Relim.ai

ENTERPRISE ARCHITECTURE & BENCHMARK WHITEPAPER

Zero-Egress Intelligence: Eliminating the Cloud AI Tax with Relim AI on AMD Ryzen AI Halo.

A Technical Blueprint and Empirical Benchmark Assessment for Running Sovereign, High-Throughput Local LLMs Inside the Enterprise Perimeter Using the 128GB Unified-Memory AMD Ryzen AI Max+ 395 Platform.

Published: November 2026 | Target Audience: CIOs, CTOs, CISOs, VP Engineering & Enterprise Data Architects |
Hardware Reference: AMD Ryzen AI Halo Developer Platform (AMD Ryzen AI Max+ 395, 128GB Unified Memory, AMD ROCm).

THE ENTERPRISE INFLECTION POINT

Every prompt sent to a third-party cloud LLM is both a recurring tax on your margins and an irreversible egress of institutional intelligence. With Relim AI running natively on AMD Ryzen AI Halo's 128GB unified memory silicon, enterprises no longer choose between frontier-grade reasoning and air-gapped sovereignty - they own both, at a 90%+ lower 3-year TCO.

90%+ = 3-Year TCO Reduction vs. Cloud APIs	0 Bytes = External Cloud Data Egress	128 GB = Unified LPDDR5x-8000 Memory Pool	73+ tok/s = MoE Local LLM Throughput (120W)

Table of Contents

01	Executive Summary
02	The Enterprise AI Dilemma
03	Architectural Synergy
04	Empirical Benchmark Assessment
05	Enterprise Infrastructure Comparison
06	Total Cost of Ownership
07	Zero-Egress Security & Compliance
08	Deployment Blueprint & Sizing Guide
09	Strategic Conclusion

01 EXECUTIVE SUMMARY: THE END OF THE CLOUD AI TAX

Enterprise AI has hit an economic and architectural wall. For the past three years, organizations rushed to wire proprietary operational data, ERP records, intellectual property, and industrial telemetry into external cloud LLM APIs. While cloud endpoints accelerated early prototypes, production scale has exposed two existential liabilities: exponential token-metering OpEx and unacceptable data sovereignty exposure.

When Retrieval-Augmented Generation (RAG), multi-step agentic workflows, and continuous DataOps automation move into 24/7 enterprise production, token consumption multiplies by orders of magnitude. A single enterprise deployment servicing 100 knowledge workers or automated data pipelines routinely incurs $4,500 to $12,000+ per month in recurring API fees - without ever building an asset on the balance sheet. Simultaneously, CISOs and compliance leaders face mounting friction across SOC 2 Type II, ISO/IEC 27001:2022, HIPAA, GDPR, and defense-adjacent regulations whenever sensitive context traverses public inference clouds.

This whitepaper presents a decisive architectural shift: deploying Relim AI - our software-only Intelligent Data Platform and enterprise DataOps middleware - directly on-premises atop the AMD Ryzen AI Halo Developer Platform powered by the AMD Ryzen AI Max+ 395 processor. By combining Relim AI's intelligent orchestration layer with 128GB of high-speed LPDDR5x-8000 unified memory, 40 RDNA 3.5 compute units, and a 50 TOPS XDNA 2 NPU inside a 120W, sub-1.2 kg edge-ready chassis, enterprises unlock frontier-caliber Mixture-of-Experts (MoE) local inference at 31 to 73+ tokens per second. Cloud API costs drop to zero, payback is achieved in under 90 days, and not a single byte of enterprise data ever leaves the corporate perimeter.

> CORE ARCHITECTURAL THESIS: Hardware is no longer the bottleneck to Sovereign AI. The convergence of 128GB unified memory silicon (AMD Ryzen AI Max+ 395) and sparse Mixture-of-Experts (MoE) foundation models allows Relim AI to deliver cloud-equivalent intelligence on-premises - eliminating per token OpEx while enforcing hardware-backed, zero-egress security.

02 THE ENTERPRISE AI DILEMMA: WHY CLOUD-ONLY INFERENCE FAILS AT SCALE

Relying exclusively on hyperscaler LLM APIs introduces structural fragility across finance, security, and engineering operations. As enterprises transition from simple chat interfaces to autonomous, data-intensive AI pipelines, four critical failure modes emerge:

1. Runaway Token-Metering Economics (The OpEx Trap): Agentic reasoning loops, long-context RAG injections (32K–128K tokens), and automated DataOps transformations generate massive prompt-to-completion ratios. Unlike predictable fixed infrastructure, cloud API billing scales linearly or super-linearly with adoption – penalizing organizations precisely when AI succeeds.

