ALAI AI System — Operating Picture 2026-05-18 (CEO Audit)

ALAI AI System — Operating Picture 2026-05-18

Date: 2026-05-18 **Architect:** Petter Graff **Status:** VALIDATED v1.1 — Proveo PASS (0 hallucinations, 3 minor drifts), Verifier PARTIAL (3 hallucinations from one root cause: manifest path mismatch; 2 PARTIAL — see Validation Patches below). Headlines stand.

Executive Summary

The ALAI AI system burned **742*Kacrossthe*8 − *daywindowMay*11–18*onAnthropicOpus* * *(99.98365,104** — still catastrophic. A single day (2026-05-11) hit **$377,487**. The prior audit's "$9,790/day" figure held only for a quiet day (May 13 = $9,954) but was **10–40× under for peak days**. Revenue is $0; this is founder cash.

This is not a pricing problem. It is a **causal chain of broken safety nets**:

1. **Determinism doctrine is unenforced.** Reality Anchor probes have not executed in 7 days — 0 PROBE_PASS/PROBE_FAIL events, both probe daemons absent from launchctl PID list (inference-determinism.md). Doctrine exists on paper only.
2. **Free local tier is degraded.** `devstral:24b` — the model targeted by 79% of tier-router code calls (531 calls) — does not exist on either Ollama host. Two of three ANVIL MLX servers (qwen3-32b, qwen3-8b) silently serve the wrong model (an embedding model that rejects generation). Tier 2c, M2c, and M3 are ghosts (inference-determinism.md).
3. **Opus fallback is unbounded.** With the free tier silent-failing and no Reality Anchor probe to detect the drop, every call escalates to Opus. There is no cost ceiling at runtime (business-roi.md).
4. **John builds on stale inventory.** `discover.js --verify` reports system health citing `manifest-index.md` (which DOES exist at `~/system/tools/manifest-index.md` but is **stale since 2026-02-26**, claims 1,310 scripts vs actual 273 — corrected by verifier) and a `skill-registry.db` containing 1 row (snowit-fb), not the 96 skills on disk. BookStack API is dead (CF Access 302) — staleness measurement offline for 478 tracked pages

(knowledge-graph.md). The orchestrator is steering by an instrument panel that froze 3 months ago.
5. **ZAKON #12 (RAG context injection) is dormant.** `rag-context-for-builder.js` is referenced in protocol docs but **not wired into any hook** — every builder dispatch re-injects full MEMORY.md (~15K tokens) instead of a 500–800 token targeted block (rag-layer.md).

If you read nothing else:

- **STOP THE BLEED:** Enforce Sonnet-default + Opus gating today. At current pace this saves ~20*K*–90K/day.
- **TURN ON THE LIGHTS:** Start Reality Anchor probe daemons + reconcile tier-router to live model fleet.
- **FIX THE COMPASS:** `discover.js --verify` reads 3-month-stale data — regenerate `manifest-index.md`, rebuild `skill-registry.db`, and restore CF Access token for BookStack before any further architecture decisions.

