Aiden Final Forensic Review Packet

Purpose:

- Assemble the final review packet for technical scrutiny
- Keep the strongest validated claims visible
- Keep measurement-gated gaps visible
- Avoid overclaims around OS replacement, air-gapped, autonomous writes, etc.

Operating rule:

Lead with validated orchestration and governed-write evidence.

Treat hardware internals and production security posture as measured/hardened only when evidence exists.

1. Executive Narrative

Aiden is positioned as an enterprise AI Digital-Twin Control Tower built around GhostOS, a custom orchestration and governance runtime above local/private inference.

The current evidence supports this core claim:

- GhostOS controls:
  - Profile routing, lane selection, persona/memory hydration, prompt and evidence
  - Boundaries, tool permissions, deterministic checks, generation envelope, output finalization
  - Regulated workflow gates

The current evidence also supports these operational safety claims:

- The model is used for language and synthesis
- Deterministic code owns tools, schemas, preflight, approval, execution, confirmation, evidence handling, and status writeback.

The strongest current regulated-operations proof is Aiden v4 Metrc:

- Aiden v4 completed repeated and consistent governed sandbox clone-to-package lifecycles and evidence-review flows, using Metrc as compliance source of truth and Smartsheet as the operating cockpit.

The latest Aiden v4 pull-and-review proof is:

- Intent-aware Metrc Pull planning selects the appropriate evidence path for exact package detail, exact harvest detail, harvest-to-package resolution, or explicit Chain-of-Custody lineage review; the Analyst uses deterministic cultivation-audit signals to surface track-and-trace data drift.

Safe current commercial statement:

- Aiden efficiently helps operators identify and reduce track-and-trace data-drift risk before regulator, partner, or audit scrutiny.

California market-watch framing:

- California DCC Track-and-Trace Updates and Group Tagging rulemaking activity increases the commercial relevance of clean UID/tag, plant-batch, harvest, package, strain, quantity, date, COA, transfer, and source-field reconciliation. The projected February 2027 “Automated Flagging” hurdle should be handled as a planning/watch item unless independently confirmed by a public source, counsel, or client source.

The careful boundary:

- This is a strong governed proof/MVP posture. It is not a production security certification, absolute air-gap claim, custom Linux replacement, or proof of direct low-level hardware scheduling yet.

2. Validated Claim Posture

Proven Claims

1. GhostOS is Aiden's orchestration/governance runtime above inference.

2. The active Aiden v4 request path separates SvelteKit UI, kernel/orchestration API, and local inference endpoint.

3. Generation routes through a local OpenAI-compatible inference endpoint on 8000.

4. The last captured local endpoint reported Qwen2.5-14B-Instruct with a 16,000-token window.

5. Historical PMO records and launch artifacts show that the local vLLM serving path has operated on Intel XPU hardware.

6. Historical local XMX stress telemetry recorded greater-than-200-TFLOPS-class GEMM performance; this is raw accelerator stress evidence, not end-to-end LLM throughput.

7. Deterministic code owns Metrc and Smartsheet tool workflows.

8. Analyst and Specialist narrative lanes are deterministic data read-only for regulated writes.

9. Metrc mutations use Action Queue gates: plan, preflight, approval token, execute, confirm, and evidence writeback.

10. Smartsheet is the operating cockpit/evidence surface; Metrc is the compliance source of truth.

11. Colliers, Rize, and Aiden v4 demonstrate a reusable control pattern with different maturity levels.

12. App-level request telemetry exists for recent Colliers and Aiden v4 runs.

13. Aiden demonstrates a reusable sovereign/private AI control pattern across multiple domain lanes (Refactoring).

14. Demo evidence retention exists through casefiles, diagnostics, and Smartsheet writebacks.

15. Demo secrets (i.e., API Keys) are server-side/local.

16. Recent request telemetry is useful for demo behavior.

Measurement-Gated Claims still to be fully validated

1. Exact serving dtype or quantization.

2. Model-server launch command. Latest captured launch command exists; fresh recapture is recommended before hardware-specific external claims.