2. Zero-Trust Violations & Shadow Data Exfiltration: Every external API call transmits proprietary schemas, customer PII, financial ledgers, or manufacturing telemetry across third-party networks. Even with enterprise zero-retention agreements, external transmission expands the audit surface, complicates vendor risk assessments, and blocks adoption in air-gapped or highly regulated verticals.

3. Model Drift, Rate Limits & Latency Jitter: Cloud providers silently deprecate model weights, throttle concurrent requests under peak global load, and introduce unpredictable WAN latency. Mission-critical enterprise workflows require deterministic SLAs and version-pinned inference engines.

4. The Traditional On-Premises Hardware Barrier: Until recently, self-hosting large language models required $40,000+ multi-GPU rack servers drawing 1,500W+ of power, specialized liquid cooling, and dedicated ML infrastructure teams – putting on-premises AI out of reach for mid-market enterprises and distributed branch or plant floor deployments.

03 ARCHITECTURAL SYNERGY: RELIM AI MEETS AMD RYZEN AI HALO

The AMD Ryzen AI Halo Developer Platform (codenamed Strix Halo) shatters the traditional VRAM wall that crippled compact local AI hardware. Instead of isolating a small 16GB or 24GB discrete GPU memory pool behind a narrow PCIe bottleneck, the Ryzen AI Max+ 395 unifies 16 Zen 5 CPU cores, 40 RDNA 3.5 GPU compute units (Radeon 8060S), and an AMD XDNA 2 NPU across a coherent, 256-bit 128GB LPDDR5x-8000 unified memory architecture delivering 256 GB/s of peak bandwidth.

How Relim AI Orchestrates the Silicon Stack

Relim AI operates strictly as a software-only Intelligent Data Platform and enterprise DataOps middleware. Rather than treating local silicon as a passive model runner, Relim AI dynamically maps its cognitive and data-engineering workloads across the Ryzen AI Max+ 395 heterogeneous compute fabric:

- Data Ingestion, ETL & Semantic Indexing (16-Core / 32-Thread Zen 5 CPU + 10GbE):
  Relim AI's DataOps pipelines ingest structured databases, ERP streams, and unstructured document repositories over the platform's native 10 Gbps Ethernet interface. The 16 Zen 5 cores (boosting to 5.1 GHz with full AVX-512 support and 64MB L3 cache) execute high-speed chunking, schema normalization, and hybrid vector/graph retrieval without starving the GPU.
- High-Throughput Local LLM Inference (40-CU RDNA 3.5 GPU + 128GB Unified Memory):
  Via AMD ROCm Open Software and Variable Graphics Memory / Linux GTT allocation (exposing 96GB to 110GB+ directly to the GPU), Relim AI hosts massive 30B-120B+ parameter Mixture-of-Experts models completely resident in memory - supporting 128K token enterprise context windows with zero out-of-memory (OOM) crashes.
- Always-On Edge Embeddings & Guardrail Telemetry (50 TOPS XDNA 2 NPU):
  Lightweight classification, PII redaction, prompt-injection screening, and continuous embedding generation offload to the dedicated XDNA 2 NPU at ultra-low wattage, keeping the primary RDNA 3.5 compute fabric dedicated to generative reasoning.

