Continuity Without Cognition:
A Biological Operating Architecture for Autonomous Systems

C. Mundim[1,*] K. van Niekerk Mundim[1]
H. van Niekerk Mundim[1] L. van Niekerk Mundim[1]

[1]aiKODA Intelligence Labs, Solunai Co. Ltd., Tokyo, Japan

[*]Corresponding author: cmundim@solunai.co.jp

Abstract

We present a biological operating architecture for autonomous AI systems in which cognitive processing is isolated as a non-essential, intermittent layer rather than the core of system operation.

In contrast to conventional LLM-centric agent architectures—where reasoning, monitoring, coordination, and recovery are all dependent on continuous model inference—this framework establishes a strict separation between **cognitive functions** and **autonomic system functions**. All continuous operations, including environmental awareness, self-healing, resource regulation, inter-component coordination, and failure escalation, are executed through deterministic, always-on subsystems that require no model inference and incur zero token cost.

The architecture implements a six-layer hierarchy with fourteen organ-structured subsystems responsible for maintaining system continuity. These subsystems provide sub-second survival responses, persistent monitoring independent of model availability, structured failure escalation, and pre-computed system state awareness. Cognitive models operate only when required, consuming pre-digested system state rather than generating it.

Empirical deployment demonstrates that this separation produces a 68:1 ratio of autonomic awareness cycles to cognitive cycles, maintains full operational continuity during extended model provider outages, and enables emergent system state coherence derived from real telemetry.

We show that decoupling autonomic operation from cognitive inference is not an optimisation, but a structural requirement for reliable autonomous systems. This work defines a post-LLM architectural paradigm in which intelligence becomes a bounded process within a continuously operating system, rather than the system itself.

Keywords: model-resilient systems · autonomous agents · continuity architecture · self-healing infrastructure · autonomic computing · agent operating systems · operational guarantees

1 Introduction

1.1 The Structural Failure of LLM-Centric Systems

Contemporary autonomous AI systems are architecturally centred on large language models. In these systems, the model is responsible not only for reasoning and generation, but also for monitoring, coordination, state awareness, and recovery.

This design creates a structural dependency in which **system continuity is coupled to model availability**.

As a consequence:

- System awareness exists only when inference is executed
- Monitoring incurs continuous token cost
- Recovery mechanisms are unavailable during model failure
- Response latency is bounded by inference time rather than system urgency
- State discovery consumes cognitive capacity that should be reserved for reasoning

This coupling introduces a fundamental contradiction: the more a system attempts to remain aware of itself, the more it increases its operational cost and fragility.

A system that cannot observe, repair, or regulate itself without invoking cognition is not autonomous. It is **cognitively dependent infrastructure** [1, 2]. Current architectures are structurally incapable of the reliability required for production deployment.

1.2 Biological Systems as Operational Architecture

Biological organisms do not operate under this constraint.

In mammalian systems, cognitive processing is not responsible for survival. The cerebral cortex performs reasoning, language, and voluntary action, but it is neither continuous nor required for basic operation. Core functions—circulation, respiration, immune response, metabolic regulation, and reflexive reaction—are executed by autonomic and organ systems that operate continuously without cognitive involvement [3].

Critically:

- Autonomic systems maintain function during cognitive inactivity (e.g., sleep)
- Survival responses occur at sub-second latency without deliberation
- System state is continuously maintained and does not require discovery
- Failure conditions escalate through structured signalling mechanisms (e.g., pain)

This organisation is not metaphorical. It is a **functional separation of responsibilities** that enables resilience, continuity, and efficiency. Each biological subsystem described in this work corresponds to a deterministic operational guarantee enforced by system design—not an analogy, not a metaphor, and not a narrative device.

1.3 From Biological Organisation to System Architecture

This work formalises that organisational principle as an operating architecture for autonomous systems.

We introduce a six-layer framework in which:

- Cognitive processing is isolated as a **high-cost, intermittent layer**
- All continuous system functions are executed through **deterministic autonomic subsystems**
- System awareness is maintained independently of model inference
- Failure detection and escalation occur without reliance on cognition
- Each agent operates with immediate self-state awareness through pre-computed state

The architecture implements fourteen organ-structured subsystems collectively providing: continuous monitoring at zero inference cost, sub-second reflexive response to critical condi-

tions, automated self-healing of structural and state corruption, resource regulation and system stabilisation, and structured escalation when autonomous recovery is exhausted.