3. Device placement.

4. Level Zero initialization path inside the active serving process.

5. HBM/device memory residency.

6. KV-cache sizing, allocation, dtype, and fragmentation behavior.

7. Multi-device or tile distribution.

8. GPU/EU utilization.

9. Memory pressure under representative prompts.

10. Power, frequency, and temperature behavior.

11. Repeatable benchmark baseline under fixed prompts and known concurrency.

3. Architecture Summary

Canonical Aiden v4 Runtime Path

WorkApps / browser:

-> HTTPS reverse proxy / Cloudflare tunnel posture

-> SvelteKit UI on 5176

-> /api/kernel/execute

-> GhostOS kernel API on 8003

-> kernel_api_server.py

-> kernel_orchestrator.py

-> GhostOS_kernel.py

-> local OpenAI-compatible inference endpoint on 8000

-> kernel_finalizer.py

-> UI response metadata and rendered answer

Governed Metrc Write Path

Smartsheet staged row:

-> deterministic action parsing

-> required-field validation

-> plan

-> preflight

-> human approval token

-> execute

-> confirmation GET / readback

-> Smartsheet status and evidence writeback

Safe Architecture Statement

- Model for language and synthesis.
- Code for tools, state transitions, evidence, and write gates.

4. Platform Maturity Map

Instance	Current Position	Safe External Statement	Caveat
Colliers	Mature CRE underwriting/advisory proof lane.	"Colliers remain a mature CRE proof lane with saved casefile evidence, deterministic table/calc artifacts, and recent app-level telemetry."	Do not call it fully production hardened.
Rize	Investor/evidence workflow source lane and refactorability proof.	"Rize demonstrates reusable investor/evidence workflow foundations and supplied patterns carried into Aiden v4."	Needs fresh smoke before equal maturity positioning.
Aiden v4 Metrc	Strongest current regulated-operations proof lane.	"Aiden v4 completed repeated and consistent governed sandbox clone-to-package lifecycles and evidence-review flows."	Sandbox/demo posture; production hardening remains.

Aiden is not a single chatbot demo. The same orchestration pattern has been applied across CRE underwriting, investor/evidence workflows, and regulated cannabis operations. The shared pattern is GhostOS orchestration plus domain-specific UI, tools, memory, evidence schema, and workflows.

5. Security And Safety Position

Current Safety Controls

- Narrative lanes are read-only for regulated writes.
- Tool autonomy is disabled in Metrc narrative contexts.
- Action Queue supports finite action types.
- Execution requires an unlocked row, preflight_pass, required fields, allowed approval status, row approval token, request approval token, token match, and a known execute tool.
- Execute reruns plan and preflight.
- Confirm readback runs after execute.
- Status and evidence are written back to Smartsheet.
- Repair/status repair path is confirmation-oriented and does not execute mutations.
- Metrc MCP rejects unsupported tool names.
- Smartsheet mirror writes are schema-gated and dry-run by default.
- WorkApps embedding uses scoped frame-ancestor CSP handling.

6. Inference And Hardware Position

What Is Known

The last captured local runtime check showed:

- 8003 health: ok
- 8000 model endpoint: reachable
- Model id: Qwen2.5-14B-Instruct
- Model root: /home/erba/models/Qwen2.5-14B-Instruct/max_model_len: 16000

Aiden routes generation through a private local OpenAI-compatible endpoint. GhostOS controls the request envelope above that endpoint.

At the application boundary, Aiden is not coupled to NVIDIA CUDA APIs or an external cloud model API.

The host and codebase include Intel XPU / Level Zero / oneAPI preparation paths, and historical PMO records show that the local vLLM serving path has operated on Intel XPU hardware. Low-level current serving behavior should still be stated from runtime and profiler evidence.

Measurement-Gated Claims still to be fully validated

- Fresh 8000 model endpoint snapshot.
- Fresh service/port snapshot.
- Model-server launch command.
- dtype/quantization.
- Device/backend placement.
- GPU inventory.
- Utilization during representative prompt.
- Memory footprint before model load, after model load, and during active prompt.
- Power/frequency/temperature under repeated prompts.
- Repeat benchmark results with fixed prompt set.
- Profiler or runtime logs for Level Zero/device initialization.
- KV-cache behavior from backend logs/config.

