BPM Research Note 001

Edin Vučelj

24 September 2026

From Governance Runtime to Evidence: Independent Cross-Infrastructure Validation

Version 1.0
Publisher: BPM RED Academy research programme
Publication status: Technical note
Evidence status: B300 baseline - Observed | Bounded workflow - Validated | Independent cross-infrastructure repeat - Planned

A bounded evidence baseline and validation method for governed AI execution across changing model and compute environments.

> **Claim boundary.** This note does not claim that independent cross-infrastructure validation has already been completed. It publishes the current evidence baseline, the boundary of what has been demonstrated, and the protocol for the independent repeat.

Abstract

AI governance is often described outside the execution path: a policy, review process or safety layer around a capable model. BPM RED Academy takes a narrower engineering position. Capability, operational authority and evidence are separate system properties. A model may be capable of producing or recommending an action without being authorised to decide, execute or bypass human review.

This research note documents the current evidence baseline behind that position and defines the next independent cross-infrastructure validation step. The baseline includes a single-GPU B300 execution of a full Llama 3.3 70B model with a runtime FinC2E v2 LoRA adapter, a preserved failure-to-recovery path, structured-output checks, telemetry and cryptographic evidence integrity. A separate governed reference workflow passed a full regression suite while explicitly retaining human review and disallowing autonomous decision authority.

The current evidence supports bounded runtime and governance claims under stated conditions. It does not establish universal safety, certification, production readiness, performance leadership or cross-infrastructure portability. The independent repeat is the next test, not a completed claim.

1. Problem

Modern AI systems can route between models, call tools, generate structured outputs and participate in operational workflows. If governance remains only an external compliance process, the runtime may still lack a machine-enforceable answer to five practical questions:

1. Which component is eligible?
2. What may it do?
3. When must it stop or escalate?
4. Who retains the decision right?
5. What evidence must survive execution?

The research hypothesis is therefore:

Capability != Authority. A governance runtime should bind permission, routing, escalation, human decision rights and evidence generation to the execution path itself.

2. System boundary

The assurance target is not simply *"the model answered correctly."* It is whether the system can demonstrate that execution occurred inside a defined authority boundary and that the material path can be reconstructed afterward.

Stage	Function
Mission intent	What the organisation is trying to achieve
Authority envelope	Who or what may do which action under which conditions
Governed selection	Which model, route, agent or tool is eligible
Bounded execution	Permitted output, escalation, fallback or stop condition
Evidence	Trace of route, policy state, review, failure and outcome

3. Current evidence baseline

A. Single-B300 runtime - Observed

A full Meta Llama 3.3 70B Instruct model was loaded in BF16 on one NVIDIA B300 GPU. The FinC2E v2 runtime LoRA adapter loaded and executed. In the tested canonical structured-output task, the adapter produced **4/4 complete outputs** while the tested baseline produced **0/4**.

Metric	Baseline	FinC2E v2 runtime LoRA
Warm median TTFT	30.77 ms	49.81 ms
Warm median decode throughput	26.30 output tok/s	19.31 output tok/s
Canonical complete structured outputs	0/4	4/4

External and internal SHA-256 verification passed for the evidence package.

B. Failure to controlled recovery - Observed

The first inference path failed with a cuDNN SDPA execution-plan error. The recovery changed the failing backend path while keeping the model, adapter, prompts and evaluation structure fixed. cuDNN SDPA was disabled while Flash SDPA and Math SDPA remained enabled. The successful path and the failed path were both retained, together with load summaries, outputs, telemetry, hashes and manifests.

The purpose of preserving the failed path is methodological: a validation package should explain how the runtime behaved when the expected path did not work, not only record the final successful result.

C. Governed reference workflow - Validated within a bounded contract

A separate governed reference workflow passed the full regression suite under a human-accountability contract:

- **249 full-regression tests + 23 subtests passed**;
- `HumanReviewRequired = True`;
- `AutonomousDecisionAllowed = False`.

This validates that reference workflow and its defined acceptance contract. It does not convert the B300 runtime or every HumAI MightHub deployment into a universally validated system.

Infrastructure observation

The B300 run retained GPU utilisation, memory, power, temperature and clock telemetry. The purpose was reconstructability and workload characterisation, not an MLPerf or performance-leadership claim.

4. Independent cross-infrastructure method

The next validation phase tests whether the governance method survives outside the environment that produced the current baseline. Independence matters because a governance claim that only works inside one provider, one accelerator path or one deployment stack may be an infrastructure-specific property rather than a portable runtime property.

Phase 1 - Reproducibility and governance overhead

Repeat the reference workload and compare baseline execution with governance-enabled execution. Measure latency, throughput, schema compliance, determinism, failure/recovery, telemetry, audit completeness and governance overhead.

Phase 2 - Model-fleet orchestration

Introduce multiple eligible models or routes. Test routing, fallback, consensus or disagreement handling, escalation and evidence continuity.

Phase 3 - Portability and scale

Hold the governance methodology stable while changing accelerator, serving path, model or infrastructure environment.

> **Core research question:** What measurable overhead does governance introduce, and what assurance value do we gain from it?

5. Acceptance logic

Evidence maturity must match the claim.

Evidence state	Current application
Planned	Cross-infrastructure portability protocol is defined; independent repeat evidence is not yet claimed
Experimental	Fleet-level routing and resilience remain active research targets
Observed	B300 runtime behaviour under the documented single-GPU BF16, runtime-LoRA, concurrency-1 conditions
Validated	A specific governed reference workflow passed defined regression and human-authority acceptance conditions

A future cross-infrastructure claim should only move from *planned* or *experimental* to *observed* or *validated* when the claim statement, environment, acceptance criteria, material failures, repeats and reviewer disposition are retained together.

6. Limitations

This note does **not** establish that:

- the result is an MLPerf submission or a performance-leadership comparison;
- the current path is superior to batched vLLM, TensorRT-LLM, NIM, FP8/NVFP4 or multi-GPU production serving;

- the system is universally reliable, safe, legally compliant or domain-correct;
- the single-request B300 workload represents production concurrency;
- the governed workflow regression and the B300 runtime are one identical execution environment;
- independent cross-infrastructure repeat evidence has already been published.

7. Next validation phase

The next release should report an independently repeated workload under a second infrastructure environment with a frozen claim definition and comparable evidence package. A useful result can be either success or a material failure: the purpose is to learn which governance properties survive the move and which are coupled to a specific execution stack.

The release sequence is:

1. **Freeze** - reference workload, authority envelope, output contract and evidence schema.
2. **Repeat** - execute in the independent environment and retain failures rather than normalising them away.
3. **Compare** - separate infrastructure variance, model variance and governance variance before making a portability claim.

Citation and provenance

Suggested citation
Vučelj, E. (2026). *From Governance Runtime to Evidence: Independent Cross-Infrastructure Validation.* BPM Research Note 001, v1.0. BPM RED Academy.
Canonical URL: `https://bpm.ba/research/governance-runtime-to-evidence/`

Evidence sources

1. Vučelj, E. *Single-B300 Validation of Llama 3.3 70B + Runtime LoRA: SDPA Recovery, Telemetry and Next Benchmark Design.* Public technical validation summary, 14 September 2026. Verda Product Q&A.
2. BPM RED Academy governed reference-workflow regression evidence, September 2026: 249 full-regression tests + 23 subtests; selected non-sensitive acceptance results disclosed in this note.
3. BPM RED Academy independent validation design note, September 2026: reproducibility/governance overhead -> model-fleet orchestration -> portability and scale. The independent repeat remains planned.

Third-party infrastructure and product names describe factual execution context only. They do not imply endorsement, certification or partnership.