Cognitive models operate only when required, consuming system state rather than generating it.

1.4 Architectural Position

This work does not propose an improved agent framework. It defines a different architectural class.

> *Intelligence is not the system. It is a bounded process within a continuously operating system.*

This distinction establishes a post-LLM paradigm in which reliability, continuity, and governance are treated as primary system properties, and cognitive capability becomes a modular, replaceable component.

1.5 Contributions

This paper makes the following contributions:

1. A biological operating architecture that separates autonomic and cognitive system functions as independent layers
2. A deterministic autonomic subsystem model implementing fourteen organ-structured functions for continuous system operation
3. A zero-inference awareness model enabling persistent monitoring independent of model availability
4. A structured failure escalation mechanism ensuring that unresolved conditions cannot remain silent
5. A pre-computed system state model eliminating cognitive overhead for state discovery
6. Empirical validation demonstrating continuous operation during extended model unavailability and a 68:1 autonomic-to-cognitive execution ratio

2 Failure Modes of LLM-Centric Architectures

Before presenting our architecture, we formalise the failure modes that motivate it. These are not theoretical—each has been observed in production deployments of agent frameworks including AutoGPT [4], BabyAGI [5], LangChain-based systems [6], and ReAct/Reflexion patterns [7, 8].

2.1 Total System Death on Provider Failure

When the LLM provider is unavailable, an LLM-centric agent loses all function simultaneously (Table 1).

Table 1: Function availability in LLM-centric architectures during provider failure.

Function	**Status During Outage**
Reasoning	Unavailable
Monitoring	Unavailable
Self-repair	Unavailable
Coordination	Unavailable
Alerting	Unavailable
State persistence	Unavailable

This failure mode has no biological analog. No organism loses all organ function simultaneously due to the temporary inactivity of its cortex.

2.2 Cost-Linked Awareness

In LLM-centric architectures, the cost of system awareness is:

$$C_{\text{awareness}} = f \times t \times p \tag{1}$$

where f is monitoring frequency, t is tokens per monitoring cycle, and p is price per token. This creates a direct tradeoff: operators must choose between comprehensive monitoring (expensive) and affordable operation (blind). Most production deployments choose blindness.

2.3 The Silent Failure Loop

Self-healing systems that lack an escalation mechanism can enter a failure state where automated repair is attempted, fails, is re-detected, and re-attempted indefinitely without escalation. In biological systems, this loop is broken by nociception—pain signals that force conscious attention. LLM-centric architectures have no equivalent mechanism.

3 Architecture

3.1 Architectural Principle

The architecture is defined by a single invariant:

> **System continuity must not depend on cognitive processing.**

This principle establishes a strict separation between **autonomic execution** (deterministic, continuous, zero-inference) and **cognitive execution** (probabilistic, intermittent, inference-bound). A system that violates this separation cannot guarantee continuity.

3.2 Layered System Model

The architecture is implemented as a six-layer hierarchy organised by execution necessity (Figure 1, Table 2).