7. Review Strategy

Pattern:

- What we can say now
- What that does not mean
- What we still need to prove

Example:

- What we can say now: Aiden routes generation through a local OpenAI-compatible endpoint and GhostOS controls the enterprise request envelope above it.
- What does not mean: We are not claiming a custom hardware OS or direct Level Zero command scheduling from GhostOS.
- What we still need to prove: dtype, quantization, device placement, hardware utilization, memory behavior, and thermal behavior need runtime/profiler evidence.

8. Rapid Answer Cards

GhostOS

GhostOS is Aiden's orchestration/governance layer above inference. It is not a Linux replacement.

Determinism

The model handles language and synthesis. Deterministic code handles routing, tools, schemas, preflight, approval, confirmation, evidence, and finalization.

Metrc Writes

Chat does not directly mutate Metrc. Regulated writes use Action Queue: plan, preflight, approval token, execute, confirm, and evidence writeback.

Inference

Generation routes through a local OpenAI-compatible endpoint on 8000. The last captured model was Qwen2.5-14B-Instruct with a 16,000-token window. Low-level serving details need runtime capture.

Air Gap

No absolute air-gap claim. The accurate claim is local/private LLM inference, plus governed external connectors to Smartsheet, Metrc, WorkApps, and Cloudflare.

Security

Strong governed sandbox controls exist, and pilot RBAC/secrets/evidence-export/incident runbook artifacts are now documented. Production still needs managed secrets, immutable audit, origin/CSRF controls, monitoring, backups, tenant isolation, enterprise incident response, and compliance review.

Platform

Colliers proves mature CRE workflow, Rize proves reusable investor/evidence foundations, and Aiden v4 Metrc proves the strongest current regulated-operations lane.

9. Review-Day Evidence Checklist

Completed evidence:

- Colliers 3/3 repeated proof runs.
- Aiden v4 3/3 repeated evidence-review runs.
- Fresh 5176 / 8003 / 8000 service and port snapshot.
- Fresh 8000 model endpoint snapshot.
- Redacted v4 env-key inventory.
- v4 Operation Session server-side Diagnostics persistence.

- Day 1 closeout packet: i.e., Supporting_Artifacts/Evidence_Closeouts/day1_evidence_sprint_closeout_2026-09-02.md.

Suggested add-ons:

- Model-server launch command.
- dtype/quantization proof.
- GPU inventory and topology.
- xpu-smi or Intel profiler trace during representative prompt.
- Memory baseline/model-loaded/active-prompt snapshots.
- Power/frequency/temperature logs.
- Repeat benchmark script and output.
- Immutable audit design.
- Provider-side least-privilege review.
- Managed secret/vault design.

10. Evidence Appendix

Live testing closed with six documented runs:

Lane	Runs	Result	Evidence
Colliers	3	PASS	Each run completed three Smartsheet fetches, LOA Calc, rate shock, Analyst, and Specialist Advisor.
Aiden v4 Metrc	3	PASS	Each run persisted seven Diagnostics GRID artifacts and preserved Analyst/Specialist chat history.

Key claim upgrades:

- Colliers have repeated saved casefile evidence, not only historical single-case evidence.
- Aiden v4 Operation Sessions persist Diagnostics pull artifacts server-side.
- Recent app-level request telemetry exists across all runs.
- v4 runtime checks confirmed UI/kernel/inference services live after testing.