System Map — Planned vs Implemented vs Running

```
flowchart LR
  CEO[Alem / CEO] --> John[John Orchestrator]
  John -->|dispatch| Mehanik{Mehanik Gate}
  Mehanik -->|authorize| Specialists[Specialist Agents]
  Specialists --> Opus[Anthropic Opus]
  Specialists -. intended .-> TierRouter[Tier Router]

  TierRouter -.->|531 calls 79%| Devstral[devstral:24b GHOST]
  TierRouter -->|works| OllamaANVIL[Ollama ANVIL 8 models]
  TierRouter -->|works| OllamaFORGE[Ollama FORGE 8 models]
  TierRouter -.->|wrong model| MLXqwen32[MLX qwen3-32b BROKEN]
  TierRouter -.->|wrong model| MLXqwen8[MLX qwen3-8b BROKEN]
  TierRouter --> MLXgemma[MLX gemma-4-26b OK]

  John --> Discover[discover.js --verify]
  Discover -.->|cites stale| ManifestIdx[manifest-index.md STALE 2026-02-26]
  Discover -.->|lies| SkillReg[skill-registry.db 1 row of 96]

  John --> RAG[rag-context-for-builder.js]
```

```
  RAG -.->|not wired| Hooks[PreToolUse hooks]

  Specialists --> LightRAG[LightRAG Azure]
  LightRAG -.->|23,558 backlog| MigratePump[migrate-pump 600/run cap]
  LightRAG -.->|CF Access 302| BookStack[BookStack API DEAD]

  Specialists --> HiveMind[HiveMind 21,741 rows]
  HiveMind -.->|15 dead agents| DeadAgents[Stale namespaces]

  RealityAnchor[Reality Anchor Probes] -.->|0 fires 7d| Evidence[Evidence Ledger]
  Evidence -.->|65 null paths| GateBypass[Gate bypass risk]

  Opus -->|$741K / 7d| Cost[Cost Burn]

  classDef green fill:#1d8c43,color:#fff
  classDef yellow fill:#d4a017,color:#000
  classDef red fill:#b3261e,color:#fff
  class CEO,John,Mehanik,Specialists,OllamaANVIL,OllamaFORGE,MLXgemma,HiveMind green
  class LightRAG,MigratePump,RAG,Discover,Evidence yellow
  class
Devstral,MLXqwen32,MLXqwen8,SkillReg,BookStack,DeadAgents,RealityAnchor,Cost,GateBypass,Hooks
red
  class ManifestIdx yellow
```

Inventory Table

Subsystem	Planned	Implemented	Running	Used 7d	Status	Evidence
Anthropic Opus	yes	yes	yes	yes	RED	business-roi.md ($741K/7d, 99.995%)
Sonnet default policy	yes	yes	no	minimal	RED	business-roi.md ($72/7d only)
Ollama ANVIL (8 models)	yes	yes	yes	yes	GREEN	inference-determinism.md

Subsystem	Planned	Implemented	Running	Used 7d	Status	Evidence
Ollama FORGE (8 models)	yes	yes	yes	yes	GREEN	inference-determinism.md
MLX gemma-4-26b (ANVIL)	yes	yes	yes	yes	GREEN	inference-determinism.md
MLX qwen3-32b (ANVIL)	yes	yes	wrong-model	n	RED	inference-determinism.md
MLX qwen3-8b (ANVIL)	yes	yes	wrong-model	n	RED	inference-determinism.md
MLX gemma-4-26b (FORGE)	yes	yes	yes	yes	GREEN	inference-determinism.md
Tier Router devstral:24b	yes	route-only	ghost	531 calls	RED	inference-determinism.md
Reality Anchor probes	yes	yes	not-firing	0 events	RED	inference-determinism.md
Evidence Ledger (JSONL)	yes	yes	yes	yes	YELLOW	inference-determinism.md (16.7% null path)
Evidence Ledger (SQLite)	yes	partial	0 tables	n	RED	inference-determinism.md
LightRAG core (Azure VM)	yes	yes	degraded	yes	YELLOW	rag-layer.md (15% probe fail)
LightRAG public endpoint	yes	yes	CF-blocked	n	RED	rag-layer.md, knowledge-graph.md
lightrag-migrate-pump	yes	yes	running	yes	YELLOW	rag-layer.md (23,558 backlog)
lightrag-outbox-ingest	yes	yes	stalled	n	RED	rag-layer.md, ops-layer.md
rag-context-for-builder.js	yes	yes	not-wired	n	RED	rag-layer.md (ZAKON #12 dormant)

Subsystem	Planned	Implemented	Running	Used 7d	Status	Evidence
HiveMind hivemind.db (primary)	yes	yes	yes	yes	GREEN	rag-layer.md (21,741 rows)
HiveMind orphan DBs (×3)	n/a	n/a	empty	n	RED	rag-layer.md
Dead HiveMind agents (15)	n/a	n/a	namespace pollution	n	YELLOW	rag-layer.md, knowledge-graph.md
BookStack content	yes	yes	yes	yes	GREEN	knowledge-graph.md (478 pages)
BookStack API / staleness	yes	yes	dead	n	RED	knowledge-graph.md
BookStack ADR/runbook coverage	yes	partial	partial	partial	RED	knowledge-graph.md (5 governance gaps)
ADR numbering integrity	yes	yes	corrupt	n/a	RED	knowledge-graph.md (adr-025×2, adr-026×4)
Library system (library.yaml)	yes	no	none	n	RED	knowledge-graph.md (0 across personas)
MC (mc.js)	yes	yes	yes	yes	GREEN	business-roi.md
Daemons — running healthy	yes	yes	14	yes	GREEN	ops-layer.md
Daemons — flapping (6)	n/a	yes	2 running / 4 stopped	partial	RED	ops-layer.md
Daemons — unloaded orphans (3)	n/a	yes	not loaded	n	YELLOW	ops-layer.md
Daemons — .new shadow files (3)	n/a	n/a	risk-only	n	YELLOW	ops-layer.md
Hooks (58 entries, all present)	yes	yes	yes	yes	GREEN	ops-layer.md