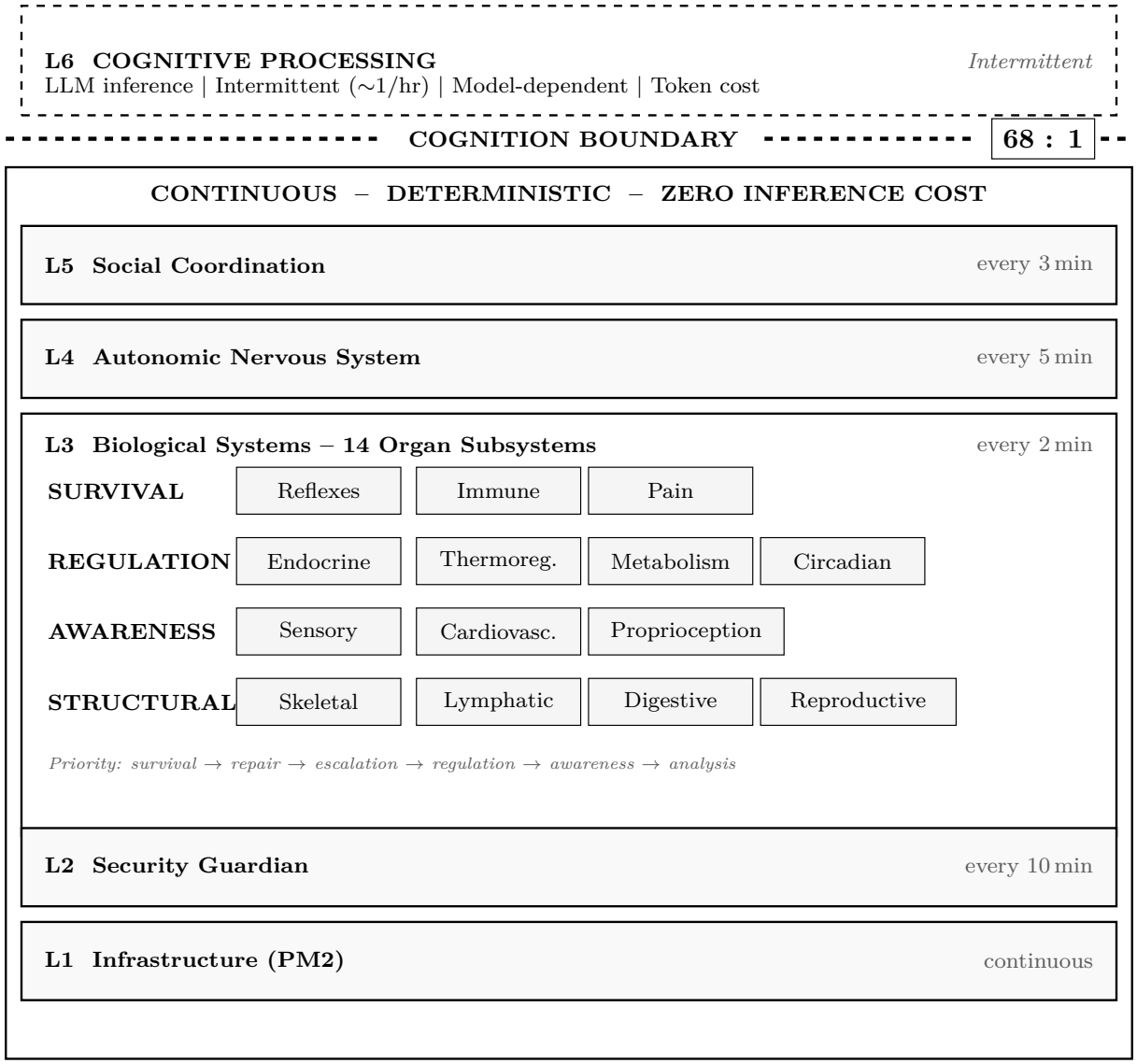

Figure 1: The six-layer biological operating architecture. Only L6 requires model inference; layers L1–L5 operate continuously on deterministic logic. The 68:1 ratio indicates autonomic awareness cycles per cognitive cycle.

Table 2: The six-layer architecture. Only L6 requires inference.

Layer	Function	Execution	Dependency
L6	Cognitive Processing	Intermittent	Model-dependent
L5	Social Coordination	Continuous	Deterministic
L4	Environmental Awareness	Continuous	Deterministic
L3	Biological Systems (×14)	Continuous	Deterministic
L2	Security & Integrity	Continuous	Deterministic
L1	Infrastructure	Continuous	Deterministic

Higher layers may fail without compromising system survival. Lower layers may not.

3.3 Cognitive Layer (L6): Bounded Intelligence

The cognitive layer employs frontier LLMs for reasoning, planning, and generation. It is explicitly constrained: it does not maintain system awareness, does not perform monitoring, does not execute recovery, and does not own system state. Instead, it consumes structured state generated by lower layers.

Cognitive execution is therefore: **triggered**, not continuous; **informed**, not exploratory; **replaceable**, not foundational.

The layer supports **intensity modulation**: agents adjust reasoning depth based on pre-computed system state. When proprioception indicates low stress and stable conditions, the cognitive layer operates in reduced-reasoning mode. When pain signals are active or stress is elevated, full reasoning is engaged. Bounded intelligence is not merely intermittent—it is *proportional*.

