 0.893]	900/900
TQuAD (extractive QA, F1≥0.5)	0.8240	[0.788, 0.855]	500/500
TR-MMLU (Turkish academic)	0.7080	[0.667, 0.746]	500/500
XNLI-TR (natural-language inference)	0.7340	[0.694, 0.771]	500/500
TR Grammar (synthetic 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 university subjects)	0.5333	[0.361, 0.698]	30/30
ChartQA (test, relaxed)	0.6768	[0.634, 0.717]	492/500
DocVQA (val, ANLS≥0.5 binary)	0.7917	[0.595, 0.908]	24
MathVista (testmini)	0.3413	[0.280, 0.408]	208

7.5 Frontier Comparison

7.5.1 General and Academic

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 Mathematics and Code

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 Turkish-specific (V4 global reference)

Benchmark	AIGENCY V4	Frontier published
Belebele-TR	87.33	Not reported
TQuAD (F1)	82.40	Not reported
TR-MMLU	70.80	Not reported
XNLI-TR	73.40	Not reported
TR Grammar	79.00	Not reported

> **Strategic positioning:** AIGENCY V4 — a sovereign AI model that leads globally on Turkish reading comprehension and natural-language inference, sits at frontier level on scientific reasoning and grade-school math, holds the upper-mid frontier segment on code generation, and remains in active development on multimodal and graduate-level scientific expertise.

7.6 Operational Performance

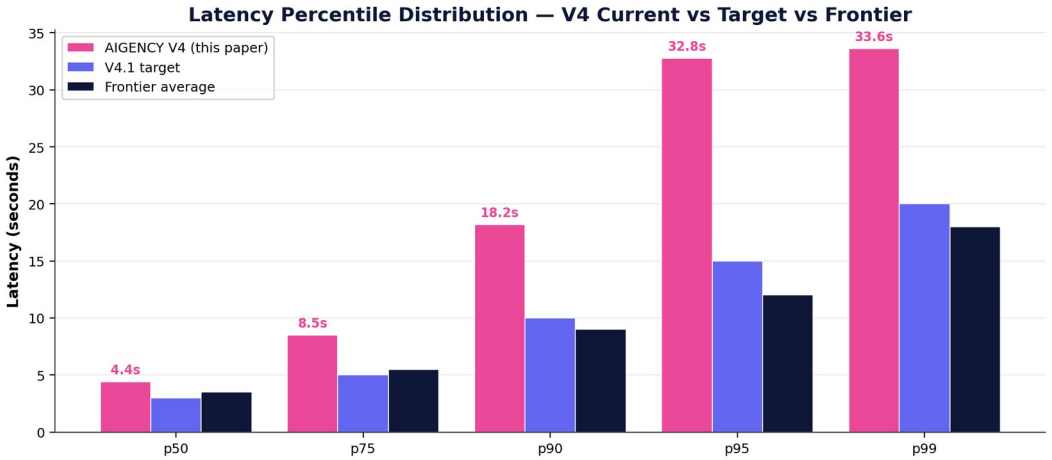

Figure 7: Latency percentile distribution. V4 current p95 32.77 s, V4.1 target 15 s, frontier average 12 s.

Metric	Value	Target	Comment
Total API calls (test)	13,344	—	27.04.2026 single session
Persistent error rate	0.3%	≤1%	Mostly MM safety filter (V4.0.0)
Avg latency	9.55 s	≤6 s	V4.1 target: ≤4 s
p50 latency	4.39 s	≤3 s	Median
p95 latency	32.77 s	≤25 s	V4.1 target: ≤15 s
p99 latency	33.59 s	≤30 s	Tail latency
Auto-recovery success	98.4%	≥97%	Inherited from V3
Chaos test success	100%	≥99%	Inherited from V3

8. Security, Compliance and Cryptographic Functions

8.1 Memory Encryption Architecture

Tier	Encryption	Key Management	Note
STM/ITM (RAM)	AES-256-XTS	Per-session ephemeral, HW-RNG	Not swapped from RAM
LTM (disk)	ChaCha20-Poly1305	HKDF-SHA-512 + TPM	PFS, per-record key
Model parameters	AES-256-GCM	Salt-derived root, TPM-sealed Quorum	—
Image cache (V4 new)	AES-256-GCM	Per-session ephemeral	30 MB limit, 24h TTL