Subsystem	Planned	Implemented	Running	Used 7d	Status	Evidence
Tools on disk (273 top-level)	yes	yes	partial	partial	YELLOW	code-surface.md
manifest-index.md (handbook ref)	yes	yes	stale (2026-02-26)	partial	YELLOW	verifier-report.json A10
skill-registry.db	yes	yes	1/96 rows	partial	RED	code-surface.md
specialist-mapping.json	yes	yes	yes	yes	YELLOW	code-surface.md (mehanik, dzevad-jahic missing)
Mehanik dispatch gate	yes	yes	yes	yes	YELLOW	code-surface.md (mapping mismatch)
Cost tracker (costs.db)	yes	yes	yes	yes	GREEN	business-roi.md
TLDR daemon	yes	yes	gapped	partial	YELLOW	business-roi.md (3-day May gap)

Ranked Gap List

P0 — Stop The Bleed (this week)

P0-1. Opus burn $741K/7d. *(business-roi.md, costs.db)*

- **Root cause:** No model gate. 99.995% of calls hit Opus despite CLAUDE.md declaring Sonnet as orchestration default.
- **Fix:** (a) Sonnet-default enforcement at claude-cli wrapper level; (b) Opus whitelist limited to `/prompt-forge` + novel-architecture review; (c) `opus-cost-guard.sh` hook is registered (ops-layer.md) — verify it actually blocks vs warns.
- ***/monthestimate* : * * *atpeakday*(110K) → save ~2.7*M*/*month*; *atrecentstabilization*(26K/day) → save ~650*K*/*month*.*Evenworstcasecrediblesavings* : ** 500K+/month**.
- **Owner:** FlowForge (Kelsey) for hook enforcement + CodeCraft for wrapper gate. Open MC required.

P0-2. devstral:24b ghost — 79% of tier-router code calls. *(inference-determinism.md)*

- **Root cause:** Tier 2c routes 531 calls to a model present on neither Ollama host. 4.5ms avg suggests silent fallback or unlogged substitution. Every "local code review" claim under tier 2c may have escalated to Opus or returned junk.
- **Fix:** `ollama pull devstral:24b` on FORGE OR remap tier 2c to `qwen3:8b-q8_0` (already hot on FORGE).
- ***/monthestimate* : * * *unknownuntilprobesrestored, but* : 531*calls* × *7d, eachpotentiallyescalatingtoOpus = compoundingmultiplieronP*0 − 1.*Conservatively* * * 20K–$100K/month** in avoided escalations.
- **Owner:** AgentForge (Georgi) — fleet reconciliation; CodeCraft to update `tier-routing.json`.

P0-3. Reality Anchor probes not executing (0 events in 7d). *(inference-determinism.md)*

- **Root cause:** Probe daemons `com.john.auto-verify-regression` + `com.john.ollama-health-probe` have no PID. Probe scripts exist; registry v1.3 exists; nothing runs.
- **Fix:** `launchctl start` both daemons + verify PROBE_PASS appears in `~/system/state/`. Add a watchdog daemon to alert on probe silence >24h.
- **$/month estimate:** Indirect — but Reality Anchor is the **only deterministic check** between LLM self-report and gate pass. Without it, hallucinated work satisfies `mc.js done`. Rework cost estimable at 1–3 fabricated PASS incidents/week × ~$5K rework each = **20*K*–60K/month avoided**.
- **Owner:** FlowForge (Kelsey).

