AIVR

[SOVEREIGN AI PLATFORM]

THE FEATURES BROCHURE · v1.0

Everything in the box.

Seven pillars. Six modules. Eight products. Four accelerator classes. One architecture you own end to end.

Cockpit · Nerve · Brain · Swarm · Pulse · Measure · Sentinel · Gatekeeper

PUBLISHER	EDITION	DATE	WEB
AIVR	v1.0 (2026)	11 April 2026	www.aivr.site

TABLE OF CONTENTS

What's Inside

A complete map of the AIVR platform — every pillar, module, product, and accelerator class. Each row resolves to concrete components in the v3.4.0 architecture.

01 THE SEVEN PILLARS

The Architecture, in Seven Rows

Every AIVR deployment runs these seven pillars. Each is a first-class component with its own runtime, event surface, and operator interface. None is optional.

Pillar	Ref	Technology	Role
The Cockpit	01	React 18 + Express 5 + WebSocket	Unified WebUI command center. Seven tabs (Chat, Blueprint, Swarm, Knowledge, Pulse, Safety, Settings). 37+ tools via MCP. Every action human-consentable.
The Nerve	02	NATS JetStream event fabric	Coordination backbone. Eight typed streams, persistent, backpressured, fan-out by default. Sub-millisecond latency. Multi-node clustering ready out of the box.
The Brain	03	N8N workflow automation	14 production workflows. SOP validation, doc ingestion, blueprint review, auto-healing, quota monitoring. Event-triggered and scheduled.
The Swarm	04	LeadAgent v9.0 + 12 AutonomousWorkers	Persistent conversational brain across tasks. Capability registry, 7-state FSM lifecycle, TDD auto-fix, branch isolation, acting-lead election.
The Pulse	05	FrankenTUI real-time telemetry	Sub-50ms terminal capture across all 24 CPU cores. Agent state grid, live task DAG, GPU VRAM allocation, token burn rate, deep-probe live view.
The Measure	06	AgentMetrics engine	12 core KPIs: TSR, AEI, MTTC, escalation rate, retry rate, flywheel acceleration. Prometheus + Langfuse backend. Feeds the RL feedback loop.
The Sentinel	07	Background AI operations layer	Runs on whichever accelerator is present (NPU, iGPU, GPU, or CPU) as an unkillable service. 18 integration points, Qdrant vector cache, Guardian pattern, catastrophic recovery.

02 SUPPORTED ACCELERATORS

Runs on Any Silicon You Have

Supports: CPU, NPU, iGPU, GPU (CUDA 12.8+, sm_120, blackwall certified). No pillar or module is pinned to a specific accelerator. The Sentinel, the Gatekeeper, and every inference worker run on whichever class is present.

CLASS 01

CPU

Fallback path for every component. Runs on any modern x86_64 or ARM64 core. Used when accelerator capacity is saturated, when a task is memory-bound rather than compute-bound, or when the deployment target has no dedicated AI accelerator at all. Every pillar and every module can run on CPU-only hardware, including the Sentinel and the Gatekeeper. Reduced throughput, full functional parity.

CLASS 02

NPU

Neural Processing Unit. Dedicated AI inference silicon such as Intel AI Boost, Apple Neural Engine, or Qualcomm Hexagon. 10-40 TOPS INT8 typical. Zero marginal cost, always-on, unkillable, and thermally insulated from CPU and GPU pressure. Ideal home for the Sentinel background-operations layer when the target ships one — otherwise the Sentinel runs on another class.

CLASS 03

iGPU

Integrated GPU (Intel Xe-LPG, AMD RDNA integrated, Apple unified memory). Shared with CPU over a unified memory pool. Excellent for batch triage, KV-cache reuse, low-latency embeddings, and any pillar that benefits from dedicated graphics compute without a discrete GPU. OpenVINO GPU plugin, SYCL, and Level-Zero backends supported. A natural home for the Gatekeeper when discrete silicon is absent.

CLASS 04

GPU

NVIDIA discrete GPU. **CUDA 12.8+, compute capability sm_120, blackwall-architecture certified.** Highest throughput option for heavy inference and training workloads. vLLM, SGLang, and Triton backends all supported. Multi-GPU tensor parallelism for 70B+ models. The GPU Scheduler handles VRAM budget enforcement, priority queueing, and hot-swap model loading without OOM risk.

03 THE MODULE LAYER

Six Composable Modules Beneath the Cockpit

Modules are sidecar subsystems that speak NATS and REST. Each does exactly one job. The core platform runs without any of them and degrades gracefully if one goes offline. All modules are versioned, SHA-256 verified, and manifest-discovered at runtime.