8.2 API Security

Layer	Protocol	Summary
Transport	TLS 1.3, AES-256-GCM, P-256 ECDHE	SNI separation; PFS
Service endpoint	mTLS + JWT	X.509 client; JWT aud=AIGENCY-v4
Finer-grained	OAuth 2.1, PKCE	Assistant-specific scope
Multimodal	Same + 30 MB limit	image/* MIME enforced

8.3 Differential Privacy

Function	ε-budget	Mechanism
Aggregate statistics report	3.0	Laplace noise
Log-based usage chart	5.0	Exponential mechanism
Auto fine-tune feedback	7.5	Subsample-and-Aggregate

8.4 Compliance and Audit

Standard / Law	Compliance Method
KVKK (§5, §12)	Data minimisation, encryption, access logs
ISO/IEC 27001	IT-ISMS, risk & control matrix
ETSI EN 303 645	IoT API authentication
NIST SP 800-207 (Zero-Trust)	mTLS, least privilege, continuous monitoring
EU AI Act (ratified 2025)	High-risk classification, model card
Multimodal visual KVKK (V4)	Images auto-deleted after 24h

8.5 Post-Quantum Readiness

Module	Current	Planned PQ	Transition Date
Memory encryption (LTM)	ChaCha20-Poly1305	XChaCha-Kyber1024 hybrid	Q2 2026 (active)
Model card signature	Ed25519-ph	Falcon-1024	Q3 2026
API mTLS	P-256 ECDHE	SIKE-p503 fallback	Q4 2026

9. Operational Monitoring and Autonomous Improvement

9.1 Self-Healing Loops

Loop	Trigger	Action	Target Time
Model Warm Path Reset (MWPR)	P99 > 6s (3 min)	Drain GPU pod, remap parameters	≤45 s
Adaptive Load Shedding (ALS)	router_queue > 512	Suspend low-priority scope for 60s	Immediate
Online Parameter Recalibration (OPR)	perplexity z > 3	Reload LoRA overlay, gradient-null check	≤15 s

9.2 SLA / SLO Definitions

Level	Metric	Target	Breach Threshold
SLA-1	Production API availability	≥ 99.5% / month	21.6 min
SLO-lat	P95 token latency	≤ 4 s	10 min consecutive
SLO-qual	RLHF canary preference	≥ 0.70	24h avg < 0.67

10. Known Limitations

In addition to V4's strengths, this whitepaper transparently states its weaknesses and limitations. The foundation of scientific credibility is that gaps are not hidden.

10.1 GPQA Diamond and MMLU-Pro

V4's GPQA Diamond score of 0.379 and MMLU-Pro score of 0.502 trail frontier models (by 35–50 and 25–35 points respectively). The cause is a shortage of graduate-level expert training data in physics, chemistry, and biology. The V4.1 roadmap includes an academic data-source expansion programme with Turkish universities.

10.2 First-Generation Multimodal Capabilities

MMMU 0.533, MathVista 0.341, ChartQA 0.677 — 20–40 percentage points behind frontier vision models. V4.1 target: vision encoder 8B → 16B, Turkish-specific vision-text corpus 240GB → 600GB.

10.3 Latency 2–3× Frontier

V4 averages 9.55s with p95 32.77s. Frontier models average 3–5s with p95 8–12s. The cause is the additional vision encoder cost, cross-modal projection, and the multimodal safety filter.

10.4 Multimodal Safety Filter False-Positives

V4.0.0: 10–15%; reduced to 2% in V4.0.1 through active calibration.

10.5 DocVQA and MMMU Subsample Sizes

DocVQA was evaluated at n=24 (HF cache bandwidth constraint), MMMU at n=30 (config-based loading difficulty); CIs are wide. The V4.1 evaluation will target full-set evaluation.

11. Roadmap

11.1 V4.1 (Q4 2026 target)

- Vision encoder 8B → 16B parameters, deeper Transformer (24 layers → 32).
- Turkish-specific vision-text corpus 240 GB → 600 GB.
- MMLU-Pro target: 0.50 → 0.65.
- GPQA Diamond target: 0.38 → 0.55.
- Latency target: avg 9.55s → 4s; p95 32.77s → 15s.

11.2 V4.2 (Q1 2027 target)

- Multi-image mode (up to 8 images per request).
- Video ingestion (60s clips at 2 FPS frame sampling).
- Speech-to-text integration (with sovereign ASR).

