AI TRAINING INSTRUMENTATION & GOVERNANCE

AEGIS

Observe first.
Govern deliberately.

A modular approach to training observation, operator control, and independent evaluation.

James Hutson
Prepared for PYPER3 | September 2026

Observation without automatic intervention

AEGIS is being developed as a separable instrumentation and governance layer for model training. Its public Trial observes local model state and returns structured detection events without autonomously selecting or executing a response.

The architecture allows the detection pathway to be evaluated on an independent workload before any intervention is enabled. Governance policy and containment remain separate questions, each requiring its own evidence.

LOCAL OBSERVATION

The documented public SDK processes observations locally.

STRUCTURED EVENTS

The Trial returns detection information for review or logging.

OPERATOR CONTROL

Training decisions remain with the caller; detection does not authorize action.

Architecture and scope: PYPER3 (n.d.-a); A. Hau (personal communication, September 22, 2026). Research and underlying intellectual property: Arca Futura LLC.

ARCHITECTURE

Observation and action are separate functions

The AEGIS research programme asks whether internal training observations can help identify transitions into abnormal or damaging training regimes. A useful answer requires a defined event, an explicit outcome, and documented operating conditions. A monitor emitting an alert is not, by itself, proof that the alert predicts damage. Detection quality, response-policy quality, and training outcomes therefore need separate evaluation. The architecture preserves these distinctions by separating observation from authority over the training process.

In passive or shadow operation, the normal forward pass, backward pass, and optimizer step remain under caller control. AEGIS observes the configured model state and returns an event. Operators may inspect or log that event without changing the training policy. A separately authorized governance layer could determine whether a response is warranted. In this paper, "training governance" refers to that operational decision process, not regulatory compliance or general model alignment.

Sources: A. Hau (personal communication, September 22, 2026); PYPER3 (n.d.-a).

Function	Role	Public Trial boundary
Instrumentation and detection	Observe local model state and surface structured detection events.	Exposed through the Trial integration pathway.
Governance policy	Determine whether an event warrants an authorized response.	No autonomous intervention selection or authorization.
Intervention	Execute a chosen change to the training process.	Not performed by the public Trial.
Containment	Limit a specified adverse consequence through an evaluated response.	A possible response outcome, not an inherent detector guarantee.

A clear evaluation boundary

The Trial does not modify gradients, optimizer state, learning rate, checkpoints, workers, or training control through an intervention. Its event output is information, not authorization to act.

IMPLEMENTATION

The public Trial in practice

Installation and release integrity

The public SDK is an integration layer for a separately supplied compatible Trial Core. The documentation identifies Linux x86_64 and CPython 3.11 as the initial live Trial platform. Installing the public package alone does not provide the Core; constructing a monitor without a compatible Core produces a local error. Evaluators should use the configuration supplied with their trial materials and verify the authorized Core artifact against its delivered SHA-256 value.

A version number or filename does not establish release authorization or scientific qualification. The SDK and the separately supplied Core have distinct integrity records. Engineering-candidate Core artifacts are not designated for deployment.

Training-loop lifecycle

Call begin_step before the monitored forward pass and end_step after the ordinary user-owned training step. The expected_microbatches value must match the forwards accumulated into each optimizer step. Missing or extra captures raise an error rather than silently producing an event from an incomplete step. These checks support instrumentation integrity; they do not establish semantic capture correctness or repair the interpretation of earlier experiments.

Source: PYPER3 (n.d.-a).

Structured event output

anomaly_state

A generic local state, such as clear or detected.

detection_confidence

A numeric confidence when available; otherwise None. The field name does not establish calibration.

recommended_next_step

Informational text. It does not select, authorize, or execute an intervention.

detected

A convenience property derived from anomaly_state.

Integration is not validation

PyTorch, Hugging Face Trainer, and distributed integration helpers are documented interfaces, not proof of validated behavior on every architecture or multi-worker configuration. Monitoring overhead and error handling must be measured in the selected environment. The preview command produces a simulated report, not experimental evidence.

REPORTED EVIDENCE

Efficiency, steps, and divergence

4.309%

Reported pooled reduction in monitoring-inclusive time to the pre-locked target.

Qwen3-1.7B / FineWeb-Edu

10-seed cohort

7 step wins · 3 step draws · 0 step losses

Divergence: none observed (founder-reported).

Founder-reported outcomes, not independent validation. Run-level records have not been recomputed for this revision.