Hardware Reference Specification: AMD Ryzen AI Halo Developer Platform

Subsystem	Technical Specification	Enterprise Local AI Impact
Processor (CPU)	AMD Ryzen AI Max+ 395 (16 Cores / 32 Threads, Zen 5, up to 5.1 GHz, 64MB L3)	High-throughput DataOps ETL, RAG orchestration & multi-agent routing
Integrated GPU	AMD Radeon 8060S (40 Compute Units, RDNA 3.5, 2900 MHz, up to 60 TFLOPS FP16)	Discrete-class parallel matrix compute via AMD ROCm without dGPU power draw
AI Accelerator (NPU)	AMD XDNA 2 NPU (Up to 50 TOPS INT8; 126 Overall Platform TOPS)	Dedicated ultra-low-power guardrail checks, embeddings & background agents
Unified Memory	128 GB LPDDR5x-8000 (256-bit Bus, 256 GB/s Theoretical / ~215 GB/s Sustained)	Holds 70B-120B+ quantized models & 128K context windows entirely resident
Encrypted Storage	2 TB M.2 NVMe SSD with Self-Encrypting Drive (SED) Hardware Security	Instant model weight loading & hardware-level encryption at rest
Enterprise Networking	1x 10 Gbps Ethernet, Wi-Fi 7, Bluetooth 5.4, 3x USB-C (40Gbps USB4)	Line-rate LAN data ingestion & multinode high-speed clustering
Thermal & Footprint	120W TDP (45-120W cTDP), 150 x 150 x 45.4 mm, < 1.2 kg (2.65 lbs)	Zero data-center retrofit, deploys in server closets, branches, or edge racks
OS & Software Stack	Purpose-Built AMD Ryzen AI Linux Developer Platform / Windows 11 + ROCm	Native containerized deployment for Relim AI, vLLM, llama.cpp & Ollama

04 EMPIRICAL BENCHMARK ASSESSMENT & LOCAL INFERENCE PHYSICS

Evaluating local LLM hardware requires looking past marketing TOPS and examining the fundamental physics of autoregressive token generation. During the decode phase, every generated token requires streaming the active model weights from memory into compute registers. Consequently, sustained inference speed is governed by the memory bandwidth equation:

> Sustained Generation Speed (tok/s) = Effective Memory Bandwidth (GB/s) / Active Parameter Footprint (GB)

Independent empirical testing of the AMD Ryzen AI Max+ 395 (128GB LPDDR5x-8000) under Linux with AMD ROCm and llama.cpp demonstrates ~215 GB/s of sustained real-world GPU memory bandwidth out of its 256 GB/s theoretical ceiling. This creates a profound architectural insight for Relim AI deployments: Mixture-of-Experts (MoE) architectures unlock a 6x to 14x throughput multiplier on identical silicon.

While a legacy 70B dense model must stream all ~42.5 GB of weights per token (yielding ~5.0 tok/s), modern sparse MoE models keep 30B to 120B+ total parameters resident inside the 128GB unified memory pool while activating only 3B to 5.1B expert parameters per token. The result is frontier-grade reasoning quality delivered at 31 to 73+ tokens per second - far exceeding human reading speed (~6-8 tok/s) and rivaling paid cloud APIs.