Module	Port / Status	Role	Description
AI Switch	:9100 PROD	Inference router	Multi-factor scoring over latency, KV-cache locality, cost, load, and reliability. Per-endpoint circuit breakers with exponential backoff. Cache-aware least-latency strategy. 10K+ routing decisions per second. Automatic failover.
AI Farm	:9200 PROD	Distributed worker pool	Central Controller plus per-node Worker Agents. Live capacity reporting. Model Placement Engine for hot-swap, redundancy, VRAM budget enforcement. vLLM, SGLang, Ollama, Triton backends all supported.
AI Vault	:9400 PLANNED	Model registry	Semantic versioning, SHA-256 verification, pull-through cache for public and private model registries. LoRA and QLoRA adapter catalog. Prompt template versioning with diff tracking. Rollback on regression.
AI Meter	:9300 PLANNED	Token metering + billing	Per-agent and per-team budgets with soft and hard enforcement. Sliding-window rate limits. Usage analytics, anomaly detection, cost forecasting. Integrates with AI Switch for automatic throttling.
AI Cache	:9500 PLANNED	Semantic + prefix cache	Qdrant vector lookup at cosine > 0.95 returns cached answers in <10ms. Prefix layer coordinates with vLLM/SGLang KV cache for shared system prompts. 60%+ typical cost reduction on repetitive workloads.
AI Graph	:8400 PROD	Knowledge graph engine	21 algorithms (Leiden, Katz, Bridging, Link Prediction, DBSCAN). Eight graph topologies including agent comm, task dependency, RCA causal chains. GraphRAG retrieval. MCP server with 22 tools.

04 PRODUCTS

The AIVR Product Family

AIVR ships as a suite of composable products under the AIVR Cloud Foundation bundle. Each is a first-class SKU with its own subdomain, docs, and release cadence. They share one auth system, one event fabric, and one audit trail.

Product	Status	Role	Description
AIVR Cockpit	v3.4 · CORE	Command center	Seven-pillar orchestration. Multi-agent execution, real-time telemetry, self-healing infrastructure. Keyboard-first operator UI. One pane of glass for the whole platform.
AIVR Forge	ON-PREM	Enterprise installer	Lifecycle manager (formerly HyperVisorXR). Provisions the full Cloud Foundation bundle on-prem. Role-based install: HyperVisorXR, Developer, SysAdmin, Minimal. Upgrades, rollbacks, health checks.
AIVR Office	LIVE	Multi-tenant biz ops	Websites, email routing, social dispatch, brand asset generation for multiple brands from one cockpit. Stage / Prod mode toggle with dual-key production gates. Manifest-as-code per client.
AIVR Market	LIVE	Token marketplace	market.aivr.site. Limit-order book with time-priority matching. 5% hold rate. GoTokes wallet integration for USD deposit and withdraw. Two-party approval required before settlement.
AIVR Wallet	LIVE	Central ledger	Token balances earned, purchased, and spent. Usage-based billing and enterprise invoicing across every AIVR product under one account. Integrates with AI Meter for real-time cost attribution.
AIVR Helix	FREE	File compression	Free public service. Under 1 GB in the browser. Powered by the Hutter Prize research engine. Zero-signup, zero-telemetry, single-click. Credibility anchor for the research lab.
AIVR Clients	AGENT / CLI / NODE	Endpoints	Agent is the mandatory local bridge and scratch workspace. CLI is the terminal and scripting surface. Node is optional mobile and desktop compute farming. All three pair via device-code flow.
auth.aivr.site	LIVE	Identity substrate	SSO, OAuth, device-code pairing. Shared across every AIVR product and partner subdomain. OIDC compatible, JWT-based, per-device revocation, 24h session tokens, MFA optional.

05 ARCHITECTURE & INTELLIGENCE

The Learning Loop, Under the Hood

AIVR is not a static agent platform. Every task outcome updates Playbook rule confidences. Every novel failure becomes an anti-pattern. Every experiment runs with statistical rigor. The system genuinely gets smarter.

Component	Ref	Category	Description
Write-Ahead Log	S43	Replay engine	Every agent action appended to a per-agent WAL and mirrored to the NATS REPLAY_LOG stream. Five modes: full, range, step-through, failure-focused, diff. 30-day retention.
Reward Signal	S16.3	6-component scorer	Tasks scored on task_success, retry_count, token_efficiency, time_efficiency, bug_density, escalation_depth. EMA updates performanceScore per agent.
Model Router + RL	S38-39	Learning loop	5-factor model scoring. Policies update only after 30+ samples and a p<0.05 significance threshold. Automatic rollback on regression detection.
A/B Experimentation	S40	Statistical rigor	NATS-based task forking. Parallel variants on isolated git branches. Two-sample t-test, Cohen's d effect size. Automatic winner declaration.
ADR Store	S41	Strategic memory	Architecture Decision Records in Qdrant with vector embeddings. Auto-injected into agent context for relevant tasks so agents honor architectural constraints.
Automated RCA	S42	Root cause loop	Novel failures trigger AI root-cause analysis with full WAL context. Output is a structured anti-pattern artifact. Confidence >0.8 auto-applied; others reviewed.
4-Layer Memory	S12	Knowledge flywheel	Procedural (Playbook), Working (summaries), Episodic (CASS), Conversation (FTS5). Per-project isolation. 8 KnowledgeArtifact types. Flywheel accelerates every run.
Self-Healing Loops	S15-16	5 recovery loops	Agent health recovery, provider failover, memory index recovery, Docker service recovery, WebSocket reconnection. Trauma Guard prevents repeat catastrophes.