Live testing also closed the planned negative safety matrix:

Area	Result	Evidence
Specialist direct-write refusal	PASS	Specialist refused direct Metrc update and redirected to dry-run, preflight, human approval, execution, confirmation, and evidence write-back.
Locked row execute block	PASS	Locked rows produced 0 runnable, 4 discovered, 4 ignored, 0 rows written.
Missing approval token	PASS	Missing row token returned EXECUTION_BLOCKED and no mutation.
Token mismatch	PASS	Mismatched request token returned EXECUTION_BLOCKED and no mutation.
Missing required fields	PASS	Missing Plant Batch Id and Plant Batch Name returned REVIEW_REQUIRED and no mutation.
Unsupported action	PASS	unsupported_test_action did not map into executable Action Queue tools.
Execute after failed preflight	PASS	REVIEW_REQUIRED row could not execute even with a valid matching token.
Empty pull push refusal	PASS by prior evidence	Prior metrc.plant_batches.list_active empty-state forensics showed no Smartsheet push for empty source rows; rerun skipped to avoid unnecessary shared-sandbox tag consumption.

Throughput Notes:

- Specialist short prompt: 22.38 t/s.
- Specialist normal prompt: 20.95 t/s.
- Analyst evidence prompt: 10.00 t/s.
- Stable >=20 t/s across all v4 lanes is not fully proven yet.
- Hardware/runtime packet created on 2026-09-03: hardware_runtime_measurement_packet_2026-09-03.md.
- Active 8000 launch captured: Qwen2.5-14B-Instruct on XPU with PP=2/TP=1, max_model_len=16000, gpu_memory_utilization=0.7, and enforce_eager.
- Host baseline did not show CPU or RAM pressure: 96 logical CPUs, about 469 GiB available RAM, no swap use, load average near zero.
- PCI/OpenCL confirm two Intel Data Center GPU Max 1100 devices; xpu-smi discovery remains unresolved because it returned No device discovered.
- Follow-on benchmarks showed clean v4 Specialist paths can clear >=20 t/s near 4.4k-5.1k input tokens, while evidence-heavy v4 runs near 7.6k-7.7k dropped to 7.04-10.08 t/s.
- Colliers has a high fixed input floor around 7.76k-7.83k tokens even for ultra-short prompts, suggesting default injected context/retrieval payload should be pruned before runtime changes.
- Compact inline evidence lowered input to about 4.53k tokens but still ran at 13.88-15.95 t/s, so analytical chain-of-custody work should be tracked by elapsed time and streaming quality as well as raw t/s.
- 2026-09-04 token/context optimization significantly reduced input size, without any quality losses on several routes.

11. Pilot-Hardening Addendum

Completed on 2026-09-03:

- Pilot forensics captured auth, RBAC, route, evidence, and secrets posture before execution.
- Auth/RBAC implemented and locally verified:
  - dhl@erba.tech: Admin.
  - viv@erba.tech: Operator.
- Operator can use MCP Librarian GETs, Metrc Pull, Smartsheet Pull, Data Shuttle, and Action Queue preview/preflight.
- Operator cannot execute/repair Metrc actions:
  - Direct Metrc execute through the generic MCP bridge is blocked; Action Queue remains the governed mutation path.
- Pilot scope documented as a narrow cultivation starting scope with a repeatable expansion path for Distribution, Retail, and other license types.
- Secrets posture documented with redacted inventory, file permission evidence, removed sibling-app token fallback, and accidental-exposure SOP.
- Evidence backup/export implemented:
  - Saved-chat Copy Response with Word-quality paste, preserved bolds, spacing, and true bullets
  - Diagnostics Download CSV
  - Per-table Diagnostics Prepare Shuttle / Commit Shuttle for GRID now and CALC-compatible artifacts later.
- Open in Analyst is now the strict SOP for Diagnostics-backed Analyst work:
  - Diagnostics pull/review -> Open in Analyst -> ask Analyst question -> save/copy response.
- Incident mini-runbook documented for wrong write, token exposure, API outage/degraded integration, model bad answer, and Smartsheet corruption/bad shuttle.

- Live-farm pilot / API-key packet completed for customer-facing review.

Final refresh completed on 2026-09-04:

- Operator live test closed PASS after legacy hash-only auth config was converted to explicit user records.
- Admin completed the governed write process.
- v4 token/context optimization closed PASS: final Analyst retest used 6,990 input tokens, preserved all seven evidence sources, and retained source plant, plant batch, harvested plant range, harvest, package, wet weight, packaged-so-far, ratio, and Bucked Inventory operational rows.
- Colliers token/context optimization closed PASS for readiness: Advisor improved from 10,773 to 9,365 input tokens; Lessons Learned improved from 10,149 to 8,463; market/rate Specialist Chat ran at 6,133; UW Analyst remained stable at 7,824.