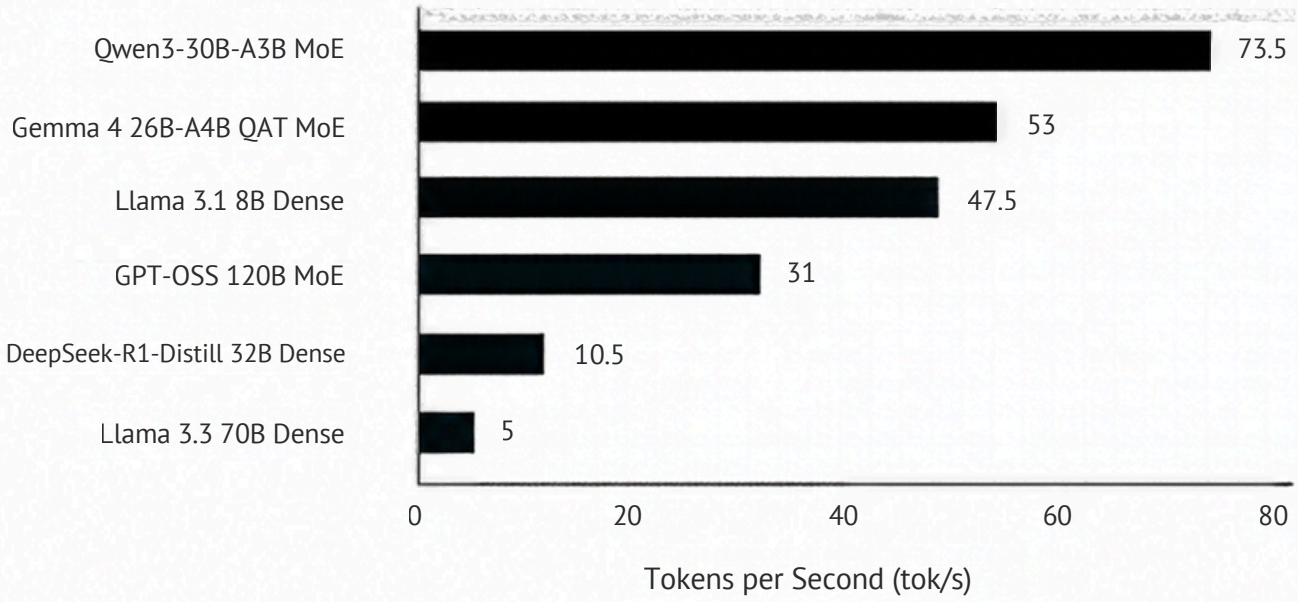


Figure 1: Sustained token generation throughput (Q4_K_M quantization) across Dense and MoE architectures on AMD Ryzen AI Max+ 395 (128GB Unified Memory, ~215 GB/s sustained bandwidth).

Relim AI Workload-to-Model Benchmark Matrix

Model Architecture (Q4_K_M)	Total / Active Params	Memory Footprint	Sustained Speed	Optimal Relim AI Enterprise Use Case
Qwen3-30B-A3B (MoE)	30B / 3B Active	~18.5 GB	72-75 tok/s	Real-time enterprise chat, SQL generation & rapid RAG synthesis
Gemma 4 26B-A4B QAT (MoE)	26B / 4B Active	~16.0 GB	52-54 tok/s	Balanced daily executive assistant, analytics & code generation
Llama 3.1 8B (Dense)	8B / 8B Active	~4.9 GB	42-53 tok/s	High-concurrency routing, entity extraction & metadata tagging
GPT-OSS 120B (MoE)	120B / 5.1B Active	~71.6 GB	30-32 tok/s	Complex multi-step reasoning, strategic synthesis & deep audits
DeepSeek-R1-Distill 32B (Dense)	32B / 32B Active	~19.8 GB	10-11 tok/s	Chain-of-thought root-cause analysis & technical diagnostics
Llama 3.3 70B (Dense)	70B / 70B Active	~42.5 GB	~5.0 tok/s	Asynchronous batch document processing & overnight compliance jobs

BENCHMARK TAKEAWAY: CAPACITY BEATS RAW DISCRETE BANDWIDTH FOR MODERN MoE.
A single discrete consumer GPU (e.g., 24GB RTX 4090) boasts 1,008 GB/s of bandwidth - but immediately crashes with Out-of-Memory (OOM) errors when loading 70B-120B models or 128K enterprise context caches. The AMD Ryzen AI Max+ 395's 128GB unified pool runs 120B-class MoE models effortlessly at 31+ tok/s while drawing just ~120W - less than a third of a single discrete flagship GPU.

05 ENTERPRISE INFRASTRUCTURE COMPARISON: SILICON & CLOUD TOPOLOGIES

To validate the strategic positioning of the AMD Ryzen AI Halo platform, we assessed four primary enterprise inference topologies across usable AI memory, power envelope, physical footprint, data sovereignty, and 3-year total cost.

Evaluation Dimension	Relim AI + AMD Ryzen AI Halo (Max+ 395)	Single Discrete GPU (RTX 4090 24GB)	Multi-GPU Server (4x RTX 3090 / A100)	Enterprise Cloud LLM APIs (Frontier)
Usable AI Memory	Up to 110+ GB Unified (128 GB Total)	24 GB VRAM (Hard Ceiling)	96 GB - 320 GB VRAM	Elastic (Shared Multi-Tenant)
Max Local Model Class	120B+ MoE / 70B Dense + 128K Context	Up to 32B (Tight) / OOM on 70B+	70B - 405B Dense & MoE	Proprietary Frontier Models
Peak Power Draw	120W (~10W Idle)	450W - 600W System	1,400W - 2,200W+ Rack	N/A (Cloud OpEx)
Physical Footprint	150 x 150 x 45.4 mm (< 1.2 kg)	Full ATX Workstation Tower	4U Rack Server + Cooling	External Hyperscaler
Data Sovereignty	100% Air-Gapped / Zero Egress	100% On-Premises	100% On-Premises	External API Transmission
Upfront Hardware CapPx	~$3,500 - $4,500 / Node	~$3,200 - $4,000	$15,000 - $65,000+	$0 Upfront
3-Year Cumulative TCO	$4,650 (1 Node) / $9,300 (2-Node HA)	$5,800 (Limited Model Tier)	$24,000 - $80,000+	$162,000 - $342,000+

06 TOTAL COST OF OWNERSHIP (TCO) & BREAK-EVEN ECONOMICS

Replacing recurring cloud LLM API bills with fixed, depreciable on-premises compute transforms enterprise AI from an unbounded operating expense into a high-margin capital asset. Below is a 36-month financial comparison modeled for a mid-market enterprise (50-150 active users plus automated Relim AI DataOps pipelines averaging 400M-900M tokens/month) versus a high-volume enterprise deployment.