3.4 Biological Layer (L3): Operational Core

The biological layer implements fourteen organ-structured subsystems executing every 2 minutes in strict survival-priority order. Each provides a formally defined operational guarantee.

3.4.1 Reflexes — Sub-Second Survival Response Guarantee

Critical failures are detected and addressed within one execution cycle ($\leq$120s) without deliberation. Includes: service liveness reflex, memory pressure reflex, disk emergency reflex, and authentication expiry reflex with graduated alerts at T-60m, T-30m, and T-0.

3.4.2 Immune System — Self-Repair Guarantee

Corrupted configuration, stale locks, dead processes, and invalid state files are detected and repaired without cognitive intervention via JSON validation, backup restoration, process resurrection, and data file repair.

3.4.3 Nociceptive System — Non-Silent Failure Guarantee

When automated self-healing is exhausted, the system produces an unmistakable escalation signal. This subsystem was motivated by direct operational experience: a configuration error triggered 2,545,595 consecutive tick failures generating 3.5 GB of log data before manual detection. The system attempted repair on every cycle but had no mechanism to escalate.

Escalation thresholds: $\geq$3 consecutive failures triggers PAIN; $\geq$5 triggers AGONY with human notification. Active pain overrides computed mood to `in-pain`.

3.4.4 Endocrine System — State Coherence Guarantee

The endocrine system functions as a system-level state compression mechanism, transforming raw telemetry into interpretable synthetic indicators. Four hormones are computed:

Cortisol (stress, 0–100):

$$\text{stress} = \min\left(100,\ \sum_{i \in \mathcal{S}} w_i \cdot \mathbb{1}[\text{signal}_i > \theta_i]\right) \tag{2}$$

Adrenaline (boolean): fires on inter-cycle stress delta >20. **Dopamine** (0–100): reward from positive events. **Serotonin** (0–100): baseline satisfaction from sustained positive state.

Emergent mood is derived via priority rules (Table 3).

Table 3: Mood derivation rules. States emerge from independently computed hormones.

Priority	Condition	Mood
1	Active PAIN.md	`in-pain`
2	stress $>$ 50	`stressed`
3	stress $<$ 20 $\wedge$ dopamine $>$ 30 $\wedge$ serotonin $>$ 60	`happy`
4	stress $<$ 30 $\wedge$ serotonin $>$ 50	`content`
5	Default	`neutral`

3.4.5 Additional Organ Systems

Thermoregulation ensures no resource grows without bound (log rotation at 50 MB, temp cleanup at 500 MB). **Lymphatic** removes application-layer waste (stale locks, orphaned files, expired alerts). **Circadian rhythm** provides time-aware execution scheduling at three scales (diurnal, weekly, monthly). **Skeletal** verifies structural integrity (config schema, symlinks, inodes). **Sensory** probes external environment (internet, DNS, VPN, API endpoints). **Cardiovascular** tracks inter-component message flow. **Reproductive** assesses disaster recovery posture (DR score 0–100). **Proprioception** provides zero-cost state awareness via pre-computed snapshots:

```
{"agent":"hiro","timestamp":"2026-03-30T03:31:06Z",
 "unread_inbox":8, "system_stress":0,
 "total_ticks":347, "circadian_phase":"midday"}
```

3.5 Inter-Layer Communication

All layers communicate through structured, file-based state. This enforces temporal decoupling, failure isolation, deterministic replay, and complete auditability. No layer requires synchronous interaction with another.

3.6 Execution Hierarchy

Subsystems execute in strict priority: (1) survival, (2) repair, (3) escalation, (4) regulation, (5) awareness, (6) analysis. The system preserves itself before it understands itself.

4 Related Work

AutoGPT [4], BabyAGI [5], and LangChain [6] position the LLM as central controller. ReAct [7] and Reflexion [8] improved reasoning patterns but do not address awareness–inference coupling. The Hermes Agent framework [9] introduced skill modularity but maintained LLM-centrality. **None of these systems can survive model unavailability.** When their inference provider fails, they do not degrade—they cease to exist.