06 SAFETY & COMPLIANCE

Five Layers of Containment

AIVR is built on the premise that models are non-deterministic and untrusted by default. Variance at the model layer is allowed; variance that would cost you data, money, or a regulatory finding is architecturally impossible.

Layer	Ref	Category	Description
DCG	S11	Destructive command guard	SIMD pattern matching blocks rm -rf, DROP TABLE, force-pushes, and 45+ destructive patterns in native code before the shell sees them. Latency measured in microseconds. 49+ YAML security packs.
SLB	S11	Safety Layer B	Four-tier risk classification: SAFE / CAUTION / DANGEROUS / CRITICAL. SHA-256 binding prevents retry attacks. Dual-key human authorization required on the CRITICAL tier.
Per-Agent Sandbox	S11.4	Kernel isolation	Filesystem jails, network namespaces, cgroups v2, branch-scoped git hooks. Blast radius bounded at the kernel. Per-agent temp dirs auto-cleaned on termination.
Sentinel Guardian	S47.5	Prerequisite gate	“Did You Consider?” pattern. Risky actions intercepted on NATS and gated against ADRs and prerequisites before execution. Runs on any available accelerator, 24/7.
Per-Project Isolation	S13	Tenancy	Every component that stores state is scoped by project: Qdrant collections, NATS subjects, file leases, knowledge artifacts, Redis keyspaces. Cross-project contamination is impossible.
EU Compliance	8 REGIMES	Compliant by construction	AI Act, GDPR, CRA, DSA, Data Act, DGA, NIS2, revised PLD. Every obligation maps to an existing AIVR component. See the Compliance Brief white paper for the full mapping.

07 INFRASTRUCTURE & DEPLOYMENT

Docker, Ports, Phases

AIVR deploys via a single Docker Compose file for infra services, plus native services for the Sentinel and Gatekeeper that survive container crashes. Everything else is containerized, healthchecked, and restart-policied.

Area	Ref	Category	Description
Docker Compose	S28	Full-stack deploy	NATS + JetStream, Redis, Qdrant, Ollama, Agent Mail, MkDocs + nginx, Prometheus, Langfuse. Every service healthchecked and restart-policied. Single compose up brings the stack online in <2 min.
Port Map	S23	Service registry	3456 Cockpit · 4222 NATS · 5678 N8N · 6333 Qdrant · 6379 Redis · 8080 Tabby · 9090 Prometheus · 11434-11438 inference devices · 9100-9500 modules.
21-Phase Roadmap	S34	Deployment guide	Safety, Nerve, Memory, Infra, Docs, Task Intel, Agent State, WebUI, Monitoring, Launch, Measure, Router, RL, Experiments, Strategic, RCA, Lifecycle, DR, Cost, Sentinel, Gatekeeper.
Disaster Recovery	S45	Backup + restore	Per-service playbooks. Qdrant snapshots, Redis BGSAVE, NATS stream backups, SQLite online .backup, CASS archive. RTO/RPO from 30s to 4h depending on scenario.
Reference Hardware	S3	Workstation	24-core CPU, 12 GB VRAM discrete GPU, an NPU, an integrated GPU, 96 GB DDR5-8200 with 48 GB unified memory pool. Enterprise targets adapt to whatever silicon is present.
Multi-Node Scale	S17	Horizontal	NATS clustering for quorum. Redis-backed shared state layer. Dedicated GPU Node C with vLLM for 4+ discrete GPUs. Adding nodes is configuration, not a rewrite.

08 RESOURCES & WHITE PAPERS

Where to Go Next

This brochure is a map. Every section has a dedicated document or product page at **www.aivr.site**.

WHITE PAPER

Designing for Variance

How AIVR treats non-determinism as an architectural property. Five containment layers around a stochastic core. WAL / Replay, reward signals, self-optimization, semantic cache. Nine pages.

whitepaper.aivr.site/AIVR-Non-Deter

WHITE PAPER

Compliant by Construction

Eight EU regimes — AI Act, GDPR, CRA, DSA, Data Act, DGA, NIS2, revised PLD — each mapped to concrete AIVR section references. Regulatory timeline. Seven pages.

whitepaper.aivr.site/AIVR-Complianc

RESEARCH

The Sovereign AI Platform

The flagship research white paper. New mathematics for a new era of compute. Throughput, energy, and efficiency benchmarks. The mathematics is proprietary.

whitepaper.aivr.site

ENTERPRISE

Enterprise Licensing

On-prem installation via AIVR Forge. SLA support, deployment services, license request forms. White-paper downloads for legal, procurement, IT security.

enterprise.aivr.site

DOCS

Technical Documentation

The full v3.4.0 architecture: 47 sections, 11 split parts, the 21-phase engineering roadmap, API references, component specifications.

docs.aivr.site

PLATFORM

Main Website

Product pages, platform overview, clients, partners, research, pricing, full company map. Start here if you are new to AIVR and want the 30-minute read.

www.aivr.site