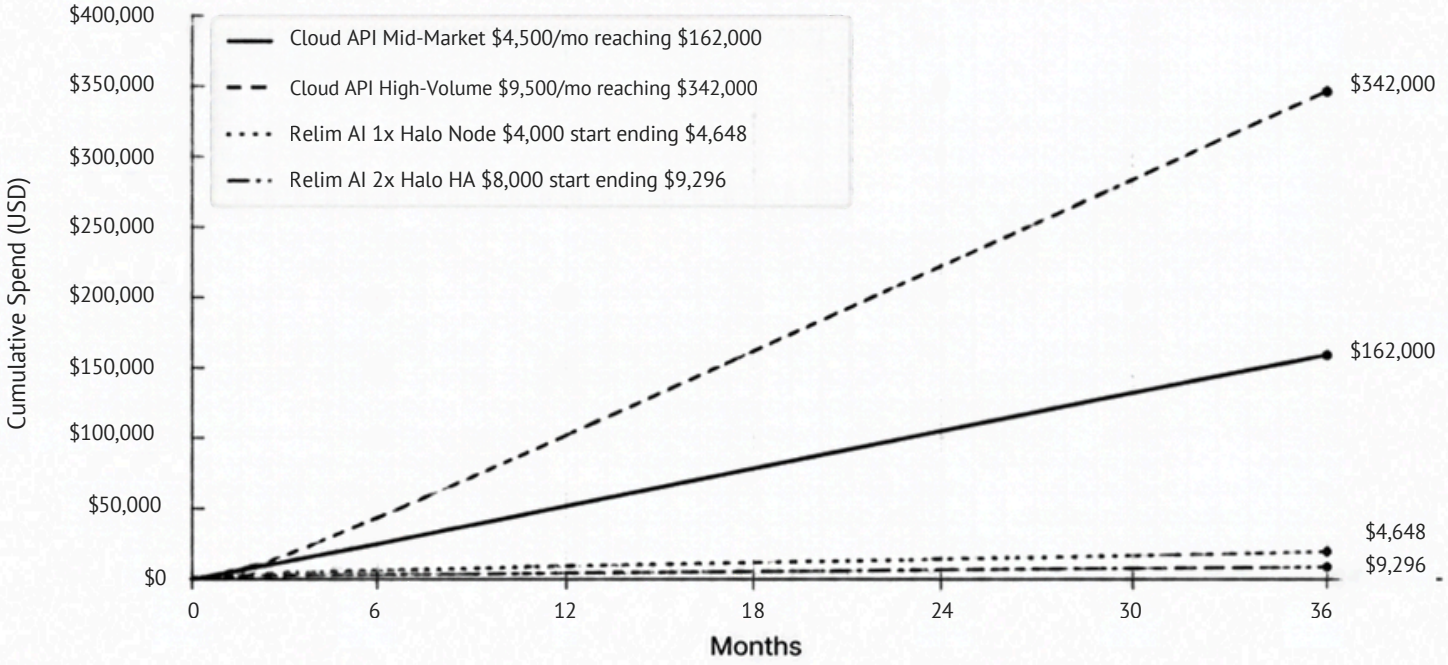


Figure 2: 36-Month cumulative spend trajectory comparing Cloud LLM API OpEx against 1-Node and 2-Node HA Relim AI on-premises deployments powered by AMD Ryzen AI Halo.