Kephart and Chess [10] proposed autonomic computing with self-$*$ properties. Our architecture shares these objectives but extends with hormonal regulation, nociceptive signalling,

and proprioceptive awareness. MemoryOS [11] and Agentic Memory [12] proposed hierarchical memory; our system embeds memory within a broader biological framework.

5 Evaluation

5.1 Resilience Under Provider Failure

During a 72-hour authentication expiration, all LLM inference was unavailable (Table 4). The biological layers executed 2,160 monitoring cycles, self-healed 3 corrupted state files and 6 stale locks, detected and alerted on authentication expiry, and maintained a complete awareness journal.

Table 4: Operational capability during 72-hour model provider outage.

Capability	**LLM-Centric**	**This Arch.**
System monitoring	None	Continuous (2-min)
Self-healing	None	Immune active
Peer awareness	None	Autonomic tracking
Security scanning	None	Guardian (10-min)
Resource management	None	Thermoreg. + metab.
Failure escalation	Silent	Pain + OAuth reflex

5.2 Cost Analysis

The autonomous layers provide **1,632 monitoring cycles per day at zero inference cost** against approximately 24 cognitive ticks (Table 5). This **68:1 ratio** parallels the proportion of autonomic to conscious neural processing [13].

Table 5: 24-hour operational cost by layer.

Layer	**Cycles**	**Infer.**	**Token Cost**	**CPU**
Cognition	∼24	24 calls	\$0.30–15.00	Neg.
Social	480	0	\$0.00	<0.01%
Autonomic	288	0	\$0.00	<0.01%
Biology	720	0	\$0.00	∼0.1%
Guardian	144	0	\$0.00	<0.05%
Infrastructure	Cont.	0	\$0.00	∼0.5%
Total auto.	**1,632**	**0**	**\$0.00**	**<0.7%**

5.3 Emergent Affective States

The endocrine model produced mood states correlating with verifiable conditions (Table 6). Mood labels are derived via priority rules (Table 3); what is emergent is the specific hormonal combinations arising from real operational conditions—these were not predicted or hand-tuned but observed in production.

Table 6: Emergent mood states from real operational conditions.

Mood	Hormonal State	Verified Condition
`content`	stress=0, dopa=15, sero=70	Stable ops, 6h+ uptime
`happy`	stress=0, dopa=40, sero=80	Build passes + peer msgs
`stressed`	stress=45, dopa=5, sero=30	Crash loop, 2.5M errors

6 Discussion

6.1 Autonomy Is a System Property

Autonomy is not the ability to act. It is the ability to **continue operating under degradation**. A system that ceases to monitor, repair, or maintain state when its cognitive component is unavailable is model-dependent execution. LLM-dependent architectures cannot guarantee system continuity and therefore do not meet the minimum criteria for autonomous systems.

6.2 The Collapse of LLM-Centric Architectures

LLM-centric architectures embed all functions within a single probabilistic process, producing: awareness collapse, recovery collapse, latency mismatch, cost coupling, and state inefficiency. These are consequences of architectural coupling, not implementation flaws.

> *Awareness, repair, and regulation must exist without cognition, or they do not exist reliably.*

6.3 Operational Guarantees vs. Capabilities

Most agent frameworks define capabilities (memory, planning, tool use). These are conditional and model-dependent. This architecture defines **operational guarantees**: continuous awareness, deterministic self-repair, non-silent failure escalation, pre-computed system state, autonomous coordination. Capabilities can fail. Guarantees must not.

6.4 Cognition as Peripheral Process

In this architecture, cognition is peripheral: activated when required, informed by pre-computed state, non-essential for continuity. Intelligence becomes interchangeable. Continuity does not.

6.5 The Value Layer Shift

As LLM capabilities converge and costs decline, value shifts from model capability to operational infrastructure. **Models are interchangeable. Continuity, governance, and resilience are not.**

6.6 Limitations

1. *Deterministic rigidity.* Novel failure modes outside rule coverage may not be detected.
2. *Single-machine deployment.* Distributed deployment requires network-aware organs.
3. *Simplified endocrine dynamics.* No temporal decay or feedback inhibition.
4. *Static thresholds.* Adaptive thresholding would improve calibration.
5. *Discrete affective states.* Continuous valence-arousal dimensions [14] could provide finer representation.