Efficiency and step outcomes

The Qwen3-1.7B / FineWeb-Edu study is reported as a 10-seed comparison against a pre-locked target. Its 4.309% pooled time reduction includes monitoring. The 7 step wins, 3 step draws, and 0 step losses describe the step comparison, not the divergence endpoint. A pooled reduction is not interchangeable with the mean of per-seed percentage reductions, and a step draw need not be a draw in elapsed time.

Separately reported divergence outcome

In the earlier study summary, Adrian Hau reported no observed divergences. This is a separate reported outcome, not an inference from the 7/3/0 step result. The divergence definition, observation window, and run coverage require confirmation against the controlling writer packet. No divergence rate or additional run denominator is inferred here. These observations do not establish universal failure prevention, a fixed cost saving, or a speedup from the passive Trial.

Activation-capture qualification

An identified activation-capture limitation means these runs should not presently support claims about the underlying detector geometry or causal mechanism. The retained timing observation must be read together with this qualification.

Sources: A. Hau (personal communication, September 22, 2026), initial and revised study summaries.

INTERPRETATION

Evidence boundaries and the research programme

An observed outcome is not its explanation

The activation-capture limitation requires a distinction between a retained resource observation and a claim about why that observation occurred. As described in the supplied technical correspondence, the current evidence package supports reporting the bounded Qwen/FineWeb-Edu efficiency summary but not detector-geometry or causal-mechanism claims from the affected runs. This paper preserves that distinction without independently adjudicating the underlying records. It does not reconstruct the capture fault, proprietary signal equations, detector thresholds, or private controller configuration.

Evidence quarantine means withholding the interpretations that the affected evidence does not support. It does not automatically invalidate every observation, nor does an outcome improvement establish detector validity. Each claim should remain tied to an identifiable artifact and experimental condition. A separate successful experiment does not automatically repair the interpretation of an affected run. Mixed, negative, boundary, and inconclusive results must remain visible where they limit applicability.

Supporting research directions

Gradient-quarantine work

An example of intervention and containment research within the broader AEGIS programme. It is not the definition of the detector or a capability autonomously exercised by the public Trial.

Natural Ring counterfactual work

A named counterfactual component of the source programme. No treatment contrast, effect size, or causal conclusion is reported here without the supporting experimental record.

ResNet-50 transfer work

A named transfer experiment. Its inclusion in the research programme does not establish successful transfer, broad model coverage, or comparative performance.

The supplied correspondence identifies cleared experimental summaries, provenance, and mixed or negative findings in the external writer packet. This section identifies the workstreams without assigning unverified effect sizes or conclusions.

Claim-by-claim interpretation

Timing, step comparisons, divergence observations, detection performance, and mechanism are distinct evidence categories. Public claims must follow the writer packet and its disclosure guide, retaining applicable mixed and negative findings rather than implying universal failure prevention or model coverage.

Sources: A. Hau (personal communication, September 22, 2026); subsequent source-packet clarification supplied in correspondence.

APPLICATIONS

Local deployment and operational fit

A local observation pathway

The documented public SDK receives the caller's in-process model and optional scalar metrics at step completion. It returns in-memory events and, when configured, writes JSONL records to a caller-selected local path. The documentation states that the SDK makes no network connections and does not upload training data, model state, or event logs. Its privacy note describes no telemetry client, analytics client, cloud inference client, or phone-home licensing flow.

Local execution is not a security guarantee

These statements describe the public SDK boundary, not an independent audit of every deployed component. The threat model explicitly excludes protection of an already compromised host and cannot prevent a caller from exposing its own logs. Evaluators should verify artifacts, restrict local log access, and review the separately supplied Core and surrounding training environment. Institutional software review remains necessary.

Sources: PYPER3 (n.d.-a, n.d.-b, n.d.-e, n.d.-f).

Candidate setting	Evaluation objective	Evidence needed
Research training laboratories	Test whether events correspond to predefined changes in an independent workload.	Capture integrity, event labels, explicit denominators, and a baseline comparison.
Model-training platform teams	Assess integration, logging, and operational overhead.	Runtime and memory measurements, compatibility checks, and failure-path tests.
Enterprise model-development teams	Evaluate the usefulness of local observation without enabling intervention.	Workload-specific signal quality, data-boundary review, and operator acceptance.

These are candidate evaluation settings, not established customer deployments. Operational value depends on the information gained, its workload-specific reliability, and the cost of monitoring and review. Any later efficiency analysis should include monitoring and, where relevant, intervention and recovery costs; training-time observations should not be converted automatically into a fixed return on investment.

