

Better Model or Better Harness? A Visual & Empirical Breakdown of the Six-Part Agent Harness (Observation, Context, Control Loop, Action, State, Verification)

Taylor Amarel · Principal Investigator · Portfolio: <https://tayloramarel.com> · Papers: <https://papers.tayloramarel.com> · Live 3D/Visual Studios: <https://compare.baltor.ai> & <https://fox.baltor.ai>

Abstract: When an AI system fails on a complex multi-hop investigation, a 14-constraint engineering specification, or a real-time 3D WebGL simulation, practitioners usually blame the model weights. Across five controlled benchmarks keeping model weights fixed (gemma4:31b , qwen3-coder:30b , Qwen2.5-7B-Instruct), we demonstrate that **+51.3 pp to +80.0 pp** of capability lift comes from the **Six-Part Agent Harness**—decomposing execution into *Observation*, *Context*, *Control Loop*, *Action*, *State*, and *Verification*.

1. THE SIX STRUCTURAL SUBSYSTEMS OF A PRODUCTION AGENT HARNESS

1. OBSERVATION (Perceptual & Evidence Grounding) Immutable SHA-256 Layer-0 source parsing, cross-document entity blocking canopies, barycentric terrain heightfields, and 17-joint RTMPose motion anchors extracted prior to generation.	2. CONTEXT (Factorized Memory: $H_{gen} \oplus H_{env}$) Compact $\leq 2,048$ B contract cards (baltor_contract_card.json) paired with domain skill/rule packs (SKILL.md , RULES.md) that eliminate middle-context constraint drop.	3. CONTROL LOOP (Recognition & Hypothesis Tree) 120-token pre-turn statutory recognition gates (unlocking 93.8% latent recall) and Protagonist vs. Antagonist competing-hypothesis trees that cut consensus echo from 78% to 14%.
4. ACTION (Typed Tools & Procedural Primitives) Taint-tracked tool scopes (tools.json) blocking CWE-918 SSRF/confusing deputy calls, plus verified 3D/physics modules (49-bone quadruped IK, shell fur, swept-AABB collision).	5. STATE (Epistemic Layers & Multi-Turn Graph) Strict separation of <i>Evidence</i> , <i>Interpretation</i> , and <i>Exploration</i> layers plus cumulative multi-turn obligation graphs holding Turn 4–5 sycophancy breaches to 0.0%.	6. VERIFICATION (Independent Gate: worker != verifier) Executable verifier.py non-regression checks, keyword-matched benign twins (95.8% recall at 4.2% false alarms), and 5-axis MOEA/D Pareto admission.

2. HOLDING THE MODEL FIXED: OBSERVED LIFT ACROSS FIVE EMPIRICAL & VISUAL SUITES

Evaluation Suite & Live Interactive Surface	Fixed Model	Raw Model (Unassisted)	Same Model + 6-Part Harness	Primary Harness Bottleneck Solved
1. 3D Foxglade & Quadruped Survival (fox.baltor.ai)	<code>gemma4:31b</code>	20.0% (2/10 specs; rigid boxes)	100.0% (10/10 specs; 49-bone rig)	Action + Observation: Heightfield normal alignment, rotary-gallop IK & bending grass
2. 3D Surface Relief, Lighting & Physics (compare.baltor.ai)	<code>qwen3-coder:30b</code>	28.1% stability (bump illusion)	100.0% swept-AABB & true geometry	Action + Verification: True lathe relief geometry, 4-pass lighting & zero tunneling
3. 14-Constraint Long-Form Specification (compare.baltor.ai)	<code>gemma4:31b</code>	42.9% (6/14; drops 8 mid-rules)	100.0% (14/14 verified)	Context + Verification: Factorized RULES.md checklist + AST/runtime verifier.py
4. DueCare & Camouflaged Alpha (α) (duecare / alpha)	<code>gemma4:31b</code>	25.0% recall / 78.0% echo	93.8% recall / 14.0% echo	Control Loop + State: Pre-turn recognition gate + 3-hop competing-hypothesis tree
5. Harmony Pareto Secure Coding (harmony.baltor.ai)	<code>gemma4:31b</code>	31.2% joint safe & functional	82.5% joint (0 PoC exploits)	Context + Verification: Factorized $H_{gen} \oplus H_{env}$ Pareto evolution