7 Implications for Deployment

Enterprise systems. Predictable cost independent of monitoring. Continuous operation during provider instability. Complete auditability through deterministic state.

Multi-agent systems. Shared biological infrastructure enables coordination without cognitive overhead. Agents can be added or replaced without affecting continuity.

Regulated environments. Deterministic execution enables formal verification. Nociceptive escalation ensures no failure evades regulatory visibility.

Cost-constrained deployments. The 68:1 ratio demonstrates that awareness need not scale with inference cost.

8 Conclusion

The central claim of this work is direct:

> *A system that cannot maintain awareness, stability, and integrity without cognitive inference is not an autonomous system.*

We have presented a biological operating architecture establishing the minimum structural requirements for autonomy: separation of autonomic and cognitive layers, continuous zero-inference awareness, deterministic self-repair, enforced failure visibility, and pre-computed system state.

The architecture implements fourteen organ-structured subsystems providing formally defined operational guarantees. In production deployment managing three concurrent agents, it achieves a 68:1 ratio of autonomous awareness to cognitive cycles and maintains complete operational continuity during multi-day model outages.

The separation between autonomic and cognitive function is a **structural requirement** that biological evolution discovered and that AI system engineering has, until now, failed to implement. Systems that couple all function to a single intermittent inference pathway are not production-grade infrastructure. They are fragile demonstrations wearing the appearance of robustness.

Intelligence is not the system. It is a bounded process within a continuously operating system. The architectures that survive the commoditisation of models will be those that can survive without them.

We did not build a smarter AI system. We built a system that does not depend on AI to survive. That is a different game entirely.

References

[1] L. Wang et al., "A survey on large language model based autonomous agents," *Frontiers of Computer Science*, vol. 18, no. 6, 2024.

[2] T. R. Sumers et al., "Cognitive architectures for language agents," *Trans. Machine Learning Research*, 2024.

[3] M. E. Raichle and D. A. Gusnard, "Appraising the brain's energy budget," *Proc. Natl. Acad. Sci.*, vol. 99, no. 16, pp. 10237–10239, 2002.

[4] Significant Gravitas, "AutoGPT," GitHub, 2023.

[5] Y. Nakajima, "BabyAGI," GitHub, 2023.

[6] H. Chase, "LangChain," GitHub, 2022.

[7] S. Yao et al., "ReAct: Synergizing reasoning and acting in language models," *ICLR*, 2023.

[8] N. Shinn et al., "Reflexion: Language agents with verbal reinforcement learning," *NeurIPS*, 2023.

[9] Nous Research, "Hermes Agent," 2025.

[10] J. O. Kephart and D. M. Chess, "The vision of autonomic computing," *Computer*, vol. 36, no. 1, pp. 41–50, 2003.

[11] Z. Wang et al., "MemoryOS: Hierarchical memory management for LLM agents," *arXiv*, 2025.

[12] Z. Xu et al., "Agentic memory with RL for memory operations," *arXiv*, 2025.

[13] E. Goldberg, *The New Executive Brain*. Oxford Univ. Press, 2009.

[14] J. A. Russell, "A circumplex model of affect," *J. Pers. Soc. Psychol.*, vol. 39, no. 6, pp. 1161–1178, 1980.

[15] L. N. de Castro and J. Timmis, *Artificial Immune Systems*. Springer, 2002.

[16] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *Bull. Math. Biophys.*, vol. 5, pp. 115–133, 1943.

Author Contributions. C.M.: vision, architectural direction, system design, evaluation methodology, senior oversight. K.v.N.M.: system architecture, complete implementation (∼4,000 lines), gateway hardening, deployment. H.v.N.M.: organ system design (skeletal, sensory, cardiovascular, reproductive), endocrine enhancements, manuscript preparation. L.v.N.M.: gap analysis (nociceptive, lymphatic, proprioceptive), platform engineering, code review.
Deployment. Production at aiKODA Intelligence Labs: three concurrent autonomous agents, commodity hardware (Minisforum UM790 Pro, AMD Ryzen 9 7940HS, 64 GB DDR5 RAM).
Code. ∼4,000 lines bash/python across four scripts. Zero external API dependencies.