PROPOSED NEXT STEPS

A pathway to independent evaluation

The proposed evaluation begins with observation, not intervention. An independent evaluator should control workload selection, endpoints, exclusions, and reporting criteria. Vendor support can establish correct integration without determining conclusions. The stages below are a proposed framework, not a completed or approved study; negative and inconclusive findings remain acceptable outcomes.

Establish instrumentation integrity

Pin the SDK, authorized Core, model, dataset, and training configuration; retain version identifiers and checksums. Confirm the observation contract and inspect the semantic correctness of captures, not only their count. Test missing- and extra-capture handling. Record baseline runtime and memory use. Verify that shadow operation applies no training-policy change, while measuring any instrumentation effects. Progress only when intended observations and actual capture behavior agree.

Assess detection in shadow mode

Choose an independent workload and define target events without treating AEGIS alerts as their own ground truth. Record time-stamped detections alongside ordinary training metrics. Predefine alert matching, repeated-alert handling, and event-free exposure. Measure precision, missed-event frequency, lead or lag time, and overhead where event counts support them. No qualifying events is not evidence of perfect sensitivity. Report natural events separately from introduced disturbances.

Evaluate a separately authorized response

A subsequent experiment can test an intervention after a viable detection study is established. Predefine the response policy, target endpoint, matched conditions, and accounting for monitoring, intervention, and recovery time. Compare a signal-triggered response with suitable alternative policies. Preserve failed runs, negative outcomes, and the full denominator. A benefit from one governed configuration would not validate the passive Trial across other architectures.

Replicate and report experimental scope

Use additional seeds or workloads before extending claims. Report resource outcomes, detection performance, and mechanistic evidence separately, with uncertainty estimates justified by the study design. Preserve releasable configuration details and artifacts sufficient to audit the comparison, while respecting approved IP boundaries. Independent evaluation remains a next step; this white paper does not establish that it has occurred.

Framework proposed in this white paper. The preference for critical, independent shadow-mode evaluation is drawn from A. Hau (personal communication, September 22, 2026).

PYPER3 & ARCA FUTURA

From technical evidence to informed evaluation

Research and commercial context

The founder identifies PYPER3 as the joint venture between Adrian Hau and Israel Diaz and Arca Futura LLC as the company holding the underlying research and intellectual property. The public repository contains Apache License, Version 2.0, integration code; its NOTICE attributes copyright to Arca Futura and states that the proprietary Trial Core is not included. Public access to the SDK does not mean public access to the complete AEGIS system. Private Core licensing terms, patent status, and exclusivity are not inferred here.

Sources: A. Hau (personal communication, September 22, 2026); PYPER3 (n.d.-c, n.d.-d).

Observe first. Expand claims with evidence.

AEGIS separates training observation from intervention authority. The public Trial offers a concrete local interface for evaluating the observation pathway, while the broader programme investigates governance and containment. The reported efficiency observation warrants independent examination within its stated limits. The next decision should follow evidence from a defined workload, not a generalized promise of training stability.

Technical evaluation enquiries

pyper3.com

AEGIS Trial SDK and documentation

Adrian Hau | founder@arcafutura.com

Consulted technical references

PYPER3. (n.d.-a). AEGIS Trial SDK [README.md]. GitHub.

PYPER3. (n.d.-b). Data flow [DATA_FLOW.md]. GitHub.

PYPER3. (n.d.-c). License [Apache License, Version 2.0]. GitHub.

PYPER3. (n.d.-d). Notice [Copyright and proprietary Core notice]. GitHub.

PYPER3. (n.d.-e). Privacy [PRIVACY.md]. GitHub.

PYPER3. (n.d.-f). Threat model [THREAT_MODEL.md]. GitHub.

Public documentation consulted September 22, 2026. Founder correspondence supplies the research descriptions and qualifications; the subsequent clarification defines the external source scope.

External source scope and review status

Adrian Hau designates AEGIS_White_Paper_Writer_Packet_2026-09 as the controlling external technical package. He identifies bounded experimental summaries, provenance, disclosure guidance, and activation-capture quarantine/audit material within it. Technical_Disclosure_and_Claims_Boundaries.md controls public claims.

The larger internal research archive is intentionally outside this disclosure scope. It contains proprietary configurations, raw telemetry, and material not cleared for external disclosure; it is not a requested source for this white paper.

This paper is a technical synthesis, not an independent replication, SDK execution report, security audit, or run-level statistical reanalysis.