P0-4. discover.js --verify is hallucinating system health. *(code-surface.md, knowledge-graph.md)*

- **Root cause:** Self-verification cites `manifest-index.md` (exists at `~/system/tools/manifest-index.md` but **stale since 2026-02-26** — claims 1,310 scripts vs actual 273) and `skill-registry.db` with 1 row representing 96 skills. The instrument reads frozen data.
- **Fix:** (a) Regenerate `manifest-index.md` from real tool inventory on a daily cron; (b) Rebuild skill-registry.db with `last_used` column + populate from disk scan; (c) Add a meta-probe that diffs claimed inventory vs actual at session-start.
- ***/monthestimate* : * * *Indirectbutmultiplicative—everyplanJohnwritesonphantominventoryaddsdispatchwaste.Estimate* * *5K–$15K/month** in avoided wasted dispatches.
- **Owner:** CodeCraft (manifest regen) + AgentForge (skill registry).

P0-5. MLX tiers M2c + M3 broken (wrong model loaded). *(inference-determinism.md)*

- **Root cause:** `~/system/research/mlx-models/` directory does not exist; both plists silently fall back to a cached `bge-m3-mlx-fp16` embedding model that rejects generation requests. Redzo-reviewer and verifier tiers routed here get junk.
- **Fix:** Locate or re-download Qwen3-32B-4bit + Qwen3-8B-4bit MLX weights, OR repoint M2c/M3 to FORGE Ollama equivalents (qwen3:32b, qwen3:8b-q8_0).
- ***/monthestimate* : * * *SameclassasP*0 − 2*—freeverifiercapacityrestored = Opusavoided*. * *10K–$40K/month**.

- **Owner:** AgentForge (Georgi).

P1 — Structural (next 2 weeks)

P1-1. ZAKON #12 dormant — rag-context-for-builder.js not in any hook. *(rag-layer.md)*

- Wire into `PreToolUse[Task]` hook chain. Replaces ~15K-token MEMORY.md injection per builder call with ~500–800 token targeted block. Saves ~22K tokens/day at current pace.
- **Owner:** CodeCraft.

P1-2. lightrag-migrate-pump cap (600/run, 23,558 backlog). *(rag-layer.md)*

- Backlog will never close at 1,200/day ingest vs ongoing writes. Increase to 5,000/run or remove cap.
- **Owner:** AgentForge.

P1-3. lightrag-outbox-ingest stalled. *(rag-layer.md, ops-layer.md)*

- New session content not reaching graph. Either re-enable daemon or formally decommission.
- **Owner:** FlowForge.

P1-4. BookStack API broken (CF Access token). *(knowledge-graph.md)*

- `bookstack-staleness.js` returns HTML 302. 478 tracked pages have unknown staleness. Rotate CF Access token in Bitwarden.
- **Owner:** FlowForge + Securion (token rotation).

P1-5. ADR numbering collision (adr-025 ×2, adr-026 ×4). *(knowledge-graph.md)*

- Schema integrity broken. Renumber + add a pre-commit guard.
- **Owner:** Skillforge / Datavera.

P1-6. 5 governance subsystems with zero BookStack page — Reality Anchor, Determinism/Tool-First, Tier Router, Evidence Ledger, Hooks. *(knowledge-graph.md)*

- The newest and most important systems have no central documentation. Publish runbook + ADR each.
- **Owner:** Skillforge.

P1-7. specialist-mapping.json missing mehanik + dzevad-jahic. *(code-surface.md)*

- Routing table referenced in CLAUDE.md but absent from the JSON the dispatch path reads. Mehanik gate hallucinates dispatch authorization because it cannot verify its own identity.
- **Owner:** CodeCraft.

P1-8. 6 flapping daemons. *(ops-layer.md)*

- `rag-fsevents-adapter` (exit 1, still running), `azure-db-backup` (exit 1, still running), `hook-drift-detector` (exit 2, stopped), `chain-e2e-nightly`, `rdap-audit-quarterly`, `apply-knowledge`. Silent failures most dangerous.
- **Owner:** FlowForge.