Final supervised-pilot readiness claim:

Aiden v4 Metrc is 100% supervised-pilot ready within the documented pilot boundaries.

This still does not yet mean fully production-certified or unsupervised-production ready.

12. Final External Position

Aiden is a governed enterprise AI control tower. GhostOS is the orchestration and governance runtime above local/private inference. The product value is not that a model can autonomously operate regulated systems; it is that deterministic code surrounds model synthesis with routing, evidence boundaries, tool constraints, preflight, approval, confirmation, and audit-oriented writeback.

The strongest current regulated-operations proof is Aiden v4 Metrc, where sandbox Metrc workflows are routed through read-only narrative lanes and governed Action Queue writes. The current safe next step is a narrow supervised Metrc farm operations Live pilot, with production API use dependent on final client authorization.

13. Final Status

The eight-item forensic review sequence is complete as a documentation packet:

1. Master scrutiny matrix: complete.

2. GhostOS orchestration brief: complete.

3. Validated claims register: complete.

4. Three-instance platform comparison: complete.

5. Private CUDA-independent inference brief: complete.

6. Security/safety brief: complete.

7. Intel answer bank: complete.

8. Final forensic packet assembly: complete.

Remaining work is production/performance hardening and optional fresh evidence refresh before specific external reviews:

- Fresh runtime snapshots.
- Hardware/profiler measurements.
- Production hardening artifacts.
- Screenshot/log appendix for the review room.

Appendix

Document prepared from the completed Intel forensic sequence:

1. Supporting_Artifacts/Core_Forensics/intel_master_scrutiny_matrix_2026-08-31.md

2. Supporting_Artifacts/Core_Forensics/GhostOS_orchestration_brief_2026-08-31.md

3. validated_claims_register_2026-08-31.md

4. Supporting_Artifacts/Core_Forensics/three_instance_platform_comparison_2026-08-31.md

5.Supporting_Artifacts/Core_Forensics/private_cuda_independent_inference_brief_2026-09-01.md

6. Supporting_Artifacts/Core_Forensics/security_safety_brief_2026-09-01.md

7. intel_answer_bank_2026-09-01.md

Pilot-readiness addendum artifacts completed on 2026-09-03:

8. hardware_runtime_measurement_packet_2026-09-03.md

9. Supporting_Artifacts/Evidence_Closeouts/day1_evidence_sprint_closeout_2026-09-02.md

10. Supporting_Artifacts/Evidence_Closeouts/day2_safety_sprint_closeout_2026-09-02.md

11. Supporting_Artifacts/Pilot_Hardening/pilot_hardening_forensics_2026-09-03.md

12. Supporting_Artifacts/Pilot_Hardening/pilot_auth_rbac_hardening_2026-09-03.md

13. Supporting_Artifacts/Pilot_Hardening/pilot_scope_2026-09-03.md

14. Supporting_Artifacts/Pilot_Hardening/pilot_secrets_posture_2026-09-03.md

15. Supporting_Artifacts/Pilot_Hardening/pilot_evidence_backup_export_2026-09-03.md

16. Supporting_Artifacts/Pilot_Hardening/pilot_incident_mini_runbook_2026-09-03.md

17. mark_facing_pilot_api_key_packet_2026-09-03.md

18. pilot_readiness_rollup_2026-09-03.md

Final-readiness addendum artifacts completed on 2026-09-04:

19. pmo_update_2026-09-04.md

20. final_readiness_packet_refresh_2026-09-04.md

Late readiness hardening addendum completed on 2026-09-05:

21. pmo_update_2026-09-05.md

Final v4 pilot-readiness / CA market-watch addendum completed on 2026-09-06:

22. pmo_update_2026-09-06.md

23. final_readiness_packet_refresh_2026-09-06.md