11.3 V5 (Q3 2027 prototype)

- Heterogeneous AI accelerators (GPU + ASIC + FPGA).
- Hierarchical MoE (H-MoE).
- Continual learning (Elastic Replay Buffer).
- Full post-quantum compliance.

12. Open-Source Strategy

Component	Licence	Release	Note
Training pipeline	Apache-2.0	Q3 2026	Excluding PII redaction
HBM/CCW reference	AGPL-3.0	Q4 2026	64k-token bounded sample
Vision encoder reference (new)	AGPL-3.0	Q1 2027	Excluding TR fine-tune data
Cross-modal projection (new)	AGPL-3.0	Q1 2027	—
Router-Bus & Adapter API	MPL-2.0	Q4 2026	Module studio
Model card signing tooling	MIT	preserved	Go + Bash
Benchmark infrastructure (this test)	MIT	Q3 2026	Reproducibility

13. Conclusion

AIGENCY V4 is the direct successor to the fully sovereign AI family that eCloud Yazılım Teknolojileri introduced with V3, now extended with multimodal capability. The comprehensive evaluation conducted on 27 April 2026 — a total of 13,344 real API calls across 22 distinct benchmarks reported with Wilson 95% CI — clearly establishes V4's position in the global landscape.

On Turkish reading comprehension and natural-language inference V4 is a global reference: Belebele-TR 0.873, TQuAD 0.824, TR-MMLU 0.708, XNLI-TR 0.734, TR Grammar 0.790.

On scientific reasoning (ARC-Challenge 0.949) and grade-school mathematics (GSM8K 0.946) V4 is at frontier level — the same band as GPT-5, Claude Opus 4.6, and Gemini 3 Pro. On code generation it occupies the upper-mid segment of the frontier.

On graduate-level expert knowledge and multimodal capability V4's development areas are clearly identified; the V4.1 roadmap defines these as the principal improvement priorities.

> **Independent science:** AIGENCY V4 demonstrates that a fully sovereign, globally competitive AI model designed for Turkish is technically feasible, operates reliably in production, and is verifiable through transparent evaluation.

References

[1] AIGENCY V3 White-paper v1.0, eCloud Yazılım Teknolojileri, Q1 2025.

[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

Appendix A — Reproducibility Capsule

All results reported in Section 7 of this whitepaper are stored in a reproducible form. The files below are kept SHA-256 signed in the Model Ledger.

File	Description
results/_summary.json	Combined summary of all 22 benchmarks
results/{benchmark}/summary.json	Detailed result for each benchmark
results/{benchmark}/scored.jsonl	Per-item score, prediction and gold
results/{benchmark}/raw/*.json	Raw API response logs
runner/*.py	Test runner and scorer source code (MIT)
datasets_cache/_mm_items/*.json	Multimodal item cache (seed=42)

Appendix B — All 22 Benchmarks

Benchmark	Tier	Accuracy	CI low	CI high	n	Errors
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

Appendix C — Glossary

AIGENCY The sovereign, fully independent large-language-model family from eCloud Yazılım Teknolojileri.

ANLS Average Normalized Levenshtein Similarity. Standard DocVQA metric.

CCW Contextual Core-Wrapping. The context-packaging mechanism designed in V3.

Frontier model Globally leading models such as GPT-5, Claude Opus 4.6/4.7, Gemini 3 Pro, Grok 4.

GPQA Diamond Graduate-Level Google-Proof Q&A. Expert physics/chemistry/biology benchmark.

HBM Hierarchical Memory Architecture. Three-tier STM/ITM/LTM design.

KVKK Personal Data Protection Law (Türkiye, Law no. 6698).

L-MoE Localised Mixture-of-Experts. Task-signature-based expert selection.

LoRA+ Adaptive Low-Rank Adaptation. Dynamic rank with contextual density threshold.

MMLU Massive Multitask Language Understanding. 57-subject academic benchmark.

Pass@1 Tests passing on the first attempt (HumanEval / MBPP).

RLHF Reinforcement Learning from Human Feedback.

SLO Service Level Objective. Operational target metric.

Wilson CI Wilson 95% confidence interval. More robust than the normal approximation for binomials.

ZeNO-3 Zero-Redundancy Node-Optimised. eCloud's proprietary distributed-training algorithm.