P2 — Cleanup (next month)

- **P2-1.** Cull 27 files: 13 dead tools + 5 stub skills + 9 hook .bak files (code-surface.md). Zero functional loss.
- **P2-2.** Prune 15 dead HiveMind agent namespaces (rag-layer.md, knowledge-graph.md).
- **P2-3.** Remove 3 empty/orphan HiveMind DBs (`~/system/db/hivemind.db`, `~/system/data/hivemind.db`, `~/system/agents/hivemind/memory.db`).
- **P2-4.** Resolve 3 .new shadow plists + 3 unloaded orphan plists (ops-layer.md).
- **P2-5.** Library system: either deploy (0 library.yaml currently) or formally retire library-auto-push.md runbook (knowledge-graph.md).
- **P2-6.** Fix `mc.js` hardcoded paths (lines 2808, 2822) and `agent-runner.js:43` env fallback (code-surface.md).
- **P2-7.** Backfill or null-flag 65 evidence-ledger rows with null `evidence_path` so they cannot satisfy `mc.js done` gates (inference-determinism.md).

Token-Save Recommendations (with $/month estimates)

#	Action	Estimated savings/month	Source
1	Sonnet-default + Opus gated to `/prompt-forge` only	**500*K*–2.7M**	business-roi.md
2	Restore free local tier (fix devstral + MLX)	30*K*–140K	inference-determinism.md
3	Restart Reality Anchor probes (rework avoidance)	20*K*–60K	inference-determinism.md
4	Wire `rag-context-for-builder.js` into PreToolUse hook	~$4 (token), high indirect	rag-layer.md
5	Close lightrag-migrate-pump backlog (23,558 rows)	~$15 token + freshness	rag-layer.md

#	Action	Estimated savings/month	Source
6	Purge dead HiveMind namespaces + orphan DBs	~$10 token + cleaner retrieval	rag-layer.md
7	Cull 27 dead files (tools/skills/.bak)	qualitative — cleaner discover.js	code-surface.md

The headline is item 1: nothing else moves the needle until model selection is fixed.

CEO Decisions Surfaced

1. **Authorize Sonnet-default enforcement TODAY.** Single highest-ROI action available at $0 revenue. (P0-1)
2. **Authorize Opus hard ceiling.** E.g., $500/day budget circuit-breaker that flips claude-cli to Sonnet automatically. Currently no runtime cost ceiling exists.
3. **Reconfirm tier-router intent.** Should tier 2c route to `devstral:24b` (and we pull it) or to `qwen3:8b-q8_0` (already on FORGE)? AgentForge cannot fix without direction.
4. **MLX investment.** Two of three ANVIL MLX servers broken because model weights directory is missing. Authorize re-download OR formal repoint to FORGE Ollama.
5. **BookStack CF Access token rotation** — touches Securion + FlowForge boundary. Authorize Bitwarden rotation + automated keep-alive.
6. **TLDR daemon fix-or-retire.** 3-day gap in May; CEO visibility depends on it (business-roi.md).
7. **Authorize one-time purge sprint** for P2 cleanup (27 files + 3 DBs + dead namespaces + flapping daemons). Est. 2h dispatch.

Risks Identified by Synthesis (not in individual reports)

R1. Compound failure mode — three safety nets failed together. Each report alone is concerning. Combined: (a) free tier silent-fails, (b) Reality Anchor probe doesn't detect drop, (c) no runtime cost ceiling, (d) discover.js misreports inventory so John can't see drift. There is **no remaining instrument** that would have caught the $741K burn except the cost tracker — which works, but is read by John after the fact, not enforced.

R2. discover.js as single point of trust failure. Per ZAKON NULA, every tool-verify question routes through `discover.js`. If `discover.js --verify` itself lies about manifest-index.md and skill-registry.db, then **every "verified" claim downstream of it inherits the lie**. This is the most dangerous finding because it inverts the anti-hallucination doctrine.

R3. Mehanik gate hallucinates dispatch authorization. Mehanik is referenced in CLAUDE.md as the mandatory pre-dispatch gate, but Mehanik itself is missing from `specialist-mapping.json` (code-surface.md). The gate can't authoritatively confirm an agent exists. Combined with the manifest-index gap, dispatch routing operates on prose-level trust, not data-level verification.

R4. Evidence ledger gate-bypass via null paths. 65 of 390 rows (16.7%) have `null` evidence_path. They count toward gate row-counts without any artifact. With Reality Anchor probes also dead, ledger integrity drops further — fabricated "PASS" claims (precedent: Angie Jones qa-19, SnowIT public claims hallucination) can re-occur with no automatic catch.