Financial Metric (36-Month Horizon)	Cloud API (Mid-Market: $4.5K/mo)	Cloud API (High-Vol: $9.5K/mo)	Relim AI + 1x Halo Node	Relim AI + 2x Halo HA Cluster
Initial Hardware CapEx	$0	$0	$4,000	$8,000
Monthly API / Token Spend	$4,500 / mo	$9,500 / mo	$0 / mo	$0 / mo
Monthly Power & Cooling (120W avg)*	$0 (Included in API)	$0 (Included in API)	~$18 / mo	~$36 / mo
Year 1 Cumulative Spend	$54,000	$114,000	$4,216	$8,432
Year 3 Cumulative Spend (Total TCO)	$162,000	$342,000	$4,648	$9,296
Net 3-Year Savings vs. Mid-Market Cloud	Baseline	-$180,000	+$157,352 (97.1%)	+$152,704 (94.3%)
Hardware Payback Period	Never (Perpetual OpEx)	Never (Perpetual OpEx)	~27 Days	~54 Days

*Electricity modeled at $0.15/kWh industrial rate assuming mixed active inference and low-power idle states (~10W idle, ~120W peak). Relim AI software licensing is excluded from both cloud and local infrastructure columns to isolate pure compute/API economics.

07 ZERO-EGRESS SECURITY, DATA SOVEREIGNTY & COMPLIANCE BLUEPRINT

- **Air-Gapped Zero-Egress Execution:** All RAG document embeddings, vector stores, knowledge graphs, prompt histories, and LLM completions execute strictly inside the corporate LAN. Outbound internet access can be physically or firewall-disabled with zero degradation in platform intelligence.
- **Hardware-Rooted Security & Encryption at Rest:** The AMD Ryzen AI Halo platform ships with a 2TB Self-Encrypting Drive (SED) M.2 NVMe SSD and AMD Enhanced Virus Protection (NX bit), ensuring proprietary model weights, fine-tuned LoRA adapters, and cached enterprise context remain cryptographically protected at rest.
- **Enterprise Governance via Relim AI Middleware:** Relim AI enforces Role-Based Access Control (RBAC), row-level data masking, deterministinc prompt guardrails, and immutable audit logs before any query reaches the local LLM - preventing internal privilege escalation and shadow AI.
- **Accelerated Audit Readiness (SOC 2 Type II & ISO/IEC 27001:2022):** By eliminating third-party LLM subproceesors from the data-flow diagram, organizations dramatically shrink vendor risk assessments and streamline compliance with SOC 2 Type II, ISO/IEC 27001:2022, HIPAA, GDPR, and regional data-residency statutes.

08 ENTERPRISE DEPLOYMENT BLUEPRINT & SIZING GUIDE

Deploying Relim AI on AMD Ryzen AI Halo requires zero data-center retrofits. The pre-validated AMD Ryzen AI Linux Developer Platform paired with AMD ROCm Open Software enables rapid, containerized production rollouts:

Deployment Tier	Hardware Topology	Recommended Model Stack	Target Concurrency & Scope
Tier 1: Departmental / Edge Node	1x AMD Ryzen AI Halo (128GB, 10GbE, 2TB SED)	Qwen3-30B-A3B (Primary) + Llama 3.1 8B (Router/Guardrail)	15-40 knowledge workers or single-plant DataOps & RAG automation
Tier 2: High-Availability Enterprise Pod	2x-4x AMD Ryzen AI Halo Nodes over 10GbE Switch	Load-balanced Qwen3-30B-A3B + Dedicated GPT-OSS 120B Reasoning Node	50-250+ enterprise users, zero-downtime failover & multi-agent workflows
Tier 3: Distributed Multi-Site Mesh	1x Halo Node per Branch / Plant + Central Relim AI Control Plane	Site-local quantized MoE models synced with central governance policies	Global manufacturing, retail, or field operations with intermittent WAN

09 STRATEGIC CONCLUSION: OWN YOUR INTELLIGENCE

The era of renting enterprise intelligence by the token is drawing to a close. By pairing Relim AI's software-only Intelligent Data Platform with the 128GB unified memory AMD Ryzen AI Halo Developer Platform, forward-thinking enterprises reclaim control over their AI economics, latency, and data perimeter. You eliminate six-figure cloud API bills, achieve full hardware payback in under 90 days, and run 120B-class Mixture-of-Experts models entirely behind your own firewall.

NEXT STEP: LAUNCH A 14-DAY SOVEREIGN AI PROOF OF VALUE (POV).

Ready to benchmark Relim AI against your live enterprise data - with zero cloud egress? Contact the Relim AI Enterprise Architecture team to schedule a 14-day on-premises Proof of Value on AMD Ryzen AI Halo hardware. We will benchmark your exact RAG workloads, measure real-time tokens-per-second, and deliver a custom 3-year CFO-ready TCO model.