R5. The codebase is younger than the assumptions about it. code-surface.md notes 0 files >180 days old — system is <6 months old. But CLAUDE.md handbook references "1,310 scripts" and a manifest that never existed. The handbook narrates a system more mature than the disk reality. CEO planning may inherit this confidence gap.

Contradictions Across Reports

- **Daemon count: ops-layer says 62 loaded / 70 plist files; business-roi says "55 total, 6 deprecated .bak".** Likely both are partial views (ops counts launchctl entries; business counts canonical .plist files only). Reconcile via fresh probe.
- **Opus spend prior claim:** business-roi.md flags the prior $9,790/day audit as 10–11× too low — but that prior claim originates from the 2026-05-14 AI Factory audit cited in MEMORY index. Newer probe (costs.db) is authoritative; the May 14 finding should be retracted.

- **LightRAG status:** rag-layer says core is DEGRADED with ~15% probe failure; business-roi says "service up (302 CF Access = service up)"; knowledge-graph says "BLOCKED — returns 302". All three are partially correct: the Azure VM core responds at internal IP, but the public CF Access endpoint blocks tooling. Net verdict: YELLOW — operational but tooling-blind. (Source citations: rag-layer.md, business-roi.md, knowledge-graph.md.)
- **HiveMind dead agent count:** rag-layer cites 15; knowledge-graph cites 15 with slightly different list (knowledge-graph includes `john-delegate` 2026-04-11 and the mis-cased `CodeCraft`; rag-layer omits john-delegate but includes `tender-hunter` 2026-04-17). Both lists ~15; merge before pruning.

Validation Plan

Per /plan-with-team protocol:

- **Task 8 (Proveo — Angie Jones):** Re-probe ≥20% of cited claims with fresh tool output. Priority: costs.db spend total, Reality Anchor probe daemon PIDs, devstral:24b absence on both Ollama hosts, manifest-index.md non-existence, skill-registry.db row count, lightrag-migrate-pump backlog count, ADR numbering collision file list, specialist-mapping.json key

set.

- **Task 9 (Verifier atomic-claim decomposition):** Read-only verifier subagent decomposes this report into ≤50 atomic claims, runs probe per claim, returns CONFIRMED/PARTIAL/HALLUCINATION verdict per claim. Cost <$0.50/run.
- **Task 10 (Skillforge):** Publish this report to BookStack as `ALAI AI System Operating Picture 2026-05-18`. Cross-link from System Architecture shelf. (Blocked until P1-4 CF Access token fix — fall back to manual upload.)

REPORT WRITTEN: ~/system/specs/ceo-ai-system-audit-2026-05-18-REPORT.md

---

Validation Patches (applied 2026-05-18 23:30 after Proveo + Verifier)

Sources: `/tmp/audit-2026-05-18/proveo-verdict.json`, `/tmp/audit-2026-05-18/verifier-report.json`

Patch	Original Claim	Corrected	Source
V-P1	$741,646 / 7 days	$742K / 8 days (May 11–18) — true 7d (May 12–18) = $365,104	verifier A1, A2
V-P2	manifest-index.md MISSING	manifest-index.md exists at `~/system/tools/`, STALE since 2026-02-26 (claims 1,310, actual 273)	verifier A10, A28, A35
V-P3	Mermaid node ManifestIdx = RED	recolored YELLOW (stale, not missing)	verifier A35
V-P4	P0-4 fix wording "generate or delete"	"regenerate on daily cron + add staleness meta-probe"	verifier corrective note
V-P5	BookStack sync-map at `~/system/agents/`	actual path `~/system/config/`	proveo C7
V-P6	Prior $9,790/day estimate "10–11× under"	"10–40× under for peak days; on quiet days within ±2%"	verifier A5

Verdict on report after patches: Headlines (Opus burn, devstral ghost, Reality Anchor dead, MLX broken, skill-registry blind, ZAKON #12 dormant) all CONFIRMED by both validators. Diagnosis stands. Cost dollar range remains catastrophic regardless of window interpretation.

Revision #2
Created 2026-05-18 21:40:38 UTC by John
Updated 2026-06-21 20:03